UNIVERSITE DE LIEGE
FACULTE DE MEDECINE VETERINAIRE
DEPARTEMENT OF ANIMAL PRODUCTIONS
Unit of Animal Genomics

Mutations germinales en Bos taurus

Germline mutations in Bos taurus

Chad Simeon Harland
Promoteur: Prof Michel Georges

THESE PRESENTEE EN VUE DE L'OBTENTION DU GRADE DE
Docteur en Sciences Vétérinaires

ANNEE ACADEMIQUE 2012-2018

Acknowledgements

My thanks to Professor Michel Georges, and Dr Carole Charlier for supervising my PhD for the last 5 years and all the support and insight they have provided during that time. In addition, the help and advice provided by Nathalie Faust, Anne Van Den Broeke, Tom Druet, Wouter Coppieters and Karim Latifa was greatly appreciated. I must also thank Bevin Harris, Richard Spelman and LIC for funding the PhD work and my time in Belgium. And finally, thanks to Keith and Maria, Vincent and others for helping keep me sane and the plentiful advice and discussion!

Abbreviations

A	Adenine
AI	Artificial insemination
APOB	Apolipoprotein B
BBB	Belgian blue breed
BP	Base pairs
C	Cytosine
CPG	Cytosine – phosphate - guanine
CNV	Copy number variant
DHF	Dutch Holstein Friesian
DNA	Deoxyribonucleic acid
dnm	De novo mutation
DSB	Double stranded break
EF	Embryo flush
ENV	Envelope protein
ERV	Endogenous retro virus
ERVK	Endogenous retro virus K
G	Guanine
GAG	Group specific antigen
GB	Gigabases
H_2O_2	Hydrogen peroxide
IBD	Identity by descent
ICM	Inner cell mass
IVF	In vitro fertilisation
IVM	In vitro maturation
INDEL	Insertion or deletion
KB	Kilobases
KG	Kilogram
LINE	Long interspersed nuclear element
LOF	Loss of function
LTR	Long terminal repeat
MAF	Minor allele frequency
MB	Megabases
MMOL	Millimolar

MOET	Multiple ovulation and embryo transfer
N_E	Effective population size
NGS	Next generation sequencing
NZDC	New Zealand dairy cattle
O_2^-	Superoxide
ORF	Open reading frame
PCR	Polymerase chain reaction
PGC	Primordial germ cell
POL	Polymerase
PRL	Prolactin gene
PRLR	Prolactin receptor
PRT	Protease
RNA	Ribonucleic acid
RNS	Radical nitrogen species
ROS	Radical oxygen species
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism
SV	Structural variant
T	Thymine
WGS	Whole genome sequence
U	Uracil
UTR	Untranslated region
UV	Ultra violet light

ABSTRACT - RÉSUMÉ 2
PREAMBLE 9
INTRODUCTION 13
1 INTRODUCTION TO GERMLINE DE NOVO MUTATIONS 15
1.1 What is de novo mutation 15
1.2 How do we estimate the rate of de novo mutation 15
2 CAUSES OF MUTATION 22
2.1 Exogenous damage 22
2.2 Endogenous damage 22
2.3 Response to damage 23
2.4 DNA replication 25
2.5 Mobile Elements 26
3 INTERACTIONS BETWEEN GAMETOGENESIS AND MUTATION 28
3.1 Mammalian developmental biology 28
3.2 Observations of mosaicism 32
4 SIGNATURES, PATTERNS AND PROPERTIES OF DE NOVO MUTATION 36
4.1 Methyl-Cytosine, C>T mutations 36
4.2 Trinucleotide patterns 36
4.3 Clustering of DNMs 38
4.4 Age and sex effects 38
5 EVOLUTION OF THE MUTATION RATE 40
5.1 Fidelity Hypothesis 40
5.2 Drift-Barrier Hypothesis 40
5.3 Evidence for evolution of the mutation rate 41
5.4 Phylogenetic versus pedigree estimates of mutation rate 42
OBJECTIVES 45
EXPERIMENTAL SECTION 49
Frequency of mosaicism points towards mutation-prone early cleavage cell divisions. 51
Evaluating the inter-individual variation in the rate and spectrum of germ-line de novo mutation and its causes in cattle. 99
A polymorphic *ERV* element that is mobilized in the germline of specific individuals causes abetalipoproteinemia and hypolipidemia in cattle by disrupting the *APOB* gene. 133
A stop-gain in the *LAMININ, ALPHA 3* gene causes recessive junctional epidermolysis bullosa in Belgian Blue cattle. 175
Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle 185
NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock 213

DISCUSSION - PERSPECTIVES 235
BIBLIOGRAPHY 253
GLOSSARY OF TERMS 265

Abstract - Résumé

Abstract

De novo mutation (dnm) in the germline is a fundamental biological process that is the source of all population genetic variation. In this thesis, we have exploited the unique population structure of cattle to select pedigrees of three (parents, proband and >1 grand-offspring) and four (plus grandparents) generations to identify and characterise germline dnm. We have also looked at the impact of recent dnms on the genetic load of the cattle population, using forward and reverse genetics to identify the causative mutations responsible for major defects in cattle populations including embryonic lethality.

In the first study, we utilised five pedigrees to identify SNP and small insertion-deletion (INDEL) dnms, assign them to a parent of origin and determine the stage of development at which they occurred. We determined the dnm rate in cattle to be ~$1.2x10^{-8}$ per base pair per generation, with 2.5 paternal dnms for each maternal dnm. We showed that 30% and 50% of the dnms in sperm and eggs respectively are mosaic in the parental DNA, occurring early in embryonic development. By simulation we show that this is incompatible with a constant mutation rate through gametogenesis and best fits a 20x higher mutation rate for the first four cell divisions of the fertilised egg. This paper is currently in review with a preprint available on BioRxiv (Harland et al. 2017a).

In a second study, we looked at the rate of dnm in the wider population and for the presence of inter-individual variation in the rate. We utilised the complete Damona dataset of 131 three generation pedigrees identifying ~7,500 dnms, confirming the previously observed degree of mosaicism and our dnm rate of $1.2x10^{-8}$ per bp per generation in a wider sample. We observe several outliers in the population, with 5-17x the average number of dnms occurring during embryo development, along with distinct mutational signatures. For one outlier pedigree, we identified two candidate causative mutations that are in the process of characterising. In addition, we detect a significant environmental effect from the use of reproductive technologies, such as in vitro fertilisation and maturation, on the rate of dnm during early embryo development. (Harland et al. 2017b, in preparation).

In the third study, we turned our attention to alternative forms of dnm. Utilising the full dataset, we identified five cases of de novo transposition of an endogenous retrovirus family K (ERVK) element, with three of the five events occurring within the germ-line of a single individual, and two of the three in the same gamete. This indicates that the ERVK family is presently active in the bovine genome with an average de novo transposition rate of ~1 event per 50 gametes, but there is strong evidence for considerable inter-individual variation. We identified ~1,600 polymorphic ERVs resulting from the activity of these elements in the cattle population. One recent de novo transposition of an ERVK element in the *APOB* gene is responsible for the lethal monogenic Cholesterol Deficiency disorder in cattle. This transposition has reached a minor allele frequency (MAF) of 2.8% in the European population (Harland et al. 2017c, in preperation).

In a fourth study, we describe a dominant deleterious missense dnm in the Prolactin gene (*PRL*) that caused heat stress, abnormal hair growth and a failure to milk in ~2,000 affected offspring of a single

bull. This set of phenotypes is opposite to those observed in 'slick' cattle, which show increased heat tolerance, short hair and potentially increased milk production. We demonstrate that the slick phenotype in Senepol cattle is due to a dominant frameshift mutation in the Prolactin receptor gene, and is an example of a beneficial variant that has undergone positive selection within a population (Littlejohn et al. 2014). The large number of affected cattle for the PRL mutation demonstrates how a rare dnm can rapidly increase in frequency within a population when it is present in an elite sire used for artificial insemination.

For examples of other recent dnms, we turned to the Belgian Blue population and investigated recessive junctional epidermolysis bullosa. This lead to the identification of a breed specific, premature stop-gain mutation in the *laminin, alpha 3* gene with a MAF of 1%, thus allowing the development of a direct genetic test for the disorder (Sartelet & Harland et al. 2015).

We then used whole genome sequences (WGS) obtained on Illumina HiSeq's to search for embryonic lethal (EL) mutations segregating in the New Zealand dairy and Belgian Blue beef cattle populations using a reverse genetic approach. We genotyped >40,000 cattle for 296 loss of function (LOF) and 3,483 potentially deleterious missense variants that were breed specific, and identified variants with a significant deficiency in homozygous mutant animals. Nine of these variants were confirmed to be EL by genotyping 200 carrier x carrier trios and demonstrating the absence of homozygous offspring. The MAF of these variants was between 1.2% and 6.6%. We estimate that 15% of the tested LOF and 6% of the missense events are EL reducing the fertility of dairy cattle (Charlier et al. 2016).

Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mullaart E, Coppieters W, Georges M. 2017a Frequency of mosaicism points towards mutation-prone early cleavage cell divisions. *bioRxiv* 079863.

Harland C, Durkin K, Artesi M, Karim L, Cambisano N, Deckers M, Tamma N, Mullaart E, Coppieters W, Georges M, Charlier C. 2017b Evidence from the bovine of major difference between individuals in the rate of de novo single nucleotide mutation. *In preparation*

Harland C, Karim L, Durkin K, Artesi M, Sartelet A, Knapp E, Tamma N, Mullaart E, Coppieters W, Georges M & Charlier C. 2017c A polymorphic element that is mobilized in the germline of specific individuals causes abetalipoproteinemia and hypolipidemia in cattle by disrupting the *APOB* gene. *In preparation*

Littlejohn MD, Henty KM, Tiplady K, Johnson T, Harland C, Lopdell T, Sherlock RG, Li W, Lukefahr SD, Shanks BC, et al. 2014. Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. *Nature Communications* **5**: 5861.

Sartelet A, Harland C, Tamma N, Karim L, Bayrou C, Li W, Ahariz N, Coppieters W, Georges M, Charlier C. 2015. A stop-gain in the laminin, alpha 3 gene causes recessive junctional epidermolysis bullosa in Belgian Blue cattle. *Animal genetics* 46: 566–570.

Charlier C, Li W, Harland C, Littlejohn M, Coppieters W, Creagh F, Davis S, Druet T, Faux P, Guillaume F. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome Research* **26**: 1333–1341.

Résumé

Les évènements mutationnels de novo au sein de la lignée germinale constituent un processus biologique fondamental, ils sont la source principale de toute nouvelle variation génétique. Au cours de cette thèse, nous avons tiré parti de la structure particulière des populations bovines (larges familles de demi-frères/sœurs) pour identifier puis caractériser les mutations de novo (dnms) dans la lignée germinale. Nous avons également étudié l'impact phénotypique de certaines de ces récentes mutations en utilisant à la fois des approches de type clonage positionnel et de génétique dite réverse pour identifier les mutations causales d'une série de caractères d'intérêt agronomique chez le bovin, allant de la fertilité au type de pelage.

Dans une première étude pilote, nous avons exploité les données de séquence 'génome entier' (WGS) de cinq pedigrees multi-générationnels pour identifier les dnms (de type SNP ou petites insertions/délétions), leur attribuer une origine parentale et déterminer le stade de développement auquel elles se sont produites. Nous avons ainsi estimé un taux moyen de mutations de novo chez le bovin à $\sim 1.2X10^{-8}$, par paire de bases et par génération, avec un biais d'origine en faveur du père de 2,5 fois. Nous avons démontré que respectivement 30% et 50% des dnms présentes dans le sperme ou les oocytes sont mosaïques dans l'ADN parental, elles sont donc apparues lors du développement embryonnaire précoce du parent. Des simulations ont montré qu'un taux de dnms constant au cours du développement ne permettait pas d'expliquer ces observations, et qu'une augmentation du taux de mutations de ~20X lors des quatre premières divisions cellulaires de l'œuf fécondé était requise. Ce manuscrit est en cours de révision et une version initiale est disponible dans BioRxiv (Harland et al., 2017a).

La seconde étude s'est attachée à renforcer et valider ces conclusions préliminaires à l'aide d'un data set étendu, ainsi qu'à quantifier d'éventuelles variations interindividuelles du taux mutationnel. Pour ce faire, un large pedigree (*Damona*) composé de 131 familles a été analysé. Un total de ~7.500 dnms ont été identifiées, caractérisées et classifiées, ce qui a confirmé le degré élevé de mosaïcisme initialement observé, ainsi que le taux moyen de mutations de novo chez le bovin. Une poignée d'individus extrêmes arborant un nombre significativement plus élevé de mutations apparues durant leur développement embryonnaire précoce ont été repérés. Chacun présente une signature mutationnelle unique. Au moins deux gènes et variations candidates sont actuellement suivis pour tenter d'expliquer ces taux mutationnels extrêmes. Enfin, la mise en évidence d'un effet significatif des technologies reproductives utilisées démontre qu'elles influencent le taux mutationnel lors du développement précoce de l'embryon qui y est soumis (Harland et al., 2017b).

La troisième étude a visé d'autres catégories d'évènements de novo, à savoir la mobilisation des éléments transposables du génome et, en particulier, les éléments rétroviraux endogènes (ERV). Exploitant le même pedigree, cinq évènements de novo de transposition d'ERV, tous issus de la famille ERVK, ont été identifiés. Remarquablement, trois de ces évènements se sont produits dans la lignée

germinale d'un même taureau et deux été transmis par le même gamète. Cela indique que la famille ERVK est encore active dans le génome bovin et on peut ainsi estimer le taux de transposition de novo à ~1 événement par 50 gamètes, avec toutefois une grande variation interindividuelle. En outre, nous avons établi un catalogue de ~1.600 ERVs polymorphes qui ségrégent dans les populations bovines étudiées. Ils résultent de l'activité récente de ces éléments. L'un de ces événements a touché la partie codante du gène *APOB*, il est responsable d'une maladie récessive létale, caractérisée pas une déficience en cholestérol sanguin. Cette mutation est présente en race laitière Holstein à une fréquence de 2,8% et est maintenant activement contre-sélectionnée (Harland et al., 2017c).

La quatrième étude décrit deux mutations de novo dominantes à effet phénotypique opposé. La première est une mutation non-synonyme dans le gène de la prolactine *(PRL)*. Cette mutation cause une diminution de la résistance à la chaleur, l'apparition d'un pelage hirsute et a un impact négatif sévère sur la production laitière. Les ~2.000 descendants atteints sont issus d'un seul taureau, mosaïque germinale pour la mutation. Cela illustre comment une mutation délétère peut très rapidement émerger lorsqu'elle se produit chez un taureau d'élite largement disséminé grâce à l'insémination artificielle. La seconde mutation introduit un décalage de la phase de lecture dans le gène du récepteur à la prolactine *(PRLR)*, elle est responsable d'une collection de phénotypes opposés aux précédents (pelage ras, résistance à la chaleur,…). C'est ici un exemple de mutation à effet bénéfique qui a été sélectionnée et fixée dans des races bovines tropicales (Littlejohn et al., 2014).

D'autres occurrences de mutations de novo délétères ont été décrites en race Blanc-Bleu belge (BBB), la plus récente étant une mutation qui introduit un codon stop prématuré dans le gène de la laminine alpha 3 *(LAMA3)*, elle y est responsable d'une épidermolyse jonctionnelle bulleuse récessive létale. Les porteurs sont maintenant identifiés en routine grâce au test génétique développé (Sartelet, Harland et al., 2015).

Enfin, des données de WGS, en populations laitière Néozélandaise et BBB, ont été exploitées dans des études dites réverses (allant du génome et de la mutation au phénotype) pour identifier et valider neuf mutations, létales pour l'embryon homozygote, et ayant des fréquences de 1,2 à 6,6%. Nous avons démontré que, collectivement, ces mutations avaient un impact négatif non-négligeable sur la fertilité de ces races. Cette information est mise à profit dans des programmes de conseil d'accouplements afin d'éviter les croisements à risque (Charlier et al., 2016).

Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mullaart E, Coppieters W, Georges M. 2017a Frequency of mosaicism points towards mutation-prone early cleavage cell divisions. *bioRxiv* 079863.

Harland C, Durkin K, Artesi M, Karim L, Cambisano N, Deckers M, Tamma N, Mullaart E, Coppieters W, Georges M, Charlier C. 2017b Evidence from the bovine of major difference between individuals in the rate of de novo single nucleotide mutation. *In preparation*

Harland C, Karim L, Durkin K, Artesi M, Sartelet A, Knapp E, Tamma N, Mullaart E, Coppieters W, Georges M & Charlier C. 2017c A polymorphic element that is mobilized in the germline of specific individuals causes abetalipoproteinemia and hypolipidemia in cattle by disrupting the *APOB* gene. *In preparation*

Littlejohn MD, Henty KM, Tiplady K, Johnson T, Harland C, Lopdell T, Sherlock RG, Li W, Lukefahr SD, Shanks BC, et al. 2014. Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. *Nature Communications* **5**: 5861.

Sartelet A, Harland C, Tamma N, Karim L, Bayrou C, Li W, Ahariz N, Coppieters W, Georges M, Charlier C. 2015. A stop-gain in the laminin, alpha 3 gene causes recessive junctional epidermolysis bullosa in Belgian Blue cattle. *Animal genetics* 46: 566–570.

Charlier C, Li W, Harland C, Littlejohn M, Coppieters W, Creagh F, Davis S, Druet T, Faux P, Guillaume F. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome Research* **26**: 1333–1341.

Preamble

De novo mutation in the germline (as opposed to the soma) is a fundamental biological process, which is responsible for the generation of most new genetic variation in the population. This variation, which is usually neutral, sometimes deleterious, and rarely beneficial, provides the substrate for natural or artificial selection. Thus, it is critical to our understanding of evolution and the natural history of species. Numerous methods have been developed to investigate the rate and properties of germline dnms, with the first methods preceding the identification of DNA as the vehicle of inheritance. The initial method utilised the existence of spontaneous occurrences of extreme phenotypes. In 1935, Haldane used spontaneous cases of haemophilia to estimate the rate of mutation for the haemophilia locus (Haldane 1935), and in later work determined that the rate of mutation in the male germ-line was up to 10x greater than in the female germ-line (Haldane 1947).

With the development of Sanger sequencing and the ability to sequence whole genomes, phylogenetic methods were utilised to estimate the long term dnm rate of species. Using an outgroup and at least two species to identify ancestral alleles at neutral shared loci, the mutations present in the two species could be identified and using paleontologically estimates of their divergence time in conjunction with the species generation length, the average long term dnm rate could be determined. This approach was utilised to estimate that the human mutation rate was $2.5x10^{-8}$ per bp per generation since the divergence from chimpanzees (Nachman and Crowell 2000).

With the advent of next generation sequencing it became possible to directly identify dnms in trios, by a variant's presence in the child but absence in the parents. Such methods have now been applied to a variety of species and have estimated the current rate of mutation in humans to be $1.2x10^{-8}$ per bp per generation (Kong et al. 2012), less than half that of the estimated phylogenetic rate. With the ability to directly and accurately detect dnms in a wide variety of species, it has now become possible to investigate other aspects of dnm aside from its rate. One area of interest is how the process of dnm interacts with differing stages of an organism's development and gametogenesis. After fertilisation, a zygote undergoes a period of accelerated cell division followed by several bottlenecks: the formation of the inner cell mass, epiblast, embryonic epiblast and primordial germ-cells. The primordial germ-cells then form either oocytes or spermatogonia depending on the gender. Oocytes enter stasis at birth and decline in number with only small numbers activating on a regular basis post puberty, while spermatogonia enter stasis until puberty at which point cell division begins again at a rate of ~23 cycles per year (Gilbert. 2000). Depending on the stage of development at which a dnm occurs, it can be shared by both the soma and germ-line, shared by multiple cells in the germ-line or specific to a single gamete. This has a significant effect on how likely a dnm is to be inherited by multiple offspring. In the case of genetic diseases, this is of critical importance for estimating the probability of multiple affected children being born to a couple. A small number of initial studies have suggested that the mutation rate is not constant through development but changes at the pre-primordial germ cell (PGC), post-PGC and post-puberty stages (Rahbari et al. 2016; Gao et al. 2016). However, experimental limitations have restricted

the ability to fully study this, with the initial studies suggesting that a minimum of four percent of all dnms occur in the first two to five cell divisions (Rahbari et al. 2016).

A dnm is the result of either a mistake during the replication of DNA or the failure to correctly repair damaged DNA. Due to the variety of different types of DNA damage that can occur, there are numerous different DNA repair mechanisms, including direct repair, base and nucleotide excision repair, double strand break repair and crosslink repair. Different mechanisms tend to leave a specific type of 'mutational signature'. These signatures have allowed the development of tools (Alexandrov et al. 2013a) to determine what mechanisms have contributed to a pool of dnms, and thus what type of DNA damage or repair mechanisms have contributed to the population of dnms. These tools were initially developed to analyse somatic mutations from cancer, due to the large numbers of mutations present and the greater associated power. These studies have identified a number of endogenous signatures of mutation such as those resulting from normal DNA replication, specific enzymes such as polymerase ε or BRCA1 and BRCA2, or exogenous factors such as aristolochic acid or aflatoxin (Alexandrov et al. 2013b; Petljak and Alexandrov 2016).

The enzymes underlying repair and replication of DNA, encoded in the genome, are themselves subject to mutation. Thus, the mutation rate itself will evolve as dnms modify the repair and replication mechanisms subjecting them to natural selection. As mutations are more likely to be deleterious and if the magnitude of the negative effects is greater than of the beneficial ones, natural selection should favour the reduction of the mutation rate until it reaches zero. However, all known organisms show substantial rates of dnm. There are two primary hypotheses as to why natural selection has not been able to drive the mutation rate to zero. The first is the fidelity cost hypothesis, under which reducing the rate of mutation is eventually limited by the physical or physiological cost of increasing DNA fidelity (Kimura 1967). Natural selection would be unable to drive the mutation rate lower once the cost of doing so is greater than the benefit. The second hypothesis is the drift-barrier hypothesis, under which the rate of dnm is determined by the power of random genetic drift and thus the effective population size of a species (Lynch 2011; Sung et al. 2012). In this model, the rate is set once the fitness gain from improving the mutation rate is less than the power of genetic drift. As the rate and properties of dnm are themselves subject to mutation and evolution, a population is likely to carry variation that modifies the mutation rate. There is some evidence of this with recent work by Harris supporting the transient existence of mutator alleles within human populations (Harris 2015; Harris and Pritchard 2017), while work by Seoighe and Scally has identified candidate mutator loci in some human populations (Seoighe and Scally 2017). Aside from this, inter-individual variation in mutation rate in mammals has been subject to little investigation due to difficulties in assigning all dnms to their germ-line of origin and differentiating between germ-line and somatic mutations. We aim to address this limitation and the timing of dnm with our study, as the addition of our third generation (grand-offspring) resolves both these issues.

Introduction

1 Introduction to germline de novo mutations

1.1 What is de novo mutation

De novo or spontaneous germline mutation is the ultimate source of all genetic variation within a population and serves as the substrate of natural and artificial selection. Since the discovery of the theory of evolution in the 19th century by Darwin and Wallace, the importance of genetic variation within a population has been recognised. Consequently, the source of existing and new variation within populations has been of considerable interest. Once evolution was combined with the laws of inheritance to form the modern synthesis, considerable work has been focused on estimating the rates of dnm and understanding its process. Initially investigated at the phenotypic level, the research has proceeded onto the molecular level after the discovery of DNA as the molecular basis of inheritance (Avery et al. 1944; Hershey and Chase 1952). From an evolutionary point of view, we can describe dnm to be genetic variation that is present in an individual and is passed onto the next generation, but which was absent from the parental gametes. From a molecular and cell biology point of view, we can describe dnm as changes to the DNA in the germ-cells of an individual. These variations or changes can take many forms: simple single nucleotide polymorphisms (SNPs), insertions or deletions of bases (INDELs), to larger or more complex rearrangements of the genome structure (structural variants, SVs) by duplications (or copy number variants, CNVs), large deletions, translocations, inversions, and the insertion of mobile genetic elements such as retrotransposons.

Thus, dnm (along with recombination) is a fundamental biological process underpinning evolution and critical to our understanding of it. De novo mutation also has a substantial medical impact. As dnms in genes are likely to be harmful, they cause rare genetic disorders by occurring either in the affected individual or its parents. They are also thought to play a critical role in more common disorders such as autism (Michaelson et al. 2012; Iossifov et al. 2014) and schizophrenia (Awadalla et al. 2010; Girard et al. 2011; Julie et al. 2011), where there is a significant excess of non-synonymous dnms observed in synaptic genes (Fromer et al. 2014). Finally, they are also important for their role in agriculture and horticulture in generating new variation, which can then be targeted by artificial selection.

1.2 How do we estimate the rate of de novo mutation

A variety of methods have been developed to study the rate and properties of dnm, with the preferred method depending on the level of technology that was available at the time. The earliest studies before the discovery of DNA were purely phenotype-based. With the discovery of DNA and the ability to sequence it, research focused on mutation rates in viruses (Drake 1993), and mutation accumulation

experiments to determine the rates of dnm at specific loci (Drake et al. 1998). Once DNA sequences became widely available, phylogenetic methods became feasible and finally, as the cost of DNA sequencing declined, whole genome and exome based methods utilising pedigrees to detect dnms has become common. It is worth noting that while the initial phenotype-based methods may have estimated the mutation rate based on any class of dnm that perturbed the phenotype, more recent studies have primarily focused on SNPs and small INDELs. This is due to the technical difficulties associated with the identification and genotyping of other classes of dnm such as SVs and large INDELs.

Phenotype-based studies

A classic example of phenotype-based studies was Haldane's investigations of dnm using haemophilia in European populations (Haldane 1935, 1947). In the first paper Haldane estimated the number of haemophiliacs living in London based on reports in medical literature and utilised this estimate to estimate the rate of mutation for haemophilia at ~$2x10^{-5}$ (Haldane 1935). In the 1947 paper Haldane utilised an improved dataset of 63 haemophiliac pedigrees from Denmark (Andreassen 1943) to estimate the effective reproductive fitness of haemophiliacs compared to non-haemophiliacs in those pedigrees. The effective mutation rate of the haemophilia locus is then estimated utilising this effective fitness and the frequency of haemophilia within the Danish population, to be $3.16x10^{-5}$. In addition, by estimating the number of women in the pedigrees who were heterozygous for haemophilia (utilising the coagulation time of their blood), an estimate of the ratio between the male and female mutation rates could be calculated, suggesting that the female mutation rate was likely less than one tenth that of males (Haldane 1947).

Phylogenetic methods

With the advent of the human genome project and the availability of sequences for multiple species, it became possible to estimate the rate of dnm using phylogenetic methods based on the variants present in DNA shared between species (Kondrashov and Crow 1993). By identifying the differences between shared neutral regions of DNA in both species and combining it with an estimate of the divergence time, the ancestral effective population size and the generation time, it is possible to calculate a dnm rate. This approach has several limitations, the first of which is that natural selection will purge even weakly deleterious variants from the population when given sufficient time. This will lead to an underestimate in the number of variants that have accumulated in the shared region over time, as any deleterious variants and those variants in perfect linkage disequilibrium with it will have been removed from the genome. Thus, it is critical to select non-functional regions of the genome, that do not contribute to an organism's evolutionary fitness, to minimise the possible effects of natural selection over large time scales. Secondly the approach is highly reliant on both the accuracy of the palaeontological estimate of the divergence time, the average generation time, and the estimated ancestral population size.

Nachman and Crowell (2000) provide a classic and robust example of this approach. They selected 18 pseudogenes shared by both humans and chimpanzees and identified 199 differences between the 18 shared sequences in the two species. This was then used to estimate a mutation rate of $2.5x10^{-8}$ per bp per generation based on an average generation time of 20 years, an ancestral population size of $1x10^4$, and a time since divergence of 5 million years. However, depending on the exact ancestral population size (10^4 or 10^5), the average generation time (20, or 25 years) and the estimated time since divergence (4.5 - 6 million years), the estimated mutation rate ranged from $1.3x10^{-8}$ to $3.4x10^{-8}$ per bp per generation (Nachman and Crowell 2000).

Direct pedigree estimates

Finally, with the advent of next generation sequencing the cost of whole genome sequencing (WGS) reached a point where it became feasible to sequence the whole genomes of multiple individuals. By sequencing trios consisting of both parents and a child it becomes possible to directly identify dnm by their presence in the child but absence in the parents (Fig. 1A) (Roach et al. 2010; Conrad et al. 2011; Kong et al. 2012; Dal et al. 2014; Keightley et al. 2014, 2015; Venn et al. 2014; Besenbacher et al. 2015; Francioli et al. 2015; Girard et al. 2016; Rahbari et al. 2016; Smeds et al. 2016; Wong et al. 2016; Maretty et al. 2017). A second method based on the sequencing of extended pedigrees, where the time to the most recent common ancestor is known or can be estimated based on the size of the identical by descent (IBD) homozygous regions, has also been developed. In this approach dnms are detected by their presence as heterozygous variants within otherwise homozygous regions (Fig. 1B) that are assumed to be IBD (Campbell et al. 2012; Palamara et al. 2015). The rate of dnm can then be estimated as a function of the number of heterozygous variants present to the total size of the homozygous regions and the time to the most recent common ancestor for that homozygous region (Fig. 1C).

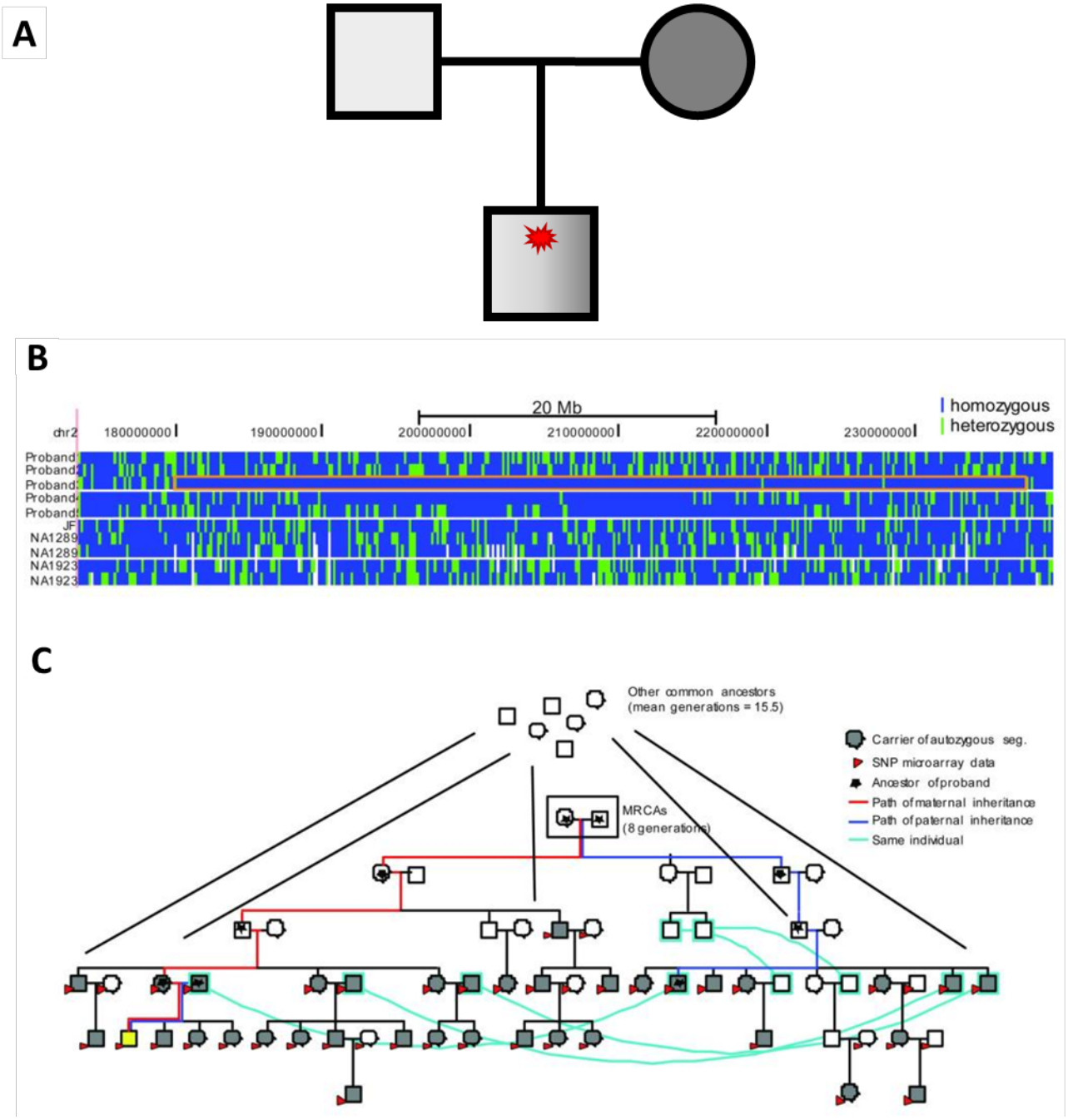


Figure 1 A: Direct dnm detection using a simple trio, the dnm (in red) is identified by its' observation in the genome of the child (proband) and its' absence in the genomes of both the sire and dam. **1B & 1C:** Identification of dnms by observation of heterozygous variants within regions of autozygosity in an extended pedigree (Campbell et al. 2012). B) A 54 Mb autozygous segment on chromosome 2 in Individual 3. Genomic coordinates (hg18) are represented horizontally, and each individual is represented vertically: the five Hutterite individuals, followed by the three European-American individuals, and then the two Yoruba. Each SNV is represented by a vertical bar colored blue if the variant is homozygous and green if it is heterozygous. The autozygous segment in Individual 3 is boxed in orange. C) Determination of the most recent common ancestor (MRCA) for this autozygous segment. The pedigree containing all the haplotype carriers of the autozygous haplotype is shown. Cyan lines connect the same individuals who are represented twice in the pedigree. Individual 3 is shown in yellow. All samples with SNP microarray data are shown with red arrows, and haplotype carriers are shown in gray. These haplotype carriers have two MRCAs (boxed) as well as additional common ancestors further

up the pedigree. The paths from these individuals to the autozygous subject are shown in red for the maternal ancestors and blue for the paternal ancestors; all ancestors of the individual are marked with a star.

There are several complicating factors for estimating the mutation rate from WGS trios. First, next generation sequencing is relatively error prone resulting in large number of false positive variants that need to be removed from the dataset. The number of false positives can be reduced by removing variants present in unrelated individuals, at the cost of losing recurrent mutations. Variant quality filters can be used to remove low confidence variants (with extreme allelic dosages) resulting from sporadic sequencing errors or misaligned sequence, at the cost, however, of increasing the number of false negatives. Second, sequence coverage and presence of collapsed repeats and structural variants in the reference genome make it difficult to estimate the exact proportion of the genome that can reliably be utilised to detect a heterozygous mutation in both the child and its parents. Third, in most studies the pedigrees are restricted to two generations and somatic tissue is sequenced, thus somatic mutations that have occurred outside of the germ-line may be present and mistaken as germ-line dnms. Ségurel et al. (2014) discuss such confounding factors in some detail. Finally, difficulties in accurately detecting and genotyping structural variants and large INDELs due to the use of short read sequencing, primarily limit these studies to SNPs and small (<10bp) INDELs. However, while the method has its limitations, it is relatively simple to perform and can accurately detect true dnms. Careful quantification of the proportion of the genome that has been analysed and of the impact of the chosen filtering strategy, allows for decent estimation of the dnm rate. To date, this approach has been applied to humans (Roach et al. 2010; Conrad et al. 2011; Kong et al. 2012; Dal et al. 2014; Besenbacher et al. 2015, 2016; Francioli et al. 2015; Girard et al. 2016; Rahbari et al. 2016; Wong et al. 2016; Maretty et al. 2017), chimpanzees (Venn et al. 2014), fish (Feng et al. 2017), flycatchers (Smeds et al. 2016), and insects (Drosophila melanogaster), (Keightley et al. 2014) and the postman butterfly (Keightley et al. 2015)). Kong et al (2012) reported one of the earliest large-scale applications of this approach. Sequencing 78 Icelandic trios, they identified 4,993 dnms for an estimated human dnm rate of $1.2x10^{-8}$ per bp per generation. They identified a 3.9:1 ratio when comparing male to female mutations and a paternal age effect of two dnms per year post-puberty. Other human studies of varying size have estimated the human mutation rate to be between $0.97x10^{-8}$ and $1.37x10^{-8}$, depending on the population. The second approach of using distant relatives, which has been utilised by Campbell et. al. (2012) and Palamara et. al. (2015), has also estimated similar rates of $1.2x10^{-8}$ and $1.6x10^{-8}$ per base pair per generation. The concordance between the rates estimated from different populations (though admittedly primarily of European descent), sample sizes and with different methods supports the robustness of the estimate. It suggests that the current human mutation rate is closer to the estimate of $1.2x10^{-8}$ than the $2.5x10^{-8}$ previously estimated from phylogenetic studies and is below the lower bound of phylogenetic estimates of $1.3x10^{-8}$ by

Nachman and Crowell (2000). Venn et al (2014) applied a similar trio-based approach, to a three-generation pedigree of chimpanzees, estimating a mutation rate of ~$1.2x10^{-8}$ with a 5.5:1 male to female ratio and a paternal age effect of 3 mutations per year. The overall mutation rate is similar to humans, but with significantly more mutations occurring in the male germ-line compared to the female. This pedigree based method has also been applied to insects, giving rates of $2.8x10^{-9}$ for *Heliconius melpomene* (postman butterfly) and $2.9x10^{-9}$ for *Drosophila melanogaster* (Keightley et al. 2014, 2015), birds with $4.6x10^{-9}$ for *Ficedula albicollis* (collared flycatcher) (Smeds et al 2016) and fish with $1.7x10^{-9}$ for *Clupea harengus* (Atlantic herring) (Feng et al. 2017). Compared to primates, these rates are between 2.6x and 9x lower per generation. However, comparing mutation rates between species purely on a per generation basis does not take into account many of the differences between species such as, differing generation times (herring ~6 years, chimpanzee 18-20 years, human 20-30 years), the number of cell divisions between fertilisation and the production of a gamete, life-cycle and reproductive processes such as spawning (atlantic herring), mammalian pregnancy (human, chimpanzee) and egg laying (flycatcher, fruitflies & Postman butterfly). If we, instead, use a per year mutation rate then the rates are approximately $0.46x10^{-9}$ for chimpanzees, $0.51x10^{-9}$ for humans, $2.3x10^{-9}$ for flycatchers and $0.28x10^{-9}$ per bp per year for herring. Perhaps a better measure of mutation rate would be a per cell division rate but exact numbers for the number of cell division per generation is hard to come by and can potentially differ greatly between reproductive strategies used by different species.

While the primary focus in recent years has been the de novo mutation rates of SNPs and small INDELs, considerable work has been undertaken in estimating the rates of structural variant formation (SV). However, differences in the definition of a SV between studies and the number of differing SV classes that can be investigated complicates work in this area.

Similar to the approaches utilised for de novo SNPs, we can detect de novo SVs via the use of trios to identify SVs present in the proband but absent in both parents. Depending on the size of the targeted SVs differing technologies can be utilised to detect the events. For the larger de novo SVs in the tens of kilobases or greater high-density microarrays can be utilised to directly identify dnms by changes in the marker intensity of the microarray variants. Istara et al (2010) utilised a microarray based approach on 386 trios identifying nine dnms for an estimated rate of 0.012 de novo SVs (≥ 62kb) per haploid genome per generation (Itsara et al. 2010). More recent work has estimated the SV dnm rate utilising trios and whole genome sequence data, in a genome of the Netherlands study of 231 trios the SV mutation rate for variants > 20bp was estimated to be 0.08 dnms per haploid genome per generation (Kloosterman et al. 2015), a similar rate of 0.098 was observed in a study of 45 trios from an autism spectrum disorder cohort (Brandler et al. 2016). This difference in dnm rate is thought to primarily due to the use of whole genome data providing a wider coverage of the genome, along with the ability to detect smaller events than is possible utilising microarray based methods. When considering only larger events (>500bp or >

100kb) the rates were 0.041 and 0.0077 de novos per haploid genome (Kloosterman et al. 2015), much closer to the 0.012 de novos per haploid genome reported from the microarray study for SVs ≥ 62kb (Itsara et al. 2010). One substantial difference between the microarray and WGS based studies is the detection of paternal to maternal bias in the origin of de novo SVs, Itsara et al. (2010) reported no bias among the dnms detected via microarray, while both WGS based studies reported strong paternal biases of 2.7:1 (Kloosterman et al. 2015), and 2:1 (Brandler et al. 2016). Considering a strong paternal bias has been observed for SNP and INDEL dnms, this may suggest that the WGS based studies provide a more accurate view of the true rate and pattern of SV dnm.

2 Causes of mutation

Mutations are the result of a failure to correctly repair damaged DNA or due to errors in its replication. The rate of mutation is thus, set by the rates and types of DNA damage, from both endogenous and exogenous sources. As well as the fidelity of the DNA repair and replication mechanisms, biological processes that are under the influence of natural selection. Thus, understanding the sources and types of damage, as well as the genes involved in its' repair and replication is of considerable importance. Different factors will cause different types of DNA damage, which in turn require a variety of ways to repair them. We can divide DNA repair mechanisms into several distinct classes, direct repair, base excision repair, nucleotide excision repair, double strand break or recombinational repair and cross-link repair (Sancar et al. 2004), each of which responds to different types of DNA damage. The interaction between the different types of DNA damage or replication error and the different repair mechanisms, gives rise to the different classes of dnm that can occur such as SNPs, INDELs and SVs.

2.1 Exogenous damage

Exogenous damage to DNA results from exposure of DNA to external factors such as electromagnetic (UV, X-ray, Gamma-ray) and particle radiation or mutagenic compounds that either directly modify the chemical structure of DNA or interact with other cellular components to generate free radicals which can modify the DNA. Ultraviolet B radiation can result in adjacent thymine bases reacting to form a pyrimidine dimer (Sinha and Häder 2002). X-rays, α, β, and γ radiation can directly cause double strand breaks (DSBs) or indirectly cause base modification or structural damage by the generation of radical species (Wiseman and Halliwell 1996; Hodgkins et al. 1996), while neutron radiation causes DSBs of DNA (Pang et al. 1998). Exposure to mutagenic compounds such as polycyclic aromatic hydrocarbons (PAH) and nitrosamines from tobacco smoke (Pfeifer et al. 2002) and alkylating agents (Tomita-Mitchell et al. 2000) such as temozolomide (Alexandrov et al. 2013a) can result in base modifications, DNA cross-linking and adduct formation. These can prevent the replication of DNA and must be removed to allow DNA and thus cell replication.

2.2 Endogenous damage

Endogenous damage of DNA results from processes that are native to the cell or due to the chemistry of DNA. This includes free radicals generated as a side product of the mitochondrial electron transport chain or the spontaneous deamination of cytosine and methyl-cytosine.
During the generation of ATP in the mitochondria, via the electron transport chain, several side reactions can occur with O_2, resulting in the formation of superoxide. A 60Kg woman is estimated to produce

between 160-320 mmol of superoxide per day, which makes this a major source of reactive oxygen species (ROS) in a cell. Secondly monoamine oxidase present in the outer mitochondrial membrane catalyses the oxidative deamination of monoamines generating hydrogen peroxide, another ROS. The generation of these radicals is reviewed in detail by Cadenas and Davies (2000). ROS can then cause oxidative damage to both proteins and DNA in the cell, resulting in base modification, base adducts, DNA crosslinking and DNA strand breaks (Cooke et al. 2003).

A second type of endogenous damage results from chemical properties of DNA. The most significant example of this would be the spontaneous deamination of cytosine to uracil. In this spontaneous reaction, a molecule of water reacts with the amine group of cytosine nucleotides in the DNA releasing an ammonia molecule in the process. The resulting uracil misspairs with guanine and will then be targeted by the DNA base excision repair mechanism, involving uracil-DNA glycosylase. This enzyme rotates the uracil base out of the helix and then removes it, leaving an abasic site which acts as a substrate for the base excision repair process. Deamination of methyl-cytosine creates a thymine nucleotide instead of the uracil (Coulondre et al. 1978). The resulting thymine guanine mismatch acts in turn as a substrate for thymine-DNA glycosylase. In a process similar to the one catalysed by uracil-DNA glycosylase, the enzyme removes the thymine nucleotide from the G:T mismatch resulting in an abasic site, which is then processed by base excision repair. Both spontaneous deaminations are exceedingly common with thousands occurring per cell per day. However, the methylated cytosine to thymine reaction is more likely to result in a dnm than the cytosine to uracil reaction, as uracil is not a canonical DNA nucleotide, while thymine is. Thus, if the wrong DNA repair mechanism acts on the G:T mismatch, there is a chance that the thymine may be kept instead of the guanine. Also, in single stranded DNA during repair or replication, the deamination of methyl-cytosine leaves an unpaired thymine base in its position. Without the misspairing between the thymine and the opposing guanine, the thymine will be retained as a valid base during the replication or repair process resulting in a C to T mutation.

2.3 Response to damage

There are effectively two ways to deal with DNA damage: either prevent it before it occurs, or repair the damage afterwards. Both processes can be observed in living organisms, where there are a variety of mechanisms and factors designed to prevent DNA damage. Two examples being UVB absorption by melanin, or the pairing of superoxide reductase and catalase, which work together to convert highly reactive O_2^- to the less reactive H_2O_2, and that in turn to water and oxygen. However, while these may reduce the amount of damage, they will not prevent it all. Thus, the DNA repair and replication mechanisms are most important to maintaining the integrity of the genome and minimising the rate of mutation.

Direct repair results in the reversal of damage to nucleotides resulting in the restoration of the original unmodified base. The most common example is the removal of O^6-methyl groups from O^6-methylguanine to restore the guanine nucleotide, catalysed by methylguanine DNA methyltransferase (Christmann et al. 2011). O^6-methylguanine pairs with thymine which would typically activate the mismatch repair system resulting in the futile excision of the thymine residue and its resynthesis. Thus, methylguanine DNA methyltransferase acts to undo the damage to the guanine, short circuiting the mismatch repair loop and allowing for correct repair of the site. A second example in humans is oxidative methyl transferase which restores 1-methyladenine and 3-methylcytosine to adenine and cytosine, respectively (Ménézo et al. 2010). In addition to these enzymes, others such as photolyase (Essen and Klar 2006) and spore photoproduct lyase (Buis et al. 2006) act to reverse UV damage to bases in non-mammalian species.

Base excision repair works on abasic sites resulting from direct DNA damage or as the product of DNA glycosylases, which have acted on oxidised, alkylated or deaminated bases or base mismatches. The abasic sites then undergo short-patch base excision repair which replaces the single missing nucleotide in a process that involves DNA Polymerase β, APE1 and DNA ligase III-XRCC1 or the alternative mechanism of long-patch repair involving APE1, DNA Polymerases δ and ε, PCNA and FEN1 resulting in the replacement of between 2-10 nucleotides (Sancar et al. 2004).

Nucleotide excision repair removes larger DNA lesions resulting from exposure to radiation or chemical mutagens as well as protein addition to DNA, though the process can also act against all simple single base lesions. A lesion is thought to be recognised by changes in the conformation of the DNA backbone. This results in excision nucleases binding and cutting the affected DNA strand 5'and 3' of the lesion. This is followed by the removal of the oligomer containing the lesion, with repair synthesis filling in the resulting gap. In humans six repair factors composed of 15 polypeptides play key roles in nucleotide excision repair (RPA, XPA, XPC, TFIIH, XPG and XPF.ECC1).

Double strand break repair acts on double strand breaks produced by reactive oxygen species, ionizing radiation, and recombination. There are two primary pathways: homologous recombination and non-homologous end-joining (Wyman and Kanaar 2006). Homologous recombination is initiated by RAD51 in eukaryotes and RecA in prokaryotes leading to strand invasion, branch migration and the formation of a Holliday junction followed by its resolution. Numerous proteins take part in the process including RAD52, RAD54, RAD55, RAD57, BRCA1, BRCA2 and the MUS81.MMS4 heterodimer.
In non-homologous end-joining the Ku70.Ku80 heterodimer binds to both ends of a double strand break then DNA-PKcs and ligase 4-XRCC4 ligate the two ends together (Wyman and Kanaar 2006). This approach is insensitive to the origin of the two ends and can potentially join them even if they originated from different chromosomes resulting in translocations between chromosomes.

Crosslink repair deals with DNA damage that has caused a crosslink between the two strands of the DNA molecule and uses a mixture of recombination and nucleotide excision repair to remove the crosslink. Different repair mechanisms share several proteins and some damage types can be repaired by multiple different mechanisms depending on the stage at which they are detected (Sancar et al. 2004).

2.4 DNA replication

In all organisms, the fidelity of DNA replication is a trade-off between two different modes of replication. High fidelity replication is needed to maintain genetic information between generations and is critical for preventing genetic diseases and disorders such as cancer from occurring. However lower fidelity replication is needed to increase the diversity in a species and is critical for allowing adaption to changing environments and as well as the development of a functional immune system. This balance between two opposing directives gives rise to a substantial collection of polymerases, repair enzymes and replication factors that act to support both opposing goals.

During DNA replication, there are several possible sources of error that can result in mutation. The first and simplest is the insertion of the incorrect nucleotide causing a mismatch. The base pairing via hydrogen bonds is not particularly accurate. In vitro unassisted base pairing results in mismatches in the range of one in ten to one in a few hundred base pairs. Polymerases therefore act to increase the specificity of base pairing via several mechanisms. These include modifying the energy difference between proper and improper pairs, providing an active site geometry that favours proper pairs and stabilising the transition states of correctly paired bases. The combination of these factors greatly reduces but does not eliminate the occurrence of mispairing. In certain cases, mispairing can result from the addition of modified base pairs or tautomers of the base.

Once an incorrect nucleotide has been inserted there are two possible outcomes. Replication can proceed, leaving the mispaired bases to be targeted by the mismatch repair mechanism, or if the polymerase has a 3’ to 5’ exonuclease function the mismatch can be removed. One side effect of a mismatch is a change in the conformation of the DNA strand. This shifts the 3’-OH group of the mismatched base from its optimal position for the addition of the next nucleotide. This misalignment slows the addition of next base sufficiently for the 5’ to 3’ exonuclease activity of some polymerases to remove the mismatched base, allowing another attempt at inserting the correct base in its position. It is thought that in polymerases with 5’ to 3’ exonuclease ability, the exonuclease activity is constantly competing with the faster polymerase activity. With mismatched bases the polymerase reaction is slowed, allowing it to be outcompeted by the usually slower 5’ to 3’ exonuclease reaction, resulting in the removal of the mismatched base. Polymerases can also cause two other types of errors, either failing to insert bases or adding extra bases which results in small deletions or insertions, respectively. In

repetitive DNA, polymerases can slip with regards to the template strand resulting in the addition or removal of multiple copies of the repetitive element.

Damaged DNA can also cause errors during DNA replication via two mechanisms, for some classes of damage the damaged base can successfully pair with an incorrect base allowing the DNA polymerase to continue past the damage. This results in the damaged DNA being paired against the incorrect base, and the consequent activation of a DNA repair mechanism to remove the damaged base (Brown 2002; Kunkel 2004).

Other classes of DNA damage will prevent the polymerase from continuing causing it to stop, resulting in a replication stall. In such cases the original polymerase can be replaced by a trans-lesion polymerase. Trans-lesion polymerases are a specific class of error prone polymerases (REV1, Pol ζ, Pol κ, Pol η, Pol ι) that allow replication of DNA past unrepaired DNA lesions that stall normal DNA replication or can cause a replication fork collapse. A feature of these polymerases is their error prone or error tolerant nature which allows them to synthesize DNA past lesions but at the cost of a higher error rate (Waters et al. 2009).

2.5 Mobile Elements

An additional source of mutation aside from DNA damage and replication errors are mobile genetic elements, which are capable of replication and movement within the genome. Major examples of these are the retrotransposons (Endogenous Retro Viruses (ERVs), Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Elements (SINEs)), which, combined, make up approximately 46% of the bovine (Adelson et al. 2009) and 42% of the human genome (Lander et al. 2001). Endogenous retroviruses consist of two long terminal repeats flanking a group of viral genes: GAG (group specific antigen), PRT (protease), POL (polymerase) and ENV (envelope protein), with the *POL* gene providing reverse transcriptase and integrase functions. These elements are thought to have evolved from retroviruses that have colonised the germ-line of a species and then lost their ability to reassemble the full viral structure (Boeke and Stoye 1997). LINEs consist of 5' and 3' UTR with two open reading frames of which ORF2 codes for an endonuclease and reverse transcriptase function. SINEs such as the Alu element consist of two similar monomers linked by an adenosine rich sequence, the left monomer contains conserved RNA Polymerase III promoter while the right ends in a poly (A) tract (Richardson et al. 2015). These elements replicate via transcription, with the resulting RNA copy being reverse transcribed and inserted randomly into the genome by an integrase which cuts the DNA generating a small overhang of approximately six base pairs. The mobile element is then inserted and the DNA repaired resulting in the creation of a micro-duplication from the overhanging bases and a copy of the retrotransposon (Coffin et al. 1997). For LINEs and ERVs the required enzymes are encoded by the element, while SINEs lack the required genes and instead rely on the LINE enzymes (Volkman and

Stetson 2014). Aside from the insertion of such a large fragment of DNA, the presence of RNA polymerase binding sites, open reading frames and poly-A signals within the element can cause interference with nearby genes (Goodier and Kazazian 2008). As a consequence retrotransposons are carefully controlled and restricted via a wide variety of mechanisms such as APOBEC induced mutagenesis, RNA silencing, DNA methylation and numerous other mechanisms (Goodier 2016).

3 Interactions between gametogenesis and mutation

One important aspect of biology that affects dnm is an organism's development and the process of gametogenesis. While the development of an organism is exceedingly complex, when considering germ-line dnm, we can restrict our focus to the single lineage of cells that give rise to the germ-line.

3.1 Mammalian developmental biology

In mammals, the life-cycle of an individual begins with the fertilisation of the haploid egg by a haploid sperm cell (Fig. 2). At this point the genome of the new zygote is formed, carrying the dnms that it has inherited from the parents (Gilbert. 2000). Any dnm that is inherited will be present in 50% of the DNA of the individual, no matter at what point they occurred during the parent's development and irrespective of the degree of mosaicism in the parental DNA. At this point the zygote then undergoes rapid cell divisions. During the first cell division, any unrepaired DNA damage inherited from the sperm or the egg, or that has occurred since fertilisation has the potential to generate a new dnm. If dnms are formed at this stage they will also be present in 50% of the DNA and will be indistinguishable from the dnms that were inherited from the sperm or egg, even though they have never been present in the parental DNA pre-fertilisation. From the second cell division onwards, any new dnms that occur in the embryo will not be present in 50% of the DNA, as they will be restricted to the descendants of the initial cell in which they occurred. If a dnm occurred at the two-cell stage it would be present in 25% of the total DNA of the embryo. If a dnm occurred at the four-cell stage it would be present in 12.5% of the total DNA. With every round of cell division, assuming synchronous cell division, the percentage of the DNA a dnm would be present in halves. In the case of asynchronous cell division where one cell line replicates at a greater rate, a new dnm will still be present in less than 50% of the DNA. But its exact proportion would be a function of which cell lineage it was present in (fast or slow), along with the number of cells currently present from both the fast and slow replicating lineages at the time of its formation. At the early stages of development, the new zygote genome is inactive. As such, the cells are entirely dependent on the enzymes, mRNAs and other factors that were provided by the oocyte (and thus are derived from the maternal genome) for all functions. This remains the case until the maternal zygote transition, when the zygotic genome becomes active and the maternally inherited RNA and proteins are degraded and replaced by newly synthesised gene products. The exact timing of the maternal to zygote transition varies between species. In mice it is at the 2 cell stage (Moore 1975; Bensaude et al. 1983), humans the 4 - 8 cell stage (Vassena et al. 2011; Braude et al. 1988), and cattle and sheep the 8 - 16 cell stage (De Sousa et al. 1998; Crosby et al. 1988). Due to this reliance on maternally inherited factors for DNA replication and repair, it is possible that these initial stages of development will be susceptible to a maternal effect. If the maternal genome carries variants that effect cell division, DNA repair or replication, these maternal products may directly affect the initial development of the zygote, even if the

alleles inherited by the zygote lack the variants responsible. This could potentially allow for a difference between the patterns of dnms that occur before the maternal zygote transition and those that occur afterwards. After the maternal zygote transition, cell replication continues with compaction of the embryo occurring at the 8 cell (mouse), 16 cell (human) or 32 cell stage (cattle) resulting in the formation of the morula. From the morula, the first major differentiation event in the embryo occurs with the formation of the blastocyst, the outer layers of the morula forming the trophoblast (which contributes to the placenta), while the inner cells form the inner cell mass (ICM) (Fig. 2). By the 64-cell stage in mice, the inner cell mass consists of ~20% of the cells (Gilbert. 2000). As the number of cells continues to grow the inner cell mass moves to one side of the blastocyst. The inner cell mass then undergoes a second differentiation step with the formation of the hypoblast and epiblast. This in turn is followed by a third differentiation step with the epiblast forming the amniotic ectoderm and embryonic epiblast. These three differentiation steps act as bottlenecks for the embryo cell lineages, with only some cells contributing to the final embryo. These bottlenecks can also affect the percentage of the DNA an embryonic dnm appears to contribute to. With the formation of the embryonic epiblast, the initial series of bottlenecks that affect all embryonic tissues is complete and individual cell lineages start to differentiate with the formation of the endoderm, ectoderm and mesoderm.

If a mutation occurs after the first cell division but before the differentiation of the primordial germ cells (PGCs) then it has a possibility to be gonosomal mosaic, being present in both the soma and germ-line at less than 50% of the total DNA. The exact percentage of the DNA it will be present in depends on the number of embryonic cells that are present at the time, the rate of division of the cell in which the mutation occurred (if asynchronous division occurs) and which cells are selected by the bottlenecks of ICM, epiblast and embryonic epiblast formation. In mice, as the formation of the ICM is thought to involve 20% of the total cells (on average 13 out of 64), there is considerable opportunity for some mutations to be present in the resulting ICM at higher or lower levels than they were in the complete zygote. Both the formation of the epiblast and embryonic epiblast offer similar opportunities. However, due to the increased number of cells existing at these later stages, the probability of significantly modifying the percentage of DNA a mutation is present in, decreases substantially. As the number of cells increases and the various cell lineages begin to form, the probability of a new mutation to be gonosomal rather than purely somatic decreases substantially. Also, the later a mutation occurs, the less likely it is that we will be able to detect that it is gonosomal mosaic.

For the germ-line the next critical step in its development is the formation of the PGCs. Between 3-40 (human, mouse) cells in the epiblast are induced to become PGCs by expressing Blimp1 and Stella (McLaren and Lawson 2005; Zheng et al. 2005). In mice, these cells do not initially divide until days 10-11 post fertilisation, when they start to migrate to the genital ridge and enter the primitive gonads. During the course of the migration, cell division resumes for the PGCs and, over the course of the next five months in humans, the number of PGCs increases to approximately three million in males and ten million in females (Mamsen et al. 2011; Gilbert. 2000).

Any dnm that occurs in PGCs or their descendants will be restricted purely to the germ-line, as there is no evidence of PGCs being converted back to form somatic tissue or vice versa. The formation of the PGCs offers one final bottleneck allowing any mutation that occurred before it to potentially increase its frequency within the germ-line. In humans if only four founding PGC are responsible for the complete germ-line then this effect can be substantial with any pre-existing mutation in an induced cell now existing in at least 25% of the initial PGCs, or 12.5% of the DNA of the germ-line. In species with greater numbers of founding PGCs this effect will be reduced. Once the PGCs have formed, the only new mutations that will be included in the germ-line are those that occur in the PGCs or their resulting gametes. All mutations that occur in the germ-line are technically germ-line mosaic mutations with three separate haplotypes present (A_{mut}, $A_{wildtype}$, B). Similar to the early stages of embryo development, any dnm that occurs in the initial PGCs will be at a relatively high allelic dosage. But the allelic dosage of the subsequent dnms that occur once the PGCs start to divide will rapidly decrease, halving each time the number of PGCs double. A consequence of this is, that unless the dnms occur during the initial selection of the PGCs or the first few divisions we are unlikely to detect them as germ-line mosaic. We will be unable to detect them as mosaic for two reasons, first the percentage of the DNA they are present in will be below our detection threshold for normal sequencing, and secondly the number of gametes they end up in is so low that we are only ever likely to see the dnm in one child, unless we sequence hundreds or thousands of children. As such the dnms that occur in an individual can be classified into five different groups. The first group are the detectable gonosomal mosaic dnms that occurred early enough in development to be present in both the germ-line and the soma and can be easily detected via sequencing. The second is undetectable gonosomal mosaic dnms that occurred before the formation of PGCs and are shared between the germ-line and soma. However, in the soma they are at very low levels or restricted to specific tissues and thus exceedingly difficult to detect. The third is detectable germ-line mosaic dnms that occurred during or immediately after the formation of the PGCs and are thus restricted to the germ-line but present in a sufficient percentage of the DNA that normal sequencing of gametic DNA will detect them, or that multiple offspring will have inherited them. The fourth is undetectable germ-line mosaic dnms that occurred in PGCs or primary spermatogonia at a stage where the percentage of the DNA they are present in is below the levels that can be easily detected by sequencing or would require thousands or millions of offspring to be sequenced before being detected in multiple offspring. The fifth and final class is germ-line non-mosaic dnms that occur in gametes or the final stages of their formation and are thus unique to a specific gamete and can only be inherited by one offspring.

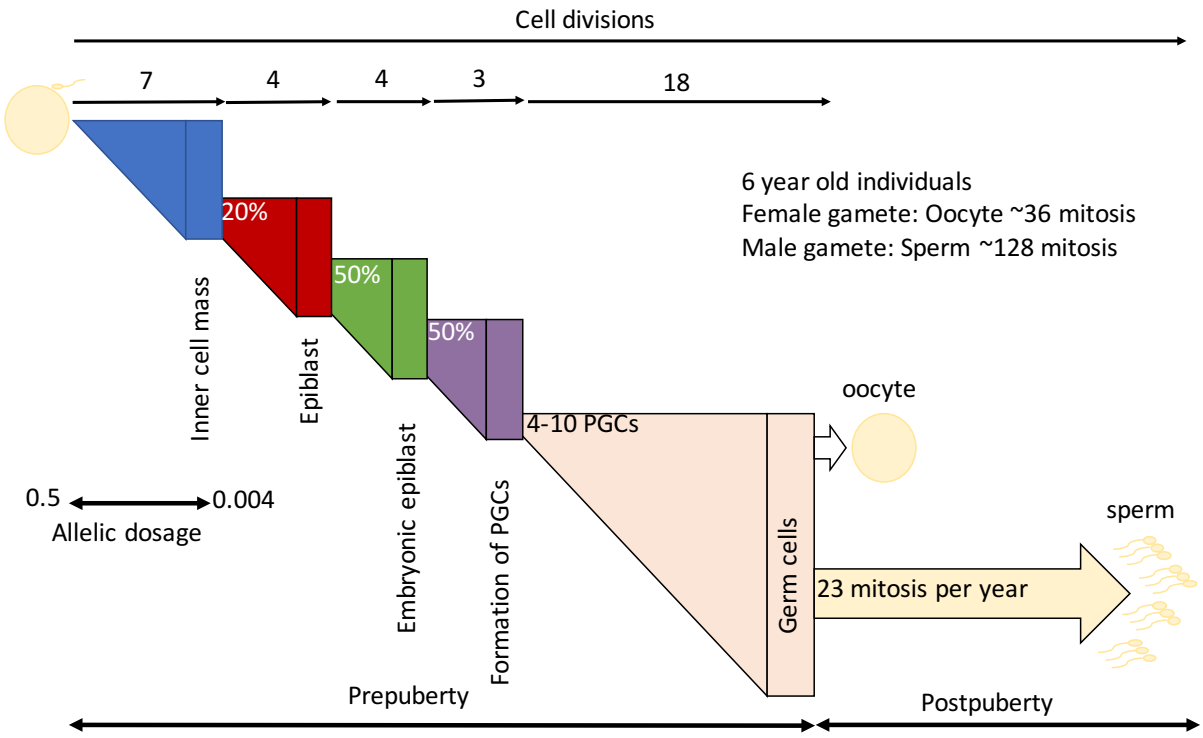


Figure 2: Schematic representation of possible bovine gametogenesis assuming synchronous cell division. After fertilisation of the oocyte the fertilised egg begins to rapidly divide for approximately seven cell divisions (blue), during this time any dnms that occur will have an allelic dosage of between 0.5 and 0.004 depending on the round of division in which they occur. After reaching a critical mass of cells, approximately 20% of the cells (in mice) form the inner cell mass (red) while the rest form extra-embryonic tissues. The inner cell mass undergoes approximately 4 additional divisions before a second division occurs forming the epiblast (green) from a fraction of the cells. The epiblast in turn undergoes approximately four rounds of replication before a fraction of the cells form the embryonic epiblast (purple). With the formation of the embryonic epiblast a small number of PGCs (~4 in humans, 40 in mice) are induced over a few additional cycles of division. Once the PGCs have formed they migrate to the primitive gonads and begin to undergo division once more (salmon). After approximately 18 additional rounds of division sufficient PGCs are available and they enter the initial phases of oogenesis or spermatogenesis, before entering a state of stasis before birth. Post puberty the final stages of oogenesis and spermatogenesis resume forming oocytes or spermatids. To maintain a constant supply of sperm the primary spermatogonia undergo approximately 23 (in human) additional rounds of mitotic division per year.

In the female germ-line, from the five-month stage to birth at nine months, the numbers of PGCs decreases to approximately 500,000 cells. The PGCs form primary oocytes, which reach the diplotene stage of the first meiosis and then halt, entering stasis until puberty. The number of primary oocytes continuously decreases until menopause. At puberty the ovarian cycle initiates, and with each cycle a few primary oocytes are activated and prepared for ovulation. With the release of the dictyate state, meiosis I completes resulting in the formation of a secondary oocyte and a polar body. The secondary oocyte in turn initiates meiosis II before pausing at metaphase II. When the oocyte is fertilised meiosis II completes, releasing an additional polar body and the fertilised mature ovum. In the maternal germ-

line there are approximately 30 mitotic cell divisions between the formation of the zygote and its oocyte (Drost and Lee 1995)(15 before the formation of the PGC and 15 after, plus two meiotic divisions).

In the male germ-line PGCs form primary spermatogonia. These enter stasis until puberty, when spermatogenesis starts. With the initiation of spermatogenesis, primary spermatogonia undergo 8 rounds of division and specialisation to generate spermatids. Of these 8 rounds, the first six are mitotic while the last two are meiotic. However, it is not until the fourth cell division, with the formation of the intermediate spermatogonia, that the cell fate is set to form a spermatid. All previous spermatogonia stages are capable of self-renewal, with each cell division being able to either create an additional cell at the same stage or differentiate to form a cell at the next stage of spermatogenesis. In humans, the full process of spermatogenesis to go from a primary spermatogonia to a spermatid takes ~65 days. With 23 cycles occurring per year, the number of cells divisions between the zygote and a gamete is a function of the number of cell divisions before the formation of the primary spermatogonia plus the number of spermatogonia cycles that have occurred post puberty. Thus, the total number of cell divisions for any spermatid can be estimated as 15 (pre-PGC) plus 20-24 (post-PGC) plus 23 (cycles) multiplied by the number years since puberty (23 x (current age – 16 years)) plus the two final meiotic divisions (Scally 2016; Drost and Lee 1995; Crow 2000). Recent work has suggested that the actual number of divisions maybe somewhat more complicated to estimate, due to observations that primary spermatogonia appear to switch between a replicative and non-replicative state. In the replicative state primary spermatogonia undergo the usual 23 cell divisions a year, however they may at some point switch to a non-replicative state under which the cell divisions do not occur. If these non-replicative spermatogonia remain in this state for some time, the average number of replications per year is less than 23. At a later stage, they may then switch back to the replicative state and start to undergo the 23 cell divisions a year. If this is the case then the number of cell divisions between post puberty may be substantially different from the expected 23 times the number of years post puberty (Scally 2016).

3.2 Observations of mosaicism

Having considered the development of mammals and their germ-line, it is clear that DNA damage or errors during DNA replication in the early stages of embryo development and gametogenesis can give rise to mosaic mutations. The question would then be how common are they?

If we look at it from a purely theoretical point of view, we expect an average of 80 dnms per generation (assuming a 30-year generation time) in humans, with 64 paternal mutations inherited from the sperm cell that has undergone approximately 380 cell divisions and 16 maternal mutations inherited from the

oocyte with 32 cell divisions. If we take the simpler case of the oocyte and assume a constant mutation rate we would expect one mutation every two cell divisions thus 7-8 mutations would be expected before the formation of the PGCs, of which two (or 12%) would be detectably gonosomal mosaic (>5% of the somatic and germ-line DNA) in the mother. If we consider the paternal germ-line, with 64 mutations in ~380 cell divisions, a constant mutation rate would give ~0.17 mutations per cell division or two gonosomal mosaic mutations of which zero to one would be detectable. This simplest paternal model has two issues. First it requires that the early embryo development differs between males and females, with a 3x higher mutation rate per cell division in the female germ-line and secondly it is incompatible with the observation by Kong et al (2012) that two dnms occur per year post-puberty during the process of spermatogenesis, for a rate of 0.087 dnm per cell division post puberty. It seems more reasonable to assume that the prepubertal embryonic development of males, and corresponding dnm rate, is similar to that of females. If this is the case then ~16 of the mutations would be expected to occur pre-puberty and ~48 occurring post puberty resulting in 0.14 mutations per cell division in spermatogenesis this is similar to the 0.12 mutations per cell division from Rahbari et al (2016) in their study of mosaicism. Assuming the early stages of embryo development, are similar between males and female then we would also expect to detect two gonosomal dnms per sperm cell. Thus, a zygote would on average inherit two gonosomal mosaic mutations from each parent, for a total of four parental mosaic mutations. These mutations could be detected as mosaic by their presence in the parental DNA at an allelic dosage of greater than 5% and less than 50%. Alternatively, if large numbers of offspring are available for an individual, it would be possible to identify gonosomal and germ-line mosaics despite their absence at detectable levels in the parental DNA by their presence in multiple offspring. This approach is advantageous in that if sufficient offspring are available it can detect mosaic mutations that are present in the parent's germ-line at very low allelic dosages (< 5%).

Looking at the literature, reports of mosaic mutations (referred to as premeiotic clusters, germinal mosaics, and mosaics) are surprisingly common in both clinical and earlier population studies. Woodruff et al. (1996) reported that 252 out of 944 screened mutations (26%) in Drosophila were mosaic and listed additional examples based on phenotype screens in nematodes, silkworms, guinea pigs, mice, rabbits, cattle and more than 80 cases in humans. As these initial studies were based on phenotypes the authors were unable to distinguish between gonosomal and germline mosaics, leaving uncertainty as to whether the mutations had occurred in early or late embryonic development. With the advent of next generation sequencing (NGS) it became possible to directly detect gonosomal mosaic mutations by their presence in the parental somatic DNA with allelic dosages of less than 50%. However, the initial NGS based studies required that a dnm be absent in the parent's DNA, directly discarding the true but detectable gonosomal mosaic dnms. Furthermore, the simple trios of father, mother and a single offspring used by most studies did not provide sufficient offspring to identify gonosomal or germ-line mosaics based on their presence in multiple offspring. As sequencing has become more common in the clinical environment, reports of confirmed mosaicism have continued to increase. Samuels and

Friedman (2015) reported 28 recent cases of mosaicism in the clinical literature. With 20 of the 28 confirmed as being gonosomal mosaic in a parent and 21 of the 28 confirmed as being mosaic in the germ-line. Since the initial large scale NGS studies of dnms in humans, a few smaller studies have directly looked for mosaic mutations. Dal et al. (2014) sequenced a quartet consisting of the two parents and a pair of monozygotic twins. They identified dnms as absent in the parents and either shared by the twins or unique to one twin. Making the assumption that mutations shared by the twins would have been inherited from the parents, while those unique to a single twin would had occurred after the splitting of the zygote. They identified 23 mutations shared by both twins, eight mutations specific to the first twin (25% of the dnms it carried), and one mutation specific to the second twin (4% of the dnms it carried). Due to the sequencing of blood and the absence of a third generation they were unable to determine whether the mutations were gonosomal mosaic rather than purely somatic mosaic. Campbell et al. (2014) screened 100 families with children affected by genomic disorders due to de novo CNVs, where the parents had initially been determined to be non-mosaic via standard clinical tests. Using a high sensitivity PCR-based CNV test they identified four cases of parental gonosomal mosaicism. They estimated that at least 4% of genomics disorders resulting from CNVs must be gonosomal mosaic and must have occurred during early embryo development of the parent, considering that the 100 families had been selected for the absence of parental mosaicism based on standard clinical tests. Rahbari et al. (2016) sequenced blood samples from three multi-sibling families identifying 739 non-mosaic dnms and 29 mosaic dnms. The mosaic dnms were identified by the presence of a small number of alternative allele reads in one parent (0.6-10%) or their presence in more than one offspring. From this they estimated that at least 3.8% of the dnms were mosaic in at least 1% of the parental blood cells. Noting that due to the design of the study they were unable to detect the earliest mosaic dnms with allelic dosages greater than 10% (first two cell divisions) or late mosaics with allelic dosages of less than 0.5% in the parents' blood cells. A recent study by Ju et al. (2017) identified 163 likely early embryonic mosaic mutations with allelic dosages of between 10% to 35% in 241 individuals, although due to the design of the study they were unable to prove their presence in the germ-line. Simulating early embryo development using an asymmetric model in which different cells divide at different rates, they determined the best fit for the level of mosaicism they had observed required 2.8 dnms per cell per cell doubling. Due to the use of an asymmetric model of embryo development this measurement is not directly comparable to those assuming symmetric division. The difficulty in comparing the two models is that a symmetric development model assumes that all cells divide at roughly the same rate, thus doubling the number of cells means every cell has undergone one additional round of cell division. In an asymmetric model of development, different cells divide at significantly different rates. Thus, when the number of cells has doubled, some cells may have undergone two divisions while others may have only undergone one or zero divisions. Excluding the estimate for the first twin in the Dal et al (2014) study and the Ju et al study (2017) the estimates of mosaicism in the three studies give a similar estimate of approximately 4% of dnms being observably mosaic in humans, which corresponds well with the

simple model described above. Both Campbell and Rahbari expect this figure of 4% to be the lower bound for the number of mosaic mutations due to the limitations of their study design.

The first twin from Dal et al (2014) however gave an estimated rate of 25% (with possible somatic contamination). For the third family in Rahbari et al (2016), 8% of the dnms were mosaic, while Ju et al (2017) estimated 2.8 dnms occurred per cell each time the total number of cells doubled during early embryo development, which - while not directly comparable to the other studies - is higher than the naive estimate of 0.5 mutations per cell division.

Taking these three points into account supports the suggestion that the overall estimate of 4% is a lower bound, and that the mutation rate during early embryo development is likely higher than that of the later stages of development. While these four studies do provide an initial look at the degree of mosaicism using modern direct DNA based methods, there experimental design limits the amount we can learn about mosaic mutations from them.

4 Signatures, patterns and properties of de novo mutation

Although dnms are by their very nature rare events, with less than one hundred being inherited by an individual from its parents, WGS and exome sequencing has made it possible to directly detect a large portion of these events. With the continuing reduction in the cost of NGS sequencing, since 2010 pedigree based dnm studies have expanded from a single quartet (Roach et al. 2010) to hundreds of trios per study (Wong et al. 2016). The increasing size of these studies has allowed them to identify a sufficiently large number of true dnms to start investigating their biological and chemical characteristics, as well as the associated mutational processes.

4.1 Methyl-Cytosine, C>T mutations

It was recognised early on that dnms observed in humans were enriched in C>T substitutions in CpG dinucleotide context (35% of observed mutations (Cooper and Youssoufian 1988) in 1.6% of genomic sequence (Josse et al. 1961)). This hypermutability of the CpG dinucleotide is thought to be due to the methylation of the cytosine residue which allows its spontaneous deamination to thymine (Coulondre et al. 1978). These early estimates are fully supported by recent NGS based dnm studies where an excess of C>T mutations at CpG sites is generally observed to be in the order of 10-20 fold higher than that expected by random mutation (Roach et al. 2010; Kong et al. 2012).

4.2 Trinucleotide patterns

While the number of identified germ-line dnms has rapidly increased, the total numbers are still relatively small. Also, studying germ-line dnms requires the sequencing of at least three genomes per 50-100 identified dnms, making them a relatively expensive source of variants for investigating mutational processes. Germline dnms are not the only type of dnm. Cancers are characterised by the presence of many somatic dnms and often carry mutations that affect some part of the DNA replication or repair pathway. Hence, a proportion of the dnms present in cancer may result from damaged DNA repair and replication mechanisms, making them a useful source insight into the consequences of damage to these mechanisms. Secondly dnms in cancer can be identified relatively easily by the sequencing of both cancerous and noncancerous tissue from the same individual. The dnms resulting from the cancer can then be identified as variants unique to the cancer sample. Finally, as the sequencing of a cancer's genome can help identify targets for treatment and provides insight into its causes, they are regularly sequenced by numerous groups. With many such cancer genomes being contributed to

public datasets such as The Cancer Genome Atlas, which contains samples from 11,000 patients (The Cancer Genome Atlas Research Network et al. 2013). The availability of such rich datasets has provided sufficient data for the application of statistical tools, designed to identify the mutational signatures resulting from specific types of DNA damage or failures of specific DNA repair and replication mechanisms (Alexandrov et al. 2013b).

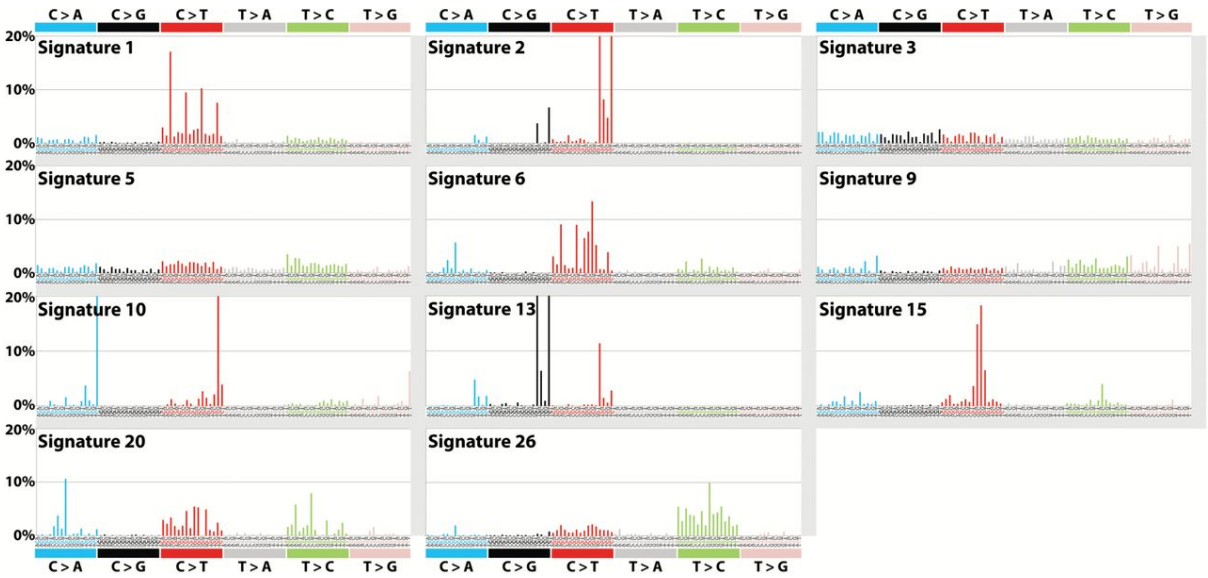


Figure 3: Endogenous signatures of mutation (Petljak and Alexandrov 2016). Each signature is displayed according to the 96 possible trinucleotide substitutions. Each of the 96 substitutions is defined by the mutation type (shown in colour at the top) and the possible 5' and 3' flanking bases. Each bar indicates the percentage of mutations attributed to the specific mutation type. All values are relative to the frequency of the specified trinucleotide in the human genome. All bars are limited to 20% even when the actual value exceeds that.

The approach undertaken by Alexandrov et al (2013b) utilises the trinucleotide pattern of a mutation. This consists of the mutated nucleotide and its 5' and 3' flanking bases and the associated complementary trinucleotide (for example TCT>TAT and complementary AGA>ATA). Based on this approach there are 96 possible unique substitutions. Mutational signatures are then defined by a unique pattern of frequencies for the 96 possible mutations (Alexandrov et al. 2013a, 2015; Petljak and Alexandrov 2016). The method described by Alexandrov et al (2013b) has resulted in the detection of 31 distinct signatures of mutation in tumours. Eleven of these are likely to result from endogenous cellular processes (Fig. 3), 7 from exogenous factors, and 13 of undetermined origin. Of the 11 endogenous signatures shown in Fig 3 (Alexandrov et al. 2015; Petljak and Alexandrov 2016; Alexandrov et al. 2013a), signatures 1 and 5 are common to all tumours and are shared with germline dnms (Rahbari et al. 2016; Ju et al. 2017). It is thought that these signatures represent the background mutational process of a replicating cell. Signature 1 shows the 10x enrichment for C>T mutations at NpCpG trinucleotides, thought to result from the spontaneous deamination of methylated-cytosine.

Signature 5 is characterized by a low level of mutations of nearly every class and may be representative of DNA damage during replication. Of the remaining classes of endogenous events signatures 2 and 13 are thought to be associated with the activity of the APOBEC family of deaminases, and signature 3 with that of the BRCA1/2 genes. Failure of DNA mismatch repair is associated with signatures 6, 15, 20 and 26, while signatures 9 and 10 are associated with the polymerases η and ε respectively. While these endogenous mutational signatures have been derived from tumour genomes, the signatures associated with them are likely to be representative of the signatures we would observe as the result of variation modifying the efficiency of the associated DNA repair and replication pathways for germ-line mutations. While the exogenous mutational signatures do not provide such a clear link to specific processes or DNA repair mechanism, they provide insight into environmental sources of DNA damage.

4.3 Clustering of DNMs

One pattern that has become noticeable with the increasing number of identified dnms is an excess of clustered mutations within 20kb (more than would be expected by chance if each mutation was an independent event) (Campbell et al. 2012; Michaelson et al. 2012; Francioli et al. 2015; Besenbacher et al. 2016). For dnms occurring within 100bp of each other it has been suggested, based on evidence from both dnms (Besenbacher et al. 2016) and population variants (Harris and Nielsen 2014), that error prone polymerases may be responsible, though a role for template switching events during replication has also been suggested (Löytynoja and Goldman 2017). For tandem mutations (pairs of adjacent mutations), DNA polymerase ζ has been suggested as a possible cause (Stone et al. 2012). For clustered mutations in the range of 10-20kb the exact mechanism remains unknown, however due to an excess in C>G events for this class it has been suggested that the error prone polymerase REV1 may play a role. It has also been suggested that clustered dnms in the 1-20kb range may be associated with maternal age and double-strand-breaks (Goldmann et al. 2017). Besenbacher et al (2016) also note that the mutational spectrum of clustered dnms differs based on the distance between them. Tandem mutations showing a different mutational spectrum compared to those separated by 2-10bp, which in turn differ from the spectrum observed for pairs at distances from 11bp to 20kb (Besenbacher et al. 2016).

4.4 Age and sex effects

In 1947 Haldane estimated that the mutation rate of the male germ-line was ~10x higher than that of the female germ-line (Haldane 1947). It has been suggested that the reason for the difference between males and females was the larger number of cell divisions that occur in the male germ-line (Miyata et al. 1987a, 1987b). As previously discussed the number of cell divisions from gamete to

gamete in the male germ-line is approximately 35 + 23(years since puberty) or 380 for a 30 year old male (Crow 2000), while for females the number of cell divisions from gamete to gamete is ~30. Thus, the number of cell divisions in the male germ-line is an order of magnitude higher than the female germ-line. A consequence of this hypothesis is that the rate of mutation on the autosomes should differ from that on the sex chromosomes (the X chromosome spends less time in the male germ-line than the autosomes, which in turn spend less time than the Y chromosome) (Miyata et al. 1987b; Drost and Lee 1995). The NGS studies have provided further support for this hypothesis. Kong et al (2012) determined that on average four times as many dnms were inherited from the father as from the mother. Venn et al (2014) showed that, in chimpanzees, the paternal mutation rate is five to six times that in the maternal germ-line. Further support comes from the strong paternal age effect identified in humans (of two additional paternal dnm per year after puberty (Kong et al. 2012), and three per year in chimpanzees (Venn et al. 2014)). This finding was confirmed by subsequent studies with larger sample sizes, though the exact size of the parental age effect varied between studies (Ségurel et al. 2014). While the initial studies found no maternal age effect (with Kong et al (2012) noting that nearly all variance in mutation rate among offspring was explained by the paternal age at birth), later studies have detected a weak maternal age effect. With larger sample sizes of 243, 693 and 816 trios compared to the initial 78 trios, a small but significant maternal age effect of 0.2-0.51 additional dnms per year has also been detected (Besenbacher et al. 2016; Goldmann et al. 2016; Wong et al. 2016). This would suggest that there is a low rate of mutation resulting from DNA damage which is independent of DNA replication and may represent the basal mutation rate of an oocyte in stasis.

5 *Evolution of the mutation rate*

The rate of dnm in an organism is a function of the fidelity of its DNA replication and repair pathways. As the corresponding enzymes are encoded in the DNA and thus subject to the effects of dnm (Mohrenweiser et al. 2003), the mutation rate will be subjected to natural selection. Variation within these pathways can thus modify the rate of dnm. As dnm are more likely to be deleterious than beneficial and due to their consequences ranging from negligible to lethal, the fitness effects of deleterious mutations are on average stronger than those of beneficial mutations. Thus, one would expect the rate of dnm to be driven to zero by natural selection. However, the rate of dnm is not zero, and varies by at least an order of magnitude between species. It appears to be associated with several population features, including a positive association with the size of the species genome (in eukaryotes) and inverse association with a species effective population size. Due to the importance of the rate of dnm, it is of considerable interest to understand what sets the lower limit on the rate of mutation for an organism. Simple simulations of small randomly mating diploid populations with two loci, one modifying reproductive fitness and the second mutations rate, suggest that the mutation rate stabilises well short of zero at a level defined by the size of the population and stability of the environment. The addition of recombination to a population in a stable environment has little or no effect on the level at which the mutation rate stabilises. In a changing environment without recombination, the mutation rate stabilises at a significantly higher level for the same population size than in a steady environment, adding recombination results in the mutation rate decreasing to the levels observed in the stable environment.

5.1 Fidelity Hypothesis

There have been two primary hypotheses put forward. Kimura suggested that the lowest achievable rate of dnm was determined by the fitness cost of increasing the fidelity of DNA repair and replication. In this fidelity cost hypothesis the rate of mutation is limited to the point where the fitness cost of reducing the mutation rate is greater than or equal to the fitness cost of the mutations that would be prevented (Kimura 1967). However, this has proven difficult to test as it is reliant on the fitness cost of increased fidelity which is difficult to estimate.

5.2 Drift-Barrier Hypothesis

The second hypothesis put forward by Lynch, argues that the lower limit of mutation rate is defined by the power of random genetic drift. Under this hypothesis, the mutation rate of a species is driven down, until it reaches the point where the gain in fitness from further reducing the mutation rate is less than the power of genetic drift for that population. Once the fitness gain is less than the power of random genetic drift, natural selection can no longer act against it (Lynch 2010, 2011; Sung et al. 2012). When applying the drift-barrier hypothesis to species with currently known rates of mutation, Lynch

argues that it can explain all current observations. He further argues that it can explain the higher mutation rates of error prone polymerases. As error-prone polymerases only act on a small number of bases, their potential effect on an organism's fitness is reduced as they have a lesser chance of damaging critical genes. Thus, their small effect on an organism's fitness means they are more susceptible to genetic drift. Due to their small effect on an organism's fitness, even a variant that gave them perfect fidelity is unlikely to have an appreciable effect on the organism's overall fitness. Thus, with no effective increase in the organism's fitness natural selection will not favour the variant leaving its' fate dependent on the genetic drift. The consequence of this is that while the error prone polymerases may be required for survival, there is little or no selection pressure to increase their fidelity. While the primary polymerases are responsible for a greater share of DNA replication and therefore have a greater effect on an organism's fitness. This greater effect on fitness places them under selective pressure and allows them to maintain higher levels of fidelity (Lynch et al. 2016). It should be noted that the drift-barrier hypothesis for the evolution of mutation rates does not rule out the fidelity cost hypothesis. It is argued that species reach the lower bound of the drift-barrier hypothesis before reaching the bound set by the fidelity cost hypothesis.

5.3 Evidence for evolution of the mutation rate

With the two hypothesises providing a theoretical framework for the evolution of the mutation rate, it is of interest to turn to the genome to look for evidence of mutator or anti-mutator phenotypes. Using publicly available 1000 genomes data (Auton et al. 2015; McVean et al. 2012), two groups have provided evidence for historical changes in the mutation rate in populations or the existence of candidate mutator loci. First, Harris utilised the initial 1000 genomes dataset to show that the frequency of TCC>TTC mutations has increased by 50% in non-singleton mutations private to European populations, compared to its frequency in mutations private to either Asian or African populations (Harris 2015). In a second study, Harris and Pritchard (2017) utilised the complete 1000 genomes dataset and re-evaluated the previous study. Analysing the allele frequency distribution of TCC>TTC mutations revealed that the enrichment peaked at a 0.6% frequency, while no enrichment was present in the rarest variants. In addition, C>T mutations on TCT, CCC, and ACC show a similar pattern of enrichment. Modelling this pattern of enrichment suggests that the increase was the result of a mutator allele that rapidly increased in frequency ~15,000 years ago before declining approximately ~2,000 years ago. The analysis was then extended to sub-populations within each continental group, identifying multiple changes in mutation spectra between closely related groups. One example of which was enrichment of ACA>AAA and TAT>TTT mutations in a subset of the Japanese population, this signature was exceedingly rare in Chinese individuals and absent in Kinh and Dai individuals. Further shifts could be observed when comparing humans to other great apes, with there being numerous differences in mutation spectrum

between the species. Taken together the diversity of these shifts in the mutational spectrum support the appearance and genetic drift of mutator phenotypes in species (Harris and Pritchard 2017). A third study by Seoighe and Scally (2017) sought to identify possible mutator loci using the 1,000 genomes data. When a mutator allele is present and active within a population it will increase the number of dnms per generation. For most haplotypes in the genome the number of dnms will not increase substantially as recombination will rapidly separate them from the mutator loci. However, haplotypes closer to the mutator will over time become enriched for dnms, as they are less likely to be separated by recombination. Thus, haplotypes in high linkage disequilibrium with the mutator loci will be enriched for rare or singleton variants, compared to the genomic background. This enrichment for rare or singleton variants can be utilised to identify loci that carry historic or potentially active mutator alleles. Applying this approach to the 1,000 genomes data identified numerous candidate loci, genes in proximity to the top 20 candidate loci were significantly enriched for both DNA repair and replication processes. For two of the loci there was additional support for a significant enrichment in the number of dnms observed in trios where one parent carried one of the haplotypes with enrichment for rare variants (Seoighe and Scally 2017).

5.4 Phylogenetic versus pedigree estimates of mutation rate

One final element of evolutionary interest has been the difference between the estimates of long-term averaged mutation rate obtained via phylogenetic methods, compared to those estimated from direct studies of dnms in trios or extended pedigrees. Indeed, the phylogenetic methods had estimated a dnm rate of ~2.5×10^{-8}, which is twice that of the average rate 1.2×10^{-8} observed in trios. This is of considerable concern as mutation rates are used as a molecular clock, to estimate the historical demographics of populations and the divergence times between species. Halving the mutation rate while keeping the other factors constant would suggest that the divergence between two species occurred twice as far back in time. Thus, the genetic estimate of the divergence time between humans and great apes would differ greatly from the palaeontological estimate. Secondly, estimates of events such as modern human and Neanderthal divergence or human migration out of Africa would be pushed back in time, leading to possible disagreements with archeologically evidence. Scally and Durbin argued that, at least for events relating to the historical demographics of modern humans, the pedigree based estimate of the human mutation rate (1.2×10^{-8}) was broadly compatible with both archaeological and mitochondrial evidence of divergence times for the human-Neanderthal split, and out of Africa and European-Asian split (Scally and Durbin 2012). They suggested that human generation time and mutation rate may have changed since the separation of human ancestors and the great apes. Further evidence for changes in mutation rate over time was obtained from investigating the genetic divergence, between 10 primate species from the old world and new world monkeys, great apes and humans. This provided evidence

supporting different rates and spectrums of mutation between the 10 lineages, suggesting that mutation rates change over evolutionary time periods (Moorjani et al. 2016). Scally recent looked at the issue further, arguing that the discrepancy between phylogenetic and pedigree estimates of mutation rate is likely due to a change in mutation rate per generation over evolutionary history due to a combination of changes to species life-history, gametogenesis and spermatogenesis (Scally 2016).

Objectives

With the advent of modern genomic technologies, our knowledge of the rates and properties of dnms has grown rapidly. This has provided insights into the evolutionary processes involved and allows us to use genetic information to investigate evolutionary history and demographics of species (Li and Durbin 2011; Schiffels and Durbin 2014). The direct estimates of the current human mutation rate have settled at around $1.2x10^{-8}$, approximately half that estimated from previous phylogenetic methods ($2.5x10^{-8}$). The differences between these methods are likely due to changes in the life-history and mutation rate of a species over evolutionary time periods, rather than to significant errors in these measurements (Ségurel et al. 2014; Scally 2016). However, the current studies in humans and other species have several limitations due to their use of simple two generation trios. This pedigree structure causes difficulties in differentiating between somatic and germ-line dnms, as well as determining the germ-line of origin of dnms (paternal, maternal or embryonic). This limits the ability to detect inter-individual variation in mutation rate and the degree of mosaicism in dnms. Furthermore, while the estimates of human mutation rate can be considered reasonably robust due to the sequencing of thousands of trios, for other species estimates have been restricted to either a single trio or a small number of trios. Finally, technical limitations of next generation sequencing have made it difficult to correctly estimate the exact percentage of the genome queried (Ségurel et al. 2014).

As such the focus of this PhD has been the utilisation of the Damona dataset, a unique dataset consisting of whole genome sequence from 743 dairy cattle (*Bos taurus*). The Damona dataset exploits the unique pedigree structure of dairy cattle to form 131, three or four generation pedigrees. Each three-generation pedigree consists of a sire, dam and child (or proband), plus an average of five grand-offspring, while four-generation pedigrees add the grand-parental generation. In addition, the sires and dams were selected so that most contribute to more than one pedigree. These multi-generational pedigrees allow us to accurately investigate dnm and overcome several of the limitations present in current human studies. Firstly, the presence of the third generation of grand-offspring allows us to confirm the inheritance of all candidate germ-line dnms, removing the issue of contamination with somatic mutations while retaining >95% of the non-mosiac dnms inherited by the proband. Secondly, we can determine the germ-line of origin for all dnms, allowing us to accurately assign events to the parent of origin. This improves our ability to detect inter-individual variation and gender differences in the rate and patterns of mutation. Thirdly, we can identify gonosomal and germ-line mosaic mutations that occurred early in the development of the proband, by the presence of complete but imperfect linkage among the grand-offspring. Finally, due to the size of the dataset and with most parents contributing to multiple trios, we can investigate the repeatability of the mutation rate in individuals and generate an accurate population level estimate of mutation rate and its variance within the population.

Considering the advantages of the Damona dataset the key objectives for this PhD were as follows.

1. To measure and characterise the maternal and paternal rates of germ-line dnm in cattle
    a. Confirm the germ-line nature of all mutations by requiring their transmission
    b. Identify the parent-of-origin for all dnms by utilising the grand-offspring to determine the originating germ-line
    c. Analyse the characteristics of mutation with regards to location, clustering, trinucleotide context, mutational signatures, recombination and transcription
2. Investigate the interactions between stages of development and rates of dnm
    a. Identify mosaic variants that have occurred within the proband rather than its parents, by identifying complete but imperfect linkage among the grand-offspring
3. Identify and quantify inter-individual variation in mutation rates
    a. Identify outliers with unusual rates and patterns of dnm
4. Identify functional variants within the cattle population
    a. Identify and develop tests for embryonic lethal variants

These objectives are primarily focused on furthering our understanding of dnm, by overcoming some of the limitations of previous studies, while providing an accurate mutation rate estimate for a third large mammalian species with a similar size genome to humans and chimpanzees. Finally, the characterisation of functional variants within the cattle population offers beneficial applied outcomes for farmers in the identification and development of tests for embryonic lethal variants.

Experimental section

Experimental Section

Study 1:

Frequency of mosaicism points towards mutation-prone early cleavage cell divisions.

Preprint: bioRxiv.org

Chad Harland, Carole Charlier, Latifa Karim, Nadine Cambisano, Manon Deckers, Erik Mullaart, Wouter Coppieters, Michel Georges

Abstract

It has recently become possible to directly estimate the germ-line de novo mutation (dnm) rate by sequencing the whole genome of father-mother-offspring trios, and this has been conducted in human, chimpanzee, birds and fish. In these studies, dnms are defined as variants that are heterozygous in the offspring while being absent in both parents. They are assumed to have occurred in the germ-line of a parent and to have been transmitted to the offspring via the sperm or oocyte. This definition assumes that detectable mosaicism in the individual in which the mutation occurred is negligible. However, instances of mosaicism are well-documented in humans and other organisms, including ruminants. We herein take advantage of the unique pedigree structure of cattle to show that mosaicism associated with dnms is a common occurrence, and that this should be considered in order to accurately estimate the mutation rate in this and possibly other species. It suggests that early cleavage cell divisions are particularly mutation-prone, and that the recurrence risk of dnm-dependent disorders in sibs may be higher than generally assumed.

Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle.

Chad Harland[1#], Carole Charlier[1#], Latifa Karim[2], Nadine Cambisano[2], Manon Deckers[2], Myriam Mni[1], Erik Mullaart[3], Wouter Coppieters[2], Michel Georges[1].

[1] Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, Belgium. [2] GIGA Genomics Platform, University of Liège, Belgium. [3] CRV, Arnhem, The Netherlands.

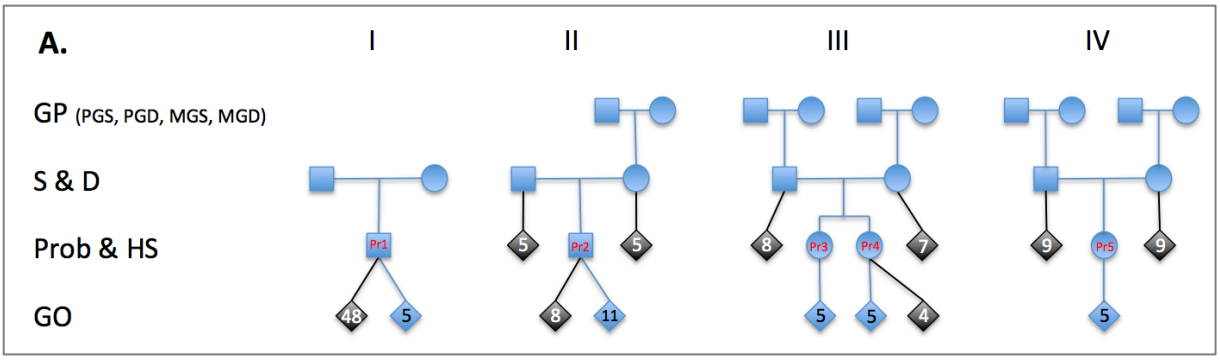
[#] *Contributed equally to this work*

Correspondence: michel.georges@ulg.ac.be

It has recently become possible to directly estimate the germ-line de novo mutation (*dnm*) rate by sequencing the whole genome of father-mother-offspring trios, and this has been conducted in human[1-5], chimpanzee[6], mice[7], birds[8] and fish[9]. In these studies, *dnm*'s are typically defined as variants that are heterozygous in the offspring while being absent in both parents. They are assumed to have occurred in the germ-line of one of the parents and to have been transmitted to the offspring via the sperm cell or oocyte. This definition assumes that detectable mosaïcism in the parent in which the mutation occurred is negligible. However, instances of detectable mosaïcism or premeiotic clusters are well documented in humans and other organisms, including ruminants[10-12]. We herein take advantage of cattle pedigrees to show that as much as ~30% to ~50% of *dnm*'s present in a gamete may occur during the early cleavage cell divisions in males and females, respectively, resulting in frequent detectable mosaïcism and a high rate of sharing of multiple *dnm*'s between siblings. This should be taken into account to accurately estimate the mutation rate in cattle and other species.

To study the process of *dnm*'s in the cattle germ-line, we sequenced the whole genome of 54 animals from four pedigrees. Grand-parents, parents and offspring (referred to as probands) were sequenced at average 26-fold depth (min = 21), and grand-offspring at average 21-fold depth (min = 10). The source of DNA was venous blood for females and sperm for males. The genome of one male proband (Pr 1) was sequenced both from semen (26-fold depth) and blood DNA (37-fold depth) (Figure 1A).

Figure 1: (A) Four pedigrees (I, II, III, IV) used for the detection of *dnm*'s. GP: grand-parents (PGS: paternal grand-sires, PGD, paternal grand-dams, MGS: maternal grand-sires, MGD: maternal grand-dams), S: sires, D: dams, Pr: probands, HS: half-sibs (of the proband), GO: grand-offspring. The five probands are labeled in red. Animals in blue were genome-sequenced at average depth of 23 and used for the detection of *dnm*'s. Animals in grey were used for confirmation by whole genome (average sequence depth of 20) or targeted sequencing (see Supplemental Methods). DNA was extracted from venous blood for females, and semen from males, except for Proband 1 for which both semen and blood DNA were analyzed.

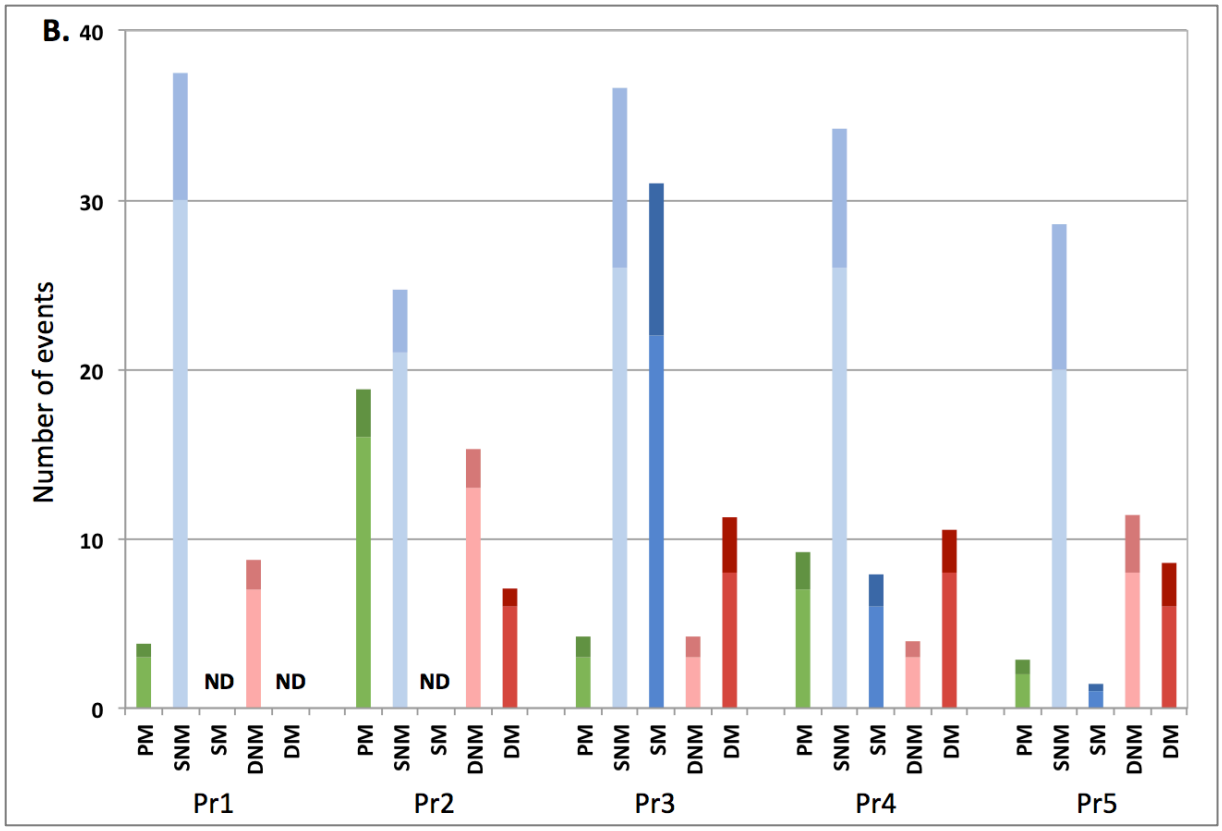


Using the standard definition, we identified 190 candidate *dnm*'s as variants that were (i) detected in a proband, (ii) absent in both parents (and grand-parents when available), (iii) transmitted to at least one grand-offspring, and (iv) not previously reported in unrelated individuals from the 1,000 Bulls project[13] (Suppl. Figure 1&2 and Suppl. Table 1). For confirmation, we developed amplicons spanning 113 candidate *dnm*'s and sequenced them at average depth of ~2,187 in the 54 animals plus 55 relatives (Figure 1A). This confirmed the genuine nature of 110/113 variants, demonstrating the excellent specificity of our bioinformatics pipeline. The three remaining ones were also detected in one of the parents (although not in the grand-parents) in the confirmation, and momentarily ignored.

We first examined what proportion of *dnm*'s detected in a proband might actually have occurred during its development rather than being inherited via the sperm or oocyte. An unambiguous distinction between the two types of *dnm*'s is their degree of linkage with either the paternal or maternal haplotype upon transmission to the next generation (i.e. the grand-offspring in Figure 1A). *Dnm*'s that have occurred in the germ-line of the sire will show *perfect* linkage with the proband's paternal haplotype in the grand-offspring (i.e. always transmitted with the paternal haplotype, never transmitted with the

maternal haplotype), while *dnm*'s that have occurred in the germ-line of the dam will show *perfect* linkage with the proband's maternal haplotype in the grand-offspring. On the contrary, *dnm*'s that have occurred during the development of the proband will be in *complete* (but imperfect) linkage with either the paternal or maternal haplotype (i.e. sometimes transmitted with the maternal haplotype, never transmitted with the paternal haplotype, or sometimes transmitted with the paternal haplotype, never transmitted with the maternal haplotype) (Suppl. Figure 1). Across the four pedigrees, 124 variants were in perfect linkage with the paternal haplotype, 32 in perfect linkage with the maternal haplotype, 10 in complete (but imperfect) linkage with the paternal haplotype and 21 in complete (but imperfect) linkage with the maternal haplotype (Figure 1B).

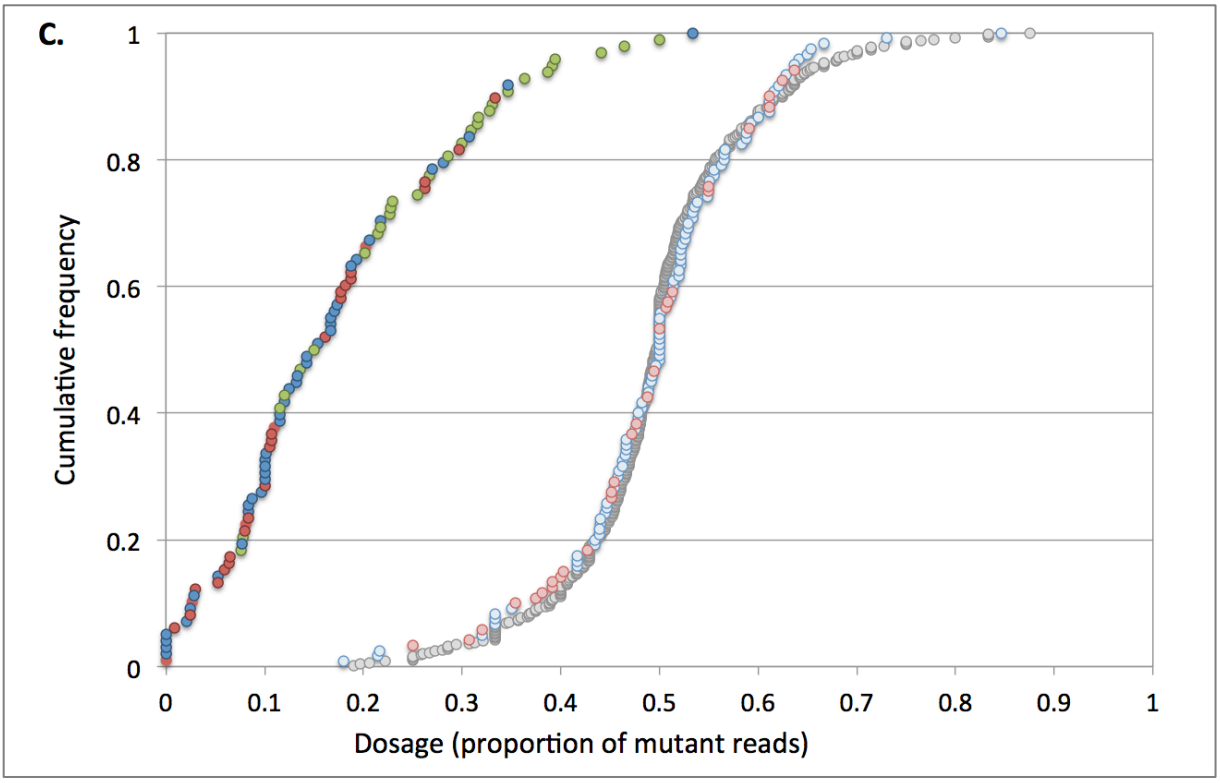
Figure 1: (B) Numbers and types of *dnm*'s detected in the five probands (Pr1, Pr2, Pr3, Pr4, Pr5). Green: PM = proband mosaic, Light blue: SNM = sire non-mosaic, Dark blue: SM = sire mosaic, Light red: DNM = dam non-mosaic, Dark red: DM = dam mosaic. For each bar, the lower light section corresponds to the actual number of detected *dnm*'s, the upper darker section to an extrapolation to the whole genome based on the estimated coverage. ND = not done (because the corresponding grand-parents were not sequenced).



If the 10+21 *dnm*'s indeed occurred during the development of the proband rather than being inherited from the sire or dam, the *dnm* dosage (defined as the proportion of reads spanning the *dnm* site that carry the mutant allele) is expected to be < 50% in the proband but equal to 50% in the grand-offspring

inheriting the *dnm*. The mean dosage was 0.26 in the proband, and 0.52 in the grand-offspring, and this difference was highly significant ($p < 10^{-6}$). The corresponding means were 0.48 and 0.49 ($p = 0.40$) for the 124+32 mutations showing perfect linkage with either the paternal or maternal haplotype (Figure 1C).

Figure 1: (C) Cumulative frequency distribution of *dnm* dosage estimated as the proportion of reads carrying the mutation. Green circles: PM mutations in the probands. Light blue circles: SNM mutations in the probands. Light red circles: DNM mutations in the probands. Dark blue circles: SM mutations in the sires. Dark red circles: DM mutations in the dams. Grey circles: corresponding PM, SNM, DNM, SM and DM mutations in the grand-offspring. The three SM and one DM variant with dosage of 0, were shared between the proband and at least one half-sib yet not detectable in the semen or blood of the corresponding parent.



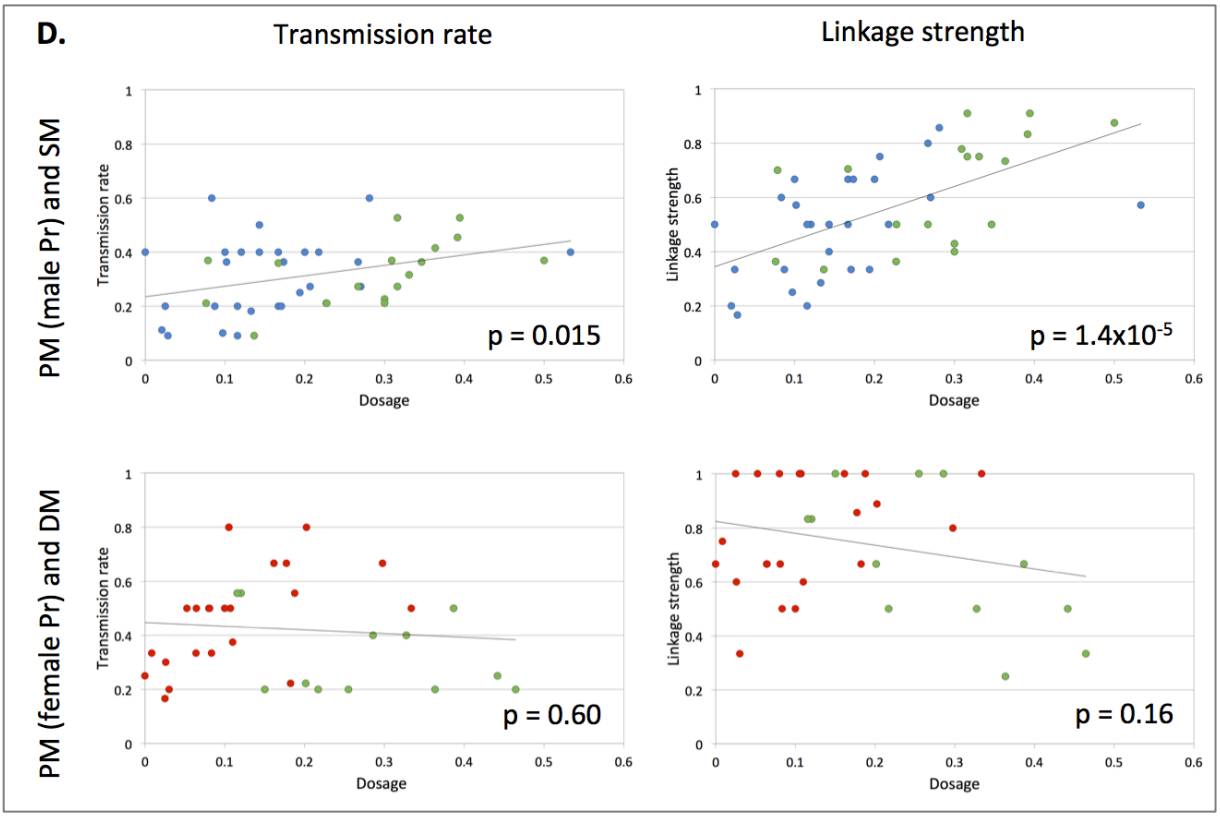
We conclude that in cattle ~17% of *dnm*'s detected in an animal using standard procedures are not inherited from the sire or dam but correspond to premeiotic clusters generated during the development of the individual. This is a lower bound, as *dnm* will only be detected and recognized as having occurred in the proband if (i) the *dnm* dosage is sufficiently high for the proband to be called heterozygote, (ii) the *dnm* is transmitted to at least one grand-offspring, and (iii) complete (but imperfect) linkage is demonstrated in the grand-offspring (Suppl. Figure 3). We will refer to this type of *dnm* as Proband-Mosaic (PM), while the others will be referred to as Sire-Non-Mosaic (SNM) (meaning that the sire is not mosaic for a mutation transmitted via his sperm), or as Dam-Non-Mosaic (DNM) (meaning that the dam is not mosaic for a mutation transmitted via her oocyte). The proportion of PM (but not SNM and

DNM) mutations differed significantly between probands ($p = 0.004$); neither differed significantly between sexes ($p > 0.30$). Of particular interest, the three PM mutations of proband 1 were detected in both sperm and blood DNA (Suppl. Table 1), indicating that they occurred early in development (see hereafter).

If detectable mosaicism for *dnm*'s is common in the individual in whom they occurred, requiring their absence in the DNA of the parents (as typically done) will cause genuine *dnm*'s to be eliminated. We took advantage of the grand-parents available in three pedigrees to recover such events as variants that were (i) absent in the grand-parents, (ii) detected in either sire or dam with a dosage significantly < 50% (Suppl. Table 1), (iii) transmitted to the proband with a dosage of ~50%, (iv) transmitted to at least one grand-offspring with a dosage of ~50%, and (v) not previously reported in unrelated individuals[13]. We will refer to these types of mutations as Sire-Mosaic (SM) and Dam-Mosaic (DM), respectively (meaning that the sire/dam is detectably mosaic for a *dnm* transmitted via the sperm or oocyte) (Suppl. Figure 1). We detected 61 such candidate events, including the 3/113 variants mentioned above (Suppl. Table 1 and Suppl. Figure 2). We developed amplicons for 34, and sequenced (average 1,498-fold depth) all 54 individuals plus 55 relatives (including ≥ 5 half-sibs of the probands) (Figure 1A). We took advantage of whole genome sequence information that became available for 27 half-sibs, to trace the inheritance of the remaining 24 candidate variants. The ensuing data indicated that 11/61 candidates were genuine *dnm*'s but occurred in the germ-line of one of the grand-parents rather than one of the parents (dosage ~50% in the sire or dam, and perfect linkage in the half-sibs). The SM/DM status was unambiguously demonstrated for 40 (dosage < 50% in the sire or dam in the confirmation, transmission to half-sibs, and complete (but imperfect) linkage) and strongly supported for the remaining 10 (dosage < 50% in the sire/dam in the confirmation or complete (but imperfect) linkage yet without transmission) (Suppl. Figure 2 and Figure 1B). Further supporting the genuine nature of the SM/DM mutations, the dosage was 0.12 on average in the corresponding parent, while being 0.51 in descendants ($p < 10^{-6}$) (Figure 1C).

For *dnm*'s that were detectably mosaic in sperm (SM and PM in male probands), allelic dosage was significantly correlated with rate of transmission ($p = 0.025$) and strength of linkage ($p = 0.0002$). These correlations were not significant for *dnm*'s that were detectably mosaic in blood (DM and PM in female probands). This suggests that the degree of mosaicism in the soma is a poor indicator of the degree of mosaicism in the germ line (Figure 1D). Accordingly, the rate of transmission of the three PM mutations of proband 1 to its 53 offspring was better predicted by their dosage in sperm than in blood (Suppl. Figure 4).

Figure 1: (D) Relationship between the *dnm* dosage (fraction of mutant reads) and the rate of transmission to offspring (left) and strength of linkage (right) for mutations that are detectably mosaic in the sperm of a male parent (upper), or in the blood of a female parent (lower). Green circles: PM mutations, blue circles: SM mutations, red circles: DM mutations. The corresponding correlations were significant in males ($p = 0.015$ and $1.4x10^{-5}$) but not in females ($p = 0.60$ and 0.16).



Considering SNM/SM and DNM/DM mutations jointly, we conclude that on average a sire is detectably mosaic (in sperm) for 29% of *dnm*'s present in a sperm cell, while a dam is detectably mosaic (in blood) for 51% of *dnm*'s present in an oocyte. These are lower bounds as we only considered *dnm*'s for which the dosage was significantly < 0.5 in the parent (condition (ii) above). These figures are possibly consistent with recent reports in the mouse (~25%)[7], but considerably larger than current estimates in human (~5%)[14]. They are certainly larger than expected assuming that the mutation rate per cell division is uniform throughout development, and it suggests that the mutation rate is higher for early cell divisions (Suppl. Figure 5). Moreover, when analyzing the transmission patterns of SM and DM mutations to the half-sibs of the proband (in whom the *dnm*'s were detected), we were struck by the fact that (i) >60% of half-sibs share at least one *dnm* with the proband, while <50% are expected ($p = 0.05$), and (ii) half-sibs sharing multiple *dnm*'s with the proband appeared surprisingly common (Suppl. Table 2). These findings also indicate that a substantial proportion of *dnm*'s must occur early in development and be present in the precursor cells common to the soma and germ line (Suppl. Figure 5).

In mammals, after fertilization, cleavage, and segregation of (i) the inner cell mass from the trophoblast, (ii) the epiblast from the hypoblast, (iii) the embryonic epiblast from the amniotic ectoderm, a small number of epiblast-derived cells located in the wall of the yolk sac in the vicinity of the allantois are induced to become primordial germ cells (PGCs). These migrate to the primitive gonad where they expand and produce >1 million gametogonia. Oogonia initiate meiosis prior to birth in females. Spermatogonia will resume mitotic divisions at puberty allowing (i) the maintenance of a pool of stem cell like spermatogonia, and (ii) sustained spermatogenesis involving ~3 additional mitotic divisions followed by meiosis (Suppl. Figure 6). We simulated the process of de novo mutagenesis in the male and female germ cell lineages assuming (i) uniform pre- and post-natal mutation rates per cell division, and (ii) 40 PGCs sampled at random from the embryonic epiblast-derived cells[15]. Pre- and post-natal mutation rates were adjusted to match the observed number of mutations per gamete (34 in sperm, 14 in oocytes). Under these conditions, we virtually never observed the level of mosaicism, nor the sharing between sibs characterizing the real data (Figure 2). We (i) increased the relative mutation rate during the early cell divisions (keeping the mutation rate per gamete constant)(10 and 20-fold increase during the first 4, 7, 11, 15 and 18 cell divisions; Suppl. Figure 6), (ii) reduced the number of induced PGCs (4, 10, or 40), and (iii) varied the relatedness between PGCs (i.e. sampled randomly amongst all embryonic epiblast-derived cells or from a sub-sector)(Suppl. Figure 6). Increasing the mutation rate during the very first cell divisions matched the real data much better (Figure 2). To quantitatively evaluate model fitting we used (i) the proportion of PM, SM and DM mutations with corresponding rate of mosaicism in sperm and soma, and (ii) the proportion of sibs sharing 0, 1, 2, ... *dnm*'s with a proband, to compute the likelihood of the data under different scenarios (see M&M). A 20-fold increased mutation rate during the first four cell divisions, combined with 4 related PGCs fitted the data best (Table 1 and Suppl. Table 2). The data were $\geq 10^{16}$ times less likely under models assuming a uniform mutation rate throughout development, and $\geq 10^{5}$ times less likely assuming an increased mutation rate passed the 7^{th} cell division (after segregation of inner cell mass and trophoblast) (Table 1 and Suppl. Table 2).

Figure 2: (A) *Dnm*'s with detectable mosaicism in sperm DNA of male probands (PM – sperm) or sires (SM – sperm), or in blood DNA of female probands (PM – blood) or dams (DM – blood), ranked by observed rate of mosaicism. Coloured lines: real data. Pr1-5: proband 1-5. Dark grey shaded area: 95% confidence interval obtained from simulations assuming uniform mutation rate per cell division and 40 unrelated PGCs. Light grey shaded area: 95% confidence interval obtained from simulations assuming 20-fold higher mutation rate during the first 4 cell divisions, and 4 related PGCs (Table 1).

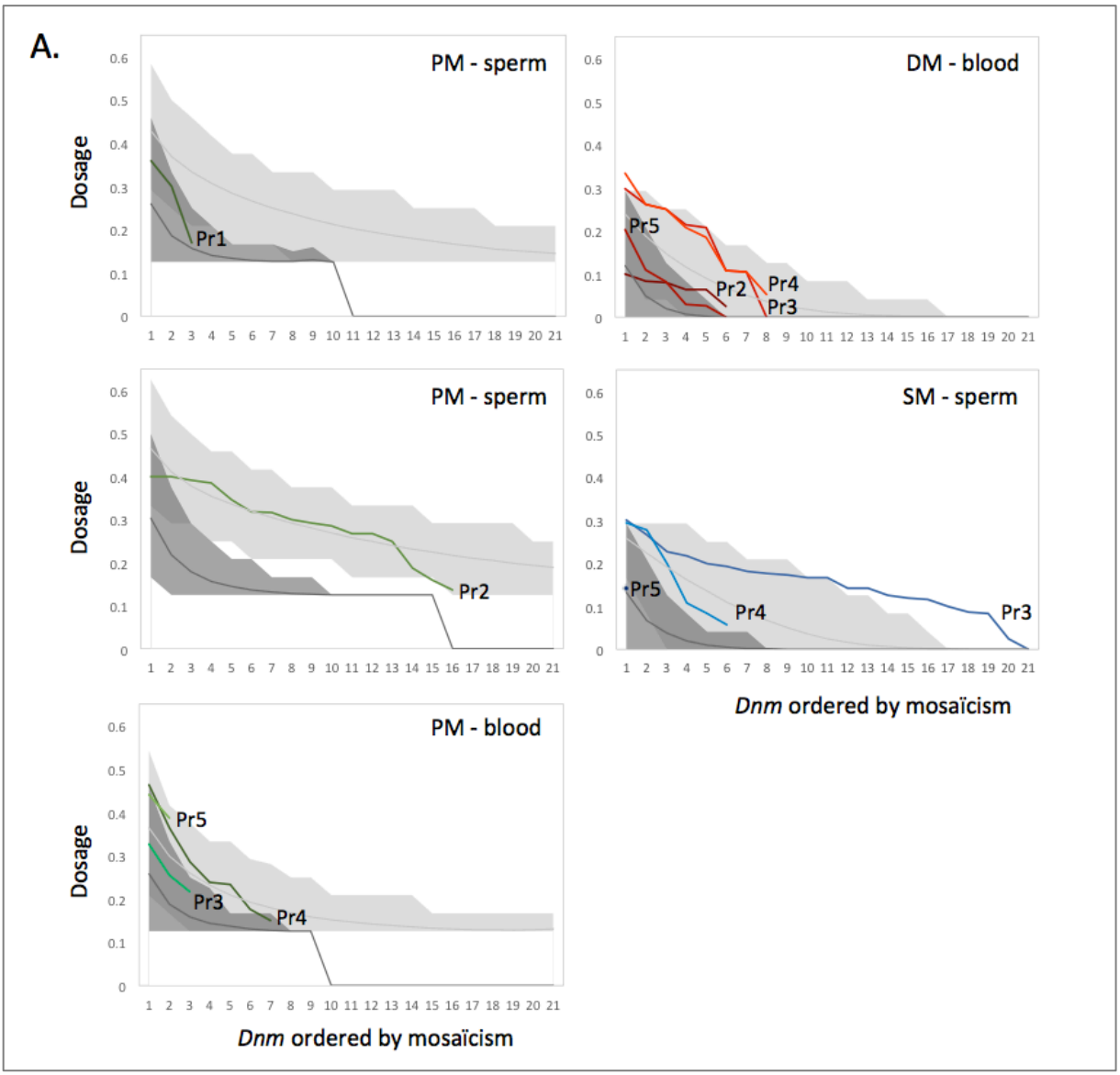


Figure 2: (B) Distribution of the proportion of half-sibs (HS) of the probands that share 0, or at least 1 (1+) of the *dnm*'s detected in the corresponding proband. Red bars: real observations for *dnm*'s transmitted by the dam (DM+DNM). Blue bars: real observations for *dnm*'s transmitted by the sire (SM+SNM). Dark grey bars: expectation under the null hypothesis of uniform prenatal mutation rate per cell division and 40 unrelated PGCs. Light grey bars: expectation under the best alternative model assuming a 20x increased mutation rate during the first 4 cell division and 4 related PGCs (Table 1).

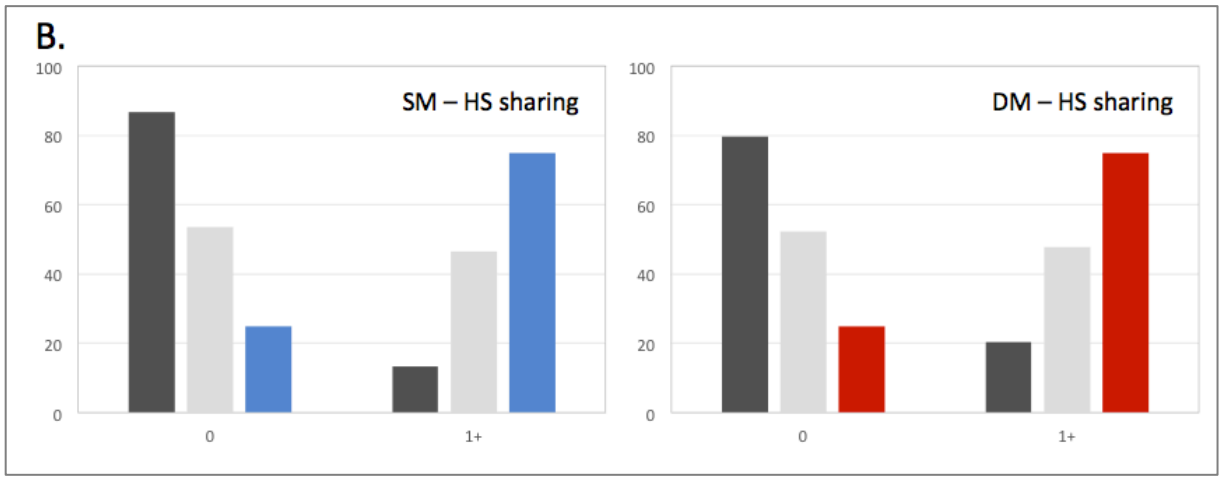


Table 1: Relative likelihood of the observations under different models of gametogenesis
The first four columns correspond to the parameters that were tested in the model: (i) the fold increase of the mutation rate (1x, 10x, 20x), (ii) during the x first cell divisions (4, 7, 11, 15, 18), (iii) the number of PGCs (4, 10, 40), and (iv) the ontogenetic relatedness of the PGCs (F(alse) or T(rue)). Log(LR) corresponds to the logarithm (base 10) of the average likelihood of the data (after 100 replications) relative to the best model (first line). Parameters are in bold when the corresponding model is the best given that parameter value. We only show results for models that are the best given at least one parameter value. Likelihoods of all models are given in Suppl. Table 3.

Fold increase mutation rate	During first x cell divisions	Number of PGCs	Related PGCs or not	Log(LR)
20x	**4**	**4**	**T**	0.00
10x	4	4	T	-0.35
20x	4	4	**F**	-0.40
10x	**7**	4	F	-1.57
20x	4	**10**	T	-1.74
20x	4	**40**	T	-3.45
10x	**11**	4	T	-5.09
20x	**15**	4	T	-7.03
10x	**18**	4	T	-9.46
1x	7	4	T	-16.17

When accounting for genome coverage, the estimated number of *dnm*'s per gamete (SNM+SM, DNM+DM) averaged 46.6 for sperm cells and 18.1 for oocytes (male/female ratio of 2.6), corresponding

to an average mutation rate of ~$1.2x10^{-8}$ per base pair per gamete. Including an estimate (from the simulations) of the number of missed SM (~3.3) / DM (~1.1) and misclassified PM mutations (~2.9 to ~8 depending on the proband), yields an average mutation rate of ~$1.17x10^{-8}$ per base pair per gamete and a male/female ratio of 2.4. The standard approach of ascertaining *dnm*'s (i.e. erroneously considering PM mutations, ignoring SM and DM mutations) would have yielded a mutation rate of $0.9x10^{-8}$ per bp per gamete, with a 2.5-fold higher mutation rate in bulls than in cows.

Two hundred twenty of the 237 identified *dnm*'s were nucleotide substitutions, and the remaining 17 were small insertion-deletions. The non-mosaic classes of mutations (SNM and DNM) were ~30-fold enriched in CpG>TpG transitions as expected. This signature was also present but less pronounced for mosaic mutations (PM, SM and DM). Mosaic mutations were ~2.6-fold enriched in C>A and/or G>T transversions, largely due to GpCpA>GpApA and TpCpT>TpApT substitutions (Figure 3). This was unlikely to be an artifact for reasons spelled out in Suppl. Figure 7. It is noteworthy that this is exactly the same mutational signature as the one recently reported for human embryonic somatic mutations[16]. In addition, this mutational signature corresponds to that reported for damage to DNA polymerases ε for which a number of missense and splice variants exist several of which are carried by individuals in this dataset (Suppl. Note 2). There was no obvious difference between the profile of *dnm*'s in the male and female germ line (data not shown). In general, *dnm*'s appeared uniformly scattered across the genome (Suppl. Figure 8).

The enrichment of C>A/G>T transversions in the mosaic mutations caused the overall Ti/Tv ratio to be 1.33, well below expectations. This was likely due to sampling variation (meaning that Ti/Tv ratios might differ between families and that we by chance sampled families at the low end), as the Ti/Tv ratio was 1.99 for 2,530 candidate *dnm*'s detected with the same bioinformatics pipeline in a follow-up study of 113 probands (excluding the ones analyzed in this work), i.e. closer to expectations and the 2.2 Ti/Tv ratio of SNPs segregating in the Holstein-Friesian dairy cattle population (MAF $\leq$ 0.01; rare allele considered to be the derived allele). However, the spectrum of the 2,530 *dnm*'s remained significantly different from the SNP spectrum, with an excess of C>A/G>T transversions in the mosaic class of mutations, an excess of C>T/G>A transitions in both mosaic and non-mosaic mutations, and a paucity of T>C/A>G transitions in both mosaic and non-mosaic mutations (Suppl. Figure 7). This could point towards recent alterations of the mutational profile in domestic cattle. It is worth noting in this regard that most analysed animals were bred using artificial insemination and/or *in vitro* embryo production. It seems unlikely that artificial insemination with frozen semen could explain the observed familial clustering of specific *dnm*'s. However, it is conceivable that in vitro maturation, fertilization and culture of oocytes and embryos affect the *dnm* rate, possibly by perturbing DNA replication. It is important to determine whether this is the case, especially as the same methods are increasingly used in human reproduction.

Figure 3: (A) SNM and DNM (i.e. *dnm*'s assumed to have occurred in the later stages of gametogenesis): fold excess or deficiency over expected for specific nucleotide substitutions when accounting for trinucleotide context. Trinucleotide frequencies were calculated from the bosTau6 reference genome, with all 96 unique events being represented and account for both the 5'>3' and 3'>5' versions of the same event. Inset: Proportion of *dnm*'s corresponding to the six possible types of nucleotide substitutions.

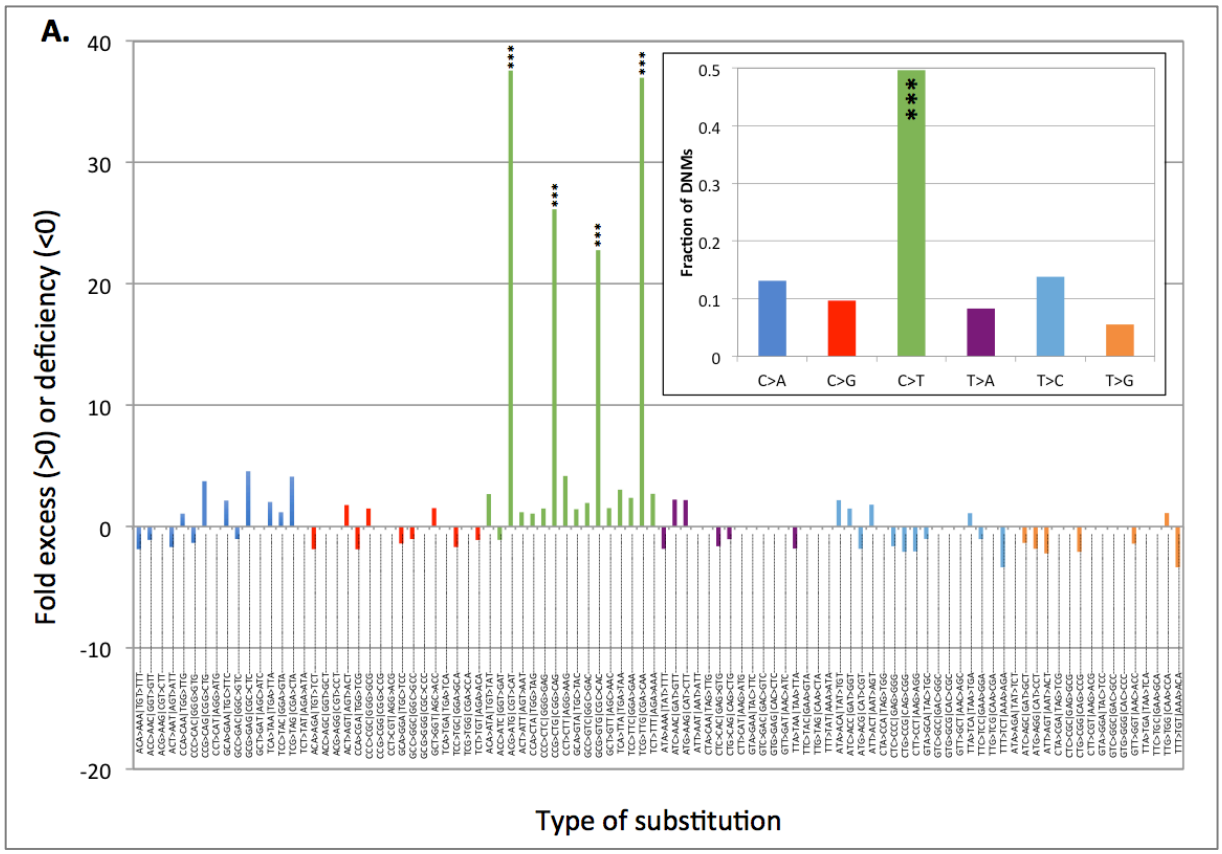
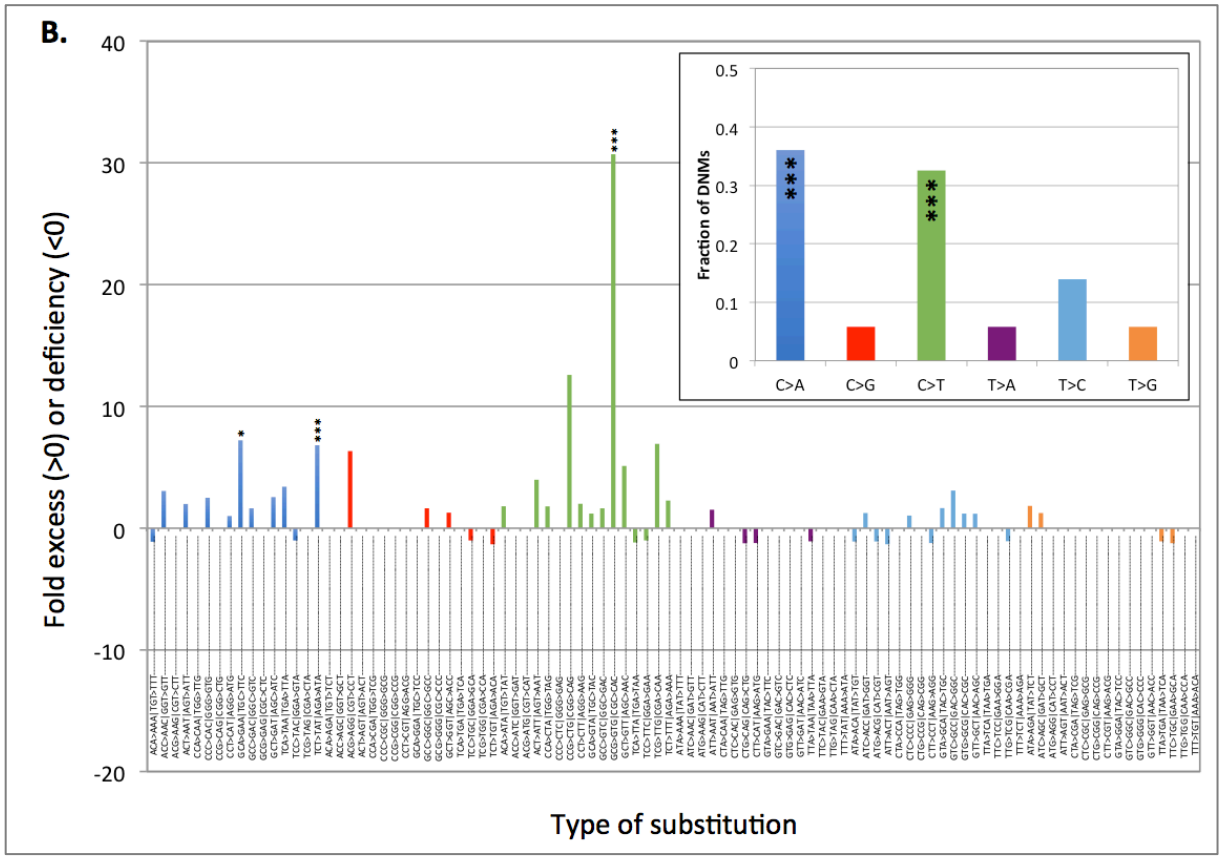


Figure 3: (B) Idem for PM, SM and DM (i.e. *dnm*'s assumed to have occurred in the early stages of gametogenesis). ***: p< 0.001, *: p< 0.05 (accounting for multiple testing by Sidak correction).



Dnm's occurring during the development of an individual, should a priori affect the maternal and paternal chromosome with equal probability. When considering the PM, SM and DM jointly, 48 mosaic mutations occurred on the maternal chromosome versus 31 on the paternal chromosome (p = 0.11). This trend suggests that the maternal and paternal chromosomes might be epigenetically distinct during early development and that this may affect their mutability.

Our work points towards the fact that direct estimates of mutation rates from sequencing families may have to be revisited, taken PM, SM and DM status into account, to obtain more accurate estimates of the mutation rate per gamete and per generation. This may affect both the overall mutation rate as well as its male/female ratio. However, our analyses suggest that the effect is likely to be modest and would, for instance, be insufficient to explain the present 2-fold discrepancy between direct and indirect estimates in human studies[17,18]. We confirmed by simulation that the rate of mosaicism does not significantly affect the rate of nucleotide substitution per generation or average fixation time[19] (Suppl. Figure 9).

Our work calls for a careful re-evaluation of the importance of mosaicism for *dnm*'s in humans. If more common than presently appreciated, the recurrence risk of *dnm*-dependent disorders in sibs may be

higher than generally assumed [11,17]. Moreover, a non-negligible proportion of true *dnm*'s may have been ignored (because they were detected at low dosage in the parents) in *dnm*-dependent searches for genes underlying inherited disorders hence reducing the potential power of such studies.

Acknowledgements

This work was funded by the DAMONA Advanced ERC project to Michel Georges. Carole Charlier is Senior Research Associate from the Fonds de la Recherche Scientifique - FNRS (F.R.S.-FNRS). Chad Harland has been funded in part by Livestock Improvement Corporation (New Zealand). We are grateful to Erik Mullaart and CRV (Arnhem the Netherlands) for providing us with the sperm and blood samples. We used the supercomputing facilities of the Consortium des Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI) funded by the F.R.S-FNRS.

Authors contributions

MG, CH, CC: designed the experiments. EM: provided samples. LK, NC, MD, WC: performed the sequencing. CH, MG, CC: analysed data. MG, CH, CC: wrote the paper.

Data availability

All sequence data will be made freely available in public databases.
Supplemental material is available online: http://www.biorxiv.org/content/early/2017/06/29/079863

References

1. Roach JC et al. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* **328**: 636–639 (2010).

2. Conrad DF et al. Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**: 712–714 (2011.).

3. Kong A et al. Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature* **488**: 471–475 (2012).

4. Campbell CD et al. Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* **44**: 1277–1281 (2012).

5. Michaelson JJ et al. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell* **151**: 1431–1442 (2012).

6. Venn O et al. Strong male bias drives germline mutation in chimpanzees. *Science* **344**:1272-1275 (2014).

7. Lindsay SJ et al. Striking differences in patterns of germline mutation between mice and humans. doi: https://doi.org/10.1101/082297

8. Smeds L, Qvarnström A, Ellegren H Direct estimate of the rate of germline mutation in a bird. *Genome Res.* **26**: 1211-1218 (2016).

9. Feng C et al. Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. *eLife*, in press (2017).

10. Woodruff RC & Thompson JN. Have premeiotic clusters of mutation been overlooked in evolutionary theory? *J. Evol. Biol.* **5:**457-464 (1992).

11. Campbell IM et al. Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet.* **31**:382-392 (2015).

12. Smit M et al. Mosaicism of Solid Gold supports the causality of a noncoding A-to-G transition in the determinism of the callipyge phenotype. *Genetics* **163**:453-456 (2003).

13. Daetwyler HD et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46**:858-865 (2014).

14. Rahbari R et al. Timing, rates and spectra of human germline mutation. *Nature Genetics* **48:** 126-133 (2016).

15. Ohinata et al. Blimp1 is a critical determinant of the germ cell lineage in mice. *Nature* **437**:207-213 (2005).

16. Ju YS et al. Somatic mutations reveal asymetric cellular dynamics in the early human embryo. *Nature* **543**: 714-718 (2017).

17. Scally A. Mutation rates and the evolution of germline structure. *Philos Trans R Soc Lond B Biol Sci* **371:** 20150137 (2016).

18. Segurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15**:47-70 (2014).
19. Woodruff RC & Thomson JN. The fundamental theorem of neutral evolution: rates of substitution and mutations should factor in premeiotic clusters. *Genetics* **125:** 333-339 (2005).

Methods

Whole genome sequencing. DNA was extracted from sperm (for males) or whole blood (females and one male) for the four families and their relatives using standard procedures. Familial relationships were confirmed by genotyping all DNAs with the 10K Illumina SNP chip. We constructed 550bp insert size whole genome Illumina Nextera PCR free libraries following the protocols recommended by the manufacturer. All samples where then sequenced on Illumina HighSeq 2000 instruments, using the 2x100bp paired end protocol by the GIGA Genomics platform (University of Liège). Data was mapped using BWA mem (version 0.7.9a-r786)[20] to the BosTau6 reference genome. Alignments were processed according to the GATK[21] best practices version 2 with PCR duplicates marked, INDEL realignment and Base Quality Score Recalibration using known sites. GATK HaplotypeCaller (version 3.4) was used according to the N+1 workflow to generate variants from the alignments. Common variants were then compared to a 10K Illumina SNP chip for each individual to confirm the identity of the library.

Detection of de novo mutations. We developed a suite of scripts to identify *dnm*'s from a vcf file produced by GATK and containing sequence information about members of four-generation pedigrees such as the ones described in Fig. 1. The first ("4_phaser_4_gen.pl") generates the linkage phase for the parents (sire and dam), the proband, and the grand-offspring. Phasing is done based on high quality variant positions and genotypes (f.i. QUAL score ≥ 50,000; PL scores ≥ 100; sequence depth ≤ 2.5x the average sequence depth). The outcome is knowledge of the grand-parental origin of the paternal and maternal chromosomes of the proband including the identification of cross-over events, as well as the grand-parental origin of the chromosomes transmitted by the proband to the grand-offspring including the identification of cross-over events. The second module ("5_de_novo_detector_4_gen.pl") identifies the candidate *dnm's per se*. It first identifies biallelic variant positions for which all grand-parents, sire, dam and proband have a genotype and sequence coverage between set limits (f.i. 10 and 60). The proportion of variants sites satisfying these depth limits was used to estimate the proportion of the genome (of the total of the 2,670,422,299 base pairs in the bovine Bostau6 build) that was explored. Candidate *dnm*'s were then identified as sites for which (i) the QUAL score was ≥ 100, (ii) the proband had genotype 0/1 with corresponding PL-scores ≥ 40, (iii) none of the grand-parents had reads with the alternate allele (AD = x,0), (iv) either the sire or the dam had no reads with the alternate allele, and (v) reads with the alternate allele were found in the grand-offspring. The same script also determines the

genotype frequencies (0/0, 0/1 and 1/1) at the corresponding position in sequenced individuals outside of the pedigree that are not descendants of the sire or the dam. The third module ("6_germline_assigner_4_gen.pl") combines the output of the first and second module to determine in which individual the *dnm* is most likely to have occurred (one of the four grand-parents, sire or dam, or proband) and on which grand-parental chromosome it occurred. All candidate *dnm*'s were manually curated using the Integrated Genome Viewer (IGV)[22] . *Dnm*'s with mutant reads in either sire or dam (even if called 0/0 by GATK) were relabelled as SM or DM, provided that the *dnm* segregated in complete (but imperfect) linkage with the paternal or maternal haplotype (respectively), in the half-sibs of the proband.

To test the corresponding pipeline, we identified 11,255 variants (of which 10,093 SNPs) for which Pr2 was heterozygous and which were not present in unrelated individuals including from the 1,000 Bulls project. The corresponding "genotype fields" of the parents and grand-offspring were modified in the vcf file such that the genotype (GT) was set at 0/0 and the unfiltered allele depth (AD) of the derived allele set at 0. The Ti/Tv ratio for the corresponding SNPs was 1.9922. The proportion of the genome explored was estimated at 76% as described above. The pipeline detected 9,325 variants (83% sensitivity) of which 8,409 SNPS (81% sensitivity). The Ti/Tv ratio amongst detected *dnm*'s was 1.9989, indicating that the pipeline did not introduce a Ti/Tv bias.

Confirmation of candidate dnm's. PCR primers were then designed for each candidate passing the quality check in IGV using BatchPrimer[23] targeting a product size of 200-1000bp with at least one primer being present in unique (non-repeat) sequence (as identified by repeatmasker). The resulting amplicons were sequenced on an Illumina MiSeq instrument using the 2x250bp paired end protocol. The sequenced amplicons were aligned to the BosTau6 reference genome using BWA mem and candidate dnms were checked in IGV and variants were called using freebayes (v1.0.2-15-g357f175)[24].

Modeling gametogenesis. (i) Data types: To compare the adequacy of the different gametogenesis models we computed the likelihood of three types of data. The first is the degree of mosaicism in the parent across *dnm*'s detected in a given gamete. Thus we may have detected *n* SM and *m* SNM *dnm*'s in a given sperm cell. The *n* SM *dnm*'s have dosages in the paternal sperm DNA of $x_1, x_2, x_3, \ldots x_n > 0$ while the *m* SNM *dnm*'s have a dosage of 0. We have three such lists for Pr3, Pr4 and Pr5. Likewise, we may have detected *n* OM and *m* ONM *dnm*'s in a given oocyte. The *n* OM *dnm*'s have dosages in the maternal blood DNA of $x_1, x_2, x_3, \ldots x_n > 0$ while the *m* ONM *dnm*'s have a dosage of 0. We have four such lists for Pr2, Pr3, Pr4 and Pr5.

The second data set consists in lists of PM *dnm*'s and their dosage ($x_1, x_2, x_3, \ldots x_n > 0$) in sperm (Pr1 and Pr2) or blood DNA (Pr3, Pr4, Pr5).

The third data type consists in the number of *dnm*'s detected in a gamete transmitted to a proband, and the numbers of those shared by the studied half-sibs of the proband. Thus, we may have detected *n* SM and *m* SNM *dnm*'s in a sperm cell (or oocyte) transmitted to a given proband, of which half-sib 1 will share x_1, half-sib 2 x_2, ..., where $x_i \leq n$.

(ii) Computing probabilities under various models of gametogenesis: For data type 1, we simulated the process of dnm in the female and male cell lineages described in Suppl. Figure 6. For the null hypothesis, the mutation rate per cell division before birth was set at an average of 0.77 (Poisson distributed), such that the number of *dnm*'s per oocyte averaged 14 (as observed). The mutation rate per cell division after birth was set at an average of 0.3 (Poisson distributed), such that the number of *dnm*'s per sperm cell averaged 34 (as observed). For alternative hypotheses, the mutation rate for the early cleavage cell divisions (4, 7, 11, 15 and 18 first cell divisions, corresponding to the different development stages in Suppl. Figure 6) was increased 10- or 20-fold when compared to the remaining prenatal cell divisions, for which the mutation rate was concomitantly reduced such that the overall number of *dnm*'s per oocyte remained unaffected (average of 14). We further tested 4, 10 and 40 induced PGCs, and unrelated or related induced PGCs as described in Suppl. Fig. 6. For all 90 possible scenarios, we determined by simulation what proportion of *dnm*'s found in a sperm cell (respectively oocyte) were characterized by a dosage in paternal sperm (respectively maternal soma) of 0-0.05, 0.05-0.10, ... These proportions were then used as probabilities in computing the likelihood of the data (i.e. a series *dnm*'s with corresponding rate of mosaicism in the parental tissue) under the corresponding model. Given our experimental design, SM and DM are only recognized as such, if (i) their dosage in the sire (SM) or dam (DM) is significantly < 0.5, and (ii) they show complete (but imperfect) linkage in the available half-sibs. We considered a fixed number of eight half-sibs in the simulations. These conditions were included in the simulations. Thus, a mutation was only considered if it satisfied these two criteria. The dosage that was considered was not the true dosage for that mutation, but the "realized" dosage assuming a sequence depth of 24.

To compute the likelihood of the second type of data, we simulated gametogenesis in exactly the same way as for data type 1. We then randomly sampled *n* gametes, where *n* corresponds to the number of GO (hence 5 for Pr1, Pr3-5, and 11 for Pr2). For all *dnm*'s in these *n* gametes, we then determined the dosage in the germ-line (Pr1, Pr2) or soma (Pr3-5). For all 90 possible scenarios, we determined by simulation what proportion of PM *dnm*'s detected in sperm of blood DNA were characterized by a dosage of 0-0.05, 0.05-0.10, ... These proportions were then used as probabilities in computing the likelihood of the data (i.e. a series PM *dnm*'s with corresponding dosage in sperm or blood) under the corresponding model. With the real data, PM mutations are only recognized as such (i) if they are transmitted to at least one of the *n* offspring, (ii) if the proband is called heterozygous for the corresponding *dnm* by GATK, and (iii) if we demonstrate complete (but imperfect) linkage in the GO. Condition (i) is achieved in the simulation by sampling *n* gametes at random. Condition (ii) and (iii)

were modelled in the simulations. We considered five and eleven GO to match the real data. Thus a mutation was only considered if it satisfied these two criteria. The dosage that was considered was not the true dosage for that mutation, but the "realized" dosage assuming a sequence depth of 24.

For data type 3, we modified the simulations in order to exactly generate a predetermined number *n* of *dnm*'s in a given "reference" gamete. Thus if in the real data an oocyte was characterized by 11 *dnm*'s (f.i. Pr3 and Pr4), we would in the simulations distribute 11 *dnm*'s across the (7+4+4+3+18) cell divisions leading to a simulated reference oocyte and track their segregation (according to their point of occurrence) across the entire germ line lineage. Under the null hypothesis of uniform prenatal mutation rate, all 36 cell divisions would have equal chance to be hit by anyone of the 11 mutations. Under the alternative hypotheses, early cleavage cell divisions would have a 10- or 20-fold higher chance than the remaining ones. Under the hypothesis of related PGCs the segregation pattern of early mutations in the germ line lineage would be concomitantly affected (see Suppl. Fig. 6). We would then samples gametes at random from the same germ line tree and count the number of mutations shared with the "reference" gamete. This would generate a frequency distribution of gametes sharing 0, 1, 2, ... , *n dnm*'s with the reference gamete. The corresponding frequencies were then used as probabilities in computing the likelihood of the data (i.e. a series half-sibs sharing 0, 1, ... , *n dnm*'s with the reference gamete transmitted to the proband) under the corresponding model.

Likelihoods of the data under the 90 tested models were then simply computed as the product of the probabilities of all *dnm*'s (data type 1 and 2) and half-sibs (data type 3) extracted from the simulations performed under the corresponding model.

(iii) Estimating the number of missed SM and DM and misclassified PM mutations: The simulations for dataset 1 allowed us to estimate the number of SM and DM mutations missed either because the "realized" dosage was to high in the parent, or because we could not demonstrate complete (but imperfect) linkage in the half-sibs. Likewise the simulations for dataset 2 allowed us to estimate the number of PM mutations that, although detected (realized dosage sufficient to be called heterozygous by GATK), were misclassified as SNM or DNM mutations because showing perfect linkage in the available GO. Under the best biological model (20x increased mutation rate during the first 4 cell divisions, 4 related PGCs, see Table 1), these numbers were: (i) average loss of 3.3 SM mutations, (ii) average loss of 1.1 lost DM mutations, (iii) average gain of 1.45 SNM and 1.45 DNM mutations for a male proband with 11 GO, (iv) average gain of 4 SNM and 4 DNM mutations for a male proband with 5 GO, and (v) average gain of 1.8 SNM and 1.8 DNM mutations for a female proband with 5 GO.

20. Li H & Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754-1760 (2009).

21. McKenna A *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20:** 1297-1303 (2010).

22.Robinson JT et al. Integrative Genomics Viewer. *Nature Biotechnology* **29**:24–26 (2011).

23. You FM *et al.* BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* **9:**253 (2008).

24. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN]. (2012) .

Supplemental Material: available online: http://www.biorxiv.org/content/early/2017/06/29/079863

Supplemental Tables

1. **Supplemental table 1**
2. **Supplemental table 2**
3. **Supplemental table 3**

Supplemental Figures

1. **Supplemental figure 1**
2. **Supplemental figure 2**
3. **Supplemental figure 3**
4. **Supplemental figure 4**
5. **Supplemental figure 5**
6. **Supplemental figure 6**
7. **Supplemental Figure 7**
8. **Supplemental Figure 8**

Supplemental Notes

1. **Supplemental Note 1**
2. **Supplemental Note 2**

Supplemental table 1: List and features of detected germ-line dnms (*dnm*'s)

PR	SEX	Dnm's in the proband			Dnm's in the sire						Dnm's in the dam						Total
		PM			SNM			SM			DNM			DM			
		TOT	PH	MH	TOT	PH	MH	TOT	PH	MH	TOT	PH	MH	TOT	PH	MH	
PR1	M	3	2	1	31			ND			5			ND			
PR2	M	16	5	11	21			ND			13			6	3	3	
PR3	F	3	0	3	26			22(2)	10(2)	12	3			8(5)	1(1)	7(4)	
PR4	F	7	3	4	26			6(2)	4(2)	2	3			8(5)	4(1)	4(4)	
PR5	F	2	0	2	20			1	1	0	8			6	3	3	
TOTAL		31	10	21	124			29(2)	15(2)	14	32			28(5)	11(1)	17(4)	244(7)
UNIQUE		31	18	41	217			27	13	14	32			23	10	13	237
PR: Proband																	
The number of dnm's that are shared by the two full sibs, Pr3 and Pr4, are given in parentheses														96	0.278	81	0.245
PH: paternal haplotype														249		249	
MH: maternal haplotype														345		330	

Supplemental table 2: Sharing of SM and DM mutations detected in the probands with their respective half-sibs.

PR2									
DM	1	36639047	P						
	6	35745447	M						
	11	28590212	P						
	11	89059789	M						
	19	29060137	M						
	X	123543408	P						

PR3												
DM	7	48504799	M									
	11	55862690	M									
	14	74925688	M									
	15	18443426	M									
	15	69251191	M									
	16	47362937	P									
	19	831054	M									
	24	32228835	M									
SM	1	149489203	P									
	3	55978642	P									
	4	96895077	M									
	4	117704053	P									
	6	75231278	P									
	6	112271783	P									
	8	102736509	M									
	9	18384195	P									
	10	89207266	P									
	11	1792493	P									
	11	66328634	M									
	14	55743851	M									
	14	74564459	M									
	15	62205610	M									
	16	11449888	M									
	16	14940359	M									
	17	65369062	M									
	21	25583086	M									
	23	9883648	P									
	24	28669200	M									
	25	7542176	P									
	27	36201351	M									

PR4												
DM	7	48504799	M									
	7	72210494	P									
	11	55862690	M									
	13	35979690	P									
	15	69251191	M									
	16	47362937	P									
	19	735040	P									
	24	32228835	M									
SM	2	75906827	M									
	4	117704053	P									
	9	77964385	M									
	16	70310323	P									
	23	9883648	P									
	X	33501758	P									

PR5										
DM	3	75872618	M							
	11	89422735	M							
	14	41813012	M							
	15	2028399	P							
	18	13124778	P							
	22	37613114	P							
SM	9	101487292	P							

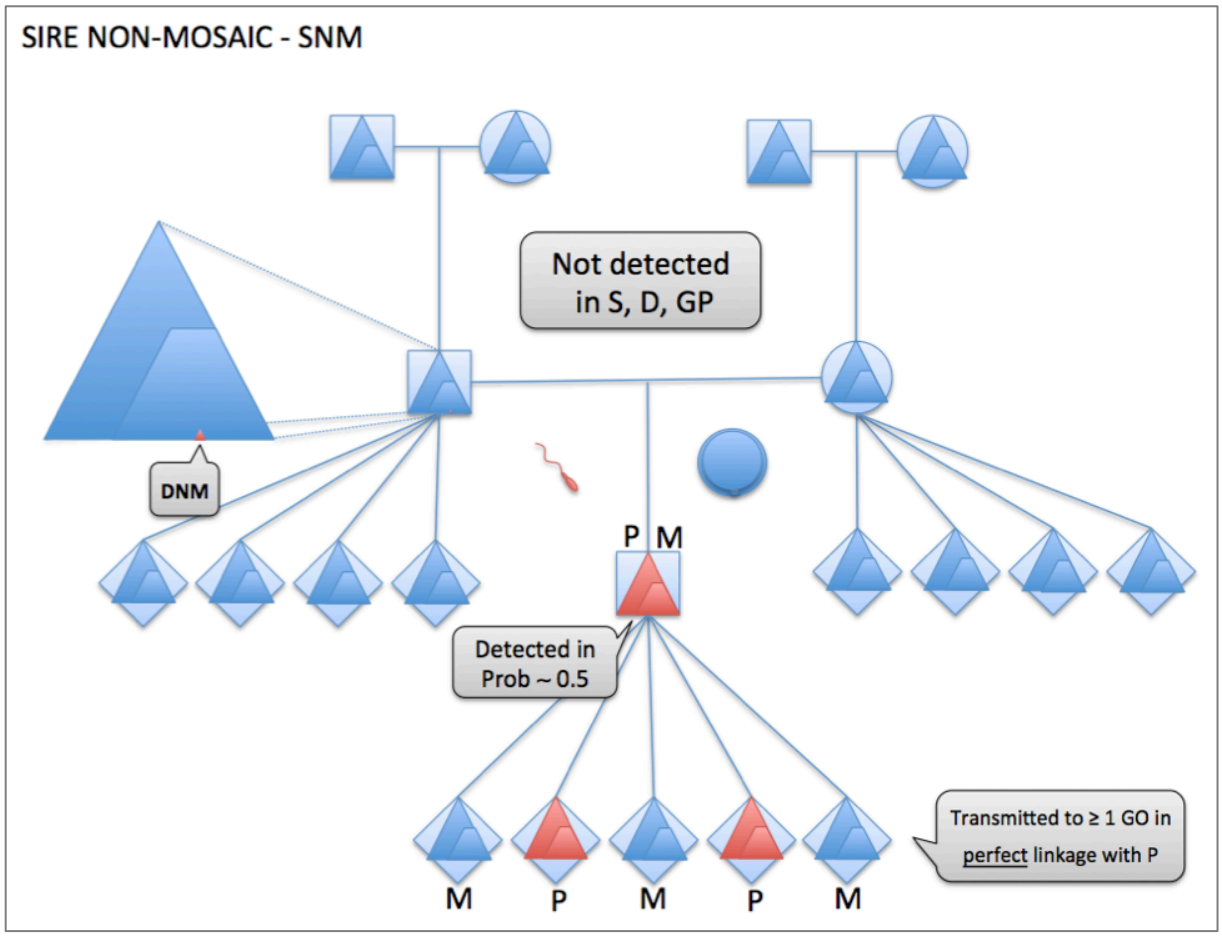
Supplemental table 3: Likelihood of (i) the degree of PM, SM and DM mosaicism and (ii) sharing of dnm's between probands and half-sibs, under different models of gametogenesis as described in suppl. fig. 6 and suppl. note 1.

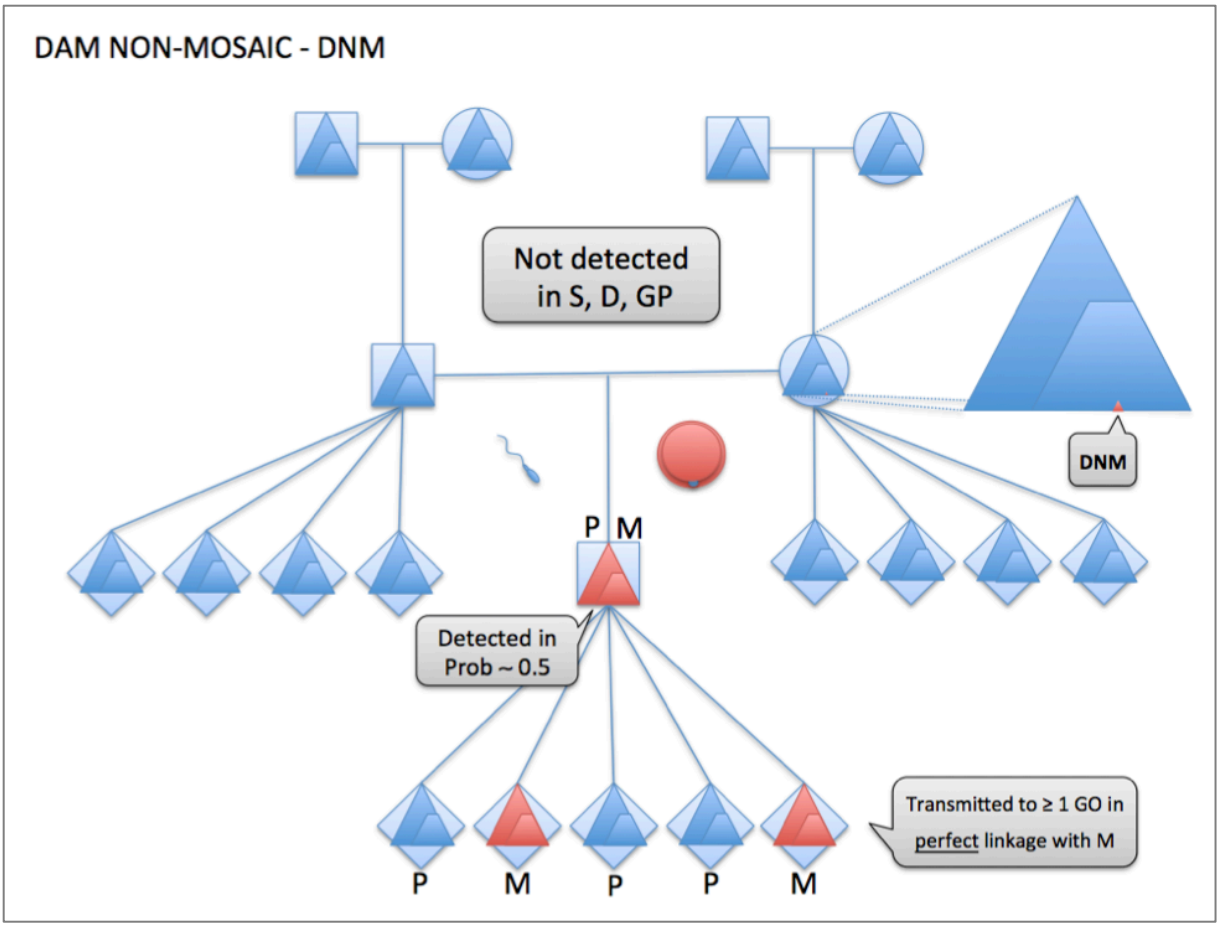
Fold increase in mutation rate	During first x cell divisions	Number of PGCs	Related PGCs or not	Log10(LR)
20x	4	4	TRUE	0.00
10x	4	4	TRUE	-0.35
20x	4	4	FALSE	-0.41
10x	4	4	FALSE	-0.47
10x	7	4	FALSE	-1.57
20x	4	10	TRUE	-1.74
10x	7	4	TRUE	-1.94
20x	7	4	TRUE	-1.94
20x	7	4	FALSE	-2.30
20x	4	10	FALSE	-2.40
20x	4	40	TRUE	-3.45
20x	7	10	TRUE	-3.47
10x	4	10	TRUE	-3.51
10x	4	10	FALSE	-4.11
20x	7	10	FALSE	-4.29
20x	4	40	FALSE	-4.49
10x	7	10	TRUE	-4.50
10x	11	4	TRUE	-5.09
20x	11	4	TRUE	-5.18
20x	11	4	FALSE	-5.57
10x	11	4	FALSE	-5.77
10x	4	40	TRUE	-5.82
10x	7	10	FALSE	-6.21
20x	7	40	TRUE	-6.39
20x	15	4	TRUE	-7.03
10x	15	4	TRUE	-7.24
10x	4	40	FALSE	-7.59
20x	11	10	TRUE	-7.78
10x	7	40	TRUE	-8.01
10x	11	10	TRUE	-8.22

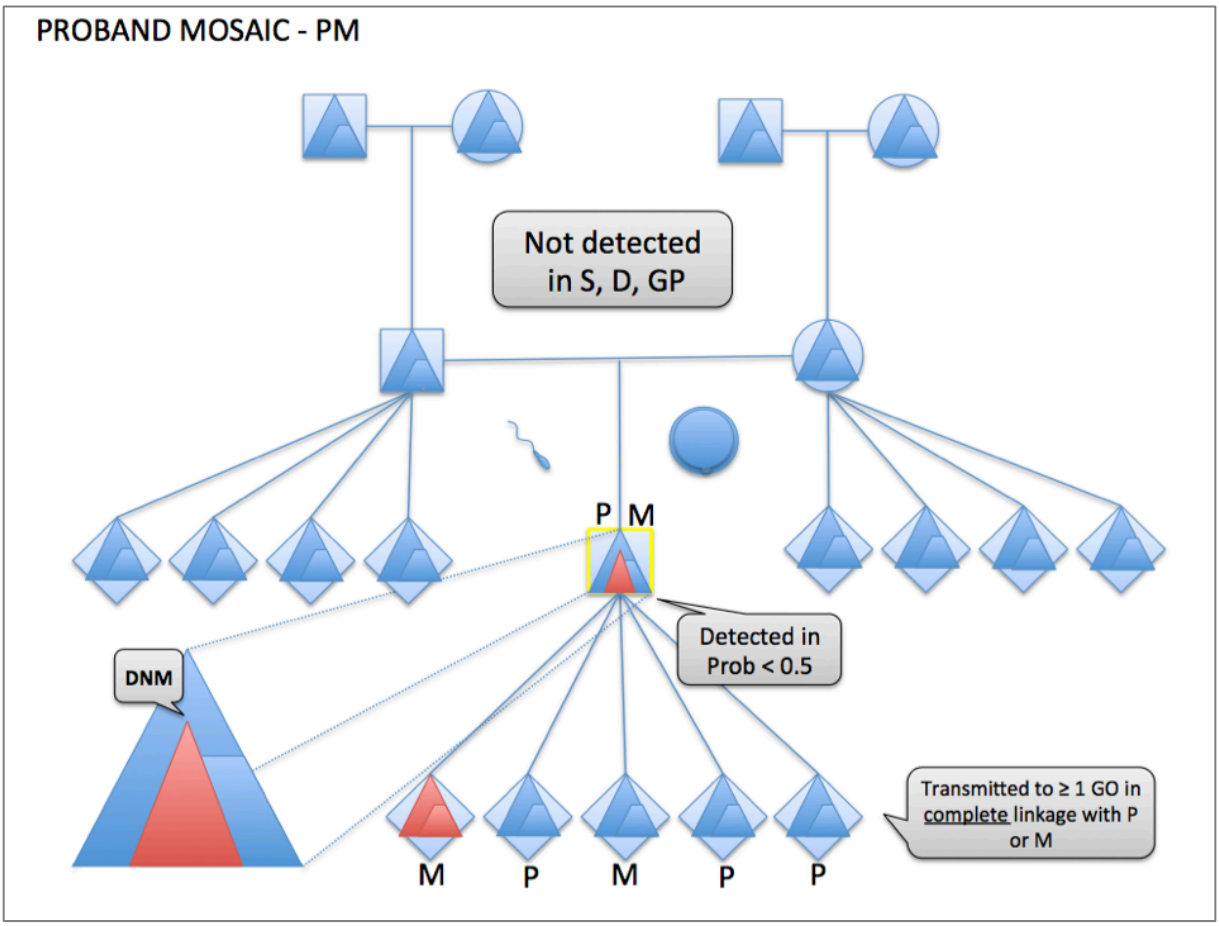
20x	7	40	FALSE	-8.68
20x	15	4	FALSE	-9.22
20x	11	10	FALSE	-9.24
10x	15	4	FALSE	-9.27
10x	18	4	TRUE	-9.47
20x	18	4	TRUE	-9.57
10x	11	10	FALSE	-10.20
10x	18	4	FALSE	-10.52
10x	7	40	FALSE	-10.55
10x	15	10	TRUE	-10.76
20x	15	10	TRUE	-10.77
20x	18	4	FALSE	-10.84
20x	11	40	TRUE	-11.69
10x	11	40	TRUE	-12.45
20x	15	10	FALSE	-13.54
10x	18	10	TRUE	-13.82
20x	18	10	TRUE	-14.09
10x	15	10	FALSE	-14.16
20x	11	40	FALSE	-14.42
20x	15	40	TRUE	-15.75
10x	11	40	FALSE	-16.10
1x	15	4	TRUE	-16.17
20x	18	10	FALSE	-16.36
10x	15	40	TRUE	-16.85
1x	7	4	TRUE	-16.91
10x	18	10	FALSE	-16.98
1x	11	4	FALSE	-17.18
1x	11	4	TRUE	-17.19
1x	4	4	FALSE	-17.29
1x	18	4	TRUE	-17.31
1x	4	4	TRUE	-17.94
1x	15	4	FALSE	-17.98
20x	18	40	TRUE	-18.08
1x	7	4	FALSE	-18.23
10x	18	40	TRUE	-18.72

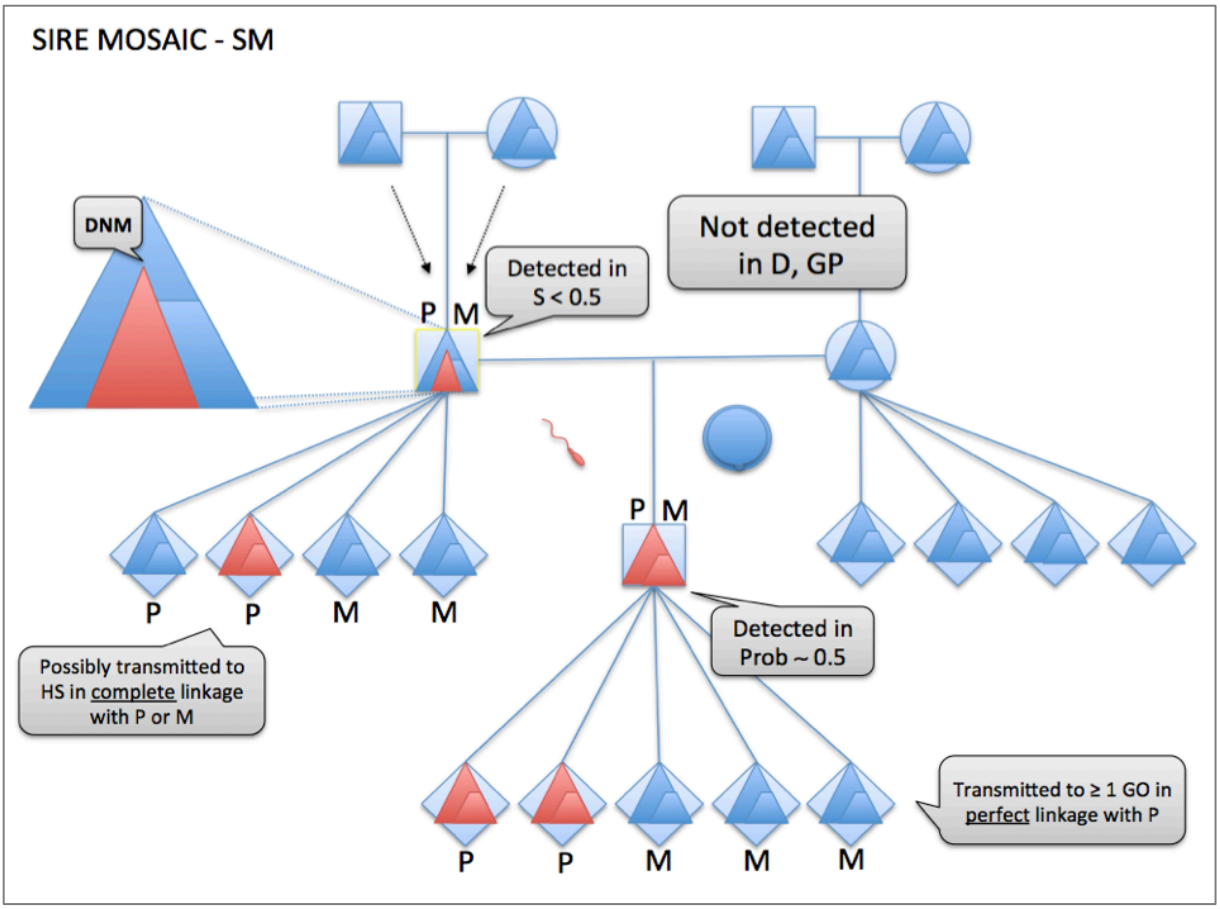
20x	15	40	FALSE	-19.22
1x	18	4	FALSE	-19.57
10x	15	40	FALSE	-20.86
10x	18	40	FALSE	-23.53
20x	18	40	FALSE	-23.56
1x	11	10	TRUE	-25.16
1x	18	10	TRUE	-25.26
1x	7	10	TRUE	-25.74
1x	4	10	TRUE	-25.89
1x	15	10	TRUE	-26.53
1x	7	10	FALSE	-28.74
1x	11	10	FALSE	-29.86
1x	4	10	FALSE	-29.93
1x	18	10	FALSE	-30.43
1x	15	10	FALSE	-31.15
1x	15	40	TRUE	-33.78
1x	18	40	TRUE	-35.76
1x	4	40	TRUE	-35.79
1x	7	40	TRUE	-35.81
1x	11	40	TRUE	-36.49
1x	4	40	FALSE	-41.91
1x	7	40	FALSE	-41.99
1x	15	40	FALSE	-42.24
1x	11	40	FALSE	-42.74
1x	18	40	FALSE	-42.79

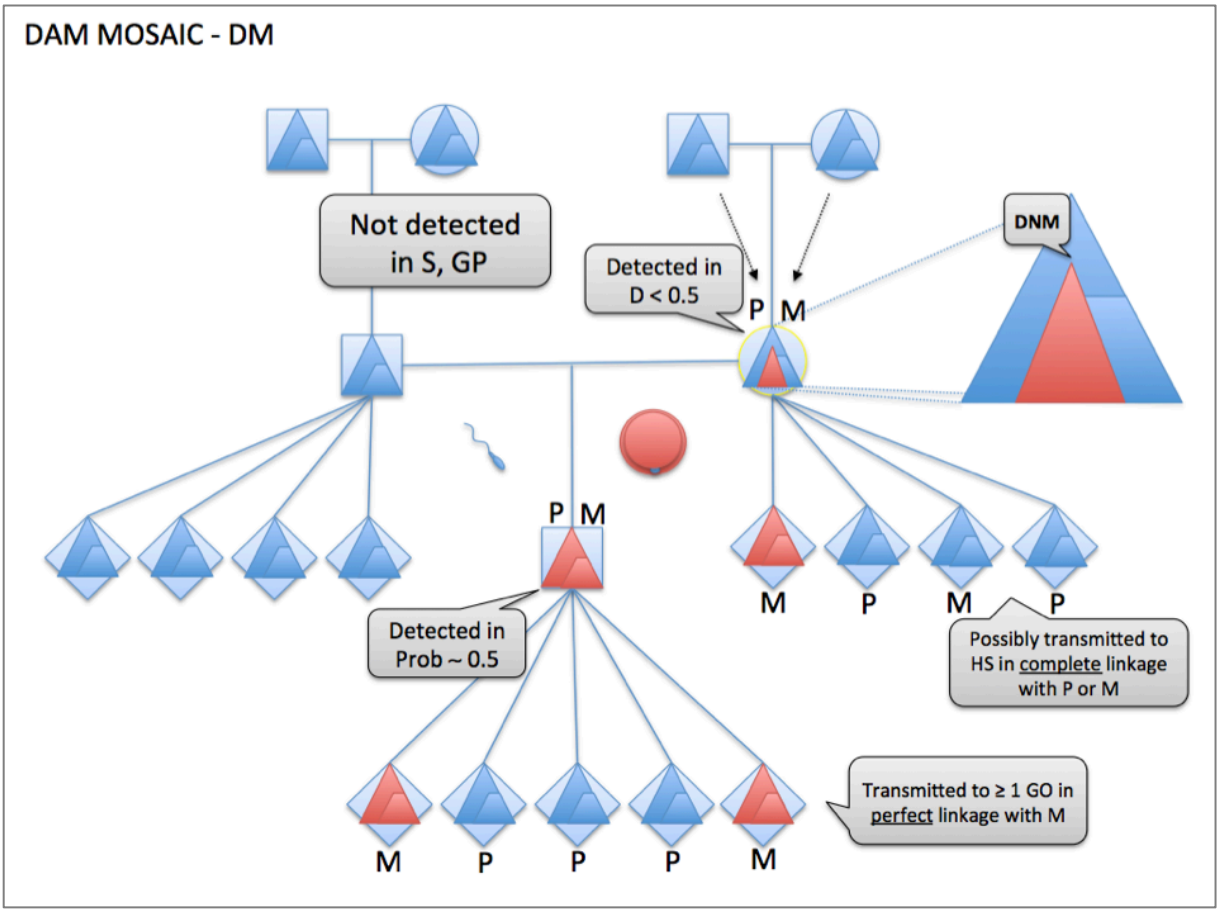
Supplemental Figure 1: Schematic representation of the five types of *dnm*'s distinguished in this work: sire non-mosaic (SNM), dam non-mosaic (DNM), proband mosaic (PM), sire mosaic (SM), and dam mosaic (DM). The individuals in the pedigree are labelled according to figure 1, i.e. GP: grandparents, S: sire, D: dam, Prob: proband, HS: half-sibs (of the proband), GO: grand-offspring. The triangles aim at illustrating the exponential increase in cell numbers by binary cell division starting from the zygote. The trapeze inside the triangle delineates the germ line, the rest constituting the soma. Cells carrying a *dnm* are labelled in red. P and M refer to the paternal and maternal origin of the haplotypes at the site of the *dnm*. The call-outs highlights the conditions used in this study to categorize *dnm*. Individuals with detectable mosaicism levels for the *dnm* are lined in yellow. By *perfect* linkage we mean that the *dnm* is either (i) <u>always</u> transmitted with the paternal, never with the maternal haplotype, or (ii) <u>always</u> transmitted with the maternal , never with the paternal haplotype. By *complete* (but imperfect) linkage, we mean that the *dnm* is either (i) <u>sometimes</u> transmitted with the paternal, never with the maternal haplotype, or (ii) <u>sometimes</u> transmitted with the maternal, never with the paternal haplotype.




DAM NON-MOSAIC - DNM
Not detected in S, D, GP
DNM
P M
Detected in Prob ~ 0.5
P M P P M
Transmitted to ≥ 1 GO in perfect linkage with M

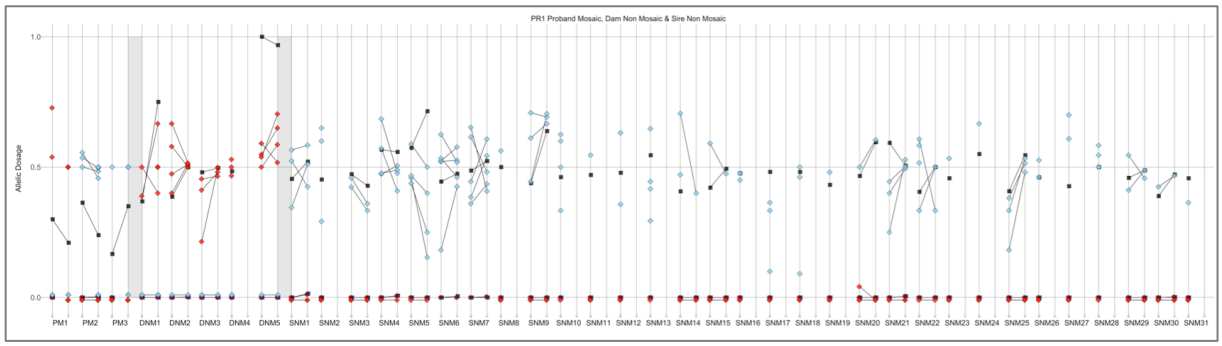

PROBAND MOSAIC - PM
Not detected in S, D, GP
P M
Detected in Prob < 0.5
DNM
M P M P P
Transmitted to ≥ 1 GO in complete linkage with P or M


SIRE MOSAIC - SM
DNM
Detected in
S < 0.5
P M
Not detected
in D, GP
P M
P P M M
Possibly transmitted to
HS in complete linkage
with P or M
Detected in
Prob ~ 0.5
P P M M M
Transmitted to ≥ 1 GO in
perfect linkage with P

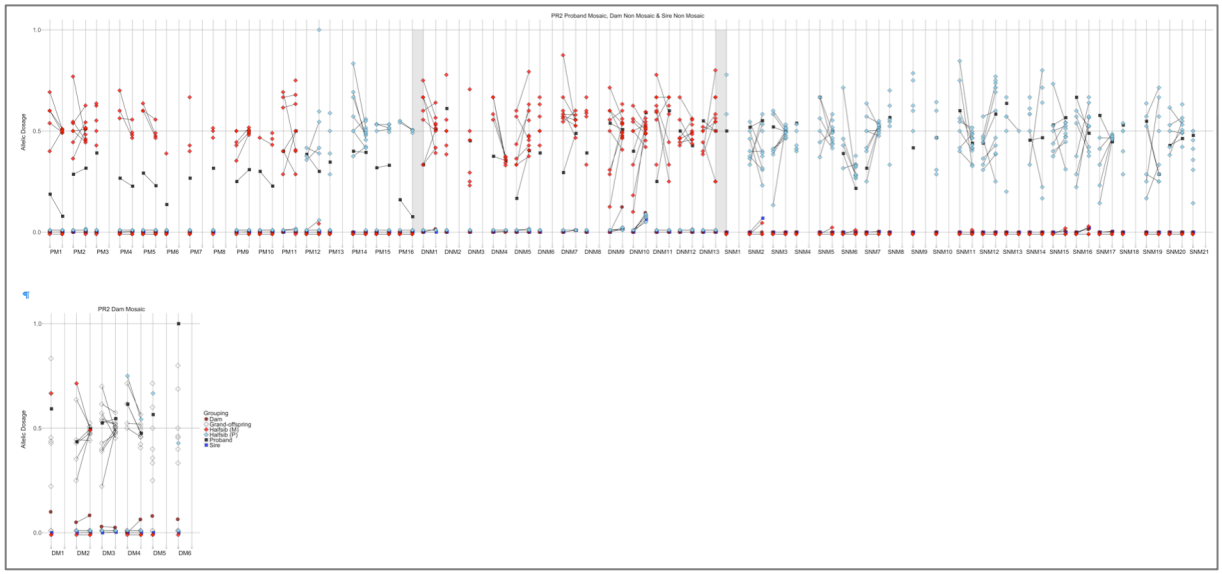

DAM MOSAIC - DM
Not detected
in S, GP
Detected in
D < 0.5
P M
DNM
P M
M P M P
Detected in
Prob ~ 0.5
Possibly transmitted to
HS in complete linkage
with P or M
M P P P M
Transmitted to ≥ 1 GO in
perfect linkage with M

Supplemental figure 2: Allelic dosage in relevant individuals for all detected *dnm*'s sorted by proband (Proband 1, Proband 2, Proband 3, Proband 4, Proband 5) and class of *dnm* within proband (upper band: PM, DNM, SNM; lower band: SM, DM). Each vertical line corresponds to a specific *dnm*. When two adjacent vertical lines are connected, the first correspond to the data from whole genome sequencing, the second to the targeted confirmation data. The mutations are labelled at the bottom of the graphs according to Supplemental Table 1. **Upper band (PM, DNM, SNM):** Black symbol: proband (circle: female; square: male); brown circle: dam; blue square: sire; red diamonds: grand-offspring inheriting the maternal haplotype of the proband at the mutation site; blue diamonds: grand-offspring inheriting the paternal haplotype of the proband at the mutation site. **Lower band (SM, DM):** Black symbol: proband (circle: female; square: male); brown circle: dam; blue square: sire; red diamonds: half-sibs (of the proband) inheriting the maternal haplotype of the sire (SM) or dam (DM) at the mutation site; blue diamonds: half-sibs (of the proband) inheriting the paternal haplotype of the sire (SM) or dam (DM) at the mutation site; white diamonds: grand-offspring.

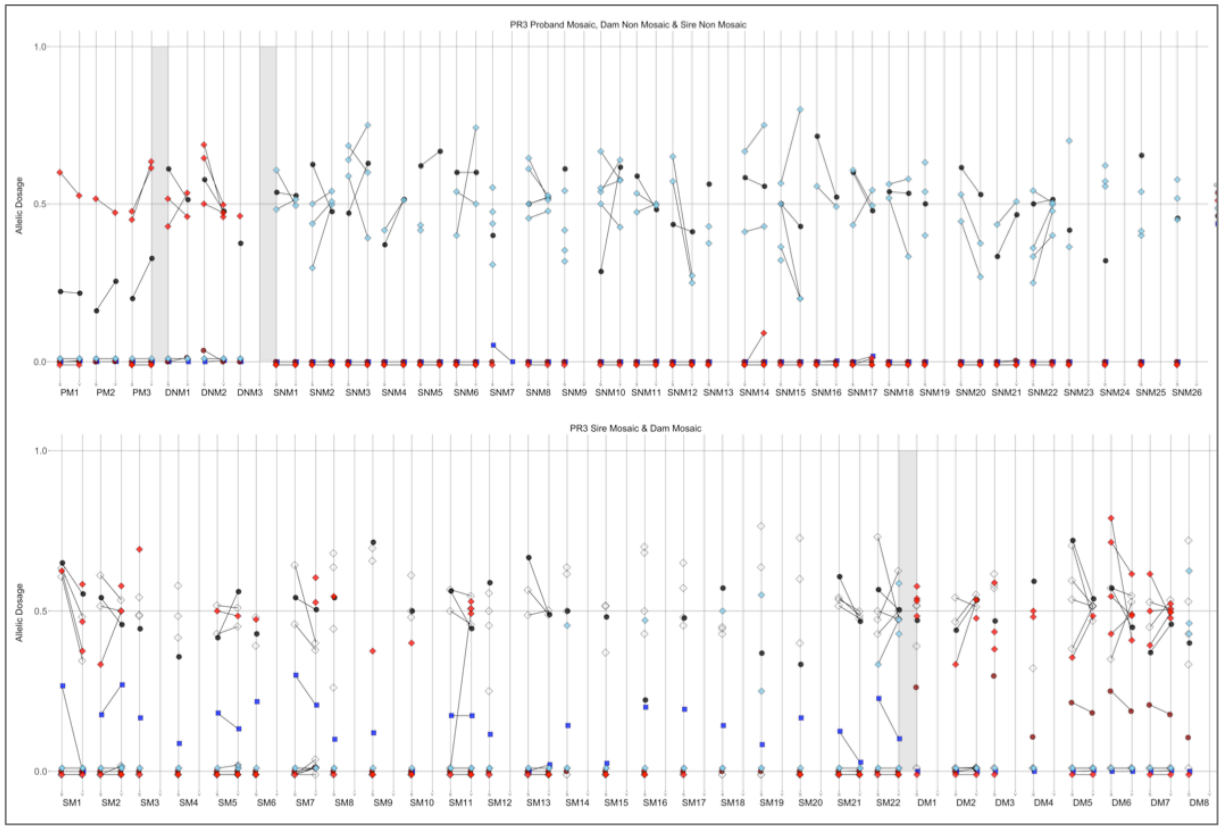
PROBAND 1



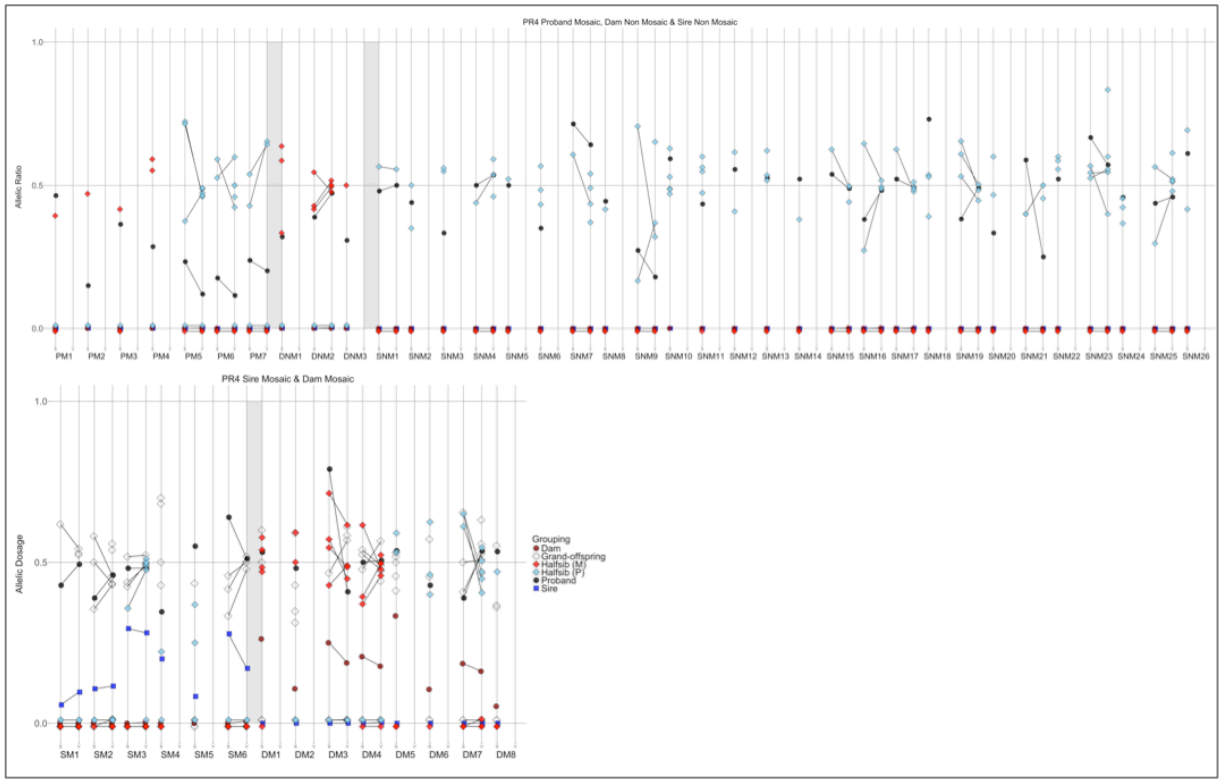
PROBAND 2



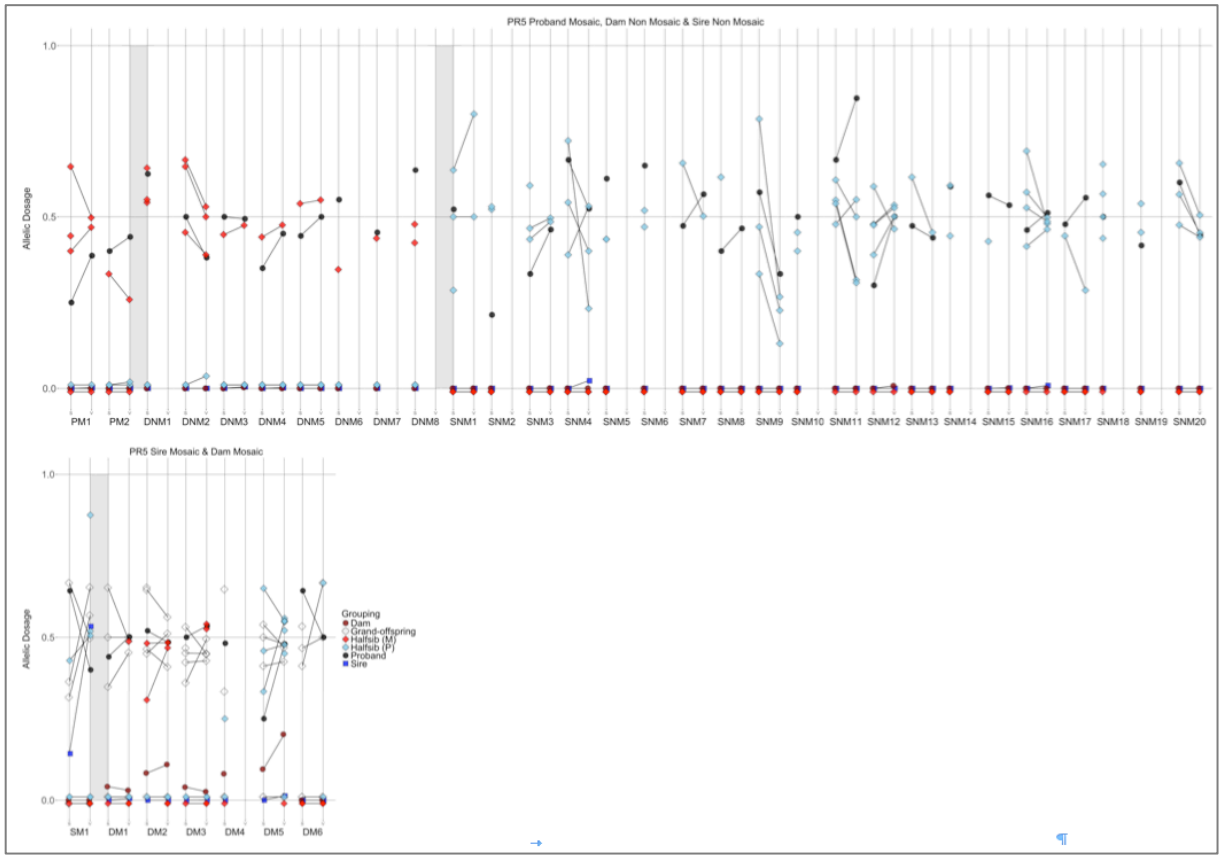
PROBAND 3



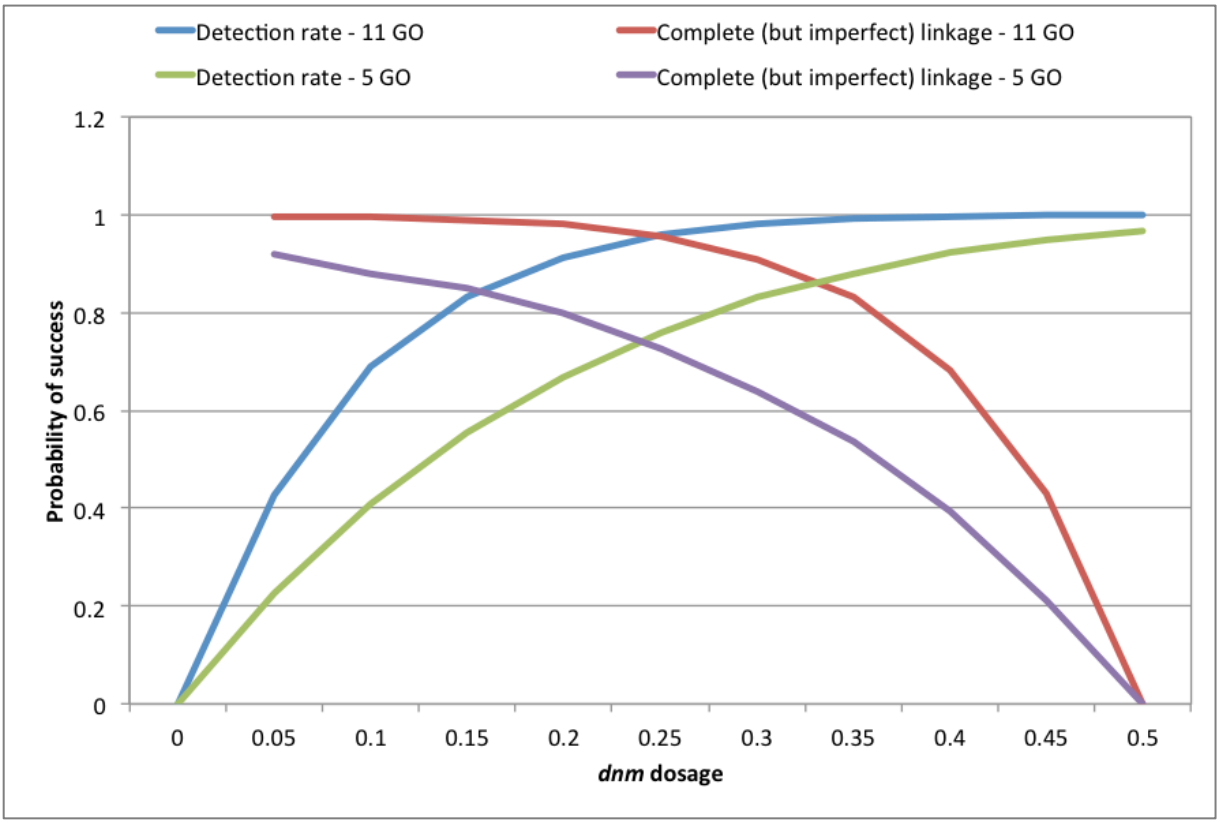
PROBAND 4



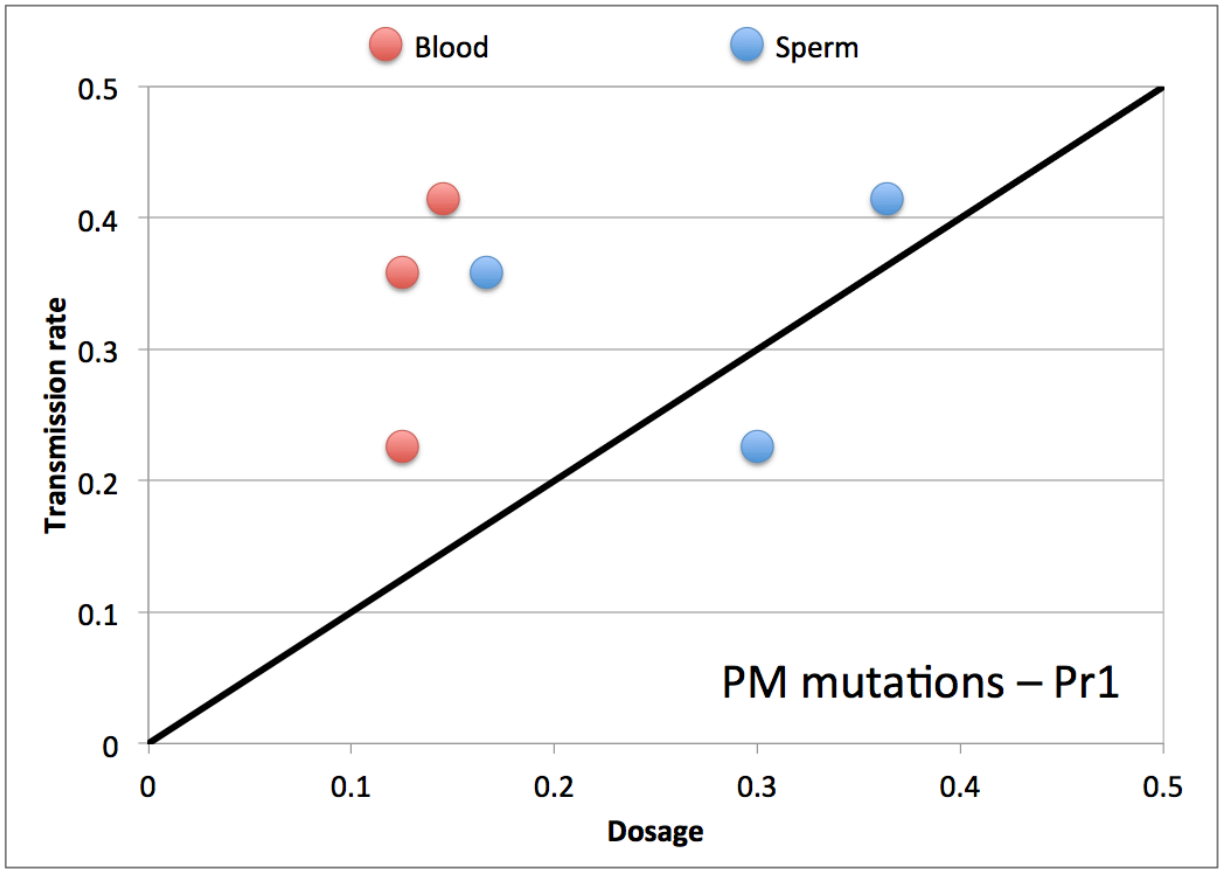
PROBAND 5


PR5 Proband Mosaic, Dam Non Mosaic & Sire Non Mosaic
Allelic Dosage
1.0
0.5
0.0
PM1 PM2 DNM1 DNM2 DNM3 DNM4 DNM5 DNM6 DNM7 DNM8 SNM1 SNM2 SNM3 SNM4 SNM5 SNM6 SNM7 SNM8 SNM9 SNM10 SNM11 SNM12 SNM13 SNM14 SNM15 SNM16 SNM17 SNM18 SNM19 SNM20
PR5 Sire Mosaic & Dam Mosaic
Allelic Dosage
1.0
0.5
0.0
SM1 DM1 DM2 DM3 DM4 DM5 DM6
Grouping
Dam
Grand-offspring
Halfsib (M)
Halfsib (P)
Proband
Sire

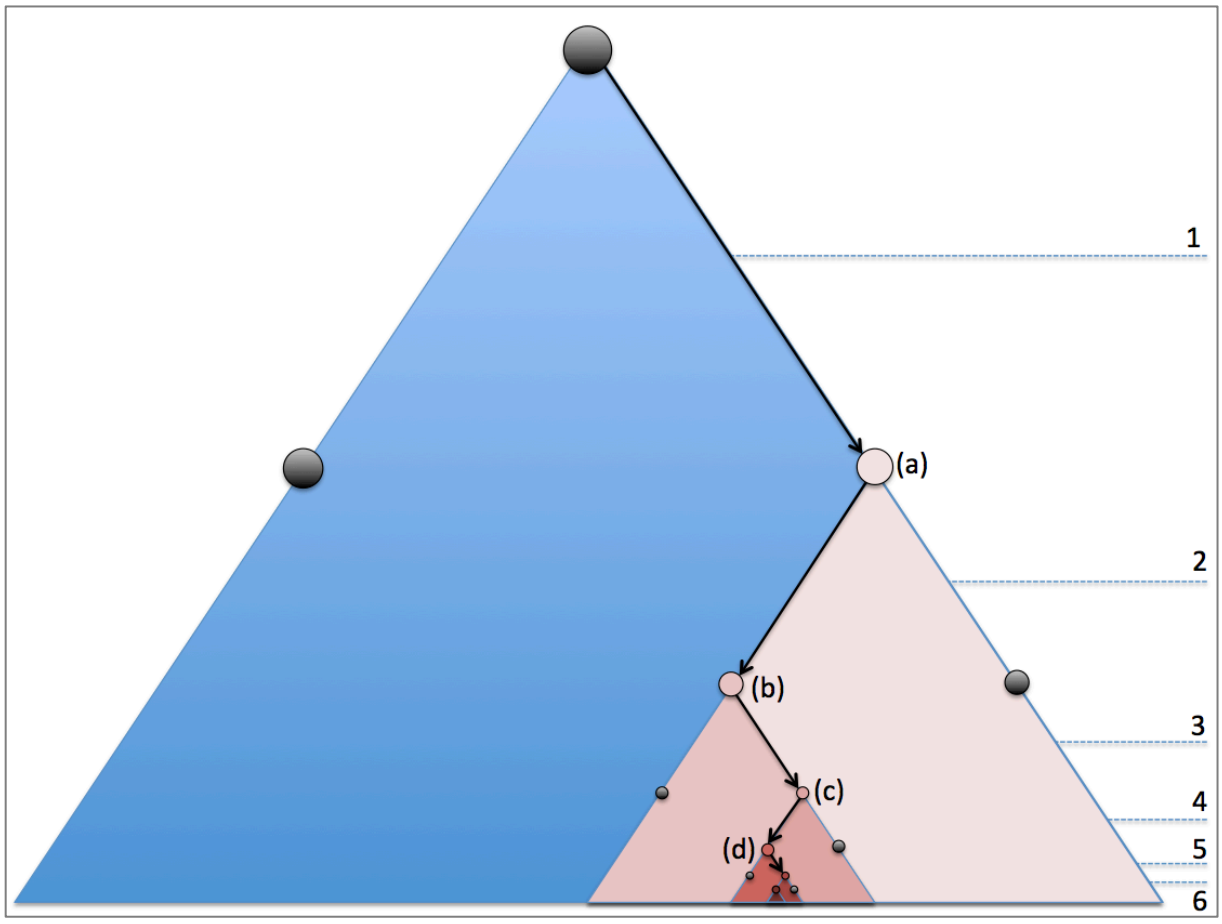
Supplemental figure 3: Probability (i) for a *dnm* to be transmitted to at least one (and hence being detected) of 5 or 11 grand-offspring, and (ii) to obtain evidence for complete (but imperfect) linkage in at least one of 5 or 11 grand-offspring, as a function of the *dnm* dosage in the proband. Probabilities were estimated by simulation (n=10,000).



Supplemental figure 4: Relationship between the dosage of three PM *dnm*'s detected in blood (red) and sperm (blue) of proband 1, and their transmission rate to the 53 corresponding grand-offspring. The average difference (absolute value) between dosage-based prediction and realized transmission rate was 0.1 for sperm and 0.2 for blood.



Supplemental figure 5: We observe (i) that sires are detectably mosaic for ~30% of *dnm's* present in a given sperm cell, while dams are detectably mosaic for ~50% of *dnm's* present in an oocyte, (ii) that >60% of half-sibs share at least one *dnm* with the proband, and (iii) half-sibs sharing multiple *dnm's* with the proband are surprisingly common. All these observations point towards a higher mutation rate during the early embryonic cell divisions. Why?



Schematic representation of a hypothetical germ line genealogy expanding by binary cell division from a founder cell. The first six cell divisions of a specific cell lineage (assumed to result in the gamete transmitted to the proband) are marked by arrows. The graph shows how a *dnm* occurring during the first cell division (a) will be shared by 50% of the gametes, during the second cell division (b) by 25% of the gametes, during the third cell division (c) by 12.5%, until after ~5 cell divisions the dosage of the *dnm* becomes so low (~3%) that it becomes practically undetectable.

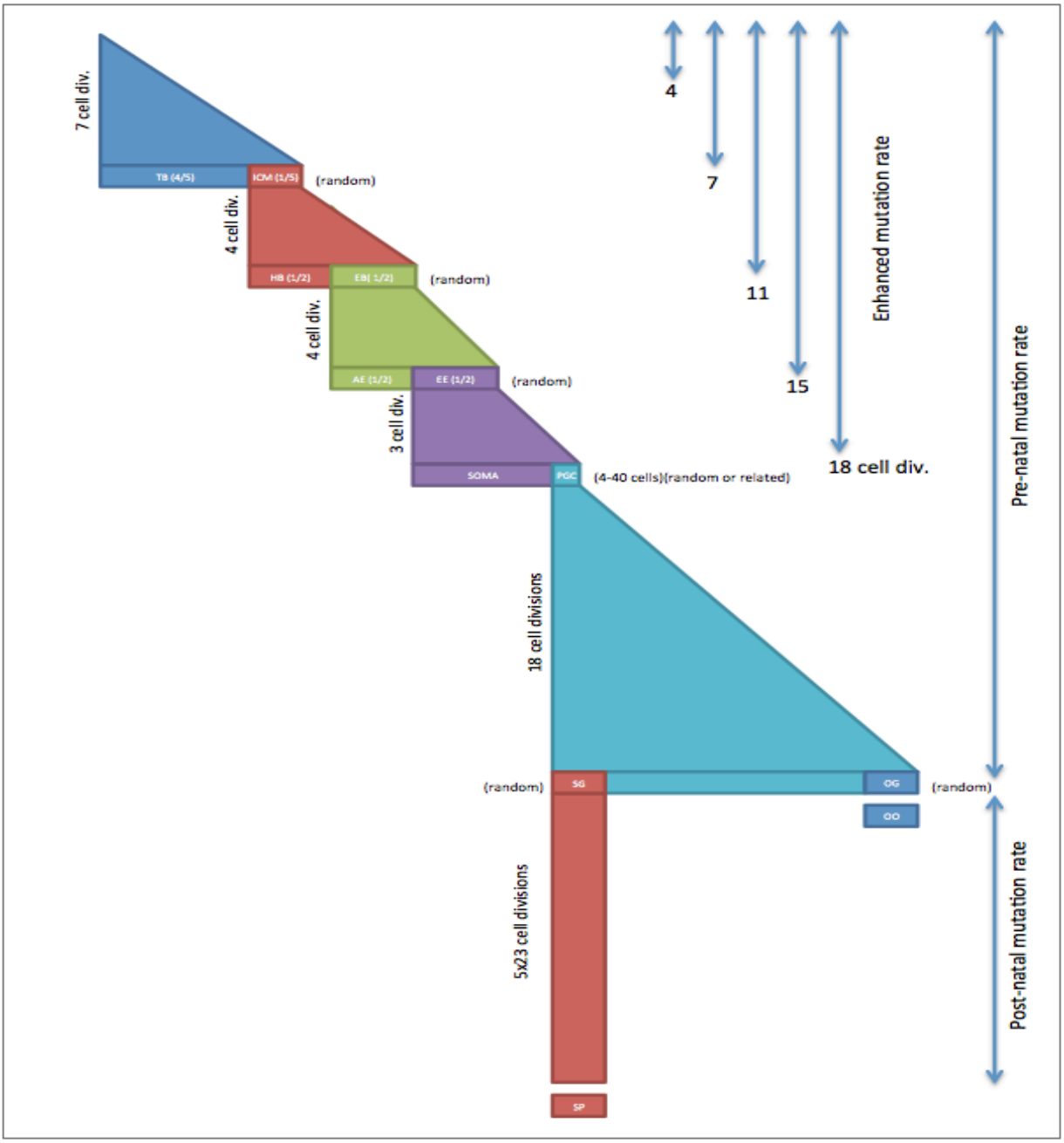
Assuming that the number of cell divisions between the zygote and an oocyte is ~22, one expects ~5/22=0.23 of *dnm* to be detectably mosaic in a female, hence considerably lower than what we observe (~50%). Assuming that the number of cell divisions between the zygote and a sperm cell is ~200, one expects under the same assumptions ~5/200=0.025 of *dnm* to be detectably mosaic in a male, hence much lower than what we observe (~30%).

The same germ line genealogy also helps us to understand why we a priori expect that the majority of half-sibs should share no *dnm* with the proband, and why sharing of multiple *dnm*'s should be exceptional, while the contrary is observed.

The easiest explanation of these findings is that founder cell(s) shared by soma and germ line have accumulated a substantial number of *dnm's* during mutation-prone early cleavage cell divisions, which will hence be shared by a large proportion of cell lineages contributing to the soma and germ line, accounting for the high degree of mosaicism and sharing of *dnm's* between sibs.

Supplementary Figure 6 re-evaluates these properties in the context of more realistic cell line genealogies.

Supplementary figure 6: Stages, with corresponding number of cell divisions, in the germ cell lineage as used in the simulations. TB: trophoblast. ICM: inner cell mass. HB: hypoblast. EB: epiblast. AE: amniotic ectoderm. EE: embryonic epiblast. PGC: primordial germ cells. OG: oogonia. OO: oocytes. SG: spermatogionia. SP: sperm cells. Numbers in parentheses refers to the fraction or number of select cells. (random) refers to random cell selection (out of the total number of cells). (related) refers to the selection of ontogenetically related cells. The basic pre-natal mutation rate was set at 0.77 on average per cell division (Poisson distributed) such that the number of dnms per oocyte matched the observations (~14).



When enhancing the mutation rate for the early cell divisions (4, 7, 11, 15 or 18), that of the remaining pre-natal cell divisions was concomitantly reduced to maintain the average number of dnms per oocyte. The male pre-natal mutation rate per cell division was assumed to be identical to the female pre-natal mutation rate per cell division. The post-natal mutation rate (applying exclusively to males) was set at an average of 0.3 per cell division (Poisson distributed) such that the total number of dnms per sperm

cells matched the observations (~34). Only one daughter cell was maintained for the post-natal cell divisions. OO and SP were produced by "meiosis" consisting in keeping the dnms in the corresponding OG and SG with 50% probability. To obtain "related" PGCs, we restricted the sampling to a "sector" of 133 (= 1% of the total) adjacent cells in the EE genealogy. The number of cell divisions for stage I (7) were from Soom et al. (1997), stage II (4), stage III (4) and stage IV (4) from McLaren and Lawson (2005), McLaren (2003), Gilbert (2000), Zheng et al. (2005), stage V (18) assuming a total number of gametogonia of 3-10 million from Mamsen et al. (2011), stage VI (5x23) from Drost and Lee (1995) and assuming an age of 5 years for the sires.

Drost JB, Lee WR. (1995). Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. *Environmental and molecular mutagenesis* **25**: 48–64.

Gilbert SF. Developmental Biology. 6th edition. Sunderland (MA): Sinauer Associates; (2000). Available from: http://www.ncbi.nlm.nih.gov/books/NBK9983/

Mamsen LS, Lutterodt MC, Andersen EW, Byskov AG, Andersen CY. (2011). Germ cell numbers in human embryonic and fetal gonads during the first two trimesters of pregnancy: analysis of six published studies. *Human Reproduction* **26**: 2140–2145.

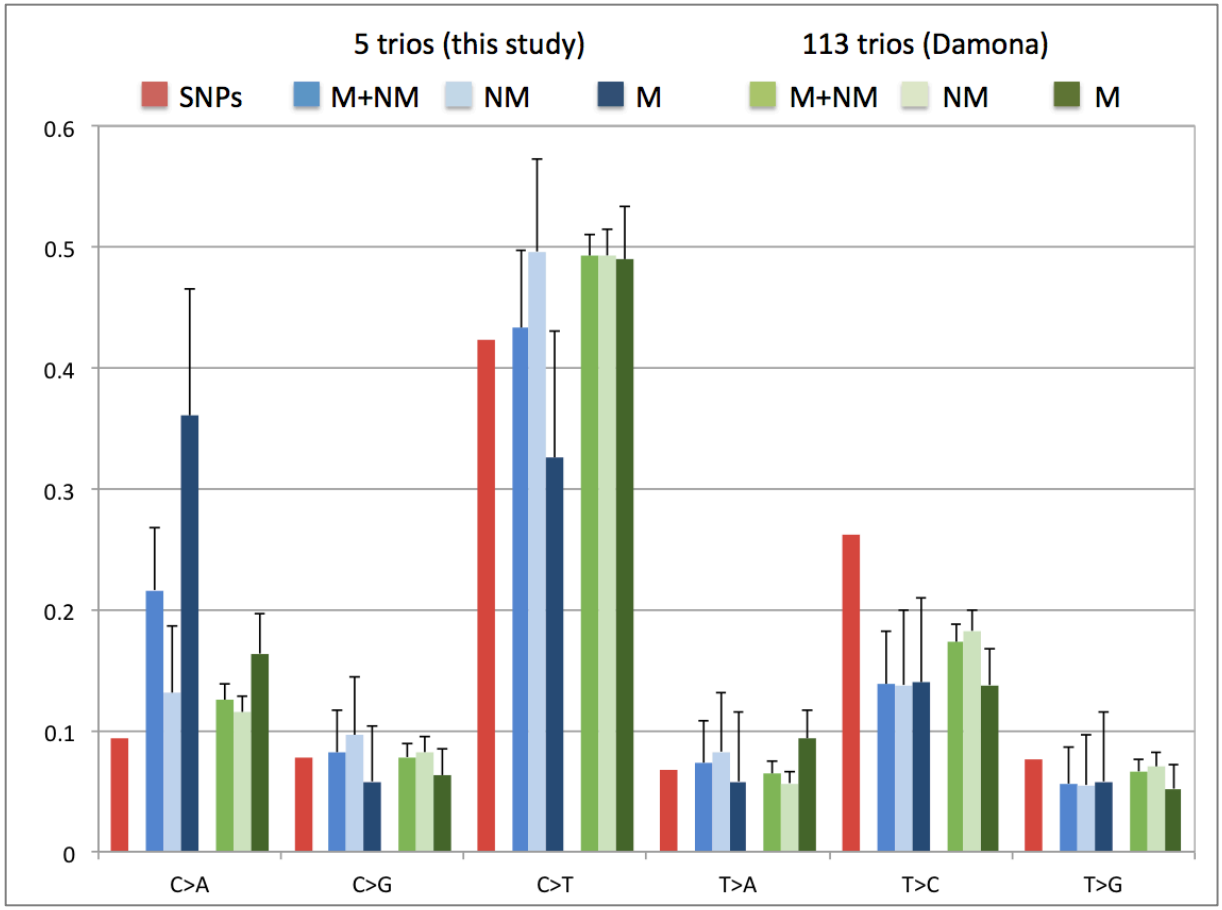
McLaren A. (2003). Primordial germ cells in the mouse. *Developmental Biology* **262**: 1–15.

McLaren A, Lawson KA. (2005). How is the mouse germ-cell lineage established? *Differentiation* **73**: 435–437.

Soom AV, Boerjan ML, Bols PE, Vanroose G, Lein A, Coryn M, de Kruif A. (1997). Timing of compaction and inner cell allocation in bovine embryos produced in vivo after superovulation. *Biology of Reproduction* **57**: 1041–1049.

Zheng et al. On the number of founding germ cells in humans. *Theoretical Biology and Medical Modelling* **2**:32 (2005).

Supplemental figure 7: Comparison of the proportion of the six possible nucleotide substitutions for (i) the 220 de novo single nucleotide substitutions detected in this study (5 probands) (*blue*), (ii) 2,530 candidate de novo single nucleotide substitutions detected in the full Damona dataset (113 other probands) (*green*), and (iii) 613,971 SNPs segregating in the Holstein-Friesian population at frequencies ≤ 0.01 (*red*). For SNPs, the rare allele was considered to be the derived allele. M: mutations with detectable mosaicism (i.e. PM, SM and DM). NM: mutations without detectable mosaicism (i.e. SNM, DNM). Error bars correspond to empirically determined upper limits of the 95% CI. The Ti/Tv ratio was 2.2 for the SNPs, 1.99 for the *dnm*'s in the full Damona dataset, and 1.33 for the 220 *dnm*'s in the present study.
Dnm's are characterized by an excess of C>A/G>T substitutions which is mainly driven by the mutations occurring early in development (PM, SM and DM).



These are unlikely to be artifacts for the following reasons:

(i) we detected a total of 50 C>A transversions. 31 of these were subjected to targeted resequencing for conformation. All 31/31 proved to be genuine in this confirmation experiment,

(ii) all reported C>A transversions were showing Mendelian transmission with perfect (PM in GO, SM in HS, DM in HS) or complete (SNM in GO, DNM in GO, SM in GO, DM in GO) linkage with the cognate parental haplotype,

(iii) the dosage

(iv) early C>A transversions were specifically enriched in the GpCpA and TpCpT trinucleotide contexts, which is exactly the same signature as recently reported for embryonic mutations in humans[1],

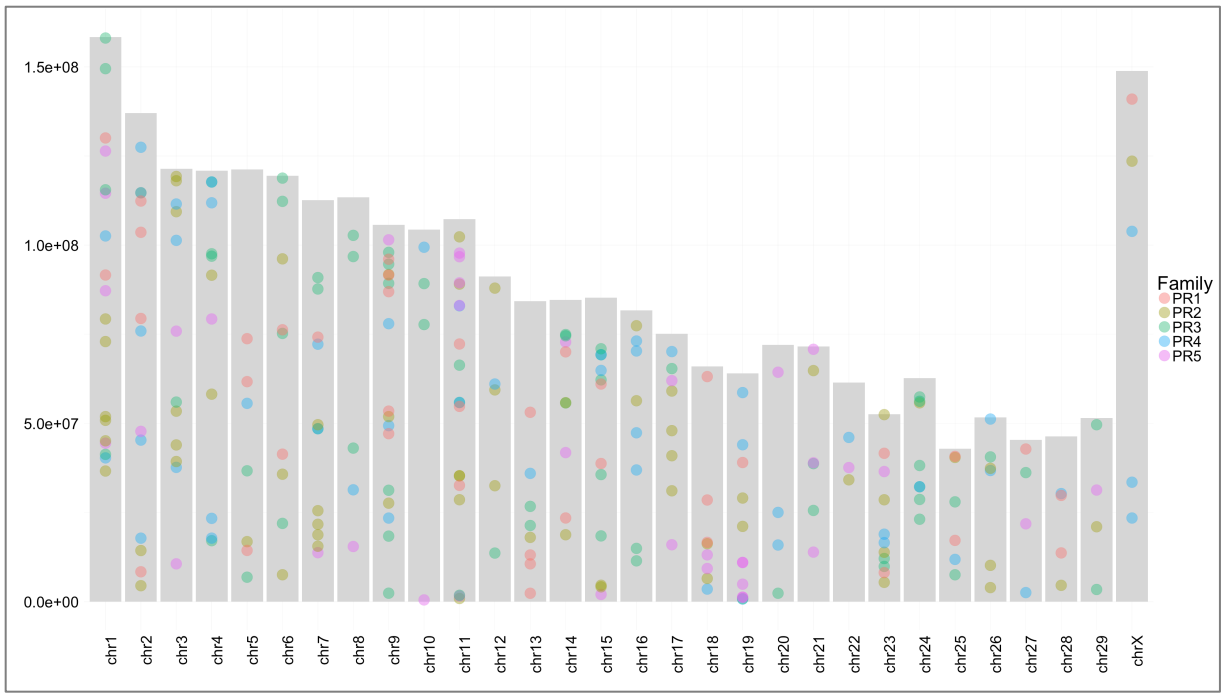
(v) only 1/50 reported C>A transversions (0/31 when considering early *dnm's* alone) occurred in the CpCpG context in which shearing-induced oxidation to 8-oxoguanine is known to preferentially occur[2],

(vi) there was no evidence for a bias towards C>A in our bioinformatics pipeline (see M&M).

1. Ju YS et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**: 714-718 (2017).

2. Costello M et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research* **41**: e67 (2013).

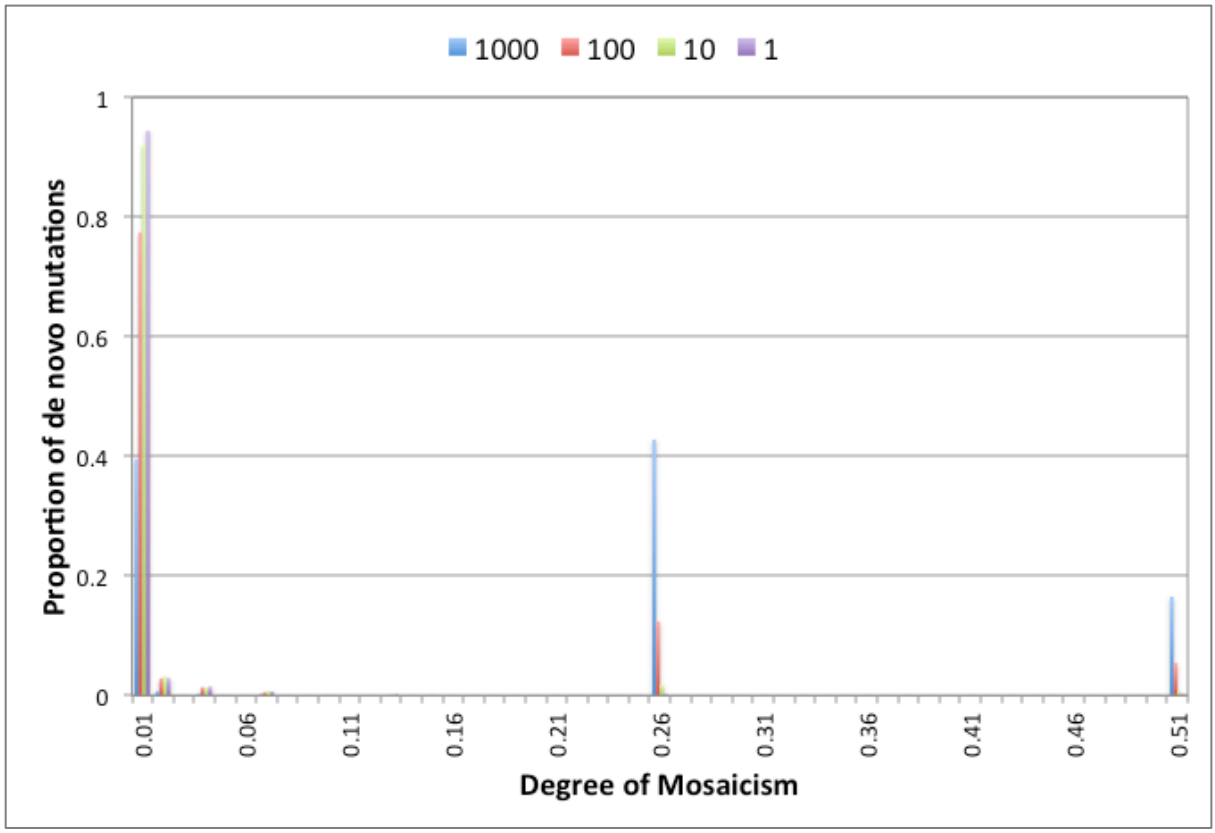
Supplemental figure 8: Chromosomal distribution of detected *dnm*'s by proband. Of the 237 *dnm*'s, one was a missense variant, one a synonymous variant, one an exonic variant in a non-coding transcript, 2 3'UTR variants, 13 upstream variants (within 5Kb upstream of a gene), 10 downstream variants, 93 intronic and the remaining intergenic.

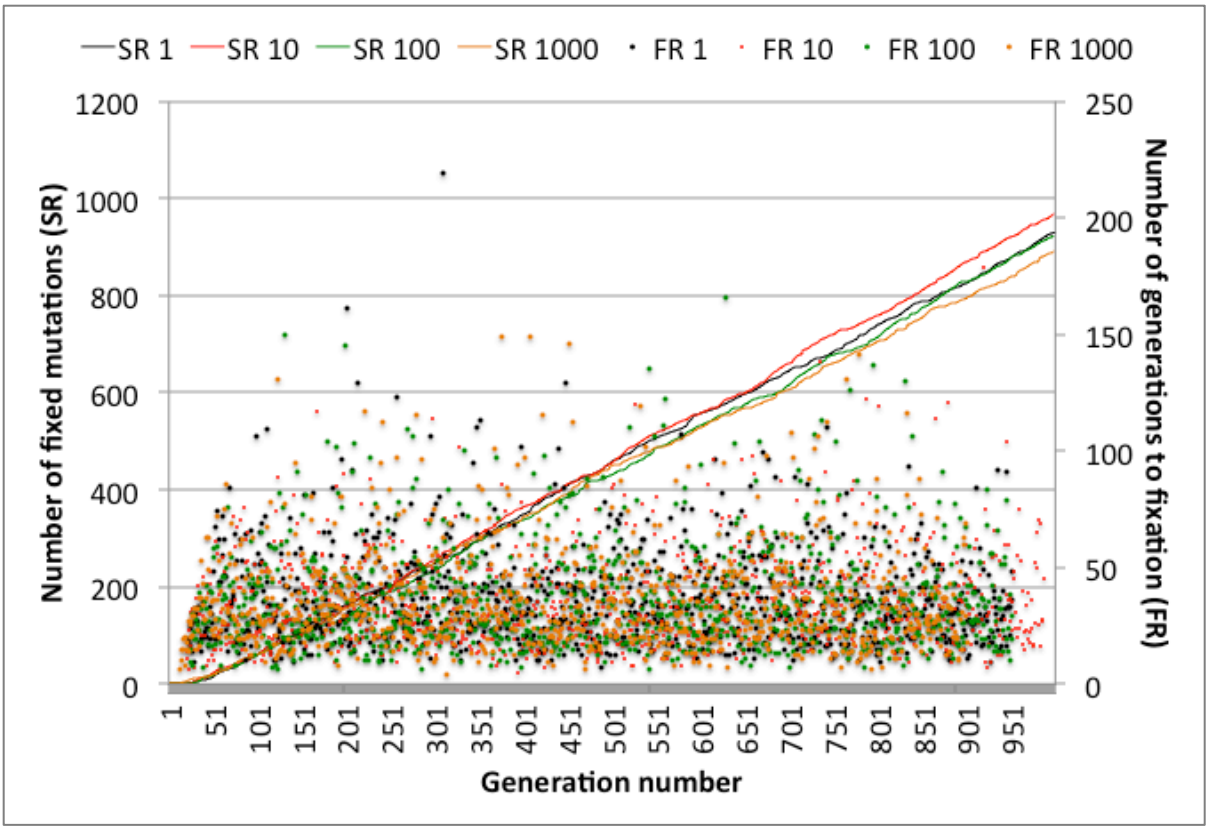


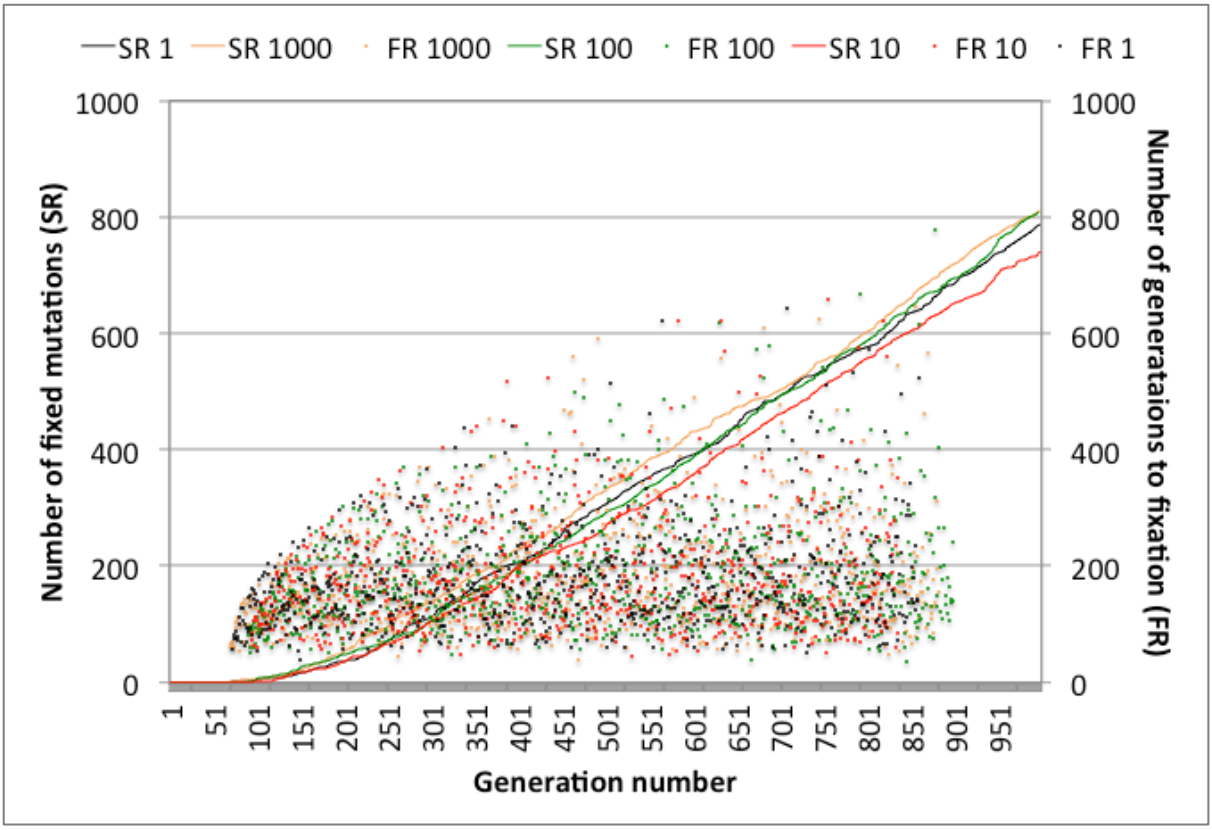
Supplemental figure 9: To evaluate the possible effect of mosaicism for dnms (*dnm*'s) on the substitution rate and time to fixation, we simulated the fate of *dnm's* under random drift in panmictic population with different effective population size (assuming equal numbers of males and females), and with various degrees of mosaicism. The mutation rate was set at one *dnm* per gamete on average. The germ line was assumed to correspond to 10 successive binary cell divisions. The mutation rate of the first two cell divisions was allowed to be the same, 10-fold, 100-fold and 1000-fold larger than for the remaining cell divisions.

A. Proportion of *dnm*'s with given degree of mosaicism in the germ-line under 4 scenarios: mutation rate of the first two cell divisions 1-fold, 10-fold, 100-fold and 1000-fold larger than for the remaining cell divisions. One can for instance see that if the mutation rate is 1000-fold larger during the first two cell divisions, the proportion of *dnm*'s with mosaicism rate of 0.50 is ~20% and of 0.25 is 40%. When the mutation rate per cell division is uniform, nearly all *dnm*'s have a mosaicism rate < 0.01.

B. Fixation rate and time to fixation of *dnm*'s in population with size 10. The lines correspond to the accumulation of fixed *dnm*'s in the population under various rates of mosaicism (SR 1, 10, 100, 1000 as in A). The dots correspond to the number of generations to fixation for individual fixed *dnm*'s under the same scenarios (FR 1, 10, 100, 1000 as in A). The fixation rate is very close to 1 per generation (corresponding to the mutation rate per gamete), and a priori independent of the rate of mosaicism. The time to fixation averaged ~37 under the four scenarios, i.e. close to 4xNe as expected.




SR 1
SR 10
SR 100
SR 1000
FR 1
FR 10
FR 100
FR 1000
Number of fixed mutations (SR)
Number of generations to fixation (FR)
Generation number
0
200
400
600
800
1000
1200
0
50
100
150
200
250
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951


SR 1
SR 1000
FR 1000
SR 100
FR 100
SR 10
FR 10
FR 1
Number of fixed mutations (SR)
Number of generataions to fixation (FR)
Generation number
0
200
400
600
800
1000
0
200
400
600
800
1000
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951

Supplemental Note 1: Gametogenesis simulations

Software, written in Scala 2.11.
Gametogenesis Simulator: Simulates de novo mutation within a full gametogenesis model when given a specific mutation rates for both early and late stages of development.
https://github.com/aeonsim/code-git/blob/master/scala-apps/simGametogensisAdvanced.scala

Sharing simulator: Simulates sharing of variants within between gamete when given a known number of de novos for embryo development and gametogenesis.
https://github.com/aeonsim/code-git/blob/master/scala-apps/simSharing_denovos.scala

Likelihoods calculator: Takes input from the gametogenesis simulator and an input of known de novos, adjusts the simulated dataset based on the likelihood of actually detecting the de novo mutations as a function of the number of grand-offspring utilised. And the probability of missing a de novo mutation with the specified allelic dosage based on random sampling of reads from a binomial distribution to simulate sequencing.
https://github.com/aeonsim/code-git/blob/master/scala-apps/likelihoodsDevel_Advanced.scala

For our simulation software we developed a model of embryo development and gametogenesis based on the bovine, human and mouse literature. The model composes of 6 stages, formation of the inner cell mass, formation of the epiblast, formation of the embryonic epiblast, PGC selection, reduction in PGCs and post puberty spermatogenesis and is shown in Suppl. Fig. 6. To the pre-puberty stages we assigned 36 cell divisions, while for the post-puberty spermatogenesis we assigned 23 cell divisions per year (human estimate). The 36 cell divisions were assigned to the differing pre-puberty stages based on the reported rates of cell division observed in embryos, during the various stages of development and the numbers of cells observed at different stages. For the bottlenecks which occur as we move from one stage of development to the next, the values were taken from the literature where possible (Inner cell mass, 21% in mice; 4 PGCs in human or 40 PGCs in mice; maximum observed PGCs in humans three million in males, ten million in females; 500,000 PGCs at birth in human females), while for unknown stages we defaulted to 50% (epiblast and embryonic epiblast formation). With a model of development, we developed two simulations with different focuses, the first fully simulates gametogenesis, while randomly adding dnms at a specified rate, tracking their numbers and allelic dosages. The second randomly places a set number of *dnms* and tracks their sharing between gametes. For both simulations we simulate a genome of 2.1Gb (signed Int datatype limit) and *dnms* are recorded in each cell as a list of random integers between 0 and 2.1 billion. This allows for the possibility of recurrence, but the chance

is exceeding low due to every value having an equal chance of being selected. For the gametogenesis simulation the focus is on generating a plausible model of dnm interacting with the process of gametogenesis. To do this we simulation the process of cell division, during each cell division we allow for the possibility of one or more *dnms* occurring, by sampling from Poisson distributions with a mean set from our desired mutation rate. The mean of the Poisson distribution was set based on our observed empirical dnm rate, to give the same average number of dnms per gamete. Three different mutation rates are supported, each drawing from different Poisson distributions, an early rate which is set for the first through X cell divisions, the standard rate which is used for X division through to birth, and the post-puberty rate which is utilised for the spermatogenesis stage. The early and standard mutation rates are set based on the number of mutations desired, and the ratio between the two stages. If a ratio of 1 is set for the early mutation rate, then the mutation rate is same for both the early and standard stages. If a ratio of 10 was set then there would be a 10x higher mutation rate in the early stages compared to the later, but the same overall number of dnms would be generated. Once the simulation starts for each cell division, each 'cell' makes an identical copy of its self, and then undergoes a mutational step drawing the number of *dnms* it will receive from the appropriate poisson distribution, the correct number of *dnms* are then created by selecting random numbers between 0 and 2.1 billion. These *dnms* are then recorded in the cell with cells being store adjacent to their most recent sister cell, and the next round of cell division occurs repeating until the required number of cell divisions has occurred. Once the number of cell divisions is equal to the number needed for a specified bottleneck (inner cellmass, epiblast, embryonic epiblast) the appropriate number of cells are randomly selected from the population and the rest discarded. A biased selection option is also available for the PGC selection stage, in this case the PGCs are selected randomly from the outer 10% cell 'lineage' or 'cell tree' structure, and thus on average are more closely related than any two random cells selected from the complete population. During the formation of the PGCs, every *dnm* in the population is tabulated and its frequency in the complete cell population is calculated to provide a 'somatic' allelic dosage for all *dnms*. Two copies of the PGCs are made one to form the male germ-line and the other the female germ-line.

Once the PGCs have undergone the required number of cell divisions to reach the maximum numbers expected from the human literature cell division ceases and the population of PGCs is subsampled down to 500,000 cells (as reported in human literature) representing the loss of oocytes during the late stages of embryo development. At this stage the female germ-line has completed the simulation and the allelic dosages of all the *dnms* present are reported with both the actual allelic dosage observed in the oocytes and the 'somatic' allelic dosages that were observed when the PGCs were selected. This somatic allelic dosage represents the allelic dosage that would be detected for a *dnm* based on the sequencing of somatic tissues, while the true allelic dosage is that of the *dnms* within the oocytes and thus the proportion of offspring that would carry the *dnm*. Half the *dnms* in each oocyte are then randomly discarded to represent the formation of the haploid oocyte and a subsample of 1000 haploid oocyte are reported with their complete list of *dnms* and the associated somatic and actual allelic dosage. For the male germ-line

each cell undergoes an additional 23 replications drawing from the post-puberty mutation rate distribution, for a specified number of years. This represents the additional cell divisions required to maintain a constant supply of sperm cells. After the specified number of additional divisions has occurred half the *dnms* are randomly discarded to generate haploid sperm cell, the true allelic dosages are reported for each *dnm* within the complete male germ-line, along with the 'somatic' allelic dosages for each sperm cell. This process is then repeated one hundred times for each set of parameters, with all results being reported.

After simulation of the gametogenesis process the output files are passed to the likelihoods program, this takes in the 'truth' output files from the simulation then revaluates it compared to the empirical data. For each dataset the 'true' *dnms* are utilised to seed a binomial distribution with a probability of success equal to the allelic dosage. We then sample from the binomial distribution to simulation the detection of each *dnm* (with it's true allelic dosage) via sequencing at 24fold depth, with a minimum detectable allelic dosage of 0.13 similar to that observed empirically from GATK HaplotypeCaller data, variants that pass this threshold are then reported with their 'new sequenced' allelic dosage. Likelihoods are then calculated and reported by comparing the spectrum of observed *dnms* to the spectrum of the simulated 'sequenced allelic dosages'. As an additional output a specific number of gametes is randomly selected and a set of all the unique *dnms* within those gametes are reported along with their allelic dosages, to provide an estimate of the number of unique *dnms* that would be observed for that number of offspring when inheritance is required.

For the sharing simulations a similar scheme is followed as described above, however instead of supplying mutation rates, we instead supply a specific number of *dnms* as observed from a single sperm or oocyte. The primary difference is that under the sharing simulations we create a 'master cell' that will receive every *dnm*. During cell division the master cell is duplicated and there is a chance for one of the specified *dnms* to be added to both cells. The duplicate master cell is then added to the non-master pool. In the next cell division, the non-master pool is duplicated without a chance of *dnm*, then the master cell duplicates with a chance of a *dnm* and the resulting duplicate master cell is added to the non-master pool. This continues as the cells enter the various bottlenecks, with the master cell always being retained, when the number of cell divisions reach the level expected for an oocyte (36), a copy of the data is output showing the percentage of oocytes that share 1 through N *dnms* with the master cell. The cell divisions then continue, to generate the sperm cells after 23 additional replications per year post-puberty, where the percentage of sperm that share 1 through N *dnms* with the master cell is reported. Similar to the gametogenesis simulation we store the cells in a linear arrangement ranging from the most recent to have split off from the master cell to the oldest split. Thus, for selecting related cells we can randomly sample our PGCs from the first 1% of the linear arrangement to select from cells that would be more closely related.

Under this approach a single cell ends up gaining all the *dnms* and thus represents the original sperm/oocyte we were testing. While the pool of non-master cells represents cells sharing a common lineage that have split off at various times. This allows us to estimate how common it is to observe gametes that share a specific number of *dnms*. Like the gametogenesis simulation we can increase the probability of a *dnm* being selected in the early stages of development, this allows us to test out how different early mutation rates, the number of PGCs selected and the selection of PGCs from a closely related cell population affects the degree of sharing observed.

The range of parameters explored in the simulations were chosen based on the underlying biological process, estimates from the empirical data or were taken from the literature. The mutation rates/numbers utilised for both simulations were taken from the average rate of mutation observed in the five initial trios. For the mutation rate in the gametogenesis simulation, we assumed that the pre-puberty mutation rate would be similar for both the male and female germ-line with the additional male specific *dnms* occurring during spermatogenesis. These mutations were then split between the early and late embryo genesis stages based on the specified early to late ratio (1x, 10x, 20x) and as a function of the number of cell divisions we considered to be early (4, 7, 11, 15 or 18). The values for the cell divisions considered to be early were derived from the literature and our embryo development model. The first value of 4 divisions was based on the reported maternal to zygote transition for cattle, while the remaining corresponded to the formation of the inner cell mass (7), epiblast, (11), embryonic epiblast (15) and the formation of the PGCs (18). These values were taken from the literature or derived from the literature by combining the rates of cell division and the time post fertilisation at which the stage of development is observed. The 10x and 20x rates for early *dnms* were selected based on the proportion of mosaic *dnms* we had observed in the five trios, compared to the proportion of non-mosaic *dnms* observed. The number of PGCs were selected based on reports in literature of that in mice up to 40 PGC form the germ-line, and from research in humans that estimate 4 initial PGCs form the germ-line, the value of 10 was chosen as an intermediate value closer to that of humans than mice.
We calculated the likelihoods for the gametogenesis scenario by splitting the allelic dosage by frequency into bins of 0.05 (i.e. 0-0.05, 0.051-0.1, …) and calculating a frequency for each allelic dosage bin based on the 100 simulations of each set of conditions and both the male and female germ-lines. For the sharing simulations the bins were based on the number of *dnms* shared between each tested gamete and the master cell/gamete (i.e. 0, 1, 2, 3, 4, …) and averaged across the 100 simulations, for each set of conditions and both germ-lines. The likelihoods when then calculated for the gametogenesis simulation by taking the empirical allelic dosages observed in our trios and multiplying out the expected frequency from the bins the *dnms* fell into from the simulation, repeating for each set of conditions. For the sharing simulation the same approach was utilised except that the number of shared *dnms* present in each observed gamete was utilised. These two likelihoods were then multiplied together to give the overall likelihood for each of the sets of conditions tested which are shown in table 3.

Supplemental Note 2: Signatures of mutation POLE

The lower than expected transition/transversion ratio for the five families is primarily due to an excess of TCT>TAT and GCA>GAA mutations. This mutational signature shows similarities to mutational signature 10 reported by Alexandrov et al (2013a), which is defined by a similar excess of TCT>TAT variants and a smaller excess of TCG>TTG, which also may be present in the early *dnms*. The TCG>TTG part of the signature is difficult to identify in a combined mutational spectrum as it is one of the spontaneous NpCpG>NpTpG classes of mutation which show a greater than ten-fold excess. This mutational signature 10 is associated in cancer with missense mutations in proofreading exonuclease domain of DNA polymerases ε (POLE) (Church et al. 2013). The exonuclease domain is thought to be responsible for a 100 fold increase in DNA replication fidelity by the excision of mispaired bases. Interestingly within the sequence dataset we observe eight missense variants in POLE with alternative allelic frequencies of up to 0.05, in addition there are three potential splice site variants with allelic frequencies of 0.05, 0.15, and 0.52. Proband 2 is heterozygous for two of the splice variants, while the sire of probands 3 and 4 is heterozygous for one and it's dam is also homozygous for that variant. Considerable work will be need to determine if these are potentially causative for the excess TCT>TAT pattern noted in the early variants. Especially considering that the signature appears to be associated with the early mosaic *dnms*, and absent or strongly reduced in late occurring *dnms*. Thus, it is difficult to suggest a simple model by which variants in POLE could result in an increased mutation rate or a shift in the mutational signature during the first four cell divisions after fertilisation, but would fail to have the same effect after the first four cell divisions. We would note that in the complete damona dataset of 131 families this mutational signature remains significant in the early mosaic classes of *dnms*, but appears to primarily be restricted to a small number of lineages.

Church DN, Briggs SEW, Palles C, Domingo E, Kearsey SJ, Grimes JM, Gorman M, Martin L, Howarth KM, Hodgson SV, et al. 2013. DNA polymerase ε and δ exonuclease domain mutations in endometrial cancer. *Hum Mol Genet* 22: 2820–2828.

Experimental Section

Study 2:

Evaluating the inter-individual variation in the rate and spectrum of germ-line de novo mutation and its causes in cattle.

In preparation

Harland C, Durkin K, Artesi M, Karim L, Cambisano N, Deckers M, Tamma N, Mullaart E, Coppieters W, Georges M, Charlier C

Abstract

To study the process of *de novo* mutation (*dnm*) in the bovine germ-line, we have sequenced the whole genomes of 743 individuals constituting 131 sire-dam-offspring trios with an average of five grand-offspring each. A first study using five pedigrees revealed the common occurrence of somatic and germ-line mosaicism for *dnms*, pointing towards mutation-prone early cleavage cell divisions (http://biorxiv.org/content/early/2016/10/09/079863). We have identified 7,498 *dnms* with an overall transition-transversion rate of 1.96, of which 3,413 are mosaic in either the proband or parents' germ-line, confirming the results of our previous study. We detect a significant environmental effect resulting from the use of reproductive technologies, such as in vitro fertilisation, on the rate of *dnm* in the early embryo. We identify four outlier animals with substantially elevated mutation rates (4.5-17-fold) during the early stages of embryo development, each with distinctive mutational signatures which differ significantly from the global spectrum. For one outlier, we identify two candidate causative mutations that may underlie the abnormal *dnm* rate, a rare homozygous deleterious event in the REV1 DNA polymerase and a maternal hemizygous deletion of the key mitochondrial gene *TFB1M*.

Introduction

De novo mutation (dnm) is a fundamental biological process, responsible for creating the majority of genetic diversity. We can define a germ-line dnm as a genetic variant that was absent in the gametes that formed an individual but present in gametes it passes to the next generation. With the advent of NGS (next generation sequencing) it has become possible to directly detect these through the whole genome sequencing (WGS) of trios consisting of both parents and an offspring. The dnms can then be identified by their absence in the parental DNA and presence in the offspring. Under these criteria it is possible to identify the germ-line dnms that occurred during for formation of the gametes in each parent. Further, for dnms that share a sequenced DNA fragment with a variant which is specific to one of the two parents it is possible to determine in which parental germ-line the dnm has occurred. Together this has allowed the estimation of the number of dnms and thus the mutation rate, as well as the ratio at which the events occur between the paternal and maternal germ-lines. This approach has been applied to numerous species at this time including humans (reviewed by Ségurel et al. 2014), chimpanzees (Venn et al. 2014), fruit flies (Keightley et al. 2014), honey bees (Yang et al. 2015), postman butterflies (Keightley et al. 2015), collared flycatchers (Smeds et al. 2016), Atlantic herring (Feng et al. 2017), and cattle (Harland et al. 2017a) allowing the estimation of the dnm rate in each. From the human studies, it has been estimated that the average human mutation is $1.2x10^{-8}$ per base pair per generation (for SNPs, small INDELs), that there is a 4:1 ratio of male to female dnms, and that there are parental age effects of ~0.5 maternal dnms per year (Wong et al. 2016) and 2 paternal dnms per year post puberty (Kong et al. 2012). Venn et al (2014) have shown that while the dnm rate in chimpanzees ($1.2x10^{-8}$ per base pair per generation) is nearly identical to that of humans, the ratio between males and females is considerably higher compared to the human (7-8x). Feng et al (2017) have provided support for the drift-barrier hypothesis (Lynch 2011; Sung et al. 2012), which predicts that mutation rate is a function of effective population size. Finally, there have been initial suggestions that the mutation rate varies during germ-line development with more dnms than expected during the initial cell divisions after fertilisation (Rahbari et al. 2016; Harland et al. 2017a).

Aside from human studies where thousands of trios have now been investigated, most studies have been limited to only a small number of families (one to five). In addition, due to the use of two generation pedigrees, only a small proportion of dnms can be assigned to the parental germ-line, which has limited the ability to detect differences in patterns of mutation and the mutation rate between individuals. While such studies have provided insight into the rates and patterns of mutation in each species, little work has yet been done on inter-individual variation at least in part due to these limitations. To further investigate this area, we herein report the results of a study estimating the cattle dnm rate in 131 sperm cells and 131 oocytes.

Results

To study the process of dnm in the bovine germ-line, we sequenced the whole genome of 743 Holstein-Friesian cattle at average depth of 12x (range: 3.9 - 63). These constituted 131 three-generation pedigrees comprising a sire–dam–offspring trio (sequenced at average 24-fold depth; range: 6 - 63) with an average of 5 grand-offspring (range 1-11) sequenced at average 6.2-fold depth (range 3.9-26.6). The 131 offspring (hereafter referred to as probands) included 40 males and 93 females. The four grand-parents were available for 4 trios, two paternal grand-parents available for 5 additional trios, and the two-maternal grand-parents for 32 additional trios. On average, 5.6 (range: 0-13) paternal, and 2.2 (range: 0-6) maternal half-sibs were available per proband. Sires were on average part of 2.3 trios (range: 1-9), and dams of 1.9 (range: 1-7). Genomic DNA was extracted from sperm for 126 males, from venous blood for 213 males, and from venous blood for all females. Sperm and blood DNA was sequenced for one male proband. This data set includes five pedigrees that have been described in a previous study (chapter 1 of this thesis).

Using a suite of public and custom-made programs (cfr. M&M), complemented by visual inspection of candidate dnm in IGV, we identified a total of 7,498 dnm in the 131 probands. We previously validated 174 of 174 tested dnms using an independent method demonstrating the excellent specificity of our pipeline (Harland et al. 2017a). Following Harland et al. (2017a), and using the criteria outlined in M&M, we classified the 7,498 dnms in 562 proband-mosaic (PM) mutations (having occurred in the germ-line of the proband), 2,795 sire-non-mosaic (SNM) (having occurred late in the germ-line of the sire), 1,522 sire-mosaic (SM) (having occurred early in the germ-line of the sire), 1,090 dam-non-mosaic (DNM) (having occurred late in the germ-line of the dam), 1,329 dam-mosaic (DM) (having occurred early in the germ-line of the dam).

Of the 7,498 dnms, there were 13 in splice sites, 35 in UTR, 70 in exons, 644 within 5kb of a gene, 2,028 in introns, with the remainder intergenic (Variant effect predictor (McLaren et al. 2016), Ensembl gene annotation, build 90). 6,487 of the dnms were single nucleotide substitutions, 254 tandem mutations, and 757 INDELs. The dnms were evenly distributed across the majority of the autosomal chromosomes, with numbers proportional to chromosome size (average of 2.78 per megabase). However, chromosomes seven and X carried significantly less dnms than expected based on their size, with an average of 2.21 per megabase for chromosome seven ($p = 0.0047$) and 2.18 per megabase for chromosome X ($p = 3x10^{-4}$) (Suppl. Fig. 1). While we are uncertain why chromosome seven has significantly less dnms than expected, the difference for chromosome X may be due to its reduced depth of sequence in males, or because the X chromosome spends less time than the autosomes in the more mutagenic male germ-line.

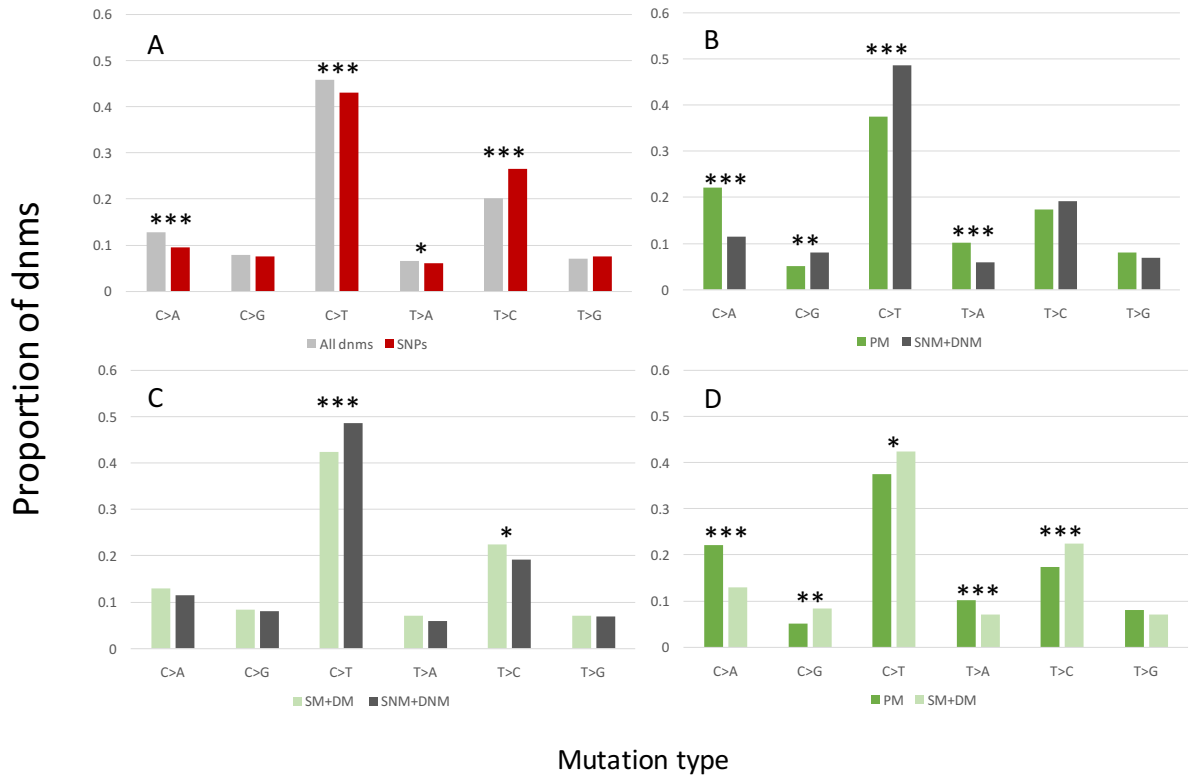


Figure 1: Comparison of the proportion of the six possible single nucleotide substitutions for (A) all dnms (light grey) vs 613,971 SNPs segregating in the Damona population at frequencies ≤ 0.01 (SNPs; red), (B) PM dnms (green) vs SNM+DNM (grey), (C) SM+DM (light green) vs SNM+DNM (dark grey), (D) PM (green) vs SM+DM (light green). Significant differences are shown by asterisks $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***. Suppl. Fig. 10 shows the combination of all five datasets.

The transition-transversion (Ts/Tv) ratio for all the dnms was 1.96, compared to the expectation of 2.1 for both rare and common variants segregating within the cattle population. The Ts/Tv ratio for the PM dnms was 1.19, for the SM and DM dnms 1.96, and for the SNM and DNM 2.1.

When considering the mutational spectrum of all dnms versus rare variants segregating in the population, we see significant differences in the proportions of the six types of single nucleotide substitutions: the dnms show significant enrichment for C>A, C>T, and T>A variants, as well as a significant deficit in the T>C variants (Fig. 1A and Suppl. Table 2).

Within the dnms we see significant differences in the mutational spectrum between the early (PM and SM+DM) and the late classes (SNM + DNM) of dnm (Fig. 1B and 1C, Suppl. Fig 1, Suppl. Table 3,). With the PM dnms showing a significant enrichment for C>A and T>A variants and a deficit of C>G and C>T variants (Fig. 1B). While the SM+DM dnms share the deficit of C>T variants but are enriched for T>C variants (Fig. 1C). The deficit of C>T variants in both PM and SM+DM is primarily driven by

a reduction in the numbers of C>T mutations at CpG sites, which is shown in the trinucleotide mutational signature in Fig. 2 (Suppl. Fig. 2, Suppl. Table 4).

When comparing the two early classes of dnm (PM and SM+DM) we see a significant enrichment for C>A and T>A variants and a deficit of C>G, C>T and T>C variants (Fig. 1D, Suppl. Fig. 3) in the PM dnms. The difference for the C>T variants is again primarily driven by a reduction in the number of C>T variants at the CpG sites as seen in the trinucleotide mutational signature (Suppl. Fig. 3).

Overall, the mutational signatures of dnms differs from that of rare segregating variants in the population. Additionally, within the dnms there appears to be a distinct difference between the mutational signatures of those dnms that occur early in development (PM and SM+DM) compared to those that occur late in development (SNM+DNM).

Finally, if we consider the patterns of mutation for PM, SM+DM and SNM+DNM, we see that the mutational signatures for PM (early) and SNM+DNM (late) appear to represent two ends of a continuum, with the SM+DM signature being partway between the two extremes (Suppl. Note 1, discusses possible explanations for these differences).

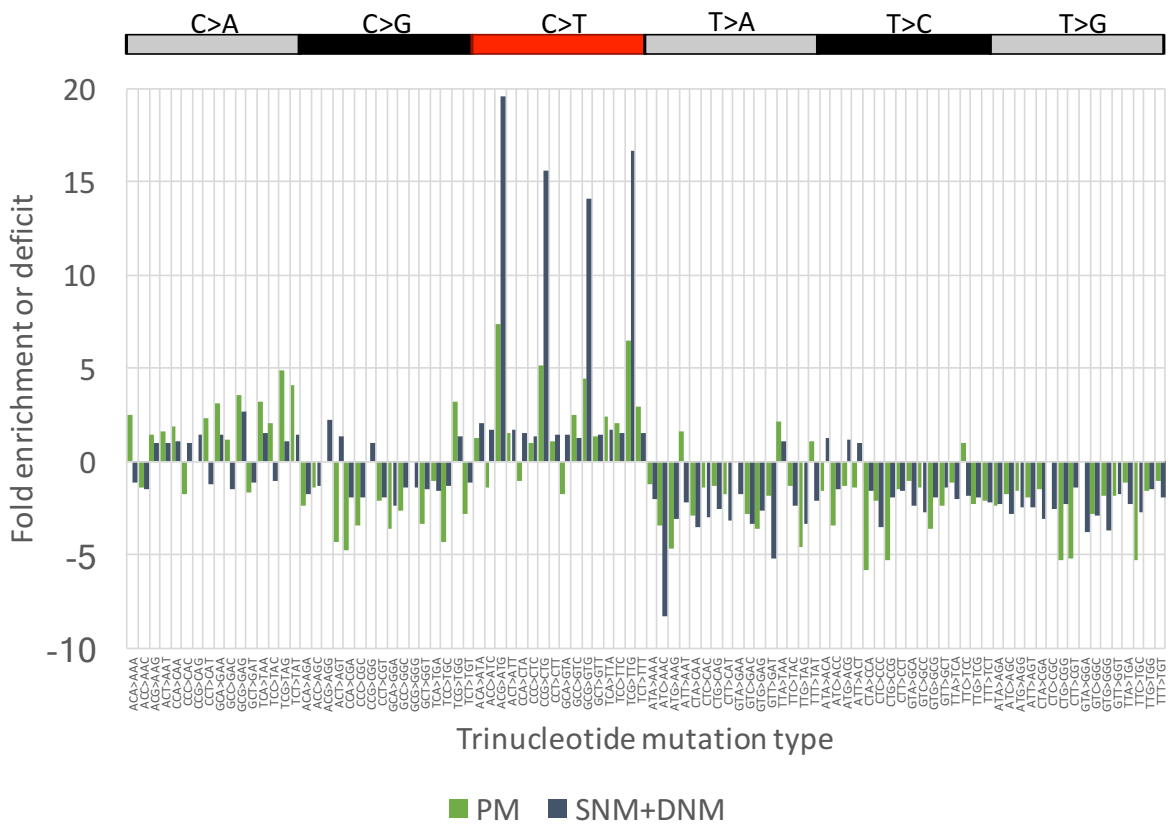


Figure 4 : A comparison of PM (green) and SNM+DNM (dark grey) dnms for the fold excess or deficiency over the expected for specific nucleotide substitutions for each of the 96 trinucleotide contexts. (Y-axis fold excess or deficiency compared to the expected proportion of such trinucleotides after adjustment for the 3 possible mutations that could occur in the bovine genome (bosTau6), X-axis shows the 96 possible trinucleotide substitutions after having combined both the 5'>3' and 3'>5' versions of each mutation type). The top bar shows the 6, single nucleotide mutational signatures for each of the blocks of trinucleotide signatures. P-values in Suppl. Table 4

The numbers of observed dnms of the five classes were adjusted using trio-specific estimates of genome coverage and sensitivity (raw data and correction factors in Suppl. Table 1). From these, we estimated that, in cattle, (i) sperm cells carry on average 41.1 dnms of which 35.3% are detectably mosaic in DNA of the sire, (ii) oocytes carry on average 22.4 dnms of which 54.9% are detectably mosaic in blood DNA of the dam, (iii) 7 dnms (9.9%) detected in the DNA of male proband, occurred during his development, (iv) 4.1 dnms (5.8%) detected in blood DNA of a female proband occurred during her development (Table 1). There was an effect of DNA source (sperm vs blood) on the number of PM mutations in males ($p = 0.018$). The difference between the number of PM dnms in male and female probands is entirely due to two male outlier probands which have 30 and 80 PM dnm mutations respectively.

These estimates confirm the results of our pilot study (Harland et al., 2017a), indicating that a large proportion of dnms are detectably mosaic in the individual in which they occurred, hence strongly suggesting that the mutation rate is considerably higher in early cleavage cell divisions than at later stages of development.

The numbers of SNM+SM and DNM+DM mutations were used to estimate the sex-averaged bovine dnm rate at $1.22x10^{-8}$ per base pair per generation (95% CI: $1.15\text{-}1.3x10^{-8}$) (i.e. very similar to human and chimpanzee), with a male-to-female ratio of 1.8:1 (i.e. considerably lower than human and chimpanzee). Note that cattle breed at a younger age and for a limited time-span, which may mitigate the age-effects observed in primates. It is noteworthy that the male to female ratio for the late occurring SNM/DNM is 2.56 ($p < 1x10^{-5}$), while the SM/DM ratio is 1.15 ($p = 1.7x10^{-4}$). While for the early PM dnms, the average number of dnms is 4.1 for both male and female probands (after removal of the outliers, see above). The reduction in the ratio of male to female dnms for the early (SM/DM) mutations, compared to that of the late (SNM/DNM), as well as the equality between the number of PM dnms occurring in the male and female germ-lines is compatible with the suggestion that they occur at a developmental stage preceding sexual differentiation. Estimating dnm rate and sex ratio using SM+DM+PM (as typically done in other studies) would have yielded a rate of $0.8x10^{-8}$ (95% CI: $0.74\text{-}0.87x10^{-8}$) and a male-to-female sex ration of 2.6:1.

Table 1: Number of dnms and supporting statistics for the five classes of dnm with PM split by proband gender. Values have been adjusted for estimated coverage of the genome and sensitivity (Suppl. Table 1). *Average*: mean number of dnms after correction per proband, *Range*: range of observed dnms in the 131 pedigrees after correction. *CV*: coefficient of variation for each class of dnm. >1 gamete: the number of individuals with multiple gametes in the population. *Repeatability*: estimated repeatability of the number of dnms observed in individuals with multiple gametes. p-value REML: significance of the individual animal effect in the REML analysis model is detailed in Suppl. Note 2.

Class	Average	Range	CV	>1 gamete	Repeatability	p-value REML
PM-M	7.0	0-91	2.2	NA	NA	NA
PM-F	4.1	0-24	1.1	NA	NA	NA
SNM	26.8	8-90	0.4	35	0.15	0.2
SM	14.4	0-41	0.6	35	0.55	$1x10^{-6}$
DNM	10.1	0-48	0.6	33	0	1
DM	12.4	0-69	1.0	33	0.77	$6x10^{-13}$

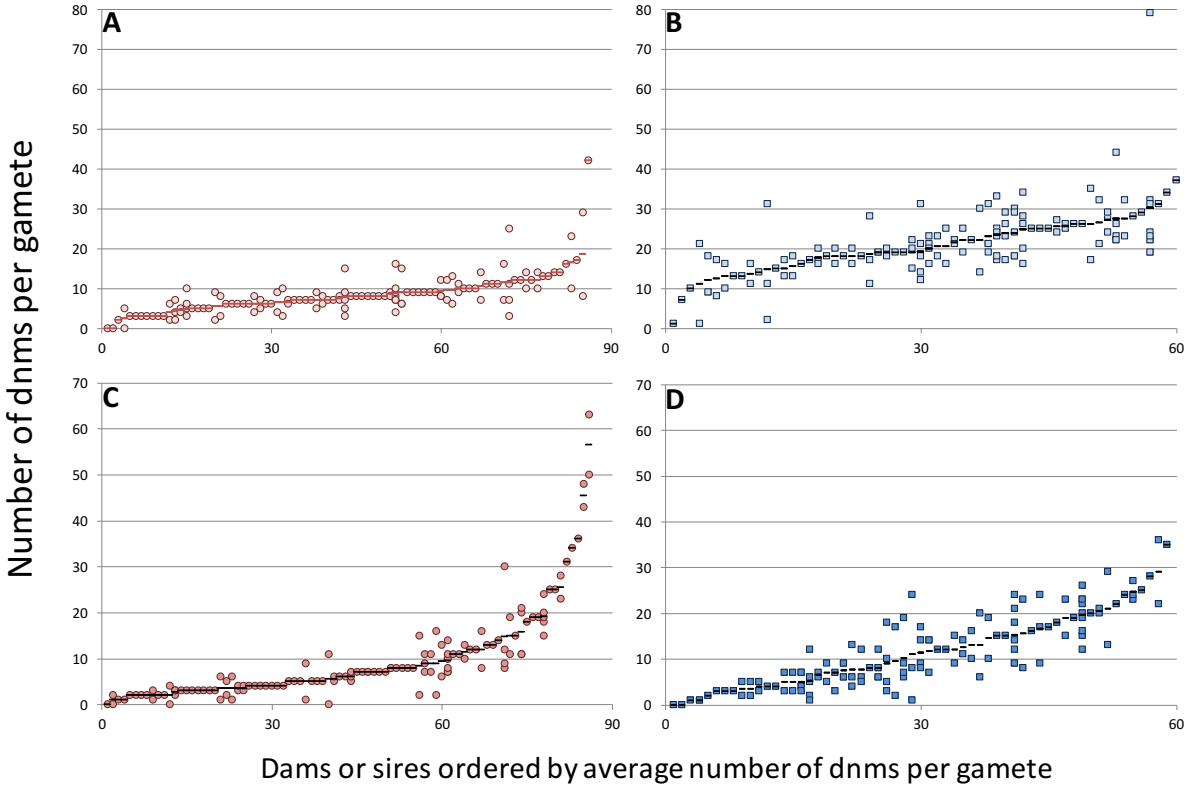


Figure 3: Number of dnms (uncorrected) by gamete and class for (A) DNM, (B) SNM, (C) DM, (D) SM. The X-axis shows each sire or dam ordered by the average number if dnms per gamete, with multiple gametes shown when present. The Y-axis indiciates the number of dnms of the specified class per gamete, the dots indicate individual oocytes (red) or sperm (blue) with the average show as a black bar.

Fig. 3 shows the observed distribution in the numbers of dnms for the 60 analysed sires (SNM and SM), and 86 analysed dams (DNM and DM). Strikingly, the coefficient of variation (mean/standard deviation) was at least twice as large for the "early" mosaic classes of mutations (PM, SM, DM) than for the "late" non-mosaic classes of mutations (SNM, DNM) (Table 1).

We first exploited the fact that dnms could be studied in multiple gametes for 39 sires and 48 dams, to estimate the "repeatability" (i.e. is there any evidence for an "animal" effect, whether due to genetics and/or permanent environmental effects?) in the number of dnms of the different classes using a restricted maximum likelihood (REML) model. The repeatability sets the upper limit for the heritability. There was no evidence for a significant sire effect on the number of SNM or dam effect on the number of DNM (Table 1). This was confirmed by the absence of a significant correlation between the number of SNM and DNM mutations between multiple gametes of the same parent (Fig. 4 A and B). There was suggestive evidence for a significant individual effect on the number of SM and DM mutations both

from the REML and correlation analyses (Table 1 and Fig. 4C and D). It should be noted, however, that a large proportion of SM and DM mutation are shared by sibs, which could unduly inflate these estimates. Work is in progress to control this effect.

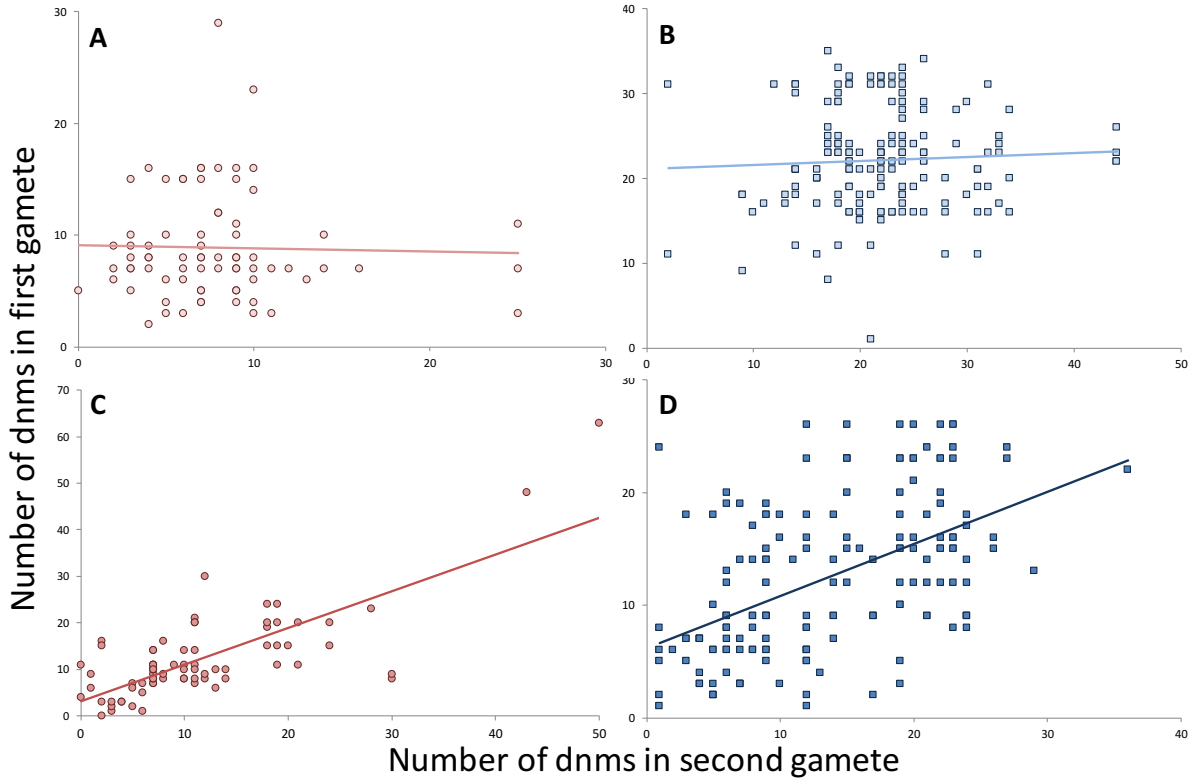


Figure 4: Number of dnms (uncorrected) and line of best fit per dnm class for sires and dams. Numbers of dnms for two gametes from the same parent shown on the y- and x-axis, respectively, for (A) DNM, (B) SNM, (C) DM, (D) SM. Red indicates oocytes, blue sperm.

Elite dairy cattle are routinely generated by relying on reproductive technologies of increasing sophistication, including (i) artificial insemination (AI), (ii) multiple ovulation and embryo transfer (MOET), and (iii) oocyte pickup, in vitro maturation, in vitro fertilisation (IVF). Accordingly, 27% of Damona probands were generated by AI, 37% by MOET and 36% by IVF (Suppl. Table 1). We tested whether the reproductive technology used to produce an individual (in this case the proband) had an effect on the "early" mutation rate reflected in the number of PM mutations. We observed a significant effect ($p = 1.9 \times 10^{-4}$) for reproductive technologies on the number of PM mutations per proband, with a mean of 1.9 PM dnms for AI, 2.7 for MOET (not significantly different, $p = 0.195$), and 4.5 for IVF (highly significant, $p = 7.35 \times 10^{-5}$) (Fig. 5 and Suppl. Table 1).

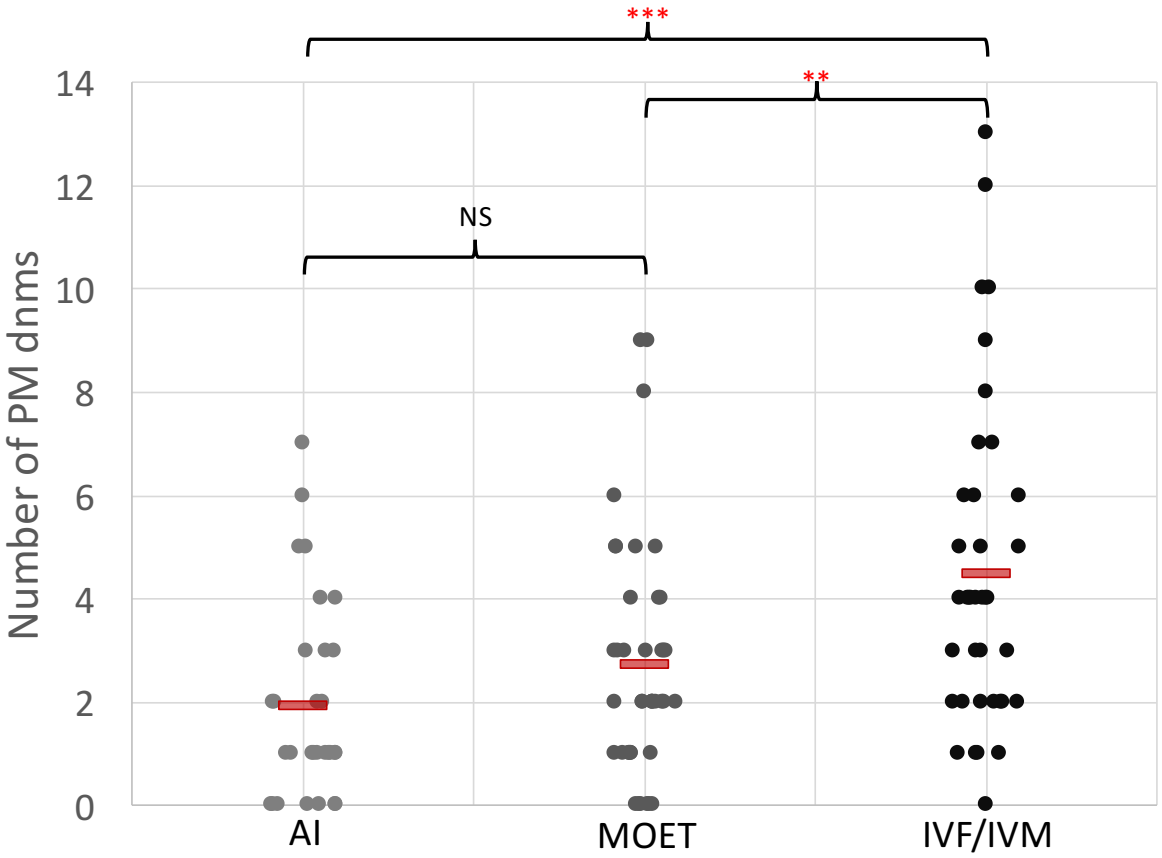


Figure 5: Distribution of PM dnms by reproductive technology utilised in the creation of the proband (AI, artificial insemination; MOET, multiple ovulation and embryo transfer; IVF/IVM, in vitro fertilisation and maturation). Red bar indicates the mean number of dnms for each technology, brackets indicate each test for significance between the three classes with NS representing not significant, ** $p < 0.01$, *** $p < 0.001$.

Examining the distribution of the number of dnms reveals at least (i) two outlier probands with, respectively, a 17-fold (9 sd above the mean, Suppl. Fig. 6) and a 6.5-fold (3 sd above the mean, Suppl. Fig. 7) excess in the number of PM mutations, and (ii) two outlier dams with, respectively, a 4.5-fold (3.2 sd above the mean, Suppl. Fig. 8) and a 5.6-fold (4.3 sd above the mean, Suppl. Fig. 9) excess in the number of DM mutations. Both dams had two offspring available for analysis in the Damona dataset. The mutational spectra differed significantly from the global spectrum for all four outliers (Fig. 6). The spectrum of proband outlier 1 was characterized by an 8-fold increase in the number of C>T transitions outside of the CpG context (Fig 6A), that of proband 2 by a six-fold excess of C>A/G>T transversions (Fig 6B), that of dam 1 by 2.1-fold excess of non-CpG C>T transitions (Fig 6C), and that of dam 2 by a 13-fold excess of tandem mutations.

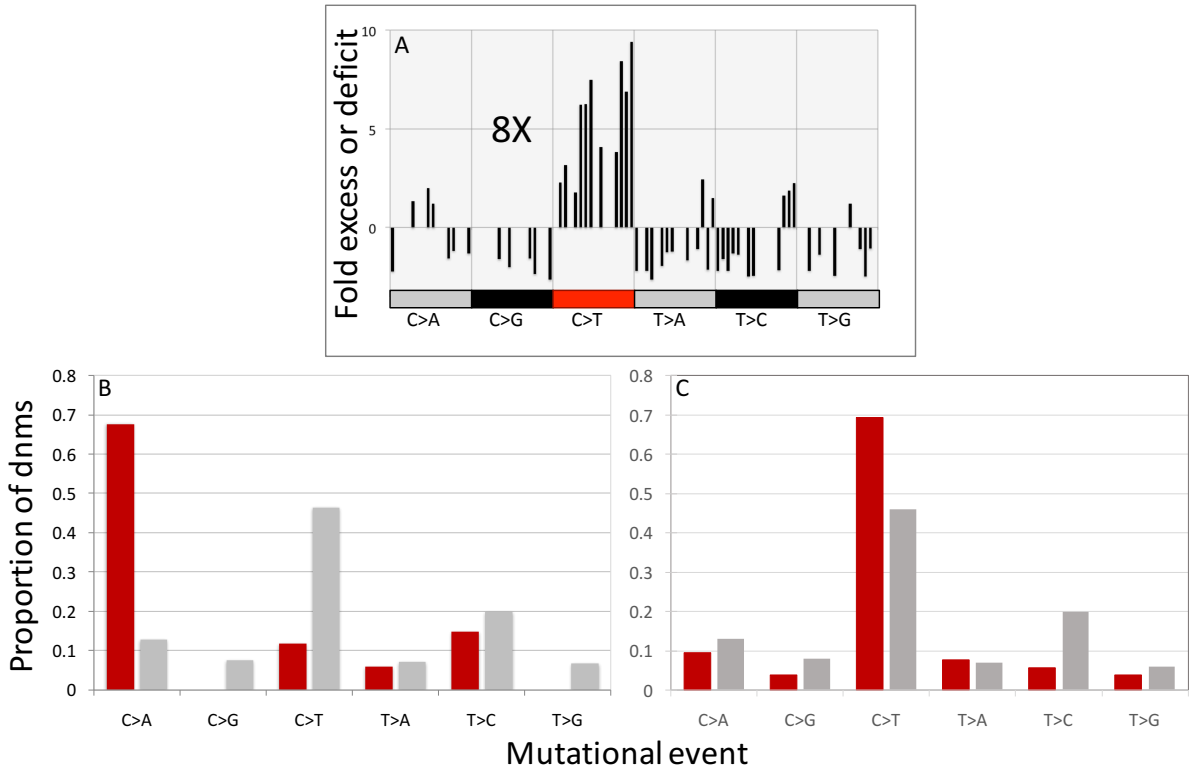


Figure 6: Mutational signatures for outliers one through three. A) Trinucleotide mutational signature (fold excess or deficiency) for outlier proband one, showing an 8x increase in C>T mutations at non-CpG sites. B) Mutational spectrum (proportion of dnms by mutation type) for outlier proband two, showing a 6x increase in C>A mutations. C) Mutational spectrum (proportion of dnms by mutation type) for outlier dam one, showing a ~2x increase in C>T mutations outside of CpG sites. For dam outlier two, no mutational spectrum is shown as its primary mutational signature is an increase in the number of tandem mutations.

Discussion

We have utilised 131 three and four generation pedigrees to identify and characterise 7,498 dnms in Dutch dairy cattle. We show that in cattle 13% of the dnms identified by the criteria used in human studies (variants absent in the parent's DNA, and heterozygous in the proband), actually occurred during the early stages of development in the proband. As a consequence of this 35% of the dnms in a sperm and 55% of dnms in an egg, would be gonosomal or germ-line mosaic in the parents. These results confirm the findings of our pilot study (Harland et al. 2017a), suggesting that the mutation rate in the early stages of embryo development is considerably higher than that at later stages. When comparing the mutational signature of the complete set of dnms to that of rare variants segregating within the population, there are significant differences between the two. With the dnms being enriched for C>A/G>T and C>T/G>A variants while having a deficit of T>C/A>G variants. Interestingly if we consider this difference from the point of the segregating variants within the population, this is similar to what we might expected from biased gene-conversion. As over time biased gene-conversion favours the C/G alleles of variants over their A/T alleles (Marais 2003; Lachance and Tishkoff 2014), and thus would reduce the relative proportions of observed C>A/G>T and C>T/G>A variants while increasing the proportion of T>C/A>C variants. Thus, it is plausible that biased gene-conversion may modify the mutation spectrum of a population over time, and may be partly responsible for the sift in mutational signatures we observe between dnms and variants segregating in the population.

In addition, we have been able to classify dnms as either occurring 'early' (PM, SM and DM) during embryo development, or as 'late' (SNM and DNM) occurring after the differentiation of the PGCs. When comparing the mutational signatures of early mosaic dnms (PM, SM and DM), with late non-mosaic dnms (SNM and DNM) we observe significant differences between their mutational spectrum. The 'early' dnms show significant excesses of C>A/G>T dnms, and deficits of C>T/G>A dnms compared to the late dnms. The difference in C>T/G>A dnms is primarily driven by a reduction in the number of C>T mutations in CpG contexts. This reduction fits well with the observation that C>T mutations at CpG sites show a clock like behaviour, due to their formation by spontaneous deamination of methyl-cytosine to thymine at CpG sites. Thus, spontaneous deamination is a time dependent process, rather than a replication dependent one (Moorjani et al. 2016) and we would expect to see an increasing proportion of CpG C>T mutations accumulate in cells over time. For the excess of C>A mutations we observe a significant increase in TCT>TAT mutations in the early mosaic classes, this is in agreement with the results of our pilot study (Harland et al. 2017a) and is similar to a pattern recently reported for early mosaic dnms in humans (Ju et al. 2017).

Additional differences between early and late occurring dnms can be observed in the ratio of dnms occurring in the male germline compared to the female. For late occurring dnms (SNM/DNM) the male to female ratio is 2.6:1, while for early dnms (SM/DM) the ratio is 1.15:1 and for the PM dnms the average number occurring in both male and female probands is the same (4.1 after account for outliers).

This strongly supports our view that the SNM and DNM mutations occur late in the germ-line development and differ from those dnms that occur early in development before sexual differentiation occurs. The excess of male (SNM) to female (DNM) dnms for the late dnms, follows the pattern reported in both humans (4:1) and chimpanzees (5.5:1) (Kong et al. 2012; Venn et al. 2014). It is thought to arise due to the additional 23 cell divisions per year, that occur after puberty, in the male germ-line during spermatogenesis (Crow 2000). The differences between humans and chimpanzees is thought to be explained in-part by increased sperm competition in chimpanzees (Venn et al. 2014), which may increase the rate of spermatogenesis or results in additional divisions during the process. Potentially the herd-based family structure of cattle, could result in comparatively lower levels of sperm competition and could partially explain the lower ratio observed in our study. A second contributing factor to the decreased male to female ratio, is that the average generation time for dairy cattle, is in the order of two to six years, compared to the 20-30 for chimpanzees or humans and thus there is substantially less time for cell divisions to accumulate in spermatogonia after puberty.

Another difference between the early and late dnms is the degree of variability within the population for these classes of events, with the early dnms having coefficients of variance that are nearly twice that of the late dnms. The four outlier individuals contribute to part of this variance, with all four being outliers for early classes of dnms (PM and DM). The first outlier had a ~17x increase in PM dnms compared to the average, with an eight-fold excess in C>T mutations at non-CpG sites. For this outlier two candidate causative variants were identified, the first being a rare deleterious missense mutation in the translesion polymerase REV1, for which the outlier was the only homozygous mutant individual observed in the Damona dataset. As a DNA polymerase and a supporting protein for DNA repair, a mutation to REV1 has a plausible means of action (Waters and Walker 2006; Waters et al. 2009). The second candidate was a large (140kb) hemizygous deletion of the critical mitochondrial *TFB1M* gene in the maternal genome but was absent from the proband. For *TFB1M* there is some evidence of a haploinsufficency phenotype associated with the gene (Koeck et al. 2011). If this is the case, then the heterozygous deletion may affect the function or efficiency of the mitochondria. During the earliest stages of development, the embryo is dependent on products of the maternal genome, until the maternal-zygote transition when its' own genome activates. Thus, variants present in the maternal genome but absent in the zygotes, could still influence the development of the embryo. A decrease in efficiency of the mitochondria during development could affect the mutation rate in at least two ways, first a decrease in efficiency could lower the cellular energy budget reducing the efficiency of DNA repair and replication. In the second case a decrease in efficiency could increase the oxidative stress the embryo was under. Oxidative stress and in the corresponding increased levels of radical oxygen species, is associated with an increase in the C>T mutation rate of DNA (Degtyareva et al. 2013). For the remaining outliers, no obvious candidate causative variant has been identified at this time, though each produces a distinct mutational signature (Fig 6). Outlier two with a 6.5fold excess of PM dnms, in particular presents a mutational signature

dominated by a six-fold increase in C>A/G>T mutations, which is an extreme case of the pattern observed for the early dnms. This raises the possibility that mutator alleles currently present in the population, may contribute to the shifted mutational spectrum for the dnms compared to rare segregating variants.

In dairy cattle, the use of reproductive technologies such as MOET and IVF/IVM has become exceedingly common for high value animals, while in humans their use is also increasing. Thus, the finding that the use of reproductive technologies has a significant impact on the mutation rate in early embryo development is of considerable importance. We note that as the intervention becomes more extreme (AI < MOET < IVF/IVM) the size of the effect increases. This may be a function of increasing level of stress an embryo is exposed to as the degree of intervention increases. For both humans and cattle, it is worth considering the potential impact a substantial increase in the number of early mosaic mutations in an individuals' germ-line would have. Especially in cases where an individual may be de novo mosaic for genetic disorder but asymptomatic or have a reduced phenotype due to the mutation only being present in a fraction of its' cells. However, for their offspring there may be a relatively large chance of inheriting the mutation and those who inherited it would full display the disorder. We would also note that our previous study (Harland et al. 2017a) has shown that the allelic dosage of a mosaic variant in the soma is a poor predictor for its' dosage within the germ-line.

The dnm rate in cattle is ~$1.2x10^{-8}$ per bp per generation when accounting for mosaic mutations (SM + SNM + DM + DNM), similar to that reported in humans and chimpanzees. Interestingly at least in cattle, there is a substantial difference of $1.2x10^{-8}$ vs $0.82x10^{-8}$ in the average mutation rate when failing to correctly account for mosaicism. If this same level of mosaicism was present in humans, correcting for it would increase the average mutation rate to ~$1.8x10^{-8}$. This value, while substantially lower than the phylogenetic estimate of $2.5x10^{-8}$, is higher than any reported in current NGS based human studies. This includes the two human studies that use heterozygous variants in regions of identical by descent autozygosity to estimate the dnm rate, which by default would include mosaic dnm mutations. These studies have respectively estimated the human dnm rate to be $1.2x10^{-8}$ (Campbell et al. 2012) and $1.6x10^{-8}$ (Palamara et al. 2015). This rate is also higher than that estimated by Rahbari et al (2016) in a study which directly identified mosaic dnms with allelic dosages up to 10% in the parent, estimating the human mutation rate was ~$1.28x10^{-8}$ after accounting for mosaicism. However, if we were to assume that mosaicism is present in humans at a higher rate than reported by Rahbari et al (2016), but at half the rate estimated in cattle, then correcting for it would give an average mutation rate of ~$1.5x10^{-8}$. This is close the rate ($1.6x10^{-8}$), reported in a study by Palamara et al. (2015), for which the experimental design would have retained mosaic mutations.

Taken together the various findings of our study suggest that, at least in cattle, early mosaic dnms contribute significantly to the rate of germ-line dnm. In addition, the process of dnm can be split into at least two parts. An 'early' stage that occurs before sexual differentiation, were dnms occur in equal numbers in both the male and female germ-line. This stage is characterised by high variability in the rate of mutation, is influenced by both environmental (reproductive technologies) and genetic factors (the outliers) and demonstrates a distinctively different mutational signature. The second stage is the late occurring dnms, these occur after sexual differentiation and the formation of the primordial germ cells, they primarily occur in the male germ-line, show a lower degree of variability in their rate and present a mutational signature that differs from both the early dnms and variants segregating in the bovine population. Thus, to fully understand the process of mutation we would argue that it is important to understand and study the interactions between dnm and the development of the germ-line in a species. Should the level of mosaicism occurring in humans, approach that we have observed in cattle it would substantially increase the estimated human mutation rate. If the level of mosaicism in humans is substantially lower than that in cattle, then it indicates even for species with similar mutation rates and genome sizes there can be substantial differences in the pattern of dnm and potentially the factors affecting the rate and patterns of mutation. Finally, we present evidence of multiple outliers with substantially increased rates of mutation and differing mutational signatures. This is potentially suggestive of the presence of mutator alleles within the bovine population, which may provide a useful means of determining the genetic factors influencing the rate of dnm.

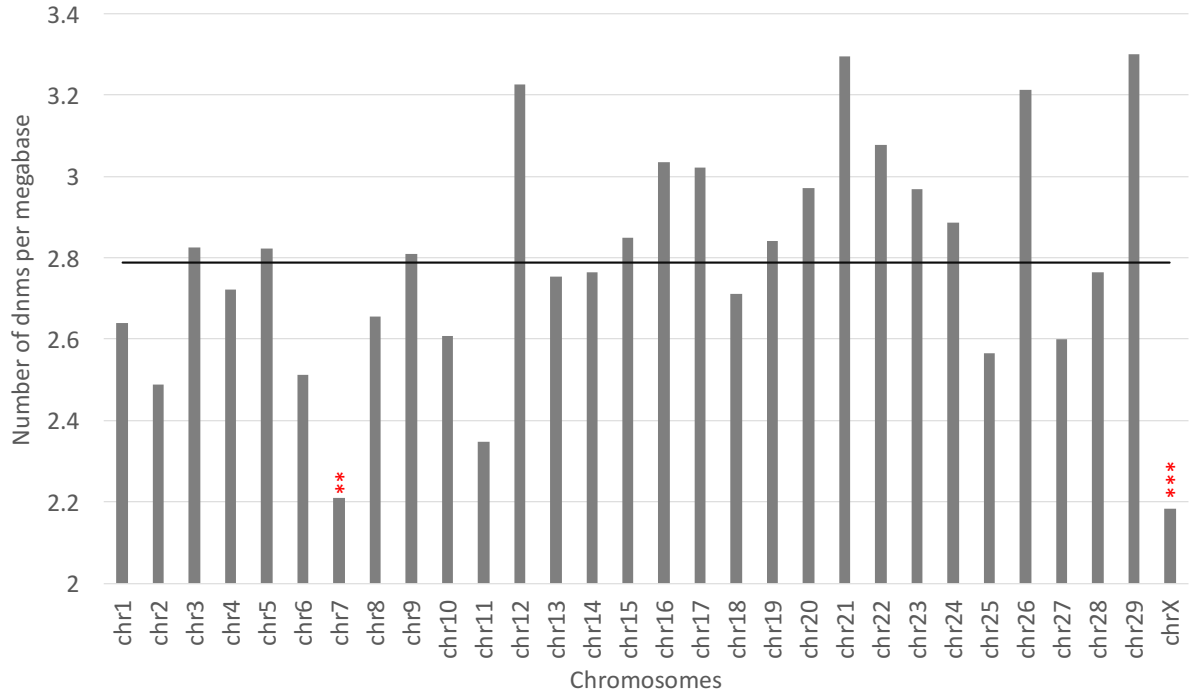
Materials and Method

Blood or sperm samples were collected for 743 Dutch Holstein Fresian cattle (*bos taurus*), which form 131 three or four generation pedigrees each consisting of at least sire, dam, proband and on an average five grand-offspring, with grand-parents were possible. For four pedigrees, all four grandparents were sequenced, 13 pedigrees have three grandparents, and 23 pedigrees have two grandparents sequenced. DNA was extracted and Illumina NextSeq 550bp whole genome libraries were constructed and sequenced on the Illumina HiSeq 2500 in 100bp paired end mode by the University of Liege, GIGA-Genomic core service. The data was prepared following the GATK best practises protocol (version 3) (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013) after alignment to the BosTau6 reference genome by BWA MEM (Li 2013). The GATK Haplotype caller (v3.4) was run following the GVCF N+1 protocol and was utilised to identify variants (bioinformatics scripts are available from: https://github.com/aeonsim/DamonaPipeline).

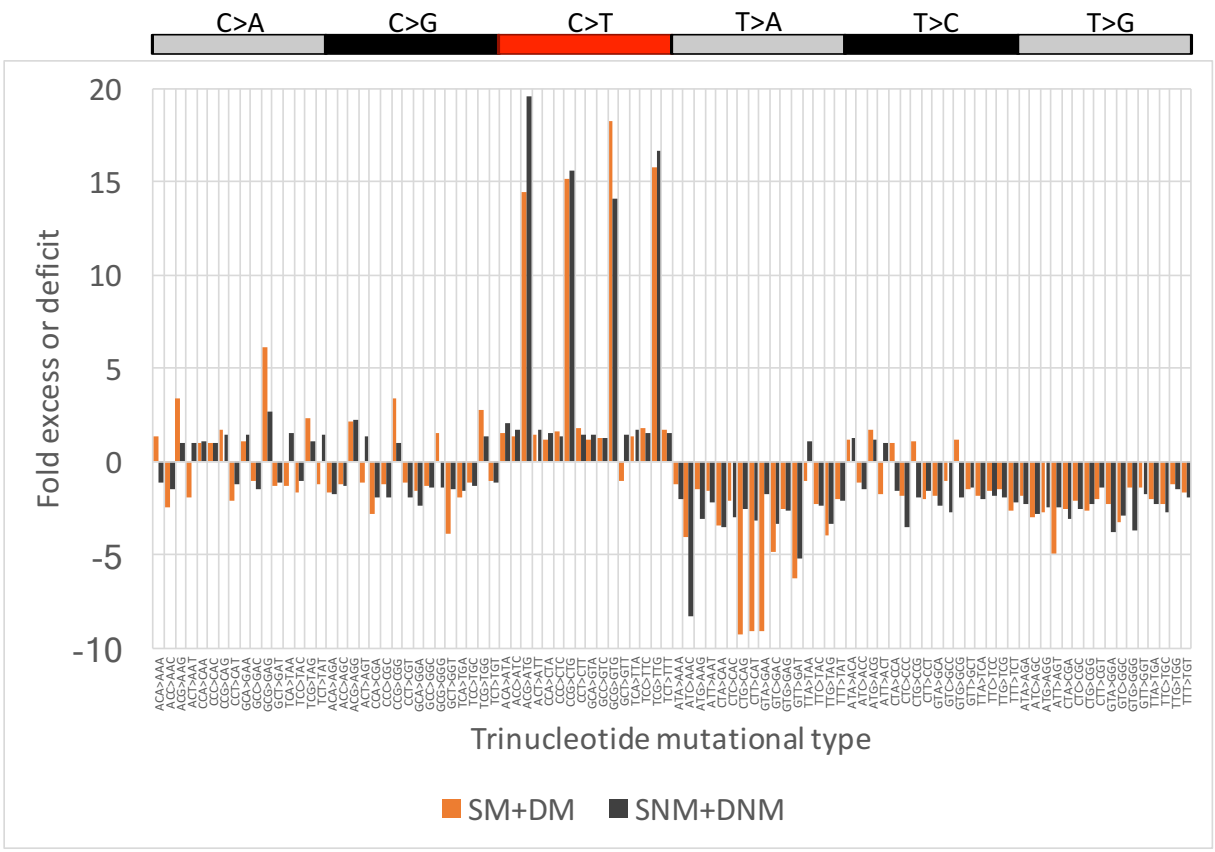
DenovoPedFilter (https://github.com/aeonsim/denovoPedFilter) was utilised to analyse the relationships between the 743 individuals identifying the 131 trios with at least one grand-offspring within the dataset and identified grandparents, parents, all descendants of the parents, proband, grand-offspring and unrelated individuals for each of the 131 trios. The genomes of all probands, grand-offspring and - where grandparents were present - parents were phased utilising Mendelian rules of inheritance and a filtered set of 7 million high confidence SNPs and INDELs (GATK VQSR 0.975, QUAL1000, MAF 0.1-0.9, TsTv 2.1) to determine parent of origin of all inherited haplotypes. Candidate dnms were then identified using all variants with a VCF QUAL > 100, and average map quality (MQ) > 50, and ten or more reads in each of the sire, dam and proband. The criteria for selecting a dnms was that the variant genotype was heterozygous in the proband (or homozygous on chrX in male probands), inherited by at least one grand-offspring, and homozygous reference in the parents, grand-parents and all unrelated individuals. Phase information from the grand-offspring inheriting the candidate dnm was used to determine the parent of origin of each selected variant. All candidate dnms where then targeted and re-genotyped using Freebayes (v1.1.0-1-gf15e66e, with the options --no-haplotypes, --no-mnps) (Garrison and Marth 2012) before being normalised and intersected with the GATK candidates. The freebayes reference and alternate read counts were then utilised to confirm the absence of the variant in grand-parents and/or parents and that the mean allelic dosage of grand-offspring carrying the variant was >0.25. The remaining candidate dnms were then assigned to one of 6 classes based on the degree of linkage between half-siblings or grand-offspring and the probability that the allelic dosage was different from 0.5 (Binomial distribution, $p < 0.05$).

The six classes and their criteria were:

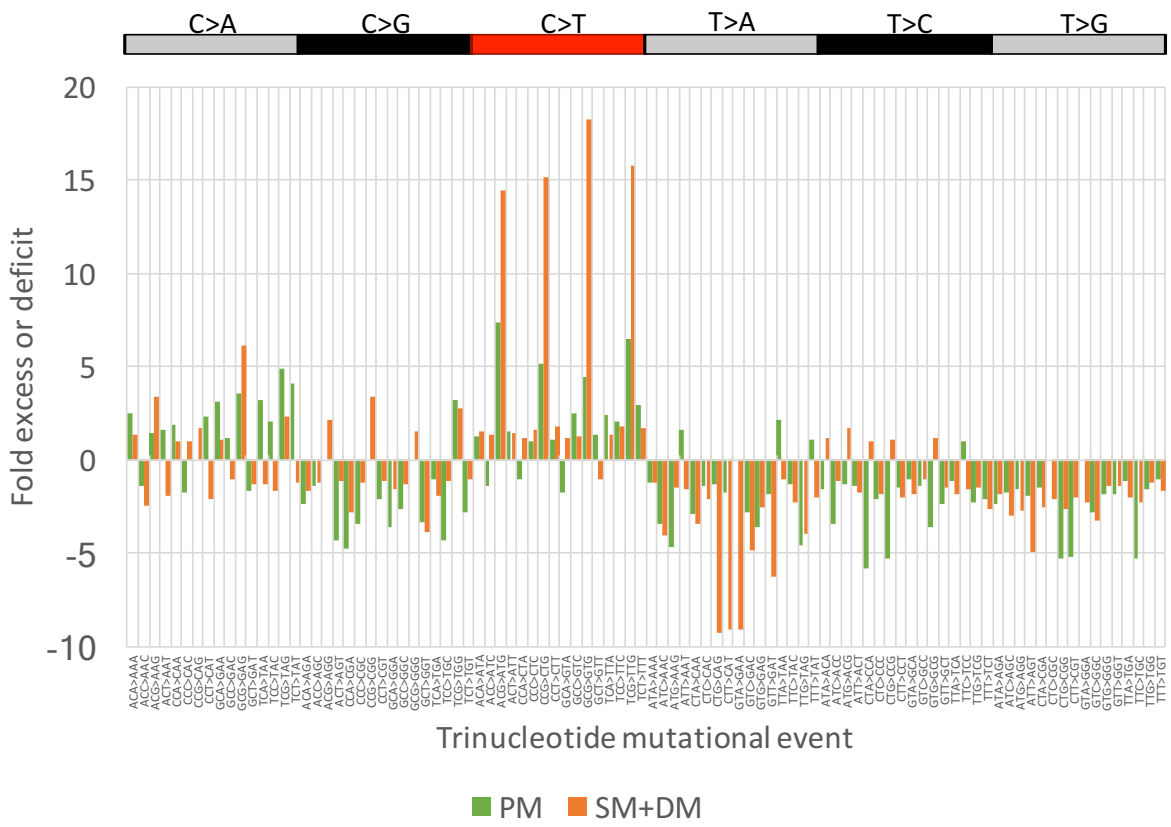
1. Proband mosaic (PM): no reads in parents, mutation not shared with half-siblings, perfect but incomplete linkage among grand-offspring and proband allelic dosage significantly different from 0.5 ($P < 0.05$, Binomial test)
2. Sire non-mosaic (SNM): no reads in parents, mutation on paternal haplotype, complete and perfect linkage in grand-offspring, variant not shared with proband half-siblings
3. Dam non-mosaic (DNM): no reads in parents, mutation on maternal haplotype, complete and perfect linkage in grand-offspring, variant not shared with proband half-siblings
4. Sire mosaic (SM): reads present in sire or variant shared with proband's paternal half-siblings, sire allelic dosage significantly different from 0.5 ($P < 0.005$, Binomial test), mutation on paternal haplotype, complete and perfect linkage in grand-offspring
5. Dam mosaic (DM): reads present in dam or variant shared with proband's maternal half-siblings, dam allelic dosage significantly different from 0.5 ($P < 0.005$, Binomial test), mutation on maternal haplotype, complete and perfect linkage in grand-offspring
6. Grandparental: parent allelic dosage not significantly different from 0.5 ($P > 0.005$, Binomial test)



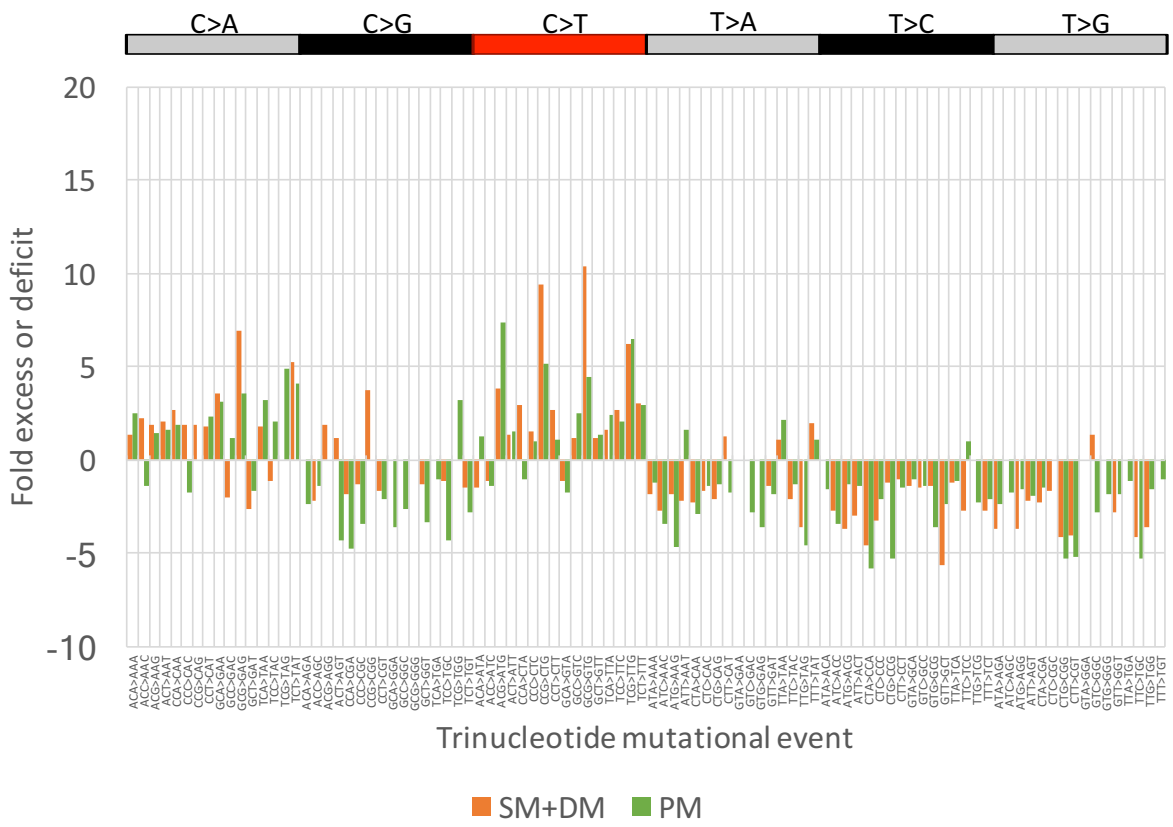
Supplementary figure 1: De novo mutation density by chromosome for the bovine reference genome (bosTau6). Bars indicate the average number of dnms per megabase for each chromosome, the horizontal black line indicates the average (2.79), red asterisk indicate significant difference from the mean (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$).



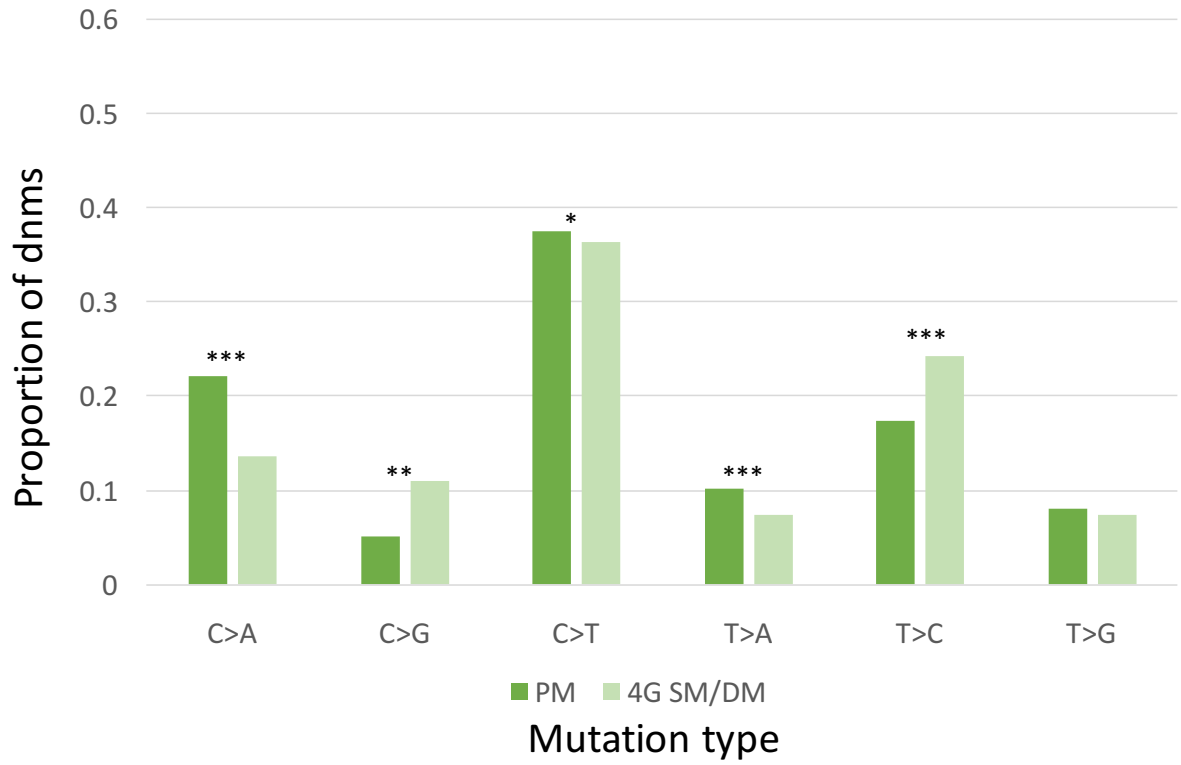
Supplementary figure 2: The trinucleotide mutational signatures of parental mosaic mutations, (SM + DM, orange) compared to NM dnms (black). The graph shows the fold excess or deficiency of observed dnms in each trinucleotide context over the expected proportions for each trinucleotide derived from the bovine reference genome (bosTau6) for each of the possible 96 trinucleotide contexts. (Y-axis fold excess or deficiency of dnms compared to reference genome, X-axis 96 possible trinucleotide substitutions, grouped by the six mutational classes (shown on the top bar)), P-values in Suppl. Table 4. Comparing the two mutational signatures there are several noticeable differences (C>A and T>A) between the early occurring parental mosaics (SM + DM dnms, occurring before formation of the primordial germ-cells) and the late occurring non-mosaic dnms (SNM + DNM, likely occurring after formation of the PGCs).



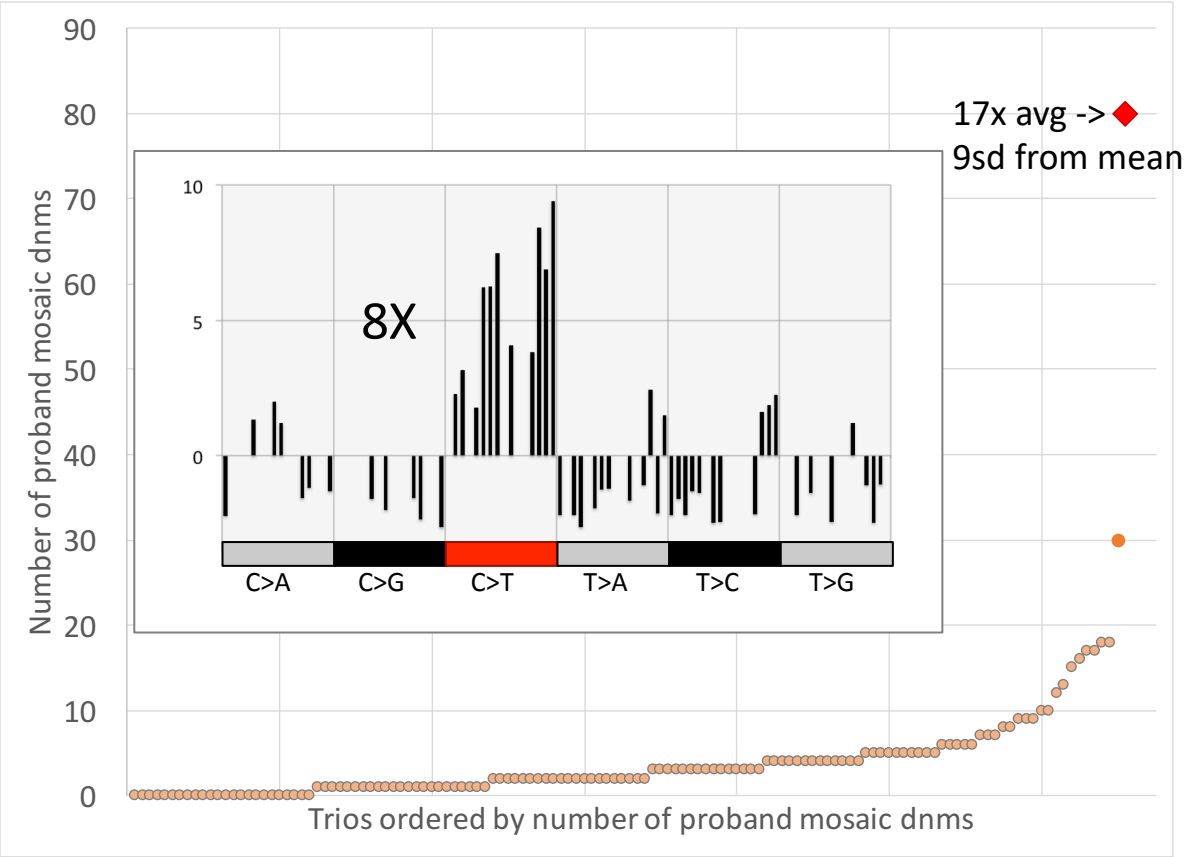
Supplementary figure 3: The trinucleotide mutational signatures of proband mosaic mutations, (PM, green) compared to the parental mosaic dnms (SM + DM, orange). The graph shows the fold excess or deficiency of observed dnms in each trinucleotide context compared to the expected proportions for each trinucleotide (derived from the bovine reference genome bosTau6) for each of the possible 96 trinucleotide contexts. (Y-axis fold excess or deficiency of dnms in each trinucleotide context compared to reference genome, X-axis 96 possible trinucleotide substitutions, grouped by the six mutational classes (shown on the top bar)), P-values in Suppl. Table 4. Noticeable differences between the parental mosaics and the proband mosaics are present, especially with regards to the C>T dnms at CpG sites. CpG C>T dnms are associated with the spontaneous deamination of methyl-cytosine to Thymine, which is a time dependent process rather than a cell division dependent mutational process. This agrees with our observation that the parental mosaic dnms generally occur later (between the 4th cell division and formation of the primordial germ-cells) than the PM dnms which occur in the first 3-4 cell divisions.



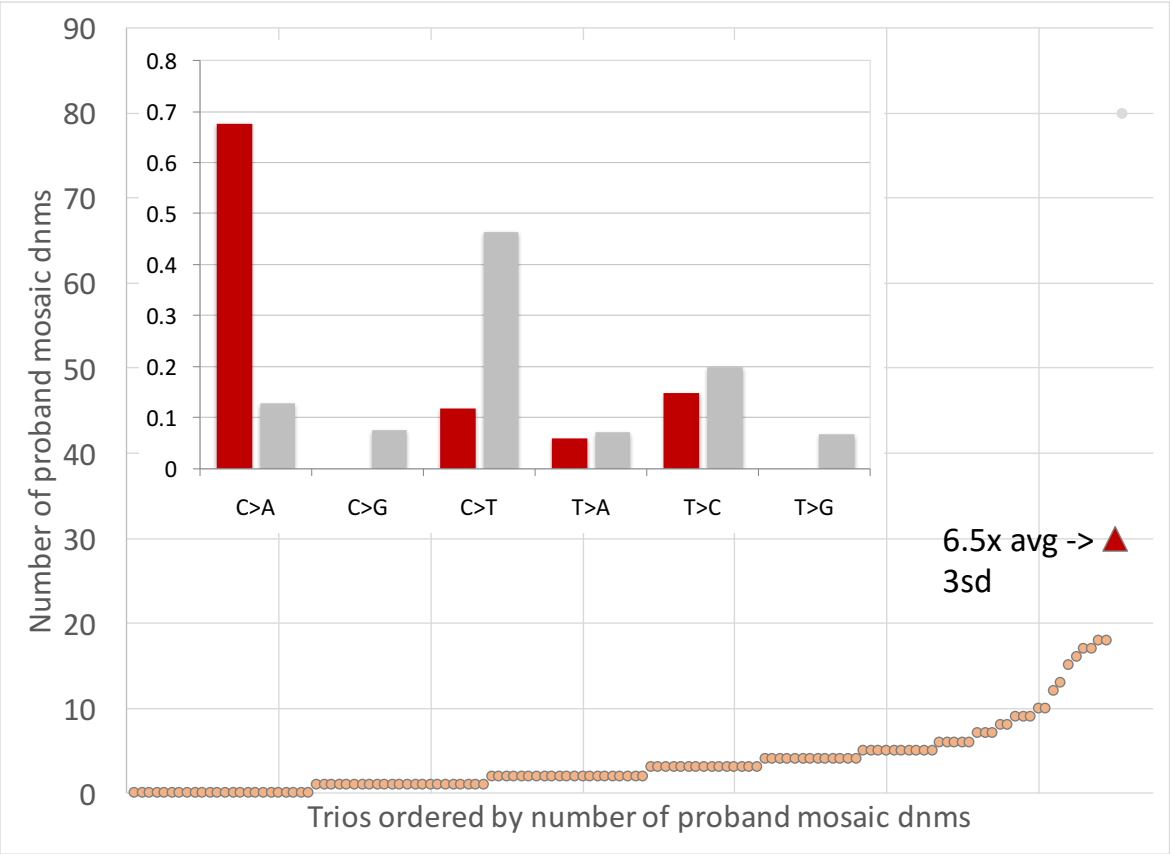
Supplementary Figure 4: The trinucleotide mutational signatures of proband mosaic mutations, (PM, green) compared to the high confidence (only using 4 generational families) parental mosaic dnms (SM + DM, orange). The graph shows the fold excess or deficiency of observed dnms in each trinucleotide context compared to the expected proportions for each trinucleotide (derived from the bovine reference genome bosTau6) for each of the possible 96 trinucleotide contexts. (Y-axis fold excess or deficiency of dnms in each trinucleotide context compared to reference genome, X-axis 96 possible trinucleotide substitutions, grouped by the six mutational classes (shown on the top bar)), P-values in Suppl. Table 4. Compared to Suppl. Fig 4 the differences between the high confidence parental mosaics and the proband mosaics is reduced especially with regards to the CpG C>T dnms, however the signatures still differ significantly. Thus part of the difference may be due to the accidental assignment of SNM or DNM dnms to the SM or DM classes, however there remains some difference, This is suggestive of a difference in the mutational processes between the PM dnms which occur in the first 3-4 cell divisions, and the parental mosaic dnms (SM/DM) which occur between the 4th cell division and the formation of the primordial germ-cells.



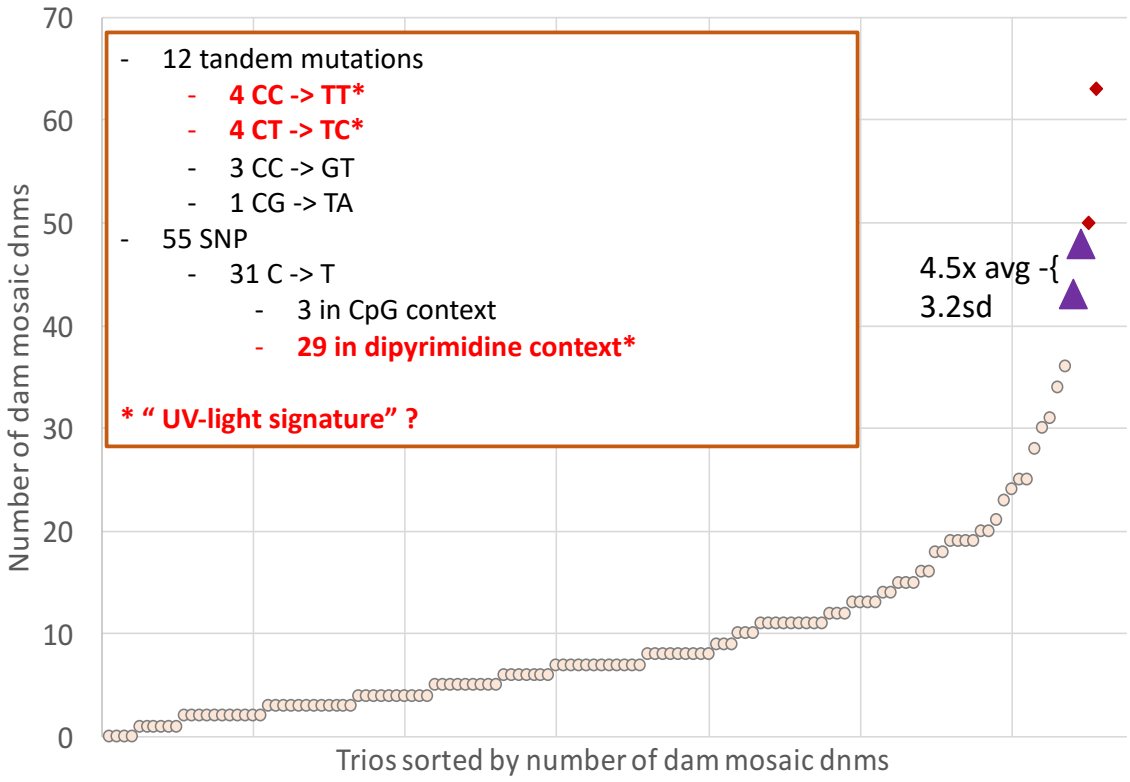
Supplementary Figure 5: Comparison of the proportion of the six possible nucleotide substitutions for PM dnms (green) compared to SM+DM dnms (pale green) from four generation pedigrees. Significant differences are shown by asterisks $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***. The Y-axis shows the proportion of each set of dnms, with the X-axis showing the six possible dnm events. Each of the six classes includes both the 5' and 3' version of each event i.e. C>A contains both C>A and G>T dnms. The significant differences observed between the two classes suggest there are differing process underlying the mutational process at the very earliest stages of embryo development (the PM dnms, first 4 cell divisions) and the slightly later parental mosaic mutations (SM/DM dnms, from cell divisions 3-4 through to formation of the primordial germ cells).



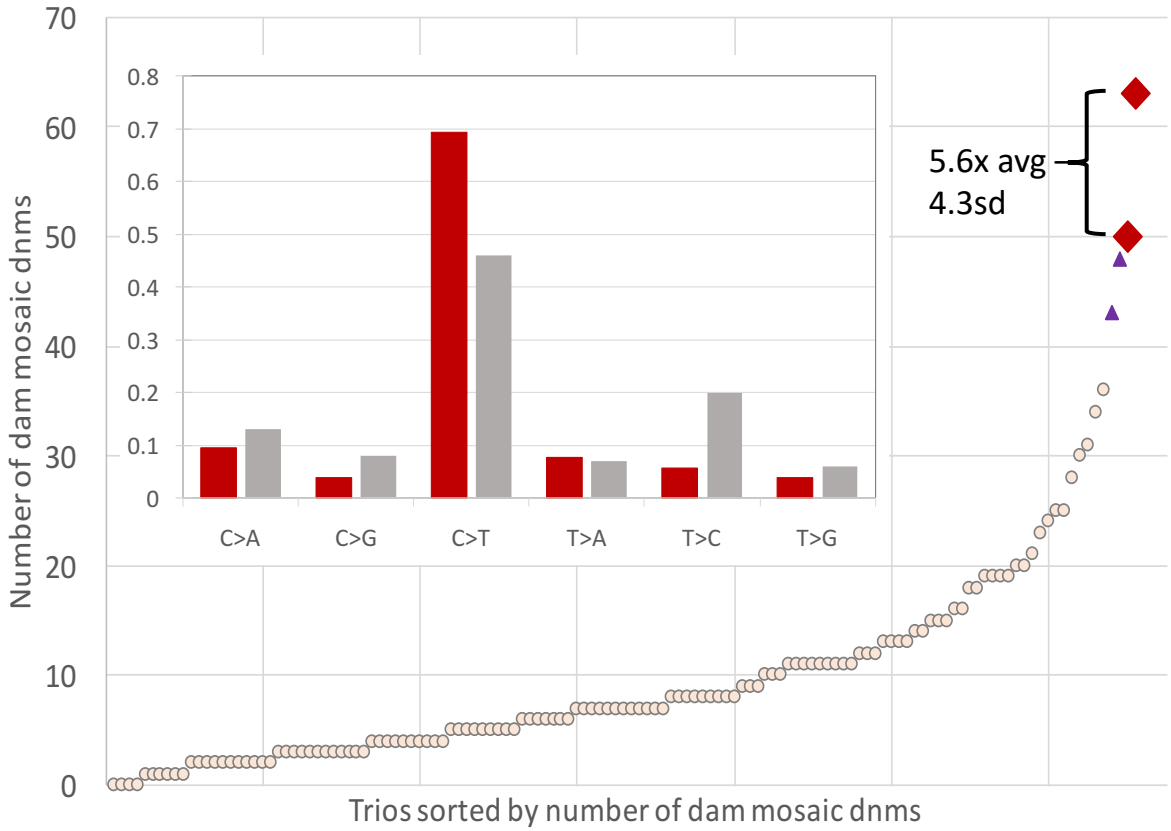
Supplementary Figure 6: Number of proband mosaic dnms (y-axis) per trio (x-axis ordered by number of PM dnms), showing outlier 1 (red diamond) with 17x the average number of PM dnms compared to the population. Outlier 2 is shown as the second orange circle, with the remaining population shown as pale orange dots. The insert graph shows the 8x increase in enrichment (y-axis, times enrichment) for C>T mutational events for outlier 1, with the 96 mutation classes grouped by the mutation event type (x axis).



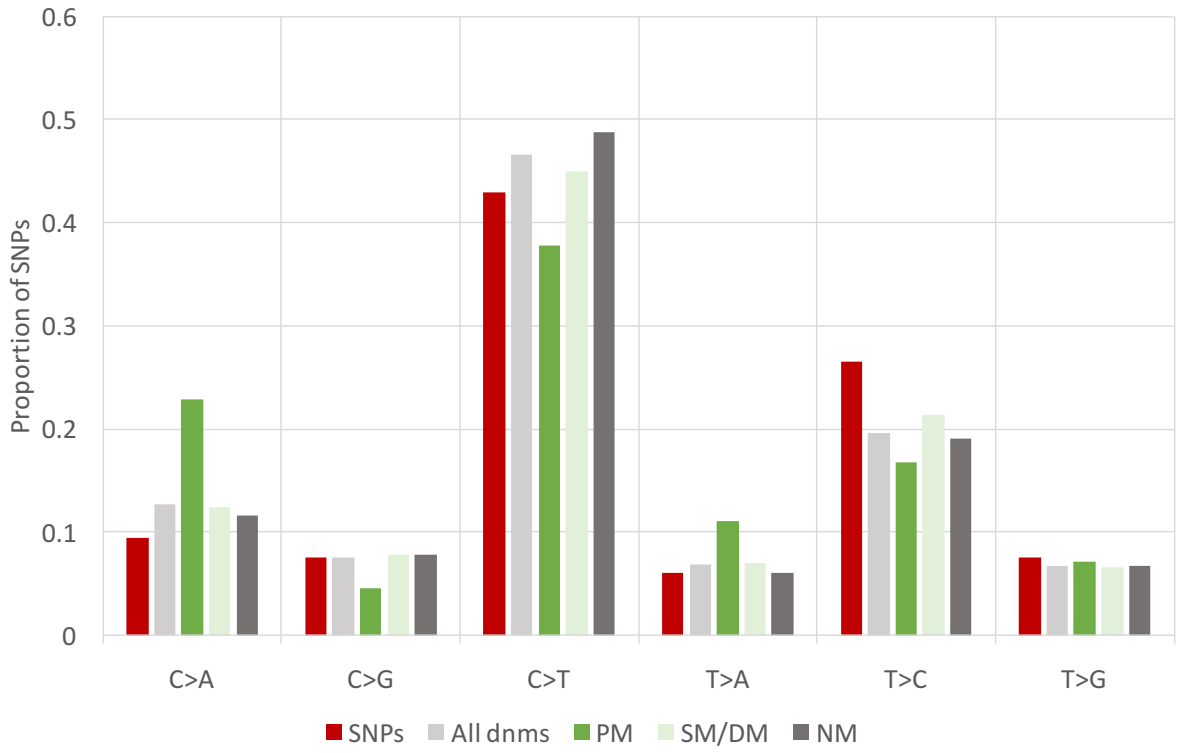
Supplementary Figure 7: Number of proband mosaic dnms (y-axis) per trio (x-axis ordered by number of PM dnms), showing outlier 2 (red triangle) with 6.5x the average number of PM dnms compared to the population. Outlier 1 is shown as the grey circle, with the remaining population shown as pale orange dots. The insert graph compares the mutational spectrum of PM dnms for outlier 2 (red) with a major increase in C>A mutations to the mutational spectrum of all the dnms (grey), y-axis is the percentage of dnms, x-axis shows mutational class.



Supplementary Figure 8: Number of dam mosaic dnms (y-axis) per trio (x-axis ordered by number of DM dnms), showing outlier 3 two offspring (purple triangles) with 4.5x the average number of DM dnms compared to the population. Outlier 4's offspring are shown as two red diamonds, with the remaining population shown as pale orange dots. The insert describes the mutational spectrum of the DM dnms observed in the two offspring of outlier 3, with a mutational signature that is similar to that reported for UV-light damage of DNA with an excess of CC>TT, CT>TC tandem mutations and dipyrimidine C>T mutations.



Supplementary Figure 9: Number of dam mosaic dnms (y-axis) per trio (x-axis ordered by number of DM dnms), showing outlier 4's two offspring (red diamonds) with 5.6x the average number of DM dnms compared to the population. Outlier 3's offspring are shown as two purple triangles, with the remaining population shown as pale orange dots. The insert graph compares the mutational spectrum of DM dnms for outlier 4 (red) with an increase in non-CpG C>T mutations to the mutational spectrum of all the dnms (grey), y-axis is the percentage of dnms, x-axis shows mutational class.



Supplementary Figure 10: Comparison of the proportion of the six possible nucleotide substitutions for (i) all the single nucleotide dnm detected in the study (All dnms; light grey), (ii) the PM dnm (PM; green), (iii) all SM + DM dnms (SM/DM; light green), (iv) all SNM + DNM SNPs (NM; dark grey), (v) 613,971 SNPs segregating in the Damona population at frequencies ≤ 0.01 (SNPs; red), P-values in Suppl. Table 2 and 3.

Supplementary Note 1: Differences between PM and SM+DM dnm classes.

This could in part be due to contamination of the SM+DM class with a small proportion of rare segregating SNPs from the three generation pedigrees, and some contamination of miss-assigned SNM+DNM dnms. When comparing the trinucleotide signatures of the PM dnms with those from our most confident set of SM+DM dnms (those from the four generation pedigrees) in Suppl. Fig. 4, we see the signatures are closer, than when comparing PM and the full SM+DM set (Suppl. Fig. 3), however there are still significant differences (Suppl. Fig. 5). A second contributing factor to these differences could be biological differences, as the PM and SM+DM dnms are different subsets of mosaic dnms. The PM dnms are identified in the proband as heterozygous variants with allelic dosages of ~0.15-0.4, and thus are likely to have occurred within the first two - three cell divisions after fertilisation. The SM+DM dnms however are identified in the parents as being homozygous reference with allelic dosages of less than 0.1-0.25 (depending on the depth of coverage) or being shared by half-sibling and thus are likely to have occurred after the first three – four cell divisions post fertilisation but early enough to still be present in the soma.

Supplementary Note 2: REML model

For the analysis of repeatability BLUPF90's renumf90 was utilised. After removal of the four outliers the mean number of dnms per gamete was fitted as a fixed effect. Two random effects were also fitted. An additive genetic effect which is specific to animals and covariance between effects depends on the additive genetic relationship. Effects are estimated by taking into account that fullsibs share 50% of their variants. A permanent environmental effect that models a random independent effect on each animal, representing something specific to the animal that affect each of his records.

Supplementary Table 1: Showing the raw data and metadata for each of the 131 pedigrees. Proband - offspring of the trio. Sire - father of the trio. Dam - mother of the trio. Proband gender – gender of the proband, F(emale) or M(ale). Proband reproductive tech – Reproductive technology used in the conception of the proband, AI (artificial insemination), MOET (multiple ovulation and embryo transfer), IVF/IVM (in vitro fertilisation and maturation). PM – unadjusted number of PM mutations that occurred during the development of the proband. SNM – unadjusted number of non-mosaic dnms inherited from the sire of the trio. SM – unadjusted number of mosaic mutations inherited from the sire. DNM - unadjusted number of non-mosaic dnms inherited from the dam of the trio. SM – unadjusted number of mosaic mutations inherited from the dam. Paternal GP – True, both grandparents on the sire's side are sequenced and included in the population. Maternal GP – True, both grandparents on the dam's side are sequenced and included in the population. Kids – number of grand-offspring sequenced for the trio. Avg kid depth – The average depth of coverage for the trios grand-offspring. Est % of genome – An estimate of the percentage of the genome (calculated from GATK SNP metadata) that has an average map quality of at least Q50 and has between 10-50 fold depth in each of the proband, sire and dam. Estimated % detected kids – the estimated percentage of heterozygous SNPs in the proband that would have been detected by one or more reads in at least one grand-offspring. True rate – the estimated dnm rate for the proband using SNM + SM + DNM + DM and correcting for the estimated fraction of the genome covered and the fraction of dnms detected by the grand-offspring. Human rate – the estimated dnm rate using criteria from the human studies of two generation pedigrees (absent in parents, present in proband), thus PM + SNM + DNM with correction for the estimated fraction of the genome covered and the fraction of dnms detected by the grand-offspring.

Proband	Sire	Dam	Proband Gender	Proband reproductive technology	PM	SNM	SM	DNM	DM	Paternal GP	Maternal GP	Kids	Avg Kid Depth	Est % of Genome	Est % detected Kids	True Rate	Humn Comp Rate
NL111116564	NL316419721	NL805719521	F	AI	6	16	13	7	23	-	-	5	5.44	93.9	0.85	1.4E-08	6.8E-09
NL140915109	NL316419721	NL805719521	F	AI	1	20	6	14	28	-	-	5	5.22	94.68	0.85	1.6E-08	8.1E-09
NL168409404	NL460942030	NL846687795	F	AI	3	21	2	9	7	-	-	5	6.52	94.71	0.92	8.4E-09	7.1E-09
NL213100979	NL780180664	NL129408912	F	AI	1	22	8	5	19	-	-	5	6.48	94.59	0.92	1.2E-08	6.0E-09
NL237429564	NL839380546	NL218568172	F	AI	7	12	2	17	3	-	YES	5	12.74	96.56	0.97	6.8E-09	7.2E-09
NL259818683	NL829877874	NL206865263	F	AI	0	1	2	0	0	-	-	2	5.45	22.95	0.53	4.6E-09	1.5E-09
NL294754005	NL839380546	NL240699372	F	AI	2	31	6	6	13	-	-	5	5.66	94.26	0.85	1.3E-08	9.1E-09
NL340920606	NL834961199	NL277557999	F	AI	2	19	10	8	2	-	-	5	5.42	95.44	0.85	9.0E-09	6.7E-09
NL340920675	NL834961199	NL288458588	F	AI	5	21	3	10	8	-	-	5	10.22	92.55	0.97	8.8E-09	7.5E-09
NL351142406	NL288458773	NL297960542	F	AI	0	9	14	3	3	YES	-	5	9.16	96.87	0.96	5.8E-09	2.4E-09
NL359643060	NL170266664	NL325763365	F	AI	3	27	3	4	2	-	-	5	5.68	91.33	0.85	8.7E-09	8.2E-09
NL386173114	NL970138118	NL226697857	F	AI	0	24	4	13	13	-	-	4	7.18	96.44	0.9	1.2E-08	8.0E-09
NL386947153	NL229469970	NL351142406	F	AI	2	31	22	3	3	-	YES	5	5.36	95.12	0.85	1.4E-08	8.3E-09
NL387028769	NL970163778	NL351128033	F	AI	1	22	19	5	16	-	-	7	11.51	95.82	0.99	1.2E-08	5.5E-09
NL389378161	NL970138118	NL321941657	F	AI	0	18	7	6	2	-	-	5	5.32	94.87	0.85	7.7E-09	5.6E-09
NL417222446	NL970138118	NL241207710	F	AI	5	30	7	10	5	-	-	5	5.68	95.73	0.85	1.2E-08	1.0E-08
NL424441748	FR2298044708	NL261840434	F	AI	1	37	3	13	3	-	-	5	6.4	93.77	0.92	1.2E-08	1.1E-08
NL442531687	NL970163778	NL351128033	F	AI	0	19	22	9	8	-	-	5	6.38	95.64	0.92	1.2E-08	6.0E-09
NL477956846	NL339291027	NL380853311	F	AI	0	16	3	6	7	-	-	1	21	96.18	0.5	1.2E-08	8.6E-09
NL106034655	NL316419721	NL816076800	M	AI	1	17	4	3	11	-	-	4	4.65	91.24	0.56	1.3E-08	7.7E-09
NL120873995	NL460508522	NL835809410	M	AI	1	17	9	6	4	-	-	5	5.84	94.77	0.85	8.4E-09	5.6E-09
NL159659261	NL460508522	NL461998375	M	AI	30	18	21	16	36	-	-	5	17.92	95.17	0.97	1.8E-08	1.3E-08
NL210903539	NL970092845	NL835486996	M	AI	0	20	20	8	4	-	-	5	6.06	92.87	0.92	1.1E-08	6.1E-09
NL246311263	NL817118723	NL129804189	M	AI	4	22	27	11	7	-	-	7	5.59	92.67	0.93	1.5E-08	8.0E-09
NL329293291	NL970119207	NL207789065	M	AI	0	26	9	6	8	-	-	5	5.42	91.74	0.85	1.2E-08	7.7E-09
NL367593805	NL170475222	NL297867801	M	AI	0	21	6	3	2	-	-	5	5.4	95.54	0.85	7.4E-09	5.5E-09
NL396647605	NL207288005	NL354090045	M	AI	4	25	29	3	4	-	-	5	21.3	91.8	0.97	1.3E-08	6.7E-09
NL482184166	NL970179333	NL424181217	M	AI	1	16	15	8	7	-	-	5	5.98	93.29	0.85	1.1E-08	5.9E-09
NL496442720	NL970163778	NL391733985	M	AI	2	23	23	8	13	-	-	5	6.68	92.61	0.92	1.5E-08	7.3E-09
NL499946960	NL970146623	NL449230705	M	AI	1	20	14	2	18	-	-	5	5.98	94.25	0.85	1.3E-08	5.4E-09
NL524213982	NL970163778	NL385310044	M	AI	0	22	20	9	2	-	-	5	5.2	94	0.85	1.2E-08	7.3E-09
NL525056357	NL355478033	NL394087241	M	AI	1	14	1	5	7	-	-	5	6.44	91.09	0.92	6.0E-09	4.5E-09
NL533957842	NL970163778	NL380853311	M	AI	3	24	15	8	7	-	-	5	6.2	94.48	0.92	1.2E-08	7.5E-09
NL535241381	NL443684324	NL467038088	M	AI	1	19	4	14	15	YES	-	5	5.58	94.79	0.85	1.2E-08	7.9E-09
NL727555328	NL396647605	NL420808451	M	AI	1	18	4	3	3	YES	-	5	25.76	94.63	0.97	5.7E-09	4.5E-09
NL856279191	NL319570393	NL319958070	M	AI	3	13	28	11	25	-	-	4	5.25	90.34	0.78	2.0E-08	7.2E-09
NL141317243	NL460508522	NL846687795	F	IVF/IVM	9	24	18	16	8	-	-	5	31.56	94.65	0.97	1.3E-08	1.0E-08
NL184750830	NL777133097	NL121357281	F	IVF/IVM	10	34	17	11	8	-	-	5	9.44	94.99	0.96	1.4E-08	1.1E-08
NL197086166	NL790545532	NL857940755	F	IVF/IVM	0	17	2	7	15	-	-	1	20.8	95.01	0.5	1.6E-08	9.5E-09
NL206910455	NL785532529	NL127584203	F	IVF/IVM	5	29	12	3	8	-	YES	5	5.66	93.16	0.85	1.2E-08	8.8E-09
NL210903041	NL785532529	NL846687795	F	IVF/IVM	1	28	6	7	11	-	-	5	5.12	91.75	0.85	1.2E-08	8.6E-09
NL215598758	NL970089335	NL140050323	F	IVF/IVM	18	28	24	14	14	-	-	7	18.64	95.15	0.99	1.6E-08	1.2E-08
NL218568172	NL970092845	NL144862979	F	IVF/IVM	0	16	6	7	4	-	-	1	40.7	92.8	0.5	1.3E-08	9.3E-09
NL227298149	NL785532529	NL127584203	F	IVF/IVM	6	24	5	7	9	-	YES	5	6.66	93.23	0.92	9.8E-09	8.1E-09
NL240698920	NL183627742	NL857940546	F	IVF/IVM	15	15	18	14	16	-	-	6	24.45	95.58	0.98	1.3E-08	8.8E-09
NL240699512	NL793715518	NL121737065	F	IVF/IVM	4	25	-	**7**	11	-	-	5	17.02	67.02	0.97	4.6E-08	1.0E-08
NL256039205	NL829877874	NL194064532	F	IVF/IVM	5	21	17	12	5	-	-	5	8.94	94.35	0.96	1.1E-08	7.9E-09
NL258237452	NL839380546	NL189787477	F	IVF/IVM	7	18	5	11	7	-	YES	5	6.74	98.29	0.92	8.5E-09	7.5E-09
NL267363937	NL159659261	NL240698920	F	IVF/IVM	3	21	23	2	11	YES	YES	5	25.56	94.95	0.97	1.2E-08	5.3E-09
NL277557634	NL112538714	NL141317243	F	IVF/IVM	4	7	0	6	7	-	YES	5	10.26	60.77	0.97	6.4E-09	5.4E-09
NL286575328	NL159659261	NL240698920	F	IVF/IVM	8	32	8	9	7	YES	YES	5	26.56	95.84	0.97	1.1E-08	9.9E-09
NL292559725	NL137409985	NL141317243	F	IVF/IVM	4	25	15	6	9	-	YES	5	5.98	95.11	0.85	1.3E-08	8.1E-09
NL298511699	NL189887001	NL240699512	F	IVF/IVM	4	25	24	9	2	-	YES	5	5.64	94.55	0.85	1.4E-08	8.9E-09
NL321193760	NL970119207	NL256039205	F	IVF/IVM	3	25	15	8	12	-	YES	5	5.86	92.46	0.85	1.4E-08	8.6E-09
NL327476193	NL141079778	NL248631985	F	IVF/IVM	4	19	21	13	11	-	-	5	5.04	95.25	0.85	1.5E-08	8.3E-09
NL330065029	NL970115159	NL215598758	F	IVF/IVM	10	23	7	6	11	-	-	5	5.62	95.32	0.85	1.1E-08	9.0E-09
NL330209906	NL777434192	NL214695128	F	IVF/IVM	7	25	15	6	8	-	YES	5	5.4	95.89	0.85	1.2E-08	8.7E-09
NL339290668	NL970115159	NL215598758	F	IVF/IVM	4	26	9	10	21	-	-	5	13.06	95.28	0.97	1.3E-08	8.1E-09
NL356407038	NL170266664	NL297960542	F	IVF/IVM	3	24	7	10	1	-	-	5	5.6	94.23	0.85	9.8E-09	8.7E-09
NL360739426	NL169100560	NL184750830	F	IVF/IVM	1	10	3	8	6	-	YES	4	5.35	94.84	0.78	6.8E-09	4.8E-09
NL366304275	NL193285101	NL325763365	F	IVF/IVM	1	18	7	8	0	YES	-	5	8	92.89	0.96	6.9E-09	5.7E-09
NL388584192	NL970137937	NL237429564	F	IVF/IVM	2	17	36	5	3	-	YES	5	7.02	95.72	0.94	1.3E-08	5.0E-09
NL392828174	NL970137937	NL340920675	F	IVF/IVM	1	35	22	12	4	-	YES	5	5.44	93.29	0.85	1.7E-08	1.1E-08
NL431724656	NL970159722	NL298511389	F	IVF/IVM	0	2	1	0	0	-	-	1	5.3	12.71	0.31	1.4E-08	9.5E-09
NL438495780	NL970159722	NL334319595	F	IVF/IVM	0	31	24	6	1	-	YES	5	5.24	91.21	0.85	1.5E-08	8.9E-09
NL473685751	NL970159722	NL395323658	F	IVF/IVM	0	11	8	5	4	-	-	1	23.3	94.23	0.5	1.1E-08	6.4E-09
NL494899498	NL970163778	NL397280456	F	IVF/IVM	2	31	26	6	3	-	-	5	5.54	95.95	0.85	1.5E-08	9.0E-09
BE119784595	NL970163778	NL389691563	M	IVF/IVM	1	32	23	4	3	-	-	5	5.44	93.12	0.85	1.5E-08	8.8E-09
NL166487864	NL460508522	NL846687795	M	IVF/IVM	2	23	12	4	14	-	-	5	5.36	94.76	0.85	1.2E-08	6.7E-09
NL167388753	NL460508522	NL846687795	M	IVF/IVM	6	25	9	8	10	-	-	5	5.38	92.97	0.85	1.2E-08	9.2E-09
NL192646453	NL780180664	NL125036074	M	IVF/IVM	13	26	12	7	6	-	-	5	5.56	96.95	0.85	1.2E-08	1.0E-08
NL194532510	NL319957882	NL817127945	M	IVF/IVM	2	22	1	9	8	-	-	6	18.25	90.89	0.98	8.4E-09	6.9E-09
NL197795743	NL780180664	NL125036074	M	IVF/IVM	5	44	6	16	1	-	-	5	5.78	96.93	0.85	1.5E-08	1.5E-08
NL240570370	NL817118723	NL857940546	M	IVF/IVM	4	15	24	10	2	-	-	5	6	93.07	0.92	1.1E-08	6.3E-09
NL277557665	NL839380546	NL174462507	M	IVF/IVM	3	21	1	12	31	-	-	5	5.58	95.23	0.85	1.5E-08	8.3E-09
NL288458773	NL137409985	NL141317243	M	IVF/IVM	17	19	23	15	11	-	YES	11	15.27	97.13	1	1.3E-08	9.8E-09
NL339291027	NL970115159	NL215598758	M	IVF/IVM	12	29	19	3	20	-	-	10	11.96	96.51	1	1.4E-08	8.5E-09
NL360454909	NL168175055	NL339290668	M	IVF/IVM	2	16	19	5	5	-	YES	5	5.42	93.78	0.85	1.1E-08	5.4E-09
NL364143450	NL168175055	NL339290668	M	IVF/IVM	4	11	10	4	6	-	YES	5	4.8	93.51	0.65	9.6E-09	5.9E-09
NL498338414	NL970163778	NL389691563	M	IVF/IVM	2	19	12	2	2	-	-	5	5.3	86.87	0.85	8.9E-09	5.8E-09
NL498338423	NL970146623	NL389691563	M	IVF/IVM	2	16	9	7	3	-	-	5	5.96	90.72	0.85	8.5E-09	6.1E-09
NL528258545	NL970163778	NL473685751	M	IVF/IVM	80	79	16	42	7	-	YES	5	6.24	95.33	0.92	3.1E-08	4.3E-08
NL865099832	NL355478033	NL396647395	M	IVF/IVM	6	30	1	3	3	-	YES	5	5.58	95.79	0.85	8.5E-09	9.0E-09
NL127584203	NL460508522	NL835487153	F	MOET	17	33	14	15	15	-	-	5	21.56	93.73	0.97	1.6E-08	1.3E-08
NL144862979	NL460508522	NL835487153	F	MOET	9	24	24	5	20	-	-	5	9.22	93.44	0.96	1.5E-08	7.9E-09
NL167741552	NL460942030	NL825215467	F	MOET	5	22	5	10	4	-	-	5	6.38	94.43	0.92	8.8E-09	8.0E-09
NL181463252	NL318340810	NL825215467	F	MOET	18	26	16	23	3	-	-	4	5.48	93.79	0.78	1.7E-08	1.7E-08
NL189787477	NL790545532	NL857940755	F	MOET	2	13	5	8	2	-	-	5	11.92	96.37	0.97	5.6E-09	4.6E-09
NL199348044	NL115467411	NL818683493	F	MOET	3	16	7	7	10	-	-	5	6.88	91.24	0.92	8.9E-09	5.8E-09
NL199348068	NL115467411	NL818683493	F	MOET	2	10	3	6	13	-	-	5	7.28	91.1	0.94	7.0E-09	3.9E-09
NL205349069	NL970089335	NL835487153	F	MOET	8	17	9	9	24	-	-	5	8.66	96.08	0.96	1.2E-08	6.9E-09
NL205349090	NL970089335	NL835487153	F	MOET	3	11	17	3	18	-	-	5	7.04	94.54	0.94	1.0E-08	3.6E-09
NL206865263	NL110314073	NL846687795	F	MOET	1	18	5	7	7	-	-	5	11.5	94.22	0.97	7.6E-09	5.3E-09
NL206865333	NL110314073	NL835487153	F	MOET	5	13	9	7	19	-	-	5	6.14	94.7	0.92	1.0E-08	5.4E-09
NL206865441	NL780180664	NL846687795	F	MOET	6	22	6	10	10	-	-	5	15.66	94.48	0.97	9.8E-09	7.8E-09
NL214695128	NL149719461	NL127584203	F	MOET	3	17	7	11	30	-	YES	5	8.86	94.97	0.96	1.3E-08	6.4E-09
NL236907247	NL149719461	NL159660078	F	MOET	2	8	6	7	19	-	-	5	23.48	95.05	0.97	8.1E-09	3.5E-09
NL237883227	NL817118723	NL194064284	F	MOET	2	20	23	10	4	-	-	5	6.12	92	0.92	1.3E-08	7.1E-09
NL240570356	NL815222426	NL121737065	F	MOET	1	17	35	12	19	-	-	5	6.58	92.96	0.92	1.8E-08	6.6E-09
NL240698944	NL183627759	NL121357281	F	MOET	4	26	25	9	8	-	-	5	5.78	93.11	0.85	1.6E-08	9.2E-09
NL266148775	NL831375195	NL215598758	F	MOET	3	14	19	5	11	-	-	5	5.86	94.93	0.85	1.1E-08	5.1E-09
NL272931244	NL197858705	NL197086166	F	MOET	2	13	2	3	2	-	YES	5	5.6	86.77	0.85	5.1E-09	4.6E-09
NL288458766	NL839380546	NL127584203	F	MOET	0	14	12	25	12	-	YES	5	7.02	95.93	0.94	1.3E-08	8.1E-09
NL296787678	NL141079778	NL240570572	F	MOET	9	19	20	9	5	-	-	5	16.92	94.55	0.97	1.1E-08	7.6E-09
NL298511389	NL970119980	NL206865263	F	MOET	0	29	3	5	11	-	YES	5	6.66	94.77	0.92	1.0E-08	7.3E-09
NL334319595	NL970115159	NL206865441	F	MOET	0	17	6	8	6	-	YES	2	22.25	94.82	0.75	9.7E-09	6.6E-09
NL334319867	NL160259407	NL206865441	F	MOET	1	18	10	8	5	-	YES	5	5.44	92.83	0.85	9.7E-09	6.4E-09
NL340920017	NL160259407	NL206865441	F	MOET	3	23	16	12	7	-	YES	5	5.48	93.69	0.85	1.4E-08	8.9E-09
NL354090045	NL170475222	NL240699512	F	MOET	0	23	9	4	5	-	YES	5	20.54	92.16	0.97	8.6E-09	5.7E-09
NL368790870	NL193285101	NL236907247	F	MOET	1	20	3	9	12	YES	YES	5	24.76	94.13	0.97	9.0E-09	6.2E-09
NL375045011	NL834961199	NL334319595	F	MOET	4	21	18	2	6	-	-	5	5.3	95.18	0.85	1.1E-08	6.2E-09
NL376245698	NL170475222	NL277557634	F	MOET	1	16	9	5	2	-	YES	5	4.88	96.42	0.65	9.6E-09	6.6E-09
NL378593582	NL834961199	NL247020663	F	MOET	2	31	5	7	4	-	-	5	6.1	92.43	0.92	1.0E-08	8.8E-09
NL386173138	NL970138118	NL226697857	F	MOET	5	18	4	6	6	-	-	4	5.43	98.2	0.78	8.3E-09	7.1E-09
NL386947122	NL117720006	NL288588900	F	MOET	5	23	11	9	1	-	-	5	6.58	93.26	0.92	9.6E-09	8.1E-09
NL395323658	NL970146623	NL296787678	F	MOET	2	34	**17**	10	48	-	YES	5	15.44	95.6	0.97	2.2E-08	9.3E-09
NL396647364	NL117720006	NL288588900	F	MOET	5	32	14	8	9	-	-	5	6.62	94.49	0.92	1.4E-08	9.7E-09
NL396647395	NL970146623	NL296787678	F	MOET	2	26	8	14	43	-	YES	5	10.3	95.18	0.97	1.8E-08	8.5E-09
NL419619493	NL207288005	NL354090045	F	MOET	1	16	13	8	3	-	YES	5	5.86	90.94	0.85	9.7E-09	6.1E-09
NL420753221	NL970138118	NL335265509	F	MOET	1	29	**3**	29	63	-	-	5	8.46	95.55	0.96	2.5E-08	1.2E-08
NL422595517	NL288458773	NL335265509	F	MOET	0	9	7	8	50	YES	-	9	21.27	95.86	1	1.4E-08	3.3E-09
NL473188508	NL970118312	NL387117788	F	MOET	2	18	12	7	1	-	-	6	6.15	89.3	0.95	8.4E-09	6.0E-09
NL485570560	NL339291027	NL420753221	F	MOET	0	20	5	3	5	-	-	1	6.8	93.13	0.39	1.7E-08	1.2E-08
NL709935540	NL482184166	NL442531687	F	MOET	0	1	0	0	1	-	-	1	34.9	8.52	0.5	8.8E-09	4.4E-09
NL124929177	NL775328514	NL839055341	M	MOET	3	28	8	7	4	-	-	5	5.08	95.14	0.85	1.1E-08	8.8E-09
NL193285101	NL780180664	NL163784142	M	MOET	16	23	8	9	6	-	-	5	23.3	96.76	0.97	9.2E-09	9.6E-09
NL256086528	NL198991265	NL194481717	M	MOET	4	18	12	6	34	-	-	5	5.38	92.26	0.85	1.7E-08	6.7E-09
NL341882275	NL170475222	NL198933717	M	MOET	3	20	6	8	25	-	-	7	5.44	96.91	0.93	1.2E-08	6.4E-09
NL443684324	NL288458773	NL395323658	M	MOET	2	18	14	7	0	YES	YES	5	8.44	98.2	0.96	7.7E-09	5.4E-09
NL523728151	NL970146623	NL298780035	M	MOET	0	28	9	7	5	-	-	5	5.76	95.05	0.85	1.1E-08	8.1E-09
NL539258462	NL331879452	NL477956846	M	MOET	0	19	20	9	5	-	YES	5	6.2	95.19	0.92	1.1E-08	6.0E-09

Supplementary Table 2: Effects and P values for mutational spectrum when comparing rare segregating variants to dnms. The effects column provides the type of effect for each tested mutational class and event with the options, significant excess (sig+), significant deficit (sig-), near significance excess (NS+), near significance deficit (NS-), no effect (NE). All p values are Sidak corrected, with a lower limit of $6x10^{-5}$ due to the 10^5 permutations used in the test.

	PM effect	PM P =	SM + DM effect	SM + DM P=	SNM + DNM effect	SNM + DNM P=	All dnm effect	All dmn P=
C>A/G>T	sig+	6.00E-05	sig+	1.74E-03	sig+	5.57E-03	sig+	6.00E-05
C>G/G>C	sig-	4.31E-03	NE	9.18E-01	NE	9.18E-01	NE	4.24E-01
C>T/G>A	NS-	5.13E-02	NE	1.00E+00	sig+	6.00E-05	sig+	6.00E-05
T>A/A>T	sig+	2.40E-04	NE	1.00E+00	NE	9.94E-01	sig+	3.21E-02
T>C/A>G	sig-	6.00E-05	sig-	6.00E-05	sig-	6.00E-05	sig-	6.00E-05
T>G/A>C	NE	9.89E-01	NE	7.20E-01	NE	7.96E-01	NE	6.63E-01

Supplementary Table 3: Effects and P values for mutational spectrum when comparing late dnms (SNM + DNM) to early dnms (PM, SM+PM). The effects column provides the type of effect for each tested mutational class and event with the options, significant excess (sig+), significant deficit (sig-), near significance excess (NS+), near significance deficit (NS-), no effect (NE). All p values are Sidak corrected, with a lower limit of $6x10^{-5}$ due to the 10^5 permutations used in the test.

	PM effect	PM P =	SM + DM effect	SM + DM P=
C>A/G>T	sig+	6.00E-05	NE	1.00E+00
C>G/G>C	sig-	4.91E-03	NE	9.19E-01
C>T/G>A	sig-	6.00E-05	sig-	3.60E-04
T>A/A>T	sig+	6.00E-05	NE	1.00E+00
T>C/A>G	NE	4.25E-01	sig+	2.46E-02
T>G/A>C	NE	9.88E-01	NE	7.17E-01

Supplementary Table 4: Differences in trinucleotide mutational signatures. Sidak corrected p-values for the trinucleotide dnm frequency differences between different dnm classe, due to the number of tests and limited dataset size power to detect differences is limited. PM (early) vs SNM+DNM (late);

SM+DM (early) vs SNM+DNM (late); PM (early) vs DM+SM (early). Red highlighting indicates significant $p < 0.05$, Yellow approaching significance $p < 0.25$, Green $p < 0.5$, white $p > 0.5$

Events	TRI-NUC	PM vs SNM+DNM P value	SM+DM vs SNM+DNM P value	PM vs SM+DM p value
C>A | G>T	ACA>AAA	0.21361	0.76814	0.99854
C>A | G>T	ACC>AAC	1	1	1
C>A | G>T	ACG>AAG	1	0.97033	1
C>A | G>T	ACT>AAT	1	1	1
C>A | G>T	CCA>CAA	1	1	1
C>A | G>T	CCC>CAC	1	1	1
C>A | G>T	CCG>CAG	1	1	1
C>A | G>T	CCT>CAT	0.30682	1	0.62916
C>A | G>T	GCA>GAA	0.87079	1	0.99994
C>A | G>T	GCC>GAC	1	1	1
C>A | G>T	GCG>GAG	1	0.99678	1
C>A | G>T	GCT>GAT	1	1	1
C>A | G>T	TCA>TAA	0.38136	1	0.24477
C>A | G>T	TCC>TAC	0.903	1	0.8528
C>A | G>T	TCG>TAG	0.96597	1	1
C>A | G>T	TCT>TAT	0.00955	0.99057	0.12752
C>G | G>C	ACA>AGA	1	1	1
C>G | G>C	ACC>AGC	1	1	1
C>G | G>C	ACG>AGG	1	1	1
C>G | G>C	ACT>AGT	0.96416	0.98952	1
C>G | G>C	CCA>CGA	1	1	1
C>G | G>C	CCC>CGC	1	1	1
C>G | G>C	CCG>CGG	1	0.96603	0.99885
C>G | G>C	CCT>CGT	1	1	1
C>G | G>C	GCA>GGA	1	1	1
C>G | G>C	GCC>GGC	1	1	1
C>G | G>C	GCG>GGG	1	1	1
C>G | G>C	GCT>GGT	1	1	1
C>G | G>C	TCA>TGA	1	1	1
C>G | G>C	TCC>TGC	1	1	1
C>G | G>C	TCG>TGG	1	1	1
C>G | G>C	TCT>TGT	1	1	1
C>T | G>A	ACA>ATA	0.99839	0.14906	1
C>T | G>A	ACC>ATC	0.99905	0.99964	1
C>T | G>A	ACG>ATG	0.08545	0.00288	1
C>T | G>A	ACT>ATT	1	1	1
C>T | G>A	CCA>CTA	1	1	1
C>T | G>A	CCC>CTC	1	1	1
C>T | G>A	CCG>CTG	0.20217	0.99976	0.9681
C>T | G>A	CCT>CTT	1	1	1
C>T | G>A	GCA>GTA	0.9942	0.99999	1
C>T | G>A	GCC>GTC	0.88298	1	0.99815
C>T | G>A	GCG>GTG	0.61524	1	0.58058
C>T | G>A	GCT>GTT	1	1	1
C>T | G>A	TCA>TTA	1	1	0.99977
C>T | G>A	TCC>TTC	1	1	1
C>T | G>A	TCG>TTG	0.4501	0.97376	0.99997
C>T | G>A	TCT>TTT	0.08545	0.44207	0.99774
T>A | A>T	ATA>AAA	1	1	1
T>A | A>T	ATC>AAC	1	1	1
T>A | A>T	ATG>AAG	1	1	1
T>A | A>T	ATT>AAT	0.2382	0.99913	0.92297
T>A | A>T	CTA>CAA	1	1	1
T>A | A>T	CTC>CAC	1	1	1
T>A | A>T	CTG>CAG	1	1	0.99993
T>A | A>T	CTT>CAT	1	1	1
T>A | A>T	GTA>GAA	1	0.87441	1
T>A | A>T	GTC>GAC	1	1	1
T>A | A>T	GTG>GAG	1	1	1
T>A | A>T	GTT>GAT	1	1	1
T>A | A>T	TTA>TAA	0.98181	1	0.99981
T>A | A>T	TTC>TAC	1	1	1
T>A | A>T	TTG>TAG	1	1	1
T>A | A>T	TTT>TAT	0.98382	0.90049	1
T>C | A>G	ATA>ACA	0.9995	0.9963	1
T>C | A>G	ATC>ACC	1	1	1
T>C | A>G	ATG>ACG	1	1	1
T>C | A>G	ATT>ACT	1	0.31412	1
T>C | A>G	CTA>CCA	1	1	0.99999
T>C | A>G	CTC>CCC	1	0.99815	1
T>C | A>G	CTG>CCG	1	0.51597	0.5922
T>C | A>G	CTT>CCT	1	1	1
T>C | A>G	GTA>GCA	1	1	1
T>C | A>G	GTC>GCC	1	0.48358	1
T>C | A>G	GTG>GCG	1	0.74566	0.9894
T>C | A>G	GTT>GCT	1	1	1
T>C | A>G	TTA>TCA	1	1	1
T>C | A>G	TTC>TCC	0.99287	1	1
T>C | A>G	TTG>TCG	1	1	1
T>C | A>G	TTT>TCT	1	1	1
T>G | A>C	ATA>AGA	1	1	1
T>G | A>C	ATC>AGC	1	1	1
T>G | A>C	ATG>AGG	1	1	1
T>G | A>C	ATT>AGT	1	1	1
T>G | A>C	CTA>CGA	1	1	1
T>G | A>C	CTC>CGC	1	1	1
T>G | A>C	CTG>CGG	1	1	1
T>G | A>C	CTT>CGT	0.99999	0.99891	1
T>G | A>C	GTA>GGA	1	1	1
T>G | A>C	GTC>GGC	1	1	1
T>G | A>C	GTG>GGG	1	0.99999	1
T>G | A>C	GTT>GGT	1	1	1
T>G | A>C	TTA>TGA	1	1	1
T>G | A>C	TTC>TGC	1	1	1
T>G | A>C	TTG>TGG	1	1	1
T>G | A>C	TTT>TGT	1	1	1

Experimental Section

Study 3:

A polymorphic *ERV* element that is mobilized in the germline of specific individuals causes abetalipoproteinemia and hypolipidemia in cattle by disrupting the *APOB* gene.

In preparation

Harland C, Karim L, Durkin K, Artesi M, Sartelet A, Knapp E, Tamma N, Mullaart E, Coppieters W, Georges M & Charlier C

Abstract

A lethal abetaliproteinemia and hypolipidemia with autosomal recessive mode of inheritance was recently described in Holstein Friesian Cattle. The corresponding locus was assigned to the 2.6 Mb chr11:74.5-77.1Mb interval by autozygosity mapping. We herein show that the causative mutation corresponds to the insertion of an endogenous retroviral ERV-K element in exon 5 of *Apolipoprotein B* (*APOB*) gene, resulting in premature transcriptional termination of the gene. We identify more than 1,000 insertion sites of this ERV element that are polymorphic in cattle. Underrepresentation and shifts towards lower allelic frequencies of genic (vs intergenic) sense (vs antisense) insertions testifies of their capacity to cause deleterious phenotypic effects. We take advantage of the Damona dataset of 743 whole genome sequences designed for the detection of dnm to estimate the average transposition rate at one event per 30 gametes. We present strong evidence that the transposition rate may be as much as 5-fold higher in specific individuals and germ-cell lineages.

Recently, multiple cases of male and female calf mortality following a 1-5 month period of failure of thrive, diarrhoea, and pathognomonic low levels of total, HDL and non-HDL cholesterol were reported in Holstein-Friesian cattle (Kipp et al. 2015). All affected animals descended from *Maughlin Storm* on paternal and maternal side, suggesting autosomal recessive inheritance. Accordingly, autozygosity mapping revealed a 2.6Mb BTA11 haplotype (74.5-77.1Mb; *bosTau6* reference genome assembly), referred to as '*Cholesterol Deficiency' (CD)* haplotype, for which all cases were homozygous. Lists of bulls carrying the *CD* haplotype were published by breeding companies (including: CRV, The Netherlands; VIT, Germany; GEN'France, France; WestGen, Canada).

The whole genome of four of these carrier bulls had been sequenced at ≥ 20-fold depth as part of the Damona project. The *Damona* project was primarily designed for the detection of dnms and involves the sequencing of 131 three generation pedigrees comprising sire, dam, offspring sequenced at ≥ 20-fold depth, plus an average of five grand-offspring sequenced at ≥ 4-fold depth. Mining the corresponding dataset in the chr11:74.5-77.1Mb interval for disruptive mutations shared heterozygous by the four carrier sires did not reveal any obvious candidate causative gene nor mutation. We noted, however, that the *APOB* gene, coding for the main apolipoprotein of chylomicrons, very low-density lipoproteins (VLDL), and low-density lipoproteins (LDL), mapped at position chr11:77.9-78.0Mb, hence closed to the published interval. More than 60 loss-of-function mutations in the *APOB* gene have been reported in humans, causing familial hypolipoproteinemia (FHBL) with symptoms that are highly reminiscent of the bovine condition. While no disruptive *APOB* mutations shared heterozygous by the four carriers could be predicted by SNP Effect Predictor (McLaren et al. 2010), visual examination of the raw BAM files in the Integrative Genomics Viewer (IGV) identified a cluster of discordant paired reads accounting collectively for ~50% of sequence depth in exon 5, and pointing towards an LTR element insertion shared by the four carriers as well as 11 more animals in the Damona dataset (Figure 1A) (Robinson et al. 2011). All of these would trace back to *Maughlin Storm*, the acknowledged introducer of the insertional mutation in the *APOB* gene in Holstein-Friesians (Kipp et al. 2015). We manually assembled two non-overlapping segments (5' and 3' boundaries) of an long terminal repeat (LTR) element from the discordant mates, and used split reads to define the breakpoints at single-base pair resolution, revealing a 6-bp target site duplication (TSD) typical of retrotransposition events (Figure 1B)(f.i. Marchi et al. 2014). We developed a PCR-based assay and genotyped ~700 of the Damona animals as well as ~500 additional Dutch Holstein animals (Suppl. Fig. 1). We confirmed the carrier status of the 15 animals identified from the sequence data and estimated the frequency of the mutation in the Dutch Holstein current population at ~4%.

To further characterize the LTR element insertion, we performed long range PCR from genomic DNA of two affected calves shown with the genotyping assay to be homozygous mutant. The ensuing gel-

purified ~7Kb amplification product was sequenced by (i) shotgun sequencing on a MiSeq instrument after generating a NexteraXT DNA library (Illumina), complemented by (ii) long single molecule sequencing on a MinION instrument (Oxford Nanopore). The complete sequence of the LTR element was assembled using custom-made scripts, showing that it corresponds to a full-length -K element (one of the 24 known bovine ERVs; (Garcia-Etxebarria and Jugo 2010)), however, with mutated (loss-of-function) open reading frames (*GAG*, *POL*, *ENV*) flanked by 1,287 bp identical LTRs (Suppl. Fig. 2, Suppl. file 1). We further extracted total RNA from the livers of the two affected calves and generated a strand-specific rRNA-depleted cDNA library that was sequenced (2x76bp paired-ends) on a HiSeq 2000 instrument. We analyzed the ensuing ~80 million bp of reads with TopHat/Cufflink and visualized the outputs with IGV, revealing complete *APOB* transcriptional shutoff downstream of the LTR insertion (Figure 1C)(Trapnell et al. 2012; Robinson et al. 2011). The RNA-Seq results were confirmed by QRT-PCRs targeting upstream (3-4) as well as downstream exons (6-7 and 7-8). The mutant mRNA was targeted via 3' RACE, showing that transcriptional termination results from the presence of a polyadenylation signal in the ERV LTR, which leads to premature formation of a poly (A) tail (Suppl. Fig.3). Provided that the truncated mRNA remains translatable, the encoded protein would only represent <3% of the full-length protein.

The previous findings resulted in the identification of an ERV-type transposable element that is presently active in the bovine germ-line. To gain some insights in the degree of activity of the corresponding LTR element, we mined the *Damona* dataset and 50 Belgian Blue Breed (BBB) whole genome sequences (each at >15-fold depth) to search for evidence of other polymorphic insertion sites in the bovine genome. We developed the *LocaTER* (Localization of Transposable Retroviral elements) pipeline exploiting three main distinctive features of element insertion seen in the case of *APOB*: (i) paired sets of discordant paired-ends (with respect to *bosTau6* reference genome assembly) mapping, respectively, to the sense strand upstream of the insertion site (end 1, set 1) and one end of the LTR (end 2, set 1), and to the antisense strand downstream of the insertion site (end 2, set 2) and the other end of the LTR (end 1, set 3), (ii) the presence of split sense and antisense reads consistently bridging the insertion site and the LTR, (iii) the presence of the signature target site duplication of >2 to <20 bp (Suppl. Fig. 4 and Mat&Met.). LTR sequences were extracted from the RepeatMasker database (http://www.repeatmasker.org/). We identified a total of 1,584 such events in the available whole genome sequences. Evidence for Mendelian inheritance (the ERV is present in the child and observed in at least one parent) was observed for 91% of polymorphic ERV insertions. 94% of events corresponded to two ERV families: ERVK (74%) and ERV1 (20%). An additional 5.5% (for a total of 99.5%) belonged to two additional classes of ERVs: MaLR and ERVL. This pattern differs markedly from that reported for ERVs in the bovine reference genome, in which there are approximately 2.4x as many MaLR and ERV1 events, and 1.7x as many ERVL events compared to ERVK (MaLR > ERV1 > ERVL > K)(Rosenkranz 2016). 75.6% of the insertions mapped to intergenic regions (expected: 75%),

and 24.4% were within genes (expected: 25%; p = 0.23). Amongst genic insertions, 133 were in sense orientation (expected: 187) and 242 in antisense orientation (expected: 187; p = $8.7x10^{-5}$). We successfully added genotyping assays for 691 ERVs on Illumina LD SNP arrays and genotyped 4,500 BBB and 5,400 HF animals. The frequency distribution was shifted towards lower values for genic versus intergenic (p 0.098), and genic sense versus antisense insertions (p = 0.53)(Figure 2). Taken together, these results show that ERVK transposition generates substantial amounts of genetic polymorphism in cattle, while the signatures of selection indicate that a sizable fraction of at least the genic sense insertions has deleterious phenotypic consequences.

Visual inspection in IGV of ERV insertion sites violating Mendelian rules revealed five occurrences of ERV elements that were absent in sire and dam, present in their offspring and transmitted to grand-offspring (Table 1). We developed PCR-based assays for these five events and confirmed all genotypes inferred from sequence data, particularly the homozygous wild-type genotypes of the parents. These findings strongly suggested that the identified insertions were all *de novo* germ-line mutations. *De novo* mutations detected in offspring (but not parents) can either have been transmitted by the sire, transmitted by the dam, or have occurred during the development of the offspring. The latter are typically characterized by (i) allelic imbalance (< 50% dnm) in the offspring but not in the grand-offspring, and (ii) imperfect (D'=1; $r^2 < 1$) linkage with either of the homologues in the grand-offspring (Supplemental Note; Harland et al. 2017). There was no evidence of allelic imbalance or incomplete linkage for either of the five *de novo* events, suggesting that they all occurred in the later stages of parental gametogenesis. We performed linkage analysis in the grand-offspring and demonstrated that four of the de novo transposition events occurred in sires, for one in a dam. This suggests that the rate of transposition is of the order of one in (4/131) sperm and one in (1/131) oocytes. Intriguingly, three of the four male transposition events occurred in the germ-line of the same bull. Moreover, two of these three *de novo* insertions were observed in the same sperm cell (Figure 3). We amplified the five full-length de novo ERV elements by long-range PCR and sequenced them as describe above (Supp. Fig 5 and supp. file 1). Assembly of the five elements revealed that all were full length ERV-K elements of ~8kb, with each element flanked by two identical LTRs. All five elements and the APOB element were unique, differing at multiple variant sites in both LTRs and ORFs. All elements carried loss-of-function mutations in essential genes. Taken together, these results indicate that the de novo ERV transpositions do not resulting from the activation of a single functional ERV in the bovine genome, but rather multiple full length mutated ERVs are being activated and transposed by enzymes supplied by a third party.

Discussion

We have identified the mutation causing CD in Holstein-Friesian cattle as a full-length insertion of ERVK in exon 5 of *APOB* leading to the premature termination of the transcript. This has allowed the development of a direct genetic test for the disorder, allowing for accurate identification of at risk sires and dams. The direct test has also been added to the Eurogenomics (http://www.eurogenomics.com/) low-density cattle microarray to allow ongoing screening of cattle in the European dairy herd. Two other groups have independently identified the causative ERVK ERV2-1 insertion, but in both cases, have incorrectly identified it as the insertion of a solo LTR of ~1,300 bp rather than a full length insertion (Menzi et al. 2016; Schütz et al. 2016).

Based on this evidence of recent ERVK activity, we developed a script to detect polymorphic ERVs in whole genome sequence data and applied it to 793 cattle from the HF and BBB breed. From this dataset, we observe evidence of selection against genic insertions which show a lower average minor allele frequency (MAF) compared to intergenic insertions. One exception to this is the ERVK BTLTR1 insertion in *AGBL4*, which has a MAF of ~0.35. Interestingly this gene has been reported to be under positive selection in dual-purpose Normande cattle (Flori et al. 2009), while other reports suggest there is a possible dominance effect on milk yield associated with the gene (Aliloo et al. 2015). The bovine reference genome was assembled from a Hereford cow, with the numbers of events in each family following the pattern of MaLR > ERVL > ERV1 > ERVK (Supl Fig 6). This differs from the pattern observed in polymorphic from BBB and DHF where the opposite pattern of ERVK > ERVL > ERV1 > MaLR is observed (Supl Fig 6A). This is suggestive of continued ERVK and ERVL activity further, with the observation of five de novo ERVK insertions, we have directly confirmed that ERVK is currently active in the genome. These five events, along with the recent APOB insertion (likely in *Maughlin Storm*), all belong to a single group: the ERV2-1-BT_LTR family of ERVK. This is direct proof that at least this family of ERVK elements are currently active in the bovine genome and can have considerable functional effects. This is further supported by the apparent selection against genic insertions in our catalogue of polymorphic insertions, and the near fixation of the ERVK insertion in *AGBL4*. With regards to the source of the recent active ERVK insertions we note that all six have mutated open reading frames making unlikely that any of the six copied elements, provided their own reverse transcriptase or integrase. Further, the SNP and INDEL difference between these recent events suggests that they are not copies of a single intact ERVK present in the bovine genome, but rather are copies of multiple different K. Taken together it would suggest that rather than a fully functional ERVK being present and replicating itself, the enzymes required for transposition are being supplied by a functional or multiple partial functional in trans. One other possible source of the enzymes may be an infection by a retrovirus closely related to the ERVK. Unlike SNP and INDEL dnms, which occur at roughly equal rates in all individuals, de novo insertions may occur in bursts. As demonstrated by three of our five de novo insertions, which occurred within the germ-line of one of the 131 individuals tested

(permutation test: $P = 6.8x10^{-5}$), in addition to which two occurred in the same sperm cell of that individual. This suggest that in cattle ERVK is generally repressed as is the case in 113 of our trios. In currently unknown circumstances, repression of the ERVK elements is perturbed allowing their transcription. The necessary enzymes are either translated or supplied by an exogenous source resulting in the reverse transcription and integration of any transcribed full length ERVK elements. With regards to the circumstances leading to the activation of the ERVK, at least for the five de novo insertions observed it did not occur during early embryo development as the insertion could not be detected via targeted PCR in the parent of origin, implying the insertion occurred late in the development of the germ-line.

Materials and Methods

Damona dataset of cattle Whole Genome Sequences (WGS)

Illumina PCR-Free libraries (100bp, paired-end) were constructed using DNA extracted from blood or sperm samples followed by shearing and size selection to obtain a mean insert size of 500bp. Sequencing was carried out on HiSeq2000 instruments. Sequencing reads were aligned to bosTau6 reference genome using BWA MEM (Li, 2013). PCR duplicates were marked with Picard software (http://picard.sourceforge.net/). Mapped bam files have undergone local realignment around INDEL and base quality score recalibration with GATK. Individual indexed libraries were sequenced on two lanes or a half lane of a HiSeq2000 to reach an average coverage of 24 (two lanes, Trio members) and 6 (half a lane, grand-offspring). In total, the dataset is composed of 743 WGS of Holstein Friesian animals from the Netherlands.

HF *Damona* dataset of ~743 whole genome sequences (WGS), 131 trios (father, mother, proband) with an average of 5 proband's offspring each.

Transcriptomic analysis

Total RNA was extracted form liver of a homozygote mutant calf using Trizol (Invitrogen) following manufacturer's instructions. A strand-specific ribosomal RNA depleted RNA-Seq library was prepared using the Illumina TruSeq Total RNA stranded kit. Sequencing (2X76bp) was carried out on a HiSeq2000 instrument to reach a total of 80 million bp. The transcriptome was analyzed using the RNASeq tool kit TopHat and Cufflinks (Trapnell et al., 2012). Mapped RNASeq reads for the mutant calf were visually evaluated in IGV (Integrative Genome Browser) (Robinson et al., 2011).

Amplification of the inserted sequence by long-range PCR and amplicon sequencing

A long range PCR (LR-PCR) across the insertion point was performed using the LongAmp™ *Taq* PCR Kit (New England BioLabs), starting from 20ng of homozygote mutant genomic DNA extracted from blood. PCR fragment analysis was performed with the QIAxcel Advanced System (Quiagen). The LR-PCR product was gel purified and 1ng of recovered DNA was used for library preparation with the Nextera XT DNA library Preparation Kit (Illumina). The indexed library was purified with AMPure XP beads and 1µl of undiluted library was quantified on an Agilent Technology 2100 Bioanalyzer using a High Sensitivity DNA chip. The libraries were pooled and 2X150bp paired reads were generated on a MiSeq instrument (Illumina). A *de novo* assembly of the complete LRPCR amplicon was obtained with the paired-end sequence assembler ABySS (Simpson et al., 2009).

Identification of polymorphic retrotransposon insertions

The *LocaTER* pipeline is designed to identify candidate polymorphic retrotransposon insertions from next generation WGS data. It requires a database of locations for fixed retrotransposons in the reference genome, Ensembl and RefSeq transcript databases, individual sorted Illumina paired end BAM files aligned with BWA MEM and a pedigree file for the population. *LocaTER* proceeds to scan individual bam files analysing each read to identify the signatures of a retrotransposon insertion. For every read in the genome it checks to determine if they are cleanly aligned (map quality 20-60), if the read is properly paired (SAM flag isProperlyPaired), is not aligned to a known retrotransposon and that the mate is aligned to a known retrotransposon of the current class. When a read is detected that matches these criteria a 1.5kb window (3x the insert size) is created starting from that read. The software then proceeds to record key information about reads within this window. It records the total number of reads within the window, identifies all improperly paired, soft and hard clipped reads recording their orientation with regards to the reference genome (5` or 3`). For hard and soft clipped reads, it analyses the read recording the exact genomic position the read clips. For the improperly paired mates it records their orientation with regards to the reference genome, the orientation with regards to the retrotransposon they are aligned to, and the family of the retrotransposon. Once the end of the window is reached the number of observed 5` and 3` improperly paired reads and the total number of clipped reads are tested to determine if they are significantly different from the genome average for a 1.5kb window. If the observations are significantly different from the genome average the window is reported along with the recorded statistics for the window. Once all individuals are analysed the data is combined and the data for each window is merged, if windows overlap by 500bp and share at least one clipping site they are merged. The merged windows are tested for significance and checked to ensure that the total number of improperly paired reads in both the 5` and 3` is compatible with the insertion of a single retrotransposon (either heterozygous or homozygous) in that window, considering the number of individuals who shared the site. The difference between the two most common split read locations is calculated to identify the likely insertion site and the size of the associated micro-duplication, sites with a difference greater than 20bp are discarded. A 1.5kb window is recalculated from the likely insertion site and all individual BAM files are reanalysed for the new windows collecting the data as described above. In addition, the number of reads that completely bridge the insertion site are determined used to estimate if the site is heterozygous (1 or more read completely bridges the 5` and 3` insertion sites) or homozygous (no reads bridge the 5` and 3` insertion sites) in the individual. The number of 5` improperly paired reads, and 3` improperly paired reads are then tested against the genome average for significance and the site retained if either are significantly different. Each site is annotated with any gene it overlaps, and the retrotransposon class with the most mates aligned to it is selected as the likely class of the new retrotransposon insertion. The pedigree for the population is then analysed to identify trios and each site is checked for any violations

of Mendelian inheritance (absent in both parents but present in the proband). Each site is then reported with the associated statistics and list of identified carriers, along with there likely genotype.

References

Garcia-Etxebarria K, Jugo BM. 2010. Genome-Wide Detection and Characterization of Endogenous Retroviruses in Bos taurus. J Virol 84: 10852–10862.

Kipp S, Segelke D, Reinhardt F, Reents R, Schierenbeck S, Wurmser C, Pausch H, Fries R, Thaller G, Tetens J, et al. 2015. A new Holstein haplotype affecting calf survival. Interbull Bulletin. https://journal.interbull.org/index.php/ib/article/view/1375 (Accessed August 17, 2017).

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio]. http://arxiv.org/abs/1303.3997 (Accessed August 20, 2017).

Marchi E, Kanapin A, Magiorkinis G, Belshaw R. 2014. Unfixed Endogenous Retroviral Insertions in the Human Population. J Virol 88: 9529–9537.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069–2070.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotech 29: 24–26.

Rosenkranz D. 2016. piRNA cluster database: a web resource for piRNA producing loci. Nucleic Acids Res 44: D223–D230.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res 19: 1117–1123.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7: 562–578.

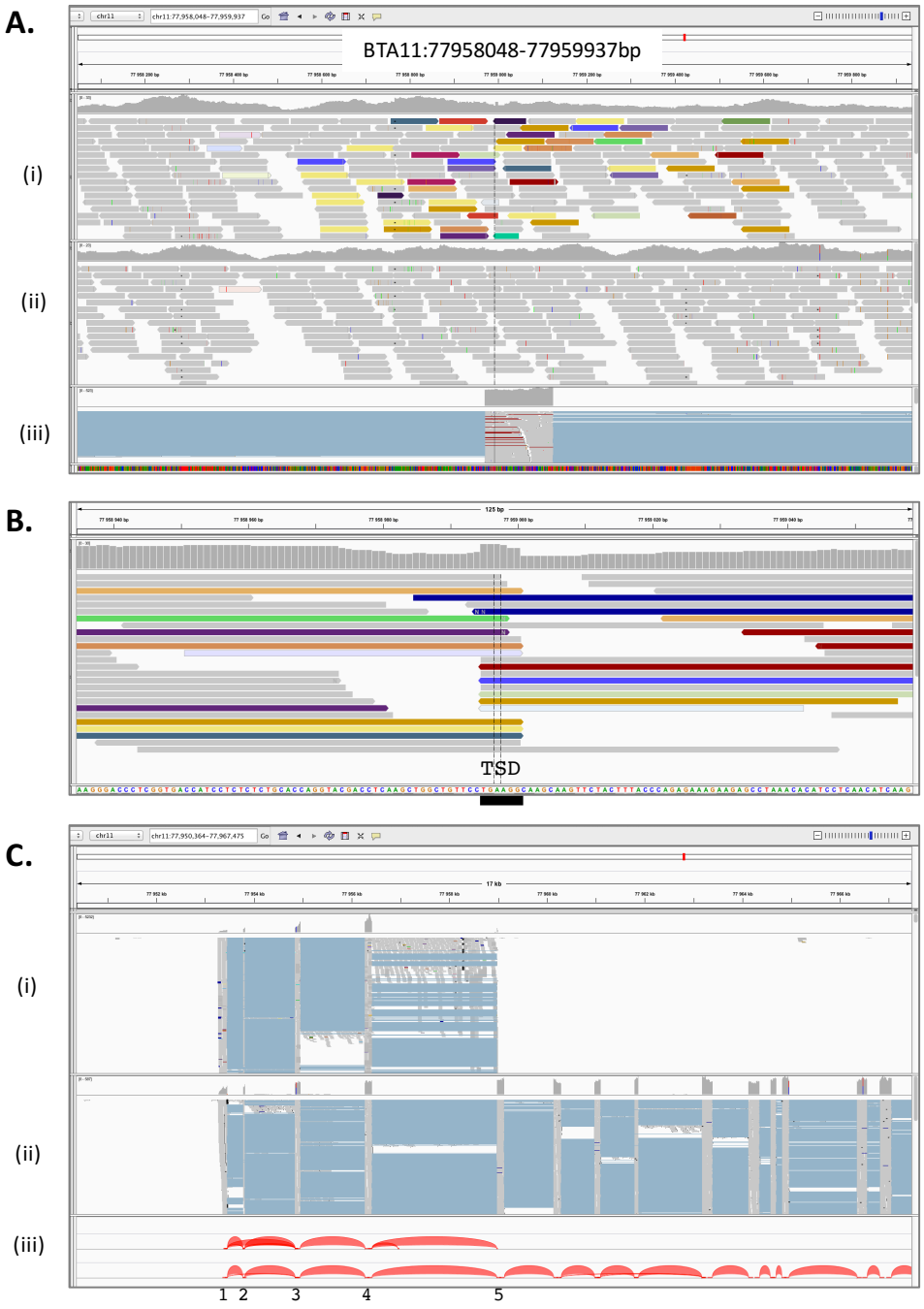


Figure 1: *APOB* insertional mutation. A. IGV screen capture of the BTA11 *APOB* surrounding exon 5 BAM files with, from top to bottom, (i) carrier animal, (ii)wild-type animal, (iii) wild-type liver RNASeq data. B. Breakpoint definition highlighting a 6 bp target site duplication (TSD). C. RNASeq data from calf livers: (i) homozygote mutant and (ii) wild-type revealing the transcriptional termination of *APOB* mutant mRNA at the insertion; (iii) intron-exon junction defined by TopHat/Cufflink for the corresponding animals, the five first *APOB* exons are numbered.

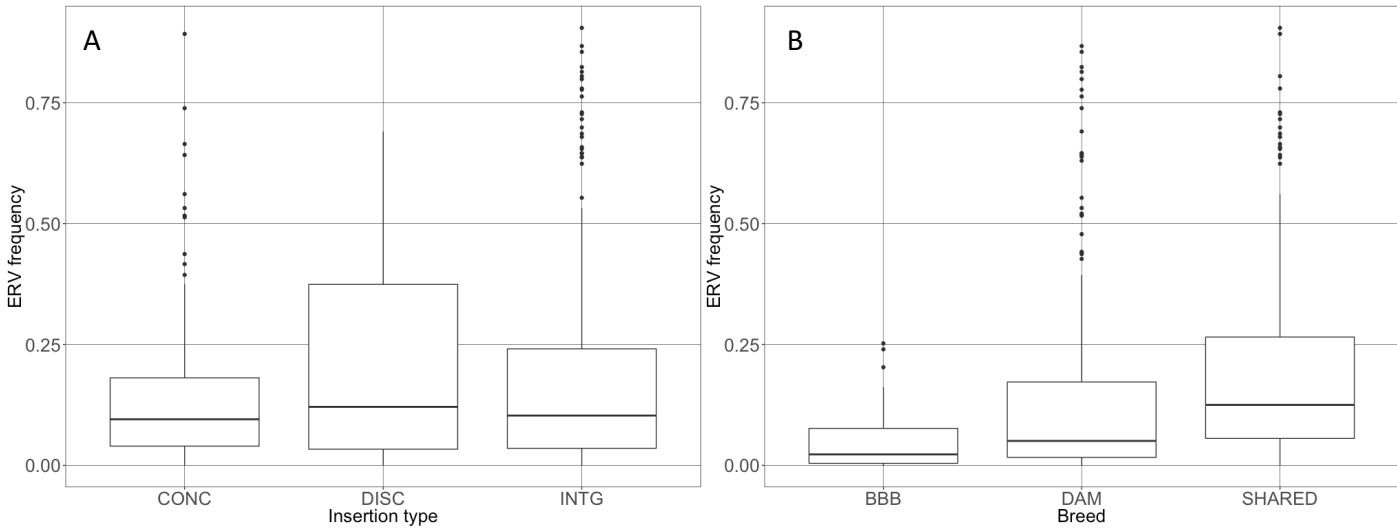


Figure 2: Frequency distribution according to their genomic localization and breed specificity. A. Concordant genic insertions (CONC), compared to discordant genic (DISC) and Intergenic (INTG). B. ERV insertion frequencies for population specific insertions, Belgian blue breed (BBB), Dutch Holstein Friesian (DAM), and ERVs present in both groups (SHARED).

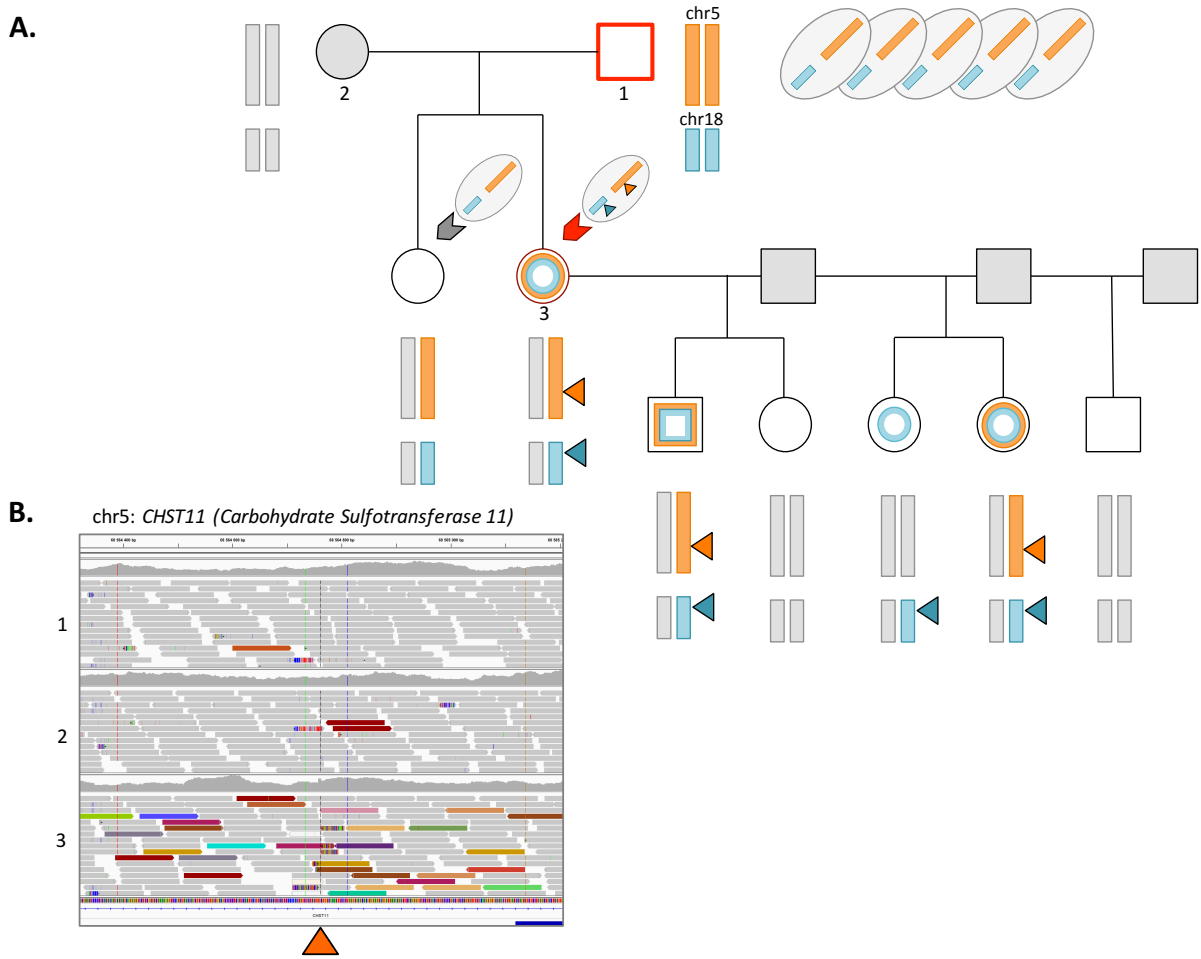
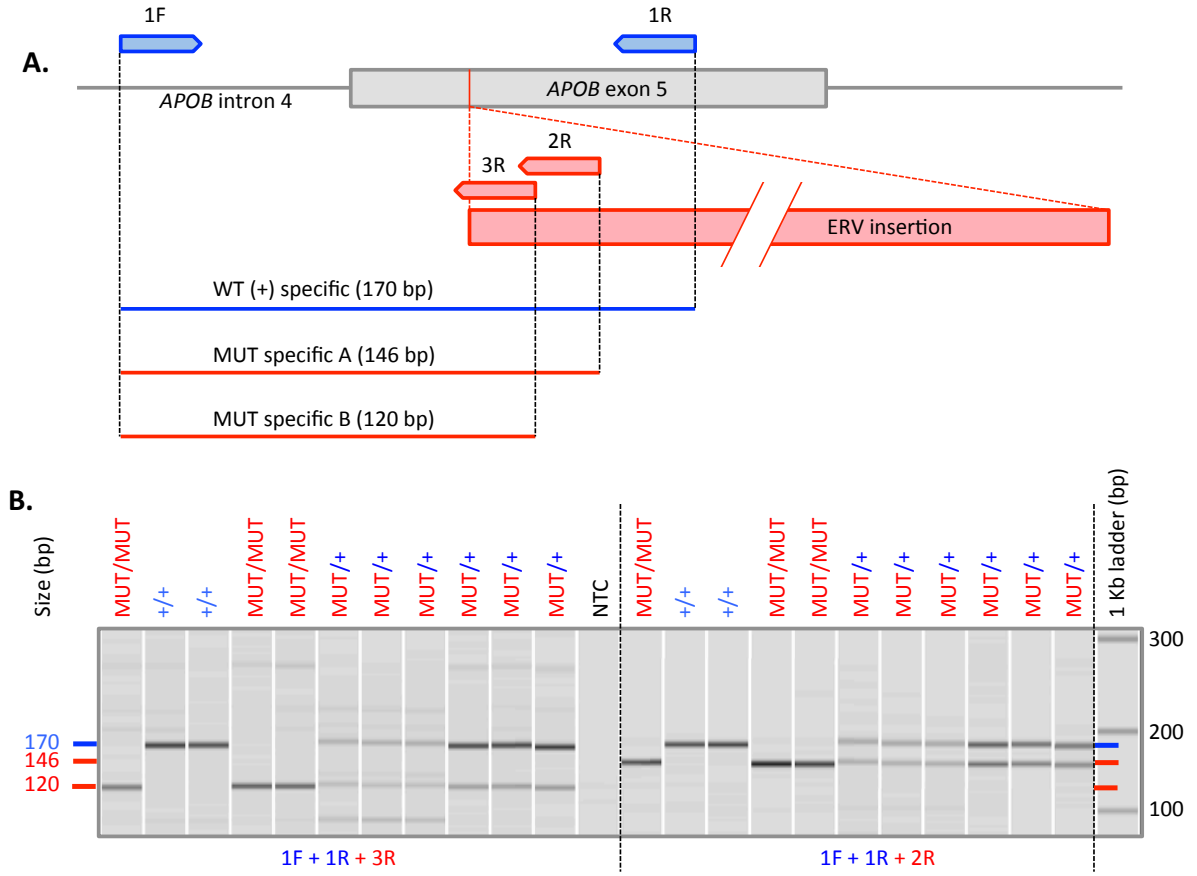


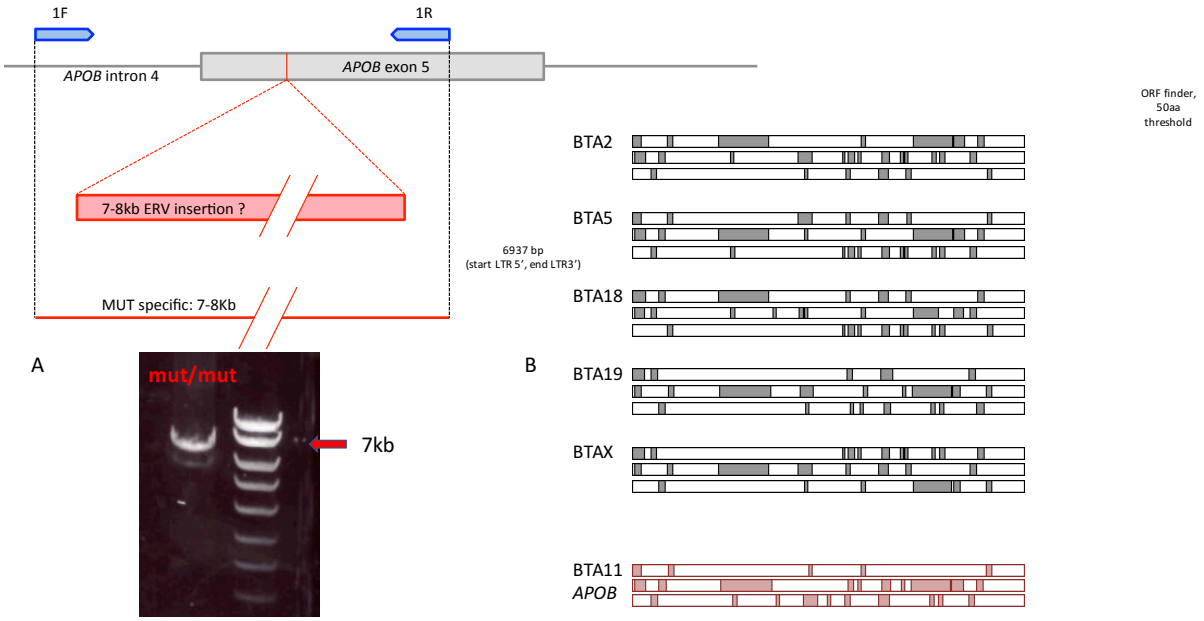
Figure 3: Birth (germ line mobilization) and familial segregation of two distinct *ERVK de novo* retroinsertions within a single sperm cell. **A.** Two *de novo* retrotransposition events (triangles), detected in the female proband '3' - respectively located on chr5 (orange) and chr18 (blue) - were shown to be absent for both parents (sire '1' and dam '2') and transmitted to at least one proband's offspring. The germline events were shown to both originate from the sire '1' (highlighted by a red square) and transmitted by a single sperm cell (red arrow). **B.** IGV screen capture of the corresponding trio (sire '1', dam '2', proband '3') for the genomic region on chromosome 5, harboring the intronic *de novo* retroinsertion within the *CHST11* gene.

Table 1: Germ-line origin, paternal (pat) or maternal (mat) and mendelian segregation of the five *de novo ERVK* events amongst offspring (x/5). The three paternal events in bold occurred in germ-line of the same sire and the two events underlined were transmitted by a single sperm cell.

Event	TSD (bp)	Localization	Orientation Gene-5' LTR	Gene	Germline	Transmission
chrX:35,311,109-35,311,114	6	exonic	concordant	*GABRQ*	mat	3/5
chr5: 68,564,763-68,564,758	6	intronic	concordant	*CHST11*	***pat***	2/5
chr2: 38,666,597-38,666,589	6	intronic	concordant	*CYTIP*	***pat***	2/5
chr18: 9,057,920-9,057,915	6	intergenic	/	/	***pat***	3/5
chr19: 4,942,306-4,942,297	6	intergenic	/	/	pat	2/5

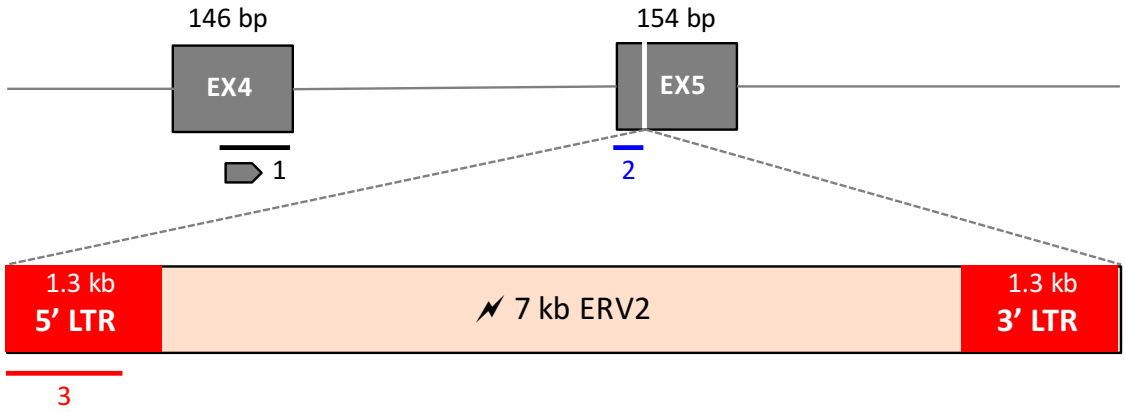


Supplemental Figure 1: PCR-based assay for carrier identification. **A.** Schematic representation of the assay design. **B.** PCR amplicons separation by QIAxcel automated capillary electrophoresis for the three *APOB* genotypes.



Supplemental Figure 2: Full-length *ERV2* (*ERVK*) insertion in *APOB* exon 5. **A.** Long-range PCR product amplified across the insertional BP (homozygote mutant calf), the gel shows the PCR product from a homozygous mutant animal with alongside a ladder with the 7kb fragment indicated (red arrow). **B.** Annotation of the ERVK full-length sequence displaying the mutated ORFs for the five de novo ERVs in grey and the *APOB* ERV in red.

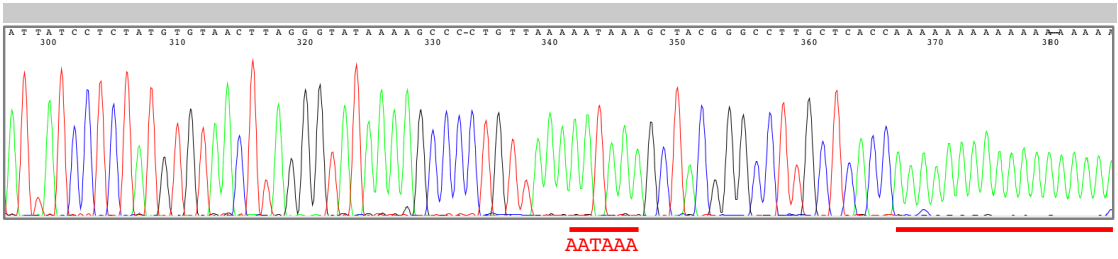
A.



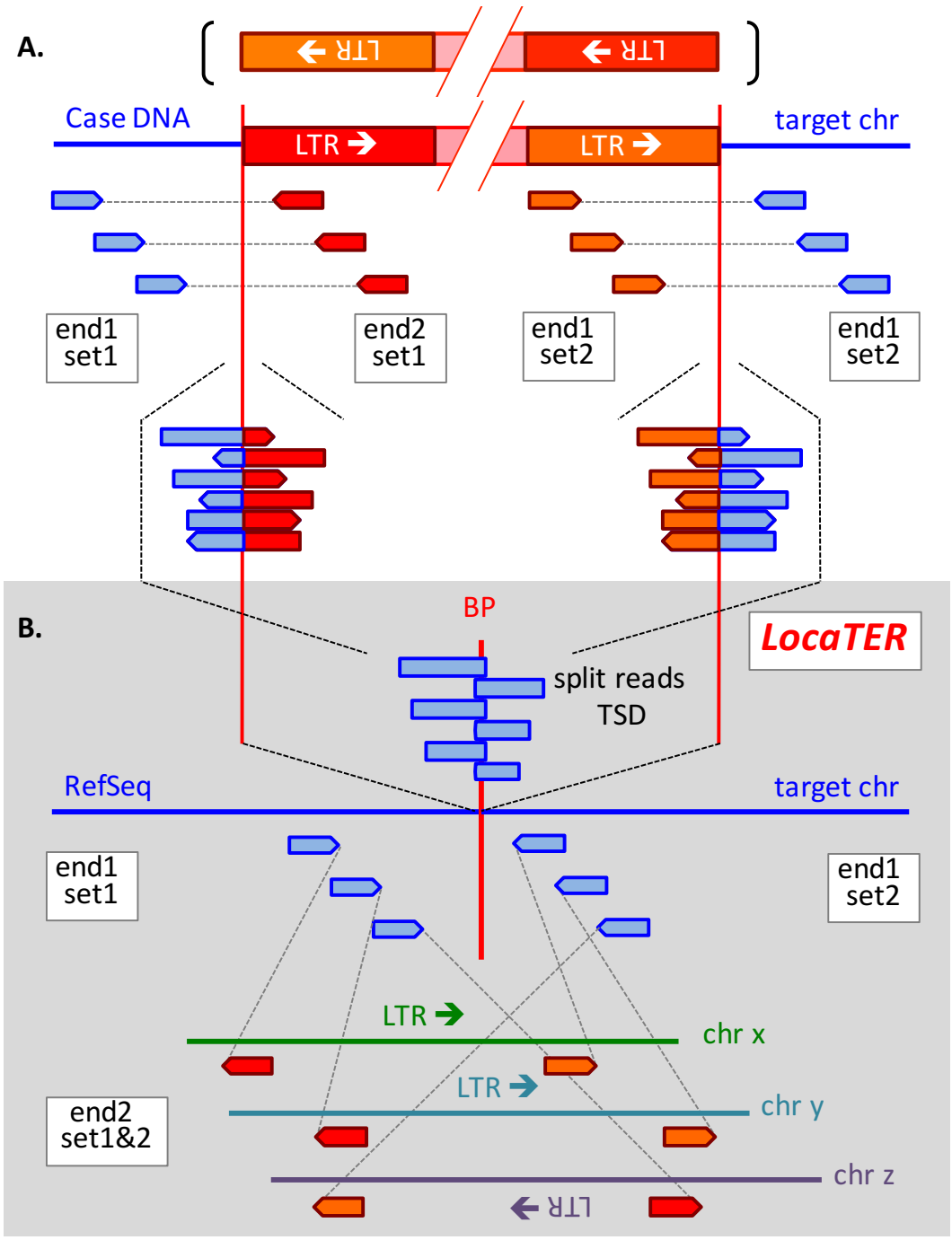
B.

AACTGCAAGGTGGAGCTGGAGGTCCCCCAACTCTGCAACTTCATCCTGAAGACGAGTGACTGCACCCTGAAAGA
GGTGGATGGCCTCAACCCTGAGGGCAAGGCCTTGCTGAAGAAGACCAAGAACTCTGAGGAGTTTGCTGCTGCGA
TGTCCAA (1)
GTACGACCTCAAGCTGGCTGTTCC**TGAAGG** (2)
TGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGG
ATCGAGCCTCAGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCAAAAAACCAGAGTCTGCCTACTTTATTG
CTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCCTACCACCGTCTCTCTCTCTCTGTGTCAAAGAGTT
AACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCCTCTGATGGCTCTCT
AACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGG
CATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAA
GGATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAA
TTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATT
CTTTCTTTGTTGTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCAAAC
CTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAG
GCCTTCTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGG
TTTTTGATAAGAATCAAAGACTGCTGACTTTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGC
CCCTGTTAAA**AATAAA**GCTACGGGCCTTGCTCACC**AAAAAAAAAAAAAAAAAAAA** (3)

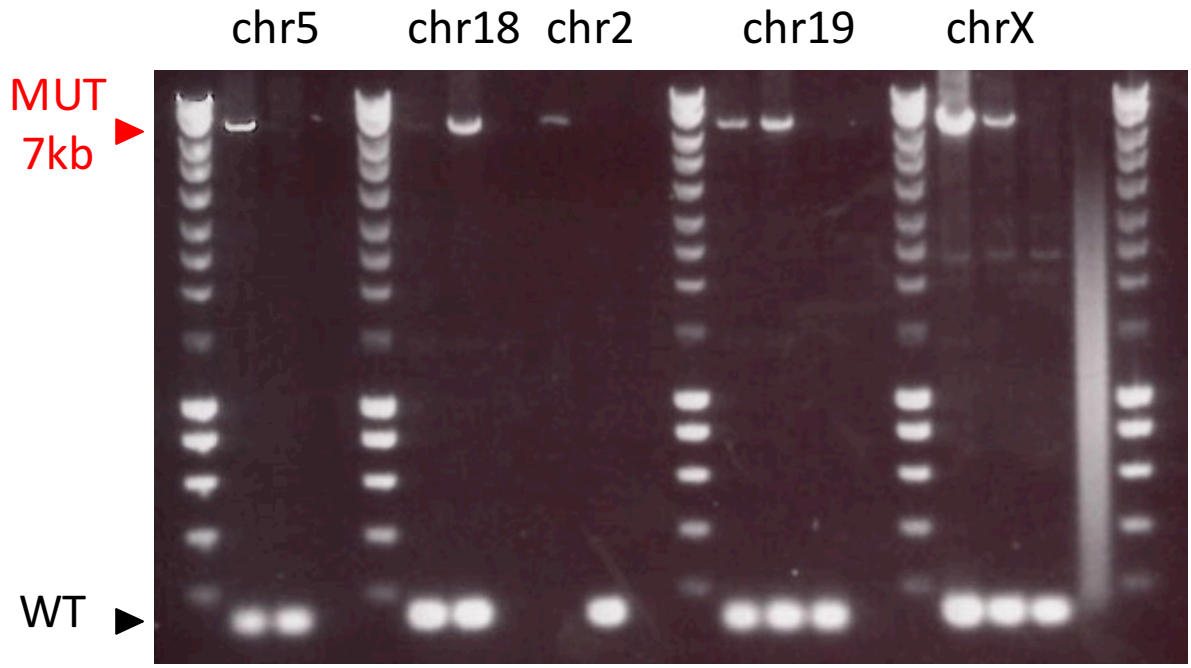
C.



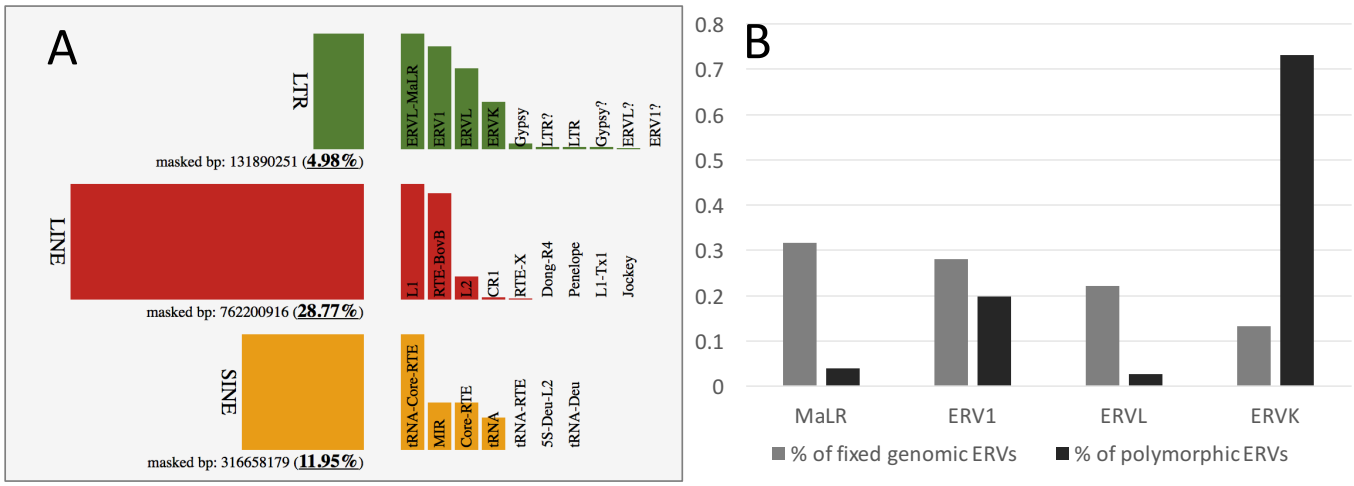
Supplemental Figure 3: Transcriptional termination of the *APOB* mutant mRNA within the 5'LTR. **A.** Structure of the gene and ERV insertion. **B.** The cDNA sequence of the *APOB* mRNA with ERV2 insertion, (1) exon 4 sequence, (2) Sequenced start of exon 5, (3) ERV2 LTR cDNA sequence with polyadenylation signal and start of poly (A) tail. **C.** Sanger sequence of the 3' RACE



Supplemental Figure 4: Schematic representation of the *LocaTER* (Localization of Transposable Endogenous Retroviral element) bioinformatics pipeline features. **(A)** Local structure of the DNA around the inserted ERV. Reads flanking the insertion but not directly overlapping the insertion sites have one read in the genomic DNA (blue), while the mate is present in the LTR of the ERV (red and orange). Some reads (blue/red and orange/blue) map directly across the insertion site (BP, breakpoint) from both the 5' and 3' ends and thus are part genomic and but ERV. **(B)** Structure of the aligned reads in the reference genome. As the polymorphic ERV is not found at this location in the reference genome the reads that map directly across the insertion point (blue/red, orange/blue from (A)) become split reads with only the genomic proportion of their DNA mapping to the reference genome. For the reads flanking the insertion site the mate in the ERV cannot be mapped as the polymorphic ERV is not present, instead the alignment software will assign them to a closely related ERV LTR that is fixed in the reference genome at some other location, resulting in the read pairs being improperly paired, due to the mate mapping to an alternative location in the genome.



Supplemental Figure 5: Validation of the five *dnm*s by PCR-base assay showing the 7kb mutant PCR product (red arrow) and the wild type PCR product (black arrow).



Supplemental Figure 6: A) Proportions of retrotransposon families in the bovine genome and their break down by sub-class (Rosenkranz 2016). **B)** Distribution of fixed genomic ERVs (grey, MaLR > ERV1 > ERVL > ERVK) compared to polymorphic ERVs (black, ERVK > ERV1 > MaLR > ERVL) detected in the Belgian Blue and Damona populations.

Supplemental file 1: Consensus sequence of the five de novo ERVs and the *APOB* ERV along with CLUSTAL O multiway alignments, the LTR 5' and 3' ends are underlined (5' TGCGGGGA, 3' GGTCCTCGGCA).

>chr2_denovo_ERVK
ATCTGCTTGGGGATACCAGGCAAGGTTTCATGAAGGAGGTAGCATTTAAGTTGGGCCAGGGAGAATGGGTAGGATTTCCACAGATAGAGGTGAA
GAAGGTAAATGTGGGCTTCCCTGGTGGCTCAGTAGTAAAGAAGCTgccNNN
TGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGGAGTCCCC
CCGGATATTCTCGAGCATTTTCCCCCAAAAAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCC
TACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCC
TCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGG
AAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTG
TTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAG
TAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGC
CAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTTCTGGCCAGAA
GATGATGTGAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACTTTGCAT
CCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATGT
CATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAGG
TGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCCC
GCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCATATA
AATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGGATTTCGAAGGTAAG
TCCCCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTGCAGACCCCCCTGTAGAATAGACAGGGAGAGA
GGAGTGGGAAGTGTTGAAAGTGTTGAAGAGTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGCTAAAA
GCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGAAAGACAGCTCTTTATAGGAGTAATTTTACAGTTATTAGG
TAAAAGAAGAATTAAAGTTAAAAAATCTGCCATTCAATCATTTCTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGACGAAGGCTCTGTT
AACTTAGATGTCTGGGAAAAAGTAGGAAAACAGTTAAAAACTTACCATGCAGAACATGGCTCAGAAAAGGTGCCTAATGACGCCTTTTCCTTAT
GGAATATTATTAGAGATGTCTTAGACCCTGCCCCAGATTCAGAAAAAGTACATCTTAAAAGGGATAATGAAGAAAATGCTGTAGTTAAACCTAC
CCCTGAACCTAAAAAAGTAACCTTTAAAGAGGAAAATGAAGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCCTGAC
TATCGGCAACTAAGAAAAATGTTAATAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGAGGAACAAGATCAGCCTTCCCCAAAACAAA
AAGAGAATTTAGGCGAGATAACAACTAAATATCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAACTTC
CCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTTTGTAAGATTTAATCACCAAACAGTTAAAGGATCTCCG
GAGCAGGAGAAGAGAAATACGGGACGCTCCCTTCCCCCCATGCCTCCTCCTCCGTGTGGGGGTAAAGGGCCTCCTCTAGGAGTTTCCCCGAGAA
GGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTCTTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGTACAA
AATAAATAAAAATATTCTCAAAAAAAAAATTAAAAAAAAAAAATGGGTTTCTCTGCATCTAAGAGAAATGTTTATTGGCTCTGTGAAATATGATT
AGAAATGTCCTGGACCCCTCTCATGAGGCTGTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCACTTTCTGCATAGA
TCATAATTTCTGCTATTGGAAAAAAAAAAAAAAAAAAGGAGGGAGAAAGTGCTTTACCTCCCGAGGAAGGAGAGGGTTTACAGGACGCTGCGGCC
GAGCGCCCTGGCGAGCTCCCTCCCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCACCGCCTCCGTTT
GTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTCAAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTGAACAGT
GGAGCTTTTAAAACAGACTGCGTGCACAATATGCAGAGCATGCTTCTTAGAGAAAACCCCCTCAGGGGGGGTCTCACCCAGGCTAGGAGAAAAG
CGAAACCGCAAGGGGATTTAGCAACAATATTAACCCCTGATGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTGGACAT
TGTAAGATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAAAAAAAAAAAATTCGGTGGTGCATGGTGTAGTAGATTCTGATTATATTG
AGAAATTAAAGTCTTGATATCACCGCCTACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTATTATCAG
ACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACAAAACATTTGGATCTAGTGATCTAGCCTTTTGGGTGCAGGAAATTACAGCTCCAAGGC
CTTTAAAAGATCTTTTAATTCCAGAGAATAAAATGCCAGGGCTATTGGACACAGGAACAGACGTCTCTTAGCATTGCTGGGAAAGACTGGCCCA
GCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAGAGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATTTAACA
ATACTTTATATATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTATGGGTATAAATAAAATATAATAAA
GGTATACCTAATTCTACTTATAGATCTATACTTAATTTGTATATAAATTAATATGTATACTATATATACATATATATACTGTATAATTAAAGGT
ATATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGATGATATATGTTCTATATACTGTATAATTGTATAT
GTGTGACATGTATTATTACAGTACACTGGTTTGTGTTGGTTGGTATTAGCTGTGTATGTGTTGTATTGCTGTAATATATATTAATTAATATATT
AATTATAAAGATTAATTCAAGTATATTGATTTATATATATTATATGTCAATAAGTTTATACTTGTTGCCATATATAAATACATATATAAATACA
TTTTACATTTAGGTATGTCTGTATACCAAAATGAGATAAGGAGGTTATACATTTATTATTTAAATTTATACAGAGACCTAAACTAAACATTAGA
CCTAAATTAACATATCCCTAGGAAAACAGCTGGACAAATAGTAAAGTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACACAGGGGCA
GACTACAATGATGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACCTCCTTACTGGAGAGTGTAAAAGG
CCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTTATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTCGTCATGT
CACTTTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGCTCTCTCCGCCGCCAGCGGCGCCCTGCACCCT
CCTTGCTGCACCCCGAGACCTAGAGCAAAGAAGTCTGTGCGGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGCAGTGCAGTCCAGACCAGC
GCGCTCCCGCCGCGGGCGACAGGAAAACGCCCCGAGCCCAAGCGGCGCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCCGTAGACG
CCTCGGCCCACAACCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCTACAGCTGCTCCTTGTTTCATCT
TGTTTAAATCCTCAATATAATGTTTGCTCCTTTGTGGTGGTGTTGTGAACAAATGTATGCTCTCATCTTGTTTAAATCCTCAATATAATGTTTG
CTCCTTTGTAATGTTACAACGTCCACCTTACCTTATGATGTAACCACTTATTGAATTATGCTCTTGCTGTATTACAACAACTGCAGGATTTAAT
GCGAACACGACGGTTTGTGGGCTTACTTATCTTGGGAATAGCAGCTTTGATAAGTGCAATTACTTCTGTTACTGTGGCAGCAATATCATTGACT
CAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGCATTGGCAACACAGGAAGCTATAGACAGAAAATTAGAGA
TGAGGGTAGACGCCCTAGAAGAAGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAGGTGAAAATGGCATTGTCCTGTCATGCTGACTA
CCGGTGGATATGTGTAACACCCCTGAAAGTAAATGACACAGATTTTGAATGGGAAAAGATTAAAAACCATATTTCAGGTATTTGGAACAGCTCT
GACATTAGCTTAGACTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCCGATTAGATTTTACTGCCGCTGGAACAGCAAATGATT
TCTTCCATACTTTCTCTAATTACATTTCAGGAAAAAATATTCTGTCTACCTTCCTCGGCTACGCTACCTTGGCTGTTTTAATTTTATTTCTAAT
AATCATTCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGCTACATCTGGCTGTTTTAAGAAATAAAAAAGGG
GGAGATGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGGAG
TCCCCCCGGATATTCTCGAGCATTTTCCCCCAAAAAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTG
TCCCCTACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAA
ACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGC
ATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAAT
GCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGT
GTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGAT
GGTGCCAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTTCTGGC

CAGAAGATGATGTGAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACTT
TGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCC
CATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGA
GAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGT
CTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGC
ATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGGCA
NNNGCTGCCTGTGATGCAGGAGATATGGGTTCAATCCCTGGATCAGAAAGATCCCTGGATAAGAAACTGGCGACCCACTCCAGTATTCTTGCCT
GGGAAATCCCATGGACAGAGGAACCTGGCAGAAATAGTCCATGGGGCTG

>ChrX_ERV_denovo
CCCCTCAGCCCTGCCACCCTAATATCAGTCTGTCTCCTGGCCAGGTTCCTATGTGCGCCTGATCCTGAGGTTCCTGGTCAAAAGGGAAGTCACC
AGTTACCTTGTGCAGATCTATTGGCCCAGCGTCCTTACCACTgtt
TGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGGAGTCCCC
CCGGATATTCTCGAGCATTTTCCCCCAAAAAAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCC
CTACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCC
CTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGG
GAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTT
GTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTA
GTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTG
CCAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTTCTGGCCAGA
AGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAATCAAAGACTGCTGACTTTGCA
TCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATG
TCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAG
GTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCC
CGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCATAT
AAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGGATTTCGAAGGTAA
GTCCCCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTGCAGACCCCCCTGTAGAATAGACAGGGAGAG
AGGAGTGGGAAGTGTTGAAAGTGTTGAAGAGTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGCTAAA
AGCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGAAAGACAGCTCTTTATAGGAGTAATTTTACAGTTATTAG
GTAAAAGAAGAATTAAAGTTAAAAAATCTGCCATTCAATCATTTTTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGACGAAGGCTCTGT
TAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGCTAAAAACTTACCATGCCGAACTTGGCTTAGAAAAGGTGCCTAATGACGCCTTTTCCTTA
TGGAATATCATTAGAGATGTCTTAGACCCTGCCCCTGATTCACAAAAAGTACATCTTAAAAGTGACAGTGAAGAAAATGCTGTAGCTAAACCTG
CCCCTAAATCTAAAAGAGTAACTTTTAAAGAGGAAAATGAAGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCCTGA
CTATCGGCAACTAAGAAAAATGTTAACAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGAGGAACAAGATCAGCCTTCCCCAAAACAA
AAAGAGAATTTAGGCGAGATAACAACTAAATATCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAACTT
CCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTTTGTAAGATTTAATCACCAAACAGTTAAAGGATCTCC
GGAGCAGGAGAAGAGAAATACGGGACGCTCCCTTCCCCCCATGCCTCCTCCTCCGTGTGGGGGTAAAGGGCCTCCTCTAGGAGTTTCCCCGAGA
AGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTCTTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGTACA
AAATAAATAAAAATATTCTCAAAAAAAAAATTAAAAAAAAAAAATGGGTTTCTCTGCATCTAAGAGAAATGTTTATTGGCTCTGTGAAATATGAT
TAGAAATGTCCTGGACCCCTCTCATGAGGCTGTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCACTTTCTGCATAG
ATCATAATTTCTGCTATTGGAAAAAAAAAAAAAAAAGGAGGGAGAAAGTGCTTTACCTCCCGAGGAAGGAGAGGGTTTACAGGACGCTGCGGCC
GAGCGCCCTGGCGAGCTCCCTCCCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCACCGCCTCCGTTT
GTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTCAAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTGAACAGT
GGAGCTTTTAAAACAGACTGCGTGCACAATATGCAGAGCATGCTTCTTAGAGAAAACCCCCTCAGGGGGGGTCTCACCCAGGCTAGGAGAAAAG
CGAAACCGCAAGGGGATTTAGCAACAATATTAACCCCTGATGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTGGACAT
TGTAAGATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAAAAAAAAAAAAAAAAATTCGGTGGTGCATGGTGTAGTAGATTCTGATTA
TATTGAGAAATTAAAGTCTTGATATCACCGCCTACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTATT
ATCAGACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACAAAACATTTGGATCTAGTGATCTAGCCTTTTGGGTGCAGGAAATTACAGCTCC
AAGGCCTTTAAAAGATCTTTTAATTCCAGAGAATAAAATGCCAGGGCTATTGGACACAGGAACAGACGTCTCTTAGCATTGCTGGGAAAGACTG
GCCCAGCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAGAGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATT
TAACAATACTTTATATATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTATGGGTATAAATAAAATATA
ATAAAGGTATACCTAATTCTACTTATAGATCTATACTTAATTTGTATATAAATTAATATGTATACTATATATACATATATATACTGTATAATTA
AAGGTATATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGATGATATATGTTCTATATACTGTATAATTG
TATATGTGTGACATGTATTATTACAGTACACTGGTTTGTGTTGGTTGGTATTAGCTGTGTATGTGTTGTATTGCTGTAATATATATTAATTAAT
ATATTAATTATAAAGATTAATTCAAGTATATTGATTTATATATATTATATGTCAATAAGTTTATACTTGTTGCCATATATAAATACATATATAA
ATACATTTTACATTTAGGTATGTCTGTATACCAAAATGAGATAAGGAGGTTATACATTTATTATTTAAATTTATACAGAGACCTAAACTAAACA
TTAGACCTAAATTAACATATCCCTAGGAAAACAGCTGGACAAATAGTAAAGTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACACAG
GGGCAGACTACAATGATGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACCTCCTTACTGGAGAGTGTA
AAAGGCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTTATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTCGT
CATGTCACTTTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGCTCTCTCCGCCGCCAGCGGCGCCCTGC
ACCCTCCTTGCTGCACCCCGAGACCTAGAGCAAAGAAGTCTGTGCGGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGCAGTGCAGTCCAGA
CCAGCGCGCTCCCGCCGCGGGCGACAGGAAAACGCCCCGAGCCCAAGCGGCGCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCCGT
AGACGCCTCGGCCCACAACCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCTACAGCTGCTCCTTGTTT
CATCTTGTTTAAATCCTCAATATAATGTTTGCTCCTTTGTGGTGGTGTTGTGAACAAATGTATGCTCTCATCTTGTTTAAATCCTCAATATAAT
GTTTGCTCCTTTGTAATGTTACAACGTCCACCTTACCTTATGATGTAACCACTTATTGAATTATGCTCTTGCTGTATTACAACAACTGCAGGAT
TTAATGCGAACACGACGGTTTGTGGGCTTACTTATCTTGGGAATAGCAGCTTTGATAAGTGCAATTACTTCTGTTACTGTGGCAGCAATATCAT
TGACTCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGCATTGGCAACACAGGAAGCTATAGACAGAAAATT
AGAGATGAGGGTAGACGCCCTAGAAGAAGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAGGTGAAAATGGCATTGTCCTGTCATGCT
GACTACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGACACAGATTTTGAATGGGAAAAGATTAAAAACCATATTTCAGGTATTTGGAACA
GCTCTGACATTAGCTTAGACTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCCGATTAGATTTTACTGCCGCTGGAACAGCAAA
TGATTTCTTCCATACTTTCTCTAACTACATTTCAGGAAAAAATATTCTGTCTACCTTCCTCGGCTACGCTACCTTGGCTGTTTTAATTTTATTT
CTAATAATCATTCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGCTACATCTGGCTGTTTTAAGAAATAAAA
AAGGGGGAGATGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTC
AGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCAAAAAAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGG
GGGCTGTCCCCTACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTTTAA
ATCCAAACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCG

AGAGGCATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAGGATTTCATTGATGA
GTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGC
CTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTG
GAAGATGGTGCCAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCT
TCTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAATCAAAGACTGC
TGACTTTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAACGCTTGG
TCTCCCCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCC
ACTCGAGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTT
ATTGGTCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTA
ATTAGCATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGGCAGG
actGTTGTTTCTTGGATATCATTTTGGATGAACTATGAATCTTCTGCAGCCAGGGTGACAGTCGGTAAGTCCTGTCCCCATCTCAGGAAAGAAC
TTGGCTCATCTTGCTCAGAAAAAAGCTGAGAGTAAGGAGCAGTGTG

>chr5_denovo_ERVK
CAAACTGTTTCTACCCGCCTGGTGTCCCCAAGTCAGCCCTCTTCATCACACTCTCGCTTGCTCACTCTCGTTCTCTTCCCCACCCCCCCAGCCC
CGGCTTAATGTTTCTGCTTCTGCCTTGCCAGTTACTTTGATTTTCNNN
TGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGGAGTCCCC
CCGGATATTCTCGAGCATTTTCCCCCAAAAAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCC
TACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCC
TCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGG
AAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTG
TTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAG
TAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGC
CAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTTCTGGCCAGAA
GATGATGTGAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACTTTGCAT
CCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATGT
CATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAGG
TGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCCC
GCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCATATA
AATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGGGATTTCGAAGGTAA
GTCCCCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTGCAGACCCCCCTGTAGAATAGACAGGGAGAG
AGGAGTGGGAAGTGTTGAAAGTGTTGAAGAGTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGCTAAA
AGCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGAAAGACAGCTCTTTATAGGAGTAATTTTACAGTTATTAG
GTAAAAGAAGAATTAAAGTTAAAAAATCTGCCATTCAATCATTTCTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGACGAAGGCTCTGT
TAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGTTAAAAACTTACCATGCAGAACATGGCTCAGAAAAGGTGCCTAATGACGCCTTTTCCTTA
TGGAATATTATTAGAGATGTCTTAGACCCTGCCCCAGATTCAGAAAAAGTACATCTTAAAAGGGATAATGAAGAAAATGCTGTAGTTAAACCTA
CCCCTGAACCTAAAAAAGTAACCTTTAAAGAGGAAAATGAAGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCCTGA
CTATCGGCAACTAAGAAAAATGTTAATAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGAGGAACAAGATCAGCCTTCCCCAAAACAA
AAAGAGAATTTAGGCGAGATAACAACTAAATATCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAACTT
CCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTTTGTAAGATTTAATCACCAAACAGTTAAAGGATCTCC
GGAGCAGGAGAAGAGAAATACGGGACGCTCCCATCCCTCCATGCCTCCTCCTCTGTGTGGGGGTAAAGGGCCTCCACTAGGAGTTTCCCCGAGA
AGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTCTTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGTACA
AAATAAATAAAAATATTCTCAAAAAAAAAATTAAAAAAAAAAATGGGTTTCTCTGCATCTAAGAGAAATGTTTATTGGCTCTGTGAAATATGAT
TAGAAATGTCCTGGACCCCTGTCATGAGGCTGTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCACTTTCTGCATAG
ATCATAATTTCTGCTATTGAAAAAAAAAAAAAAAAGGAGGGAGAAAGTGCTTTACCTCCCGAGGAAGGAGAGGGTTTACAGGACGCTGCGGCCG
AGCGCCCTGGCGAGCTCCCTCCCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCACCGCCTCCGTTTG
TGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTCAAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTGAACAGTG
GAGCTTTTAAAACAGACTGCGTGCACAATATGCAGAGCATGCTTCTTAGAGAAAACCCCCTCAGGGGGGGTCTCACCCAGGCTAGGAGAAAAGC
GAAACCGCAAGGGGATTTAGCAACAATATTAACCCCTGATGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTGGACATT
GTAAGATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAAAAAAAAAAAAATTCGGTGGTGCATGGTGTAGTAGATTCTGATTATATT
GAGAAATTAAAGTCTTGATATCACCGCCTACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTATTATCA
GACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACGAAACATTTGGATCTAGTGATCTAGCCTTTTGGGTGCAGGAAATTACAGCTCCAAGG
CCTTTAAAAGATCTTTTAATTCCAGAGAATAAAATGCCAGGGCTATTGGACACAGGAACAGACGTCTCTTAGCATTGCTGGGAAAGACTGGCCC
AGCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAGAGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATTTAAC
AATACTTTATATATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTATGGGTATAAATAAAATATAATAA
AGGTATACCTAATTCTACTTATAGATCTATACTTAATTTGTATATAAATTAATATGTATACTATATATACATATATATACTGTATAATTAAATG
TATATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGATGATATATGTTCTATATACTGTATAATTGTATA
TGTGTGACATGTATTATTACAGTACACTGGTTTGTGTTGGTTGGTATTAGCTGTGTACGTGTTGTATTGCTGTAATATATATTAATTAATATAT
TAATTATAAAGATTAATTCAAATATATTGATTTATATATATTATATGTCAATAAGTTTATACTTGTTGCCATATATAAATACATATATAAATAC
ATTTTACATTTAGGTATGTCTGTATACCAAAATGAGATAAGGAGGTTATACATTTATTACTTAAATTTATACAGAGACCTAAACTAAACATTAG
ACCTAAATTAACATGTCCCTAGGAAAACAGCTGGACAAATAGTAAAGTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACACAGGGGC
AGACTACAATGATGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACCTCCTTACTGGAGAGTGTAAAAG
GCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTTATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTCGTCATG
TCACATTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGCTCTCTCCGCCGCCAGCGGCGCCCTGCACCC
TCCTTGCTGCACCCCGAGACCTAGAGCAAAGAAGTCTGTGCGGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGCAGTGCAGTCCAGACCAG
CGCGCTCCCRCCGCGGGCGACAGGAAAACGCCCCGAGCCCAAGCGGCGCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCCGTAGAC
GCCTCGGCCCACAACCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCTACAGCTGCTCCTTGTTTCATC
TTGTTTAAATCCTCAATATAATGTTTGCTCCTTTGTGGTGGTGTTGTGAACAAATGTATGCTCTCATCTTGTTTAAATCCTCAATATAATGTTT
GCTCCTTTGTAATGTTACAACGTCCACCTTACCTTATGATGTAACCACTTATTGAATTATGCTCTTGCTGTATTACAACAACTGCAGGATTTAA
TGCGAACACGACGGTTTGTGGGCTTACTTATCTTGGGAATAGCAGCTTTGATAAGTGCAATTACTTCTGTTACTGTGGCAGCAATATCATTGAC
TCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGCATTGGCAACACAGGAAGCTATAGACAGAAAATTAGAG
ATGAGGGTAGACGCCCTAGAAGAAGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAGGTGAAAATGGCATTGTCCTGTCATGCTGACT
ACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGACACAGATTTTGAATGGGAAAAGATTAAAAACCATATTTCAGGTATTTGGAACAGCTC
TGACATTAGCTTAGACTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCCGATTAGATTTTACTGCCGCTGGAACAGCAAATGAT
TTCTTCCATACTTTCTCTAATTACATTTCAGGAAAAAATATTCTGTCTACCTTCCTCGGCTACGCTACCTTGGCTGTTTTAATTTTATTTCTAA
TAATCATTCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGCTACATCTGGCTGTTTTAAGAAATAAAAAAGG
GGGAGATGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGGA

GTCCCCCCGGATATTCTCGAGCATTTTCCCCCAAAAAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCT
GTCCCCTACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTTTAAATCCA
AACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGG
CATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAA
TGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGG
TGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGA
TGGTGCCAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTTCTGG
CCAGAAGATGATGTGAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACT
TTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAACGCTTGGTCTCC
CCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCG
AGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGG
TCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAG
CATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGGCANNNATTTTCCATGTGTGGGTTACCTG
AAAAACAAAACTGGTTTTCCTGCCCTGCCTTTTGAAAATGATGGGCTAATCAGAAGGCCCTAAGACCATGTATGGAGATGAAAAGATGATTACC
AGGTTGGAAGATTTGAATTAAAA

>chr18_denovo_ERVK
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTGGTGCTGTCGATTCCGTTTGCAGTCGTCTGTTT
AACCTAGGGTGCAAGAGAAAACAGGAGATTTATNNN
TGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGGAGTCCCC
CCGGATATTCTCGAGCATTTTCCCCCCAAAAAAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCC
CCTACCACCGTCTCTCTCTCTCTGTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACC
CCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATG
GGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCT
TGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTT
AGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTTGTAAACCCATTACACATCAGCCCTATAGGAATGCAATTTTATCTTTGGAAGAT
GGTGCCAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTTCTGGC
CAGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACTT
TGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCC
CATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGA
GAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGT
CTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGC
ATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGGGATTTCGAA
GGTAAGTCCCCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTGCAGACCCCCCTGTAGAATAGACAGG
GAGAGAGGAGTGGGAAGTGTTGAAAGTGTTGAAGAGTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAG
CTAAAAGCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGAAAGACAGCTCTTTATAGGAGTAATTTTACAGTT
ATTAGGTAAAAGAGGAATTAAAGTTAAAAAATCTGCCATTCAATCATTTCTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGACGAAGGC
TCTGTTAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGTTAAAAACTTACCATGCAGAACATGGCTCAGAAAAGGTGCCTAATGACGCCTTTT
CCTTATGGAATATTATTAGAGATGTCTTAGACCCTGCCCCAGATTCAGAAAAAGTACATCTTAAAAGGGATAATGAAGAAAATGCTGTAGTTAA
ACCTACCCCTGAACCTAAAAAAGTAACCTTTAAAGAGGAAAATGAAGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCG
CCTGACTATCGGCAACTAAGAAAAATGTTAATAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGAGGAACAAGATCAGCCTTCCCCAA
AACAAAAAGAGAATTTAGGCGAGATAACAACTAAATATCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGA
AACTTCCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTTTGTAAGATTTAGTCACCAAACAGTTAAAGGA
TCTCCGGAGCAGGAGAAGAGAAATACGGGACGCTCCCATCCCTCCATGCCTCCTCCTCTGTGTGGGGGTAAAGGGCCTCCACTAGGAGTTTCCC
CGAGAAGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTCTTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAG
GTACAAAATAAATAAAAATATTCTCAAAAAAAAAAAAAAAAAAAAAATGGGTTTCTCTGCATCTAAGAATATACAAATGTTTATTGGCTCTGTGAA
ATATGATTAGAAATGTCCTGGACCCCTGTCATGAGGCTGTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCACTTTC
TGCATAGATCATAATTTCTGCTATTGGAAAAAAAAAAAAAAAAAGGAGGGAGAAAGTGCTTTACCTCCCGAGGAAGGAGAGGGTTTACAGGACGCT
GCGGCCGAGCGCCCTGGCGAGCTCCCTCCCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCACCGCCT
CCGTTTGTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTCAAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTG
AACAGTGGAGCTTTTAAAACAGACTGCGTGCACAATATGCAGAGCATGCTTTTTAGAGAAAACCCCCTCGGGGGGGTCTCACCCAGGCTAGGAG
AAAAGCGAAACTGCAAGGGGATTTAGCAACAATATTAACCCCTGATGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTG
GACATTGTAAGATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAAAAAAAAAAAAAATTCGGTGGTGCATGGTGTAGTAGATTCTGAT
TATATTGAGAAATTAAAGTCTTGATATCACCGCCTACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTA
TTATCAGACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACGAAACATTTGGATCTAGTGATCTAGCCTTTTGGGTGCAGGAAATTACAGCT
CCAAGGCCTTTAAAAGATCTTTTAATTCCAGAGAATAAAATGCCAGGGCTATTGGACACAGGAACAGACGTCTCTTAGCATTGCTGGGAAAGAC
TGGCCCAGCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAGAGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACA
TTTAACAATACTTTATATATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTATGGGTATAAATAAAATA
TAATAAAGGTATACCTAATTCTACTTATAGATCTATACTTAATTTGTATATAAATTAATATGTATACTATATATACATATATATACTGTATAAT
TAAATGTATATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGATGATATATGTTCTATATACTGTATAAT
TGTATATGTGTGACATGTATTATTACAGTACACTGGTTTGTGTTGGTTGGTATTAGCTGTGTACGTGTTGTATTGCTGTAATATATATTAATTA
ATATATTAATTATAAAGATTAATTCAAATATATTGATTTATATATATTATATGTCAATAAGTTTATACTTGTTGCCATATATAAATACATATAT
AAATACATTTTACATTTAGGTATGTCTGTATACCAAAATGAGATAAGGAGGTTATACATTTATTACTTAAATTTATACAGAGACCTAAACTAAA
CATTAGACCTAAATTAACATGTCCCTAGGAAAACAGCTGGACAAATAGTAAAGTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACAC
AGGAGCAGACTACAATGATGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACCTCCTTACTGGAGAGTG
TAAAAGGCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTTATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTC
GTCATGTCACATTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGCTCTCTCCGCCGCCAGCGGCGCCCT
GCACCCTCCTTGCTGCACCCCGAGACCTAGAGCAAAGAAGTCTGTGCGGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGCAGTGCAGTCCA
GACCAGCGCGCTCCCGCCGCGGGCGACAGGAAAACGCCCCGAGCCCAAGCGGCGCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCC
GTAGACGCCTCGGCCCACAACCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCTACAGCTGCTCCTTGT
TTCATCTTGTTTAAATCCTCAATATAATGTTTGCTCCTTTGTGGTGGTGTTGTGAACAAATGTATGCTCTCATCTTGTTTAAATCCTCAATATA
ATGTTTGCTCCTTTGTAATGTTACAACGTCCACCTTATCTTATGATGTAACCACTTATTGAATTATGCTCTTGCTGTATTACAACAACTGCAGG
ATTTAATGCGAACACGACGGTTTGTGGGCTTACTTATCTTGGGAATAGCAGCATTGATAAGTGCAATTACTTCTGTTACTGTGGCAGCAATATC
ATTGACTCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGCATTGGCAACACAGGAAGCTATAGACAGAAAA
TTAGAGATGAGGGTAGACGCCCTAGAGGAAGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAGGTGAAAATGGCATTGTCCTGCCATG
CTGACTACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGAGACAGATTTTGAATGGGAAAAGATTAAAAACCATATTTCAGGTATTTGGAA
CAGCTCTGACATTAGCTTAGACTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCTGATTAGATTTTACTGCCGCTGGAACAGCA

AATGATTTCTTCCATACTTTCTCTAACTACATTTCAGGAAAAAATATTCTGTCTACCTTCCTCGGCTACGCTACCTTGGCTGTTTTAATTTTAT
TTCTAATAATCATTCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGCTACATCTGGCTGTTTTAAGAAATAA
AAAAGGGGGAGATGCGGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCC
TCAGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCCAAAAAAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTAC
TGGGGGCTGTCCCCTACCACCGTCTCTCTCTCTCTGTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTT
TAAATCCAAACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCC
TCGAGAGGCATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAGGATTTCATTGA
TGAGTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGT
GGCCTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTTGTAAACCCATTACACATCAGCCCTATAGGAATGCAATTTTA
TCTTTGGAAGATGGTGCCAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGT
AGGCCTTCTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAA
GACTGCTGACTTTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAAC
GCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTA
AGACCCACTCGAGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTCTTGTCTTGA
AGTTTTATTGGTCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATA
TCCCTAATTAGCATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGGCANNNATTTATGCTAAG
TTCCACCTAGATACGTACTGTGGGTTTTGTTCCTTTGTAGGAGTCAACCAGCAGAAACAGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNN

>chr19_denovo_ERVK
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCTTGTGTCCCAGTTACCAGGTTAACCTAACCTAATTTGTGGGGGCAAGGGTCNNN
TGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGGAGTCCCC
CCGGATATTCTCGAGCATTTTCCCCCAAAAAAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCC
CTACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCC
CTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGG
GAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTT
GTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTA
GTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTTGTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATG
GTGCCAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTTCTGGCC
AGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACTTT
GCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCC
ATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAG
AAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTC
TCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCA
TATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGGATTTCGAAGG
TAAGTCCCCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTGCAGACCCCCCTGTAGAATAGACAGGGA
GAGAGGAGTGGGAAGTGTTGAAAGTGTTGAAGAGTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGCT
AAAAGCAAAATCTTTTACTATACTAATTTTCTGACATGGGTAATACTGAATCAAATGAAAGACAGCTCTTTATAGGAGTAATTTTACAGTTAT
TAGGTAAAAGAAGAATTAAAGTTAAAAAATCTGCCATTCAATCATTTTTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGACGAAGGCTC
TGTTAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGCTAAAAACTTACCATGCCGAACTTGGCTTAGAAAAGGTGCCTAATGACGCCTTTTCC
TTATGGAATATCATTAGAGATGTCTTAGACCCTGCCCCTGATTCACAAAAAGTACATCTTAAAAGTGACAGTGAAGAAAATGCTGTAGCTAAAC
CTGCCCCTAAATCTAAAAGAGTAACTTTTAAAGAGGAAAATGAAGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCC
TGACTATCGGCAACTAAGAAAAATGTTAACAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGAGGAACAAGATCAGCCTTCCCCAAAA
CAAAAAGAGAATTTAGGCGAGATAACAACTAAATATCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAA
CTTCCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTTTGTAAGATTTAATCACCAAACAGTTAAAGGATC
TCCGGAGCAGGAGAAGAGAAATACGGGACGCTCCCTTCCCCCCATGCCTCCTCCTCCGTGTGGGGGTAAAGGGCCTCCTCTAGGAGTTTCCCCG
AGAAGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTCTTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGT
ACAAAATAAATAAAAATATTCTCAAAAAAAAAATTAAAAAAAAAATGGGTTTCTCTGCATCTAAGAGAAATGTTTATTGGCTCTGTGAAATATG
ATTAGAAATGTCCTGGACCCCTCTCATGAGGCTGTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCACTTTCTGCAT
AGATCATAATTTCTGCTATTGGAAAAAAAAAAAAAAAAAGGAGGGAGAAAGTGCTTTACCTCCCGAGGAAGGAGAGGGTTTACAGGACGCTGCG
GCCGAGCGCCCTGGCGAGCTCCCTCCCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCACCGCCTCCG
TTTGTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTCAAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTGAAC
AGTGGAGCTTTTAAAACAGACTGCGTGCACAATATGCAGAGCATGCTTCTTAGAGAAAACCCCCTCAGGGGGGGTCTCACCCAGGCTAGGAGAA
AAGCGAAACCGCAAGGGGATTTAGCAACAATATTAACCCCTGATGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTGGA
CATTGTAAGATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAAAAAAATTCGGTGGTGCATGGTGTAGTAGATTCTGATTATATTGA
GAAATTAAAGTCTTGATATCACCGCCTACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTATTATCAGA
CAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACAAAACATTTGGATCTAGTGATCTAGCCTTTTGGGTGCAGGAAATTACAGCTCCAAGGCC
TTTAAAAGATCTTTTAATTCCAGAGAATAAAATGCCAGGGCTATTGGACACAGGAACAGACGTCTCTTAGCATTGCTGGGAAAGACTGGCCCAG
CTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAGAGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATTTAACAA
TACTTTATATATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTATGGGTATAAATAAAATATAATAAAG
GTATACCTAATTCTACTTATAGATCTATACTTAATTTGTATATAAATTAATATGTATACTATATATACATATATATATACTGTATAATTAAAGG
TATATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGATGATATATGTTCTATATACTGTATAATTGTATA
TGTGTGACATGTATTATTACAGTACACTGATTTGTGTTGGTTGGTATTAGCTGTGTACGTGTTGTATTGCTGTAATATATATTAATTAATATAT
TAATTATAAAGATTAATTCAAATATATTGATTTATATATATTATATGTCAATAAGTTTATACTTGTTGCCATATATAAATACATATATAAATAC
ATTTTACATTTAGGTATGTCTGTATACCAAAATGAGATAAGGAGGTTATACATTTATTACTTAAATTTATACAGAGACCTAAACTAAACATTAG
ACCTCAATTAACATGTCCCTAGGAAAACAGCTGGACAAATAGTAAAGTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACACAGGAGC
AGACTACAATGATGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACCTCCTTACTAGAGAGTGTAAAAG
GCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTTATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTCGTCATG
TCACATTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGCTCTCTCCTCCGCGCCGCCATCGGCGCCCCA
CACCCTCCTTGCTGCACCCCGCGACCTAGAGCAAGGAAGTCTGTGCGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGGAGTGCAGTCCAGA
CCAGCGCGCTCCCGCCGGCGACAGGAAAACGCCCGGAGCCCAAGCGGCGGCGGCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCCG
TAGACGCCGCGGCCCACAGCCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCTACAGCTGCTCCTTGTT
TCATCTTGTTTAAATCCTCAATATAATGTTTGCTCCTTTGTAATGTTACAACGTCCACCTTATCTTATGATGTAACCACTTATTGAATTATGCT
CTTGCTGTATTACAACAACTGCAGGATTTAATGCGAACACGACGGTTTGTGGGCTTACTTATCTTGGGAATAGCAGCTTTGATAAGTGCAATTA
CTTCTGTTACTGTGGCAGCAATATCATTGACTCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGCATTGGC
AACACAGGAAGCTATAGACAGAAAATTAGAAATGAGGGTAGATGCCCTAGAGGAAGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAG
GTGAAAATGGCATTGTCCTGTCATGCTGACTACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGAGACAGATTTTGAATGGGAAAAGATTA

AAAACCATATTTCAGGTATTTGGAACAGCTCTGACATTAGCTTAGACTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCCGATT
AGATTTTACTGCCGCTGGAACAGCAAATGATTTCTTCCATACTTTCTCTAACTACATTTCAGGAAAAAATATTCTGTCTACCTTCCTCGGCTAC
GCTACCTTGGCTGTTTTAATTTTATTTCTAATAATCATTCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGC
TACATCTGGCTGTTTTAAGAAATAAAAAAGGGGGAGATGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCG
GCATACGCAAAGGCGGGGATCGAGCCTCAGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCAAAAAAACCAGAGTCTGCCTACTTTATTGCT
TTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCCTACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAA
AGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCAT
ACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAA
TAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGT
GGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTTGTAAACCCATTACACAT
CCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCT
TGTCAAGAGTCATAAAATGTTAGTAGGCCTTCTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAA
CCTTGGTTTTTGATAAGAACCAAAGACTGCTGACTTTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATA
AAGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGC
TTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTA
AGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCT
ATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGG
TCCTCGGCANNNAGGGTCCATTCAAATAGAGGTGAAAATATTGCCTTATAGAGCCAATCTTTGTGAAAGCAAGNNNNNNNNNNNNNNNNNNNN
NNN

>chr11_APOB
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGAGCCATAGGCTAGGTGGCTTTGCAGCCACCTANGCCTATGGCTC
ACGGTAGAGAAGGGACCCTCGGTGACCATCCTCTCTCTGCACCAGGTACGACCTCAAGCTGGCTGTTCCTGAAGG
TGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGGAGTCCCC
CCGGATATTCTCGAGCATTTTCCCCCAAAAAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCC
TACCACCGTCTCTCTCTCTCTGTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCC
TCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGG
AAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTG
TTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAG
TAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTTGTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGG
TGCCAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTTCTGGCCA
GAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAATCAAAGACTGCTGACTTTG
CATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCA
TGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGA
AGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCT
CCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCAT
ATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGGGATTTCGAAGG
TAAGTCCCCAACCCCATTCCCCCAGTGTTGAGTTTCGGGACGGATAGGACCCCACTTAGGGTGCTGCAGACCCCCCTGTAGAATAGACAGGGA
GAGAGGAGTGGGAAGTGTTGAAAGTGTTGAAGAGTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGCT
AAAAGCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGAAAGACAGCTCTTTATAGGAGTAATTTTACAGTTAT
TAGGTAAAAGAGGAATTAAAGTTAAAAAATCTGCCATTCAATCATTTCTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGACGAAGGCTC
TGTTAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGTTAAAAACTTACCATGCAGAACATGGCTCAGAAAAGGTGCCTAATGACGCCTTTTCC
TTATGGAATATTATTAGAGATGTCTTAGACCCTGCCCCAGATTCAGAAAAAGTACATCTTAAAAGGGATAGTGAAGAAAATGCTGTAGTTAAAC
CTACCCCTGAACCTAAAAAAGTAACCTTTAAAGAGGAAAATGAAGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCC
TGACTATCGGCAACTAAGAAAAATGTTAATAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGAGGAACAAGATCAGCCTTCCCAAAA
CAAAAAGAGAATTTAGGCGAGATAACAACTAAATATCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAA
CTTCCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTTTGTAAGATTTAGTCACCAAACAGTTAAAGGATC
TCCGGAGCAGGAGAAGAGAAATAGGGGACGCTCCCATCCCTCCATGCCTCCTCCTCTGTGTGGGGGTAAAGGGCCTCCACTAGGAGTTTCCCCG
AGAAGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTCTTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGT
ACAAAATAAATAAAAATATTCTCAAAAAAAAAAAAAATAAATAAAAAAAAAAAATGGGTTTCTCTGCATCTAAGAATATACAAATGTTTATTGGCT
CTGTGAAATATGATTAGAAATGTCCTGGACCCCTCTCATGAGGCTGTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTC
CAGTTTCTGCATAGATCATAATTTCTGCTATTGGAAAAAAAAAAAAAAAAGGAGGGAGAAAGTGCTTTACCTCCCGAGCAAGGAGAGGGTTTACA
GGACGCTGCGGCCGAGCGCCCTGGCGAGCTCCCTCCCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCC
ACCGCCTCCGTTTGTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTCAAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAG
TTTTTTGAACAGTGGAGCTTTTAAAACAGACTGCGTGCACAATATGCAGAGCATGCTTCTTAGAGAAAACCCCCTCAGGGGGGGTCTCACCCAG
GCTAGGAGAAAAGCGAAACCGCAAGGGGATTTAGCAACAATATTAACCCCTGATGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCC
TACCTTTGGACATTGTAAGATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAAAAAAAAAAAAATTCGGTGGTGCATGGTGTAGTAG
ATTCTGATTATATTGAGAAATTAAAGTCTTGATATCACCGCCTACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTT
TTACCTTATTATCAGACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACGAAACATTTGGATCTAGTGATCTAGCCTTTTGGATGCAGGAAA
TTACAGCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAGAGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATT
TAACAATACTTTATATATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTATGGGTATAAATAAAATATA
ATAAAGGTATACCTAATTCTACTTATAGATCTATACTTAATTTGTATATAAATTAATATGTATACTATATATACATATATATATACTGTATAAT
TAAATGTATATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGATGATATATGTTCTATATACTGTATAAT
TGTATATGTGTGACATGCATTATTGCAGTACATTGGTTTGTGTTGGTTGGTATTAGCTGTGTATGTGTTGTATTGCTGTAATATATATTAATTA
ATATATTAATTATAAAGATTAATTCAAATATATTGATTTATATATATTATATGTCAATAAGTTTATACTTGTTGCCATATATAAATACATATAT
AAATACATTTTACATTTAGGTATGTCTGTATACCAAAATGAGATAAGGAGGTTATACATTTATTACTTAAATTTATACAGAGACCTAAACTAAA
CATTAGACCTAAATTAACATGTCCCTAGGAAAACAGCTGGACAAATAGTAAAGTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACAC
AGGGGCAGACTACAATGATGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACCTCCTTACTGGAGAGTG
TAAAAGGCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTTATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTC
GTCATGTCACTTTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGCTCTCTCCGCCGCCAGCGGCGCCCT
GCACCCTCCTTGCTGCACCCCGAGACCTAGAGCAAAGAAGTCTGTGCGGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGCAGTGCAGTCCA
GACCAGCGCGCTCCCGCCGCGGGCGACAGGAAAACGCCCCGAGCCCAAGCGGCGCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCC
GTAGACGCCTCGGCCCACAACCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCTACAGCTGCTCCTTGT
CTCATCTTGTTTAAATCCTCAATATAATGTTTGCTCCTTTGTAATGTTACAATGTCCACCTTATCTTATGATGTAACCACTTATTGAATTATGC
TCTTGCTGTATTACAACAACTGCAGGATTTAATGCGAACACGACGGTTTGTGGGCTTACTTATCTTGGGAATAGCAGCTTTGATAAGCGCAATT
ACTTCTGTTACTGTGGCAGCAATATCATTGACTCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGCATTGG
CAACACAGGAAGCTATAGACAGGAAATTAGAAATGAGGGTAGATGCCCTAGAGGAAGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAA

GGTGAAAATGGCATTGTCCTGTCATGCTGACTACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGAGACAGATTTTGAATGGGAAAAGATT
AAAAACCATATTTCAGGTATTTGGAACAGCTCTGACATTAGCTTAGACTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCCGAT
TAGATTTTACTGCCGCTGGAACAGCAAATGATTTCTTCCATACTTTCTCTAACTACATTTCAGGAAAAAATATTCTGTCTACCTTCCTCGGCTA
CGCTACCTTGGCTGTTTTAATTTTATTTCTAATAATCATTCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAG
CTACATCTGGCTGTTTTAAGAAATAAAAAAGGGGGAGATGCGGGGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTC
GGCATACGCAAAGGCGGGATCGAGCCTCAGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCAAAAAACCAGAGTCTGCCTACTTTATTGCT
TTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCCTACCACCGTCTCTCTCTCTCTGTGTCAAAGAGTTAACTTACAGCTCCAATTAATAA
AGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCCTGCAGCTATGCAT
ACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAA
TAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGT
GGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTTGTAAACCCATTACACAT
CCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCAAACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCT
TGTCAAGAGTCATAAAATGTTAGTAGGCCTTCTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAGAAA
CCTTGGTTTTTGATAAGAATCAAAGACTGCTGACTTTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATA
AA
GCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGAACCCACCACGCTT
ACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAG
TGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTAT
CCTCATTCTATTGTTCTCTATATCCCTAATTAGCATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTC
CTCGGCA TGAAGGCAAGCAAGTTCTACTTTACCCAGAGAAAGAAGAGCCTAAACACAGCCTCAACATNNNNNNNNNNNNNNN

```
CLUSTAL O(1.2.1) multiple sequence alignment


chr11_APOB            -------------------NNNNN--------N-NNNNNNNNN-NNNNNNNNNNNNNNNN
chr19_denovo_ERVK     ----------------------------------------NNNN-NNNNNNNNNNNNNNNN
ChrX_ERV_denovo       ------------------------------------------------------------
chr18_denovo_ERVK     ------------------NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
chr2_denovo_ERVK      ------------ATCTGCTTGGGGATACCAGGCAAGGTTTCATGAAGGAGGTA-GCATTT
chr5_denovo_ERVK      CAAACTGTTTCTACCCGCCTGGTGTCCCCAAGTCAGCCCTCTTCATC-------ACACTC


chr11_APOB            NNNNN--NNNNNNNNNNNNNGAGCCATAGGCTAGGTGGCTTTGCAGCCACCTANGCCTAT
chr19_denovo_ERVK     NNNNNNNNNNNNNNNNNCTTGT-----------------------------------------
ChrX_ERV_denovo       --------------------CCCC-------------TCAGCCCTGCCACCCTAATATCA
chr18_denovo_ERVK     NNNNNNNNNNNNNNNNNNNTGGTGC-------------------------------------
chr2_denovo_ERVK      AAGTTGGGCCAGGGAGAATGGGTA-------------GGATTT------CCACAGAT---
chr5_denovo_ERVK      TCGCTTGCTCAC----TCTCGTTC-------------TCTTCCCCACCCCCCCAGCC---


chr11_APOB            GG-----------CTCA-----CGGTAGAGA---------AGGGACCCTCGGTGACCATC
chr19_denovo_ERVK     -----------GTCCCAGTTACCAG-------------------------------------
ChrX_ERV_denovo       GTCTGTCTCCTGGCCAGGTTCCTATGTGCGCCTGATCCTGAGGTTCCTGGTCAAAAGGGA
chr18_denovo_ERVK     ----------TGTC-----------------------------GATTCCGTT------TGC
chr2_denovo_ERVK      --------AGAGGT-----------------------------GAAGAAGGT------AAA
chr5_denovo_ERVK      --------CC-GGC-----------------------------TTAATGT------TTC


chr11_APOB            CTCTCTCTG---CACCAGGTACGACCTCAAGCTGGCTGT-------TCCTGAAGGTGCGG
chr19_denovo_ERVK     ---------GTTAACCTAACC-TAATTTGTGGGGGCA---------AGGGTCNNNTGCGG
ChrX_ERV_denovo       AGTCACCAGT--TACCTTGTG-CAGATCTATTGGCCCAGCGTCCTTACCACTgttTGCGG
chr18_denovo_ERVK     AGTCGTCTGTTTAACCTAGGG-TG-CAAGAGAAAACAGGA-----GATTTATNNNTGCGG
chr2_denovo_ERVK      TGT-----GGGCTTCCCTGGT-GG-CTC-AGTAGTAAAGA-----AGCTgccNNNTGCGG
chr5_denovo_ERVK      TGC------TTCTGCCTTGC-------C-AGTT--ACTTT-----GATTTTCNNNTGCGG
                                    **  .          :     .                  *****

chr11_APOB            GGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGC
chr19_denovo_ERVK     GGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGC
ChrX_ERV_denovo       GGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGC
chr18_denovo_ERVK     GGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGC
chr2_denovo_ERVK      GGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGC
chr5_denovo_ERVK      GGAGCCGGTGAGGCATTCCACTCGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGC
                      ************************************************************

chr11_APOB            AAAGGCGGGATCGAGCCTCAGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCA--AA
chr19_denovo_ERVK     AAAGGCGGGATCGAGCCTCAGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCAA-AA
ChrX_ERV_denovo       AAAGGCGGGATCGAGCCTCAGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCAA-AA
chr18_denovo_ERVK     AAAGGCGGGATCGAGCCTCAGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCCAAAA
chr2_denovo_ERVK      AAAGGCGGGATCGAGCCTCAGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCA--AA
chr5_denovo_ERVK      AAAGGCGGGATCGAGCCTCAGGAGTCCCCCCGGATATTCTCGAGCATTTTCCCCCA--AA
                      ************************************************************.  **

chr11_APOB            AAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTG
chr19_denovo_ERVK     AAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTG
ChrX_ERV_denovo       AAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTG
chr18_denovo_ERVK     AAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTG
chr2_denovo_ERVK      AAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTG
chr5_denovo_ERVK      AAACCAGAGTCTGCCTACTTTATTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTG
                      ************************************************************

chr11_APOB            TCCCCTACCACCGTCTCTCTCTCTCTGTGTCAAAGAGTTAACTTACAGCTCCAATTAATA
chr19_denovo_ERVK     TCCCCTACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATA
ChrX_ERV_denovo       TCCCCTACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATA
chr18_denovo_ERVK     TCCCCTACCACCGTCTCTCTCTCTCTGTGTCAAAGAGTTAACTTACAGCTCCAATTAATA
chr2_denovo_ERVK      TCCCCTACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATA
chr5_denovo_ERVK      TCCCCTACCACCATCTCGCTCTCTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATA
                      ************.**** **  **** :*********************************

chr11_APOB            AAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGC
chr19_denovo_ERVK     AAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGC
ChrX_ERV_denovo       AAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGC
chr18_denovo_ERVK     AAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGC
chr2_denovo_ERVK      AAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGC
chr5_denovo_ERVK      AAGTTCCTGGGCAATTAGGAGTGTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGC
                      ************************************************************

chr11_APOB            CTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCT
chr19_denovo_ERVK     CTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCT
```

```
ChrX_ERV_denovo         CTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCT
chr18_denovo_ERVK       CTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCT
chr2_denovo_ERVK        CTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCT
chr5_denovo_ERVK        CTGACAAGTTTACCCGGACTCCTGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCT
                        ************************************************************

chr11_APOB              CGAGAGGCATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAA
chr19_denovo_ERVK       CGAGAGGCATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAA
ChrX_ERV_denovo         CGAGAGGCATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAA
chr18_denovo_ERVK       CGAGAGGCATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAA
chr2_denovo_ERVK        CGAGAGGCATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAA
chr5_denovo_ERVK        CGAGAGGCATGGGAAGCTTAAGATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAA
                        ************************************************************

chr11_APOB              CTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTC
chr19_denovo_ERVK       CTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTC
ChrX_ERV_denovo         CTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTC
chr18_denovo_ERVK       CTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTC
chr2_denovo_ERVK        CTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTC
chr5_denovo_ERVK        CTGTCAGAATAAACTAGTAAAGGATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTC
                        ************************************************************

chr11_APOB              ACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAAT
chr19_denovo_ERVK       ACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAAT
ChrX_ERV_denovo         ACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAAT
chr18_denovo_ERVK       ACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAAT
chr2_denovo_ERVK        ACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAAT
chr5_denovo_ERVK        ACATCCCCTGAATTGTATCCTTGAATATGTATCAATTAATAGTGGGTATGTAGAAAAAAT
                        ************************************************************

chr11_APOB              AAGTAGTGGCCTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTT
chr19_denovo_ERVK       AAGTAGTGGCCTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTT
ChrX_ERV_denovo         AAGTAGTGGCCTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTT---
chr18_denovo_ERVK       AAGTAGTGGCCTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTT
chr2_denovo_ERVK        AAGTAGTGGCCTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTT---
chr5_denovo_ERVK        AAGTAGTGGCCTTGGTGTTAGTAACTTTAGACCCTTAAGGTAATAAATTCTTTCTTT---
                        ************************************************************

chr11_APOB              GTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCA
chr19_denovo_ERVK       GTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCA
ChrX_ERV_denovo         GTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCA
chr18_denovo_ERVK       GTAAACCCATTACACATCAGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCA
chr2_denovo_ERVK        GTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCA
chr5_denovo_ERVK        GTAAACCCATTACACATCCGCCCTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCA
                        ******************.*****************************************

chr11_APOB              AACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAG
chr19_denovo_ERVK       AACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAG
ChrX_ERV_denovo         AACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAG
chr18_denovo_ERVK       AACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAG
chr2_denovo_ERVK        AACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAG
chr5_denovo_ERVK        AACCTTGAAATAATTACTCTTAGAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAG
                        ************************************************************

chr11_APOB              TCATAAAATGTTAGTAGGCCTTCTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGT
chr19_denovo_ERVK       TCATAAAATGTTAGTAGGCCTTCTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGT
ChrX_ERV_denovo         TCATAAAATGTTAGTAGGCCTTCTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGT
chr18_denovo_ERVK       TCATAAAATGTTAGTAGGCCTTCTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGT
chr2_denovo_ERVK        TCATAAAATGTTAGTAGGCCTTCTGGCCAGAAGATGATGTGAATCACCTAAACCATTTGT
chr5_denovo_ERVK        TCATAAAATGTTAGTAGGCCTTCTGGCCAGAAGATGATGTGAATCACCTAAACCATTTGT
                        ****************************************.*******************

chr11_APOB              ATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAATCAAAGACTGCTGACT
chr19_denovo_ERVK       ATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACT
ChrX_ERV_denovo         ATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAATCAAAGACTGCTGACT
chr18_denovo_ERVK       ATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACT
chr2_denovo_ERVK        ATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACT
chr5_denovo_ERVK        ATACGATACATTTGCAGGAAAGAAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACT
                        ******************************************* ***************

chr11_APOB              TTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAA
chr19_denovo_ERVK       TTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAA
ChrX_ERV_denovo         TTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAA
chr18_denovo_ERVK       TTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAA
chr2_denovo_ERVK        TTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAA
chr5_denovo_ERVK        TTGCATCCCCTATTATCCTCTATGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAA
                        ************************************************************
```

```
chr11_APOB            AGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCT
chr19_denovo_ERVK     AGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCT
ChrX_ERV_denovo       AGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCT
chr18_denovo_ERVK     AGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCT
chr2_denovo_ERVK      AGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCT
chr5_denovo_ERVK      AGCTACGGGCCTTGCTCACCAACGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCT
                      ************************************************************

chr11_APOB            GAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGA
chr19_denovo_ERVK     GAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGA
ChrX_ERV_denovo       GAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGA
chr18_denovo_ERVK     GAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGA
chr2_denovo_ERVK      GAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGA
chr5_denovo_ERVK      GAGTCTCCATCTGGAGCGCGGAACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGA
                      ************************************************************

chr11_APOB            CCCACTCGAGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGC
chr19_denovo_ERVK     CCCACTCGAGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGC
ChrX_ERV_denovo       CCCACTCGAGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGC
chr18_denovo_ERVK     CCCACTCGAGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGC
chr2_denovo_ERVK      CCCACTCGAGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGC
chr5_denovo_ERVK      CCCACTCGAGAAGGTGTCTAGGGTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGC
                      ************************************************************

chr11_APOB            CTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAG
chr19_denovo_ERVK     CTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAG
ChrX_ERV_denovo       CTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAG
chr18_denovo_ERVK     CTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAG
chr2_denovo_ERVK      CTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAG
chr5_denovo_ERVK      CTACGTAAGTGGTGCAAACTTCTTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAG
                      ************************************************************

chr11_APOB            CTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCT
chr19_denovo_ERVK     CTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCT
ChrX_ERV_denovo       CTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCT
chr18_denovo_ERVK     CTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCT
chr2_denovo_ERVK      CTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCT
chr5_denovo_ERVK      CTACTCAGCTTCTTTTCTCCACTGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCT
                      ************************************************************

chr11_APOB            CTATATCCCTAATTAGCATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGG
chr19_denovo_ERVK     CTATATCCCTAATTAGCATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGG
ChrX_ERV_denovo       CTATATCCCTAATTAGCATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGG
chr18_denovo_ERVK     CTATATCCCTAATTAGCATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGG
chr2_denovo_ERVK      CTATATCCCTAATTAGCATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGG
chr5_denovo_ERVK      CTATATCCCTAATTAGCATATAAATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGG
                      ************************************************************

chr11_APOB            ATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGGGATTTCGAAGGTAAGTCC
chr19_denovo_ERVK     ATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGG-ATTTCGAAGGTAAGTCC
ChrX_ERV_denovo       ATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGG-ATTTCGAAGGTAAGTCC
chr18_denovo_ERVK     ATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGGGATTTCGAAGGTAAGTCC
chr2_denovo_ERVK      ATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGG-ATTTCGAAGGTAAGTCC
chr5_denovo_ERVK      ATCAGCCGGGGCTGGTCCTCGGCAGGTGGCGCCCGATACAGGGATTTCGAAGGTAAGTCC
                      ****************************************** ****************

chr11_APOB            CCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTG
chr19_denovo_ERVK     CCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTG
ChrX_ERV_denovo       CCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTG
chr18_denovo_ERVK     CCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTG
chr2_denovo_ERVK      CCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTG
chr5_denovo_ERVK      CCAACCCCATTCCCCCAGTGTTGAGTTTTCGGGACGGATAGGACCCCACTTAGGGTGCTG
                      ************************************************************

chr11_APOB            CAGACCCCCCTGTAGAATAGACAGGGAGAGAGGAGTGGGAAGTGTTGAAAGTGTTGAAGA
chr19_denovo_ERVK     CAGACCCCCCTGTAGAATAGACAGGGAGAGAGGAGTGGGAAGTGTTGAAAGTGTTGAAGA
ChrX_ERV_denovo       CAGACCCCCCTGTAGAATAGACAGGGAGAGAGGAGTGGGAAGTGTTGAAAGTGTTGAAGA
chr18_denovo_ERVK     CAGACCCCCCTGTAGAATAGACAGGGAGAGAGGAGTGGGAAGTGTTGAAAGTGTTGAAGA
chr2_denovo_ERVK      CAGACCCCCCTGTAGAATAGACAGGGAGAGAGGAGTGGGAAGTGTTGAAAGTGTTGAAGA
chr5_denovo_ERVK      CAGACCCCCCTGTAGAATAGACAGGGAGAGAGGAGTGGGAAGTGTTGAAAGTGTTGAAGA
                      ************************************************************

chr11_APOB            GTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGC
chr19_denovo_ERVK     GTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGC
ChrX_ERV_denovo       GTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGC
chr18_denovo_ERVK     GTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGC
```

```
chr2_denovo_ERVK       GTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGC
chr5_denovo_ERVK       GTTAGAGAGAAAAAACGATTTCTAAAAAGGCTGACAAAAAAGGCCTAATCTTTTGATAGC
                       ************************************************************

chr11_APOB             TAAAAGCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGA
chr19_denovo_ERVK      TAAAAGCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGA
ChrX_ERV_denovo        TAAAAGCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGA
chr18_denovo_ERVK      TAAAAGCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGA
chr2_denovo_ERVK       TAAAAGCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGA
chr5_denovo_ERVK       TAAAAGCAAAATCTTTTACTATACTTAATTTTCTGACATGGGTAATACTGAATCAAATGA
                       ************************************************************

chr11_APOB             AAGACAGCTCTTTATAGGAGTAATTTTACAGTTATTAGGTAAAAGAGGAATTAAAGTTAA
chr19_denovo_ERVK      AAGACAGCTCTTTATAGGAGTAATTTTACAGTTATTAGGTAAAAGAAGAATTAAAGTTAA
ChrX_ERV_denovo        AAGACAGCTCTTTATAGGAGTAATTTTACAGTTATTAGGTAAAAGAAGAATTAAAGTTAA
chr18_denovo_ERVK      AAGACAGCTCTTTATAGGAGTAATTTTACAGTTATTAGGTAAAAGAGGAATTAAAGTTAA
chr2_denovo_ERVK       AAGACAGCTCTTTATAGGAGTAATTTTACAGTTATTAGGTAAAAGAAGAATTAAAGTTAA
chr5_denovo_ERVK       AAGACAGCTCTTTATAGGAGTAATTTTACAGTTATTAGGTAAAAGAAGAATTAAAGTTAA
                       ********************************************.*************

chr11_APOB             AAAATCTGCCATTCAATCATTTCTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGA
chr19_denovo_ERVK      AAAATCTGCCATTCAATCATTTTTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGA
ChrX_ERV_denovo        AAAATCTGCCATTCAATCATTTTTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGA
chr18_denovo_ERVK      AAAATCTGCCATTCAATCATTTCTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGA
chr2_denovo_ERVK       AAAATCTGCCATTCAATCATTTCTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGA
chr5_denovo_ERVK       AAAATCTGCCATTCAATCATTTCTTTCATTTGTACAAGAGCACTGTCCCTGGTTTCCAGA
                       ********************** ************************************

chr11_APOB             CGAAGGCTCTGTTAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGTTAAAAACTTACCA
chr19_denovo_ERVK      CGAAGGCTCTGTTAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGCTAAAAACTTACCA
ChrX_ERV_denovo        CGAAGGCTCTGTTAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGCTAAAAACTTACCA
chr18_denovo_ERVK      CGAAGGCTCTGTTAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGTTAAAAACTTACCA
chr2_denovo_ERVK       CGAAGGCTCTGTTAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGTTAAAAACTTACCA
chr5_denovo_ERVK       CGAAGGCTCTGTTAACTTAGATGTCTGGGAAAAAGTAGGAAAACAGTTAAAAACTTACCA
                       ********************************************* **************

chr11_APOB             TGCAGAACATGGCTCAGAAAAGGTGCCTAATGACGCCTTTTCCTTATGGAATATTATTAG
chr19_denovo_ERVK      TGCCGAACTTGGCTTAGAAAAGGTGCCTAATGACGCCTTTTCCTTATGGAATATCATTAG
ChrX_ERV_denovo        TGCCGAACTTGGCTTAGAAAAGGTGCCTAATGACGCCTTTTCCTTATGGAATATCATTAG
chr18_denovo_ERVK      TGCAGAACATGGCTCAGAAAAGGTGCCTAATGACGCCTTTTCCTTATGGAATATTATTAG
chr2_denovo_ERVK       TGCAGAACATGGCTCAGAAAAGGTGCCTAATGACGCCTTTTCCTTATGGAATATTATTAG
chr5_denovo_ERVK       TGCAGAACATGGCTCAGAAAAGGTGCCTAATGACGCCTTTTCCTTATGGAATATTATTAG
                       ***.****:***** ************************************** *****

chr11_APOB             AGATGTCTTAGACCCTGCCCCAGATTCAGAAAAAGTACATCTTAAAAGGGATAGTGAAGA
chr19_denovo_ERVK      AGATGTCTTAGACCCTGCCCCTGATTCACAAAAAGTACATCTTAAAAGTGACAGTGAAGA
ChrX_ERV_denovo        AGATGTCTTAGACCCTGCCCCTGATTCACAAAAAGTACATCTTAAAAGTGACAGTGAAGA
chr18_denovo_ERVK      AGATGTCTTAGACCCTGCCCCAGATTCAGAAAAAGTACATCTTAAAAGGGATAATGAAGA
chr2_denovo_ERVK       AGATGTCTTAGACCCTGCCCCAGATTCAGAAAAAGTACATCTTAAAAGGGATAATGAAGA
chr5_denovo_ERVK       AGATGTCTTAGACCCTGCCCCAGATTCAGAAAAAGTACATCTTAAAAGGGATAATGAAGA
                       **********************:****** ****************** ** *.******

chr11_APOB             AAATGCTGTAGTTAAACCTACCCCTGAACCTAAAAAAGTAACCTTTAAAGAGGAAAATGA
chr19_denovo_ERVK      AAATGCTGTAGCTAAACCTGCCCCTAAATCTAAAAGAGTAACTTTTAAAGAGGAAAATGA
ChrX_ERV_denovo        AAATGCTGTAGCTAAACCTGCCCCTAAATCTAAAAGAGTAACTTTTAAAGAGGAAAATGA
chr18_denovo_ERVK      AAATGCTGTAGTTAAACCTACCCCTGAACCTAAAAAAGTAACCTTTAAAGAGGAAAATGA
chr2_denovo_ERVK       AAATGCTGTAGTTAAACCTACCCCTGAACCTAAAAAAGTAACCTTTAAAGAGGAAAATGA
chr5_denovo_ERVK       AAATGCTGTAGTTAAACCTACCCCTGAACCTAAAAAAGTAACCTTTAAAGAGGAAAATGA
                       *********** *******.*****.** ******.****** *****************

chr11_APOB             AGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCCTGACTATCG
chr19_denovo_ERVK      AGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCCTGACTATCG
ChrX_ERV_denovo        AGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCCTGACTATCG
chr18_denovo_ERVK      AGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCCTGACTATCG
chr2_denovo_ERVK       AGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCCTGACTATCG
chr5_denovo_ERVK       AGTCGAAATCGTAGTTAAACCTGAAGAAAATGAGAAAAACGAAGACCCGCCTGACTATCG
                       ************************************************************

chr11_APOB             GCAACTAAGAAAAATGTTAATAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGA
chr19_denovo_ERVK      GCAACTAAGAAAAATGTTAACAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGA
ChrX_ERV_denovo        GCAACTAAGAAAAATGTTAACAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGA
chr18_denovo_ERVK      GCAACTAAGAAAAATGTTAATAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGA
chr2_denovo_ERVK       GCAACTAAGAAAAATGTTAATAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGA
chr5_denovo_ERVK       GCAACTAAGAAAAATGTTAATAGCCGTGACTACTCAGGAACAACCTAAAGATAGGGATGA
                       ******************* ****************************************

chr11_APOB             GGAACAAGATCAGCCTTCCCCAAAACAAAAAGAGAATTTAGGCGAGATAACAACTAAATA
```

```
chr19_denovo_ERVK       GGAACAAGATCAGCCTTCCCCAAAACAAAAAGAGAATTTAGGCGAGATAACAACTAAATA
ChrX_ERV_denovo         GGAACAAGATCAGCCTTCCCCAAAACAAAAAGAGAATTTAGGCGAGATAACAACTAAATA
chr18_denovo_ERVK       GGAACAAGATCAGCCTTCCCCAAAACAAAAAGAGAATTTAGGCGAGATAACAACTAAATA
chr2_denovo_ERVK        GGAACAAGATCAGCCTTCCCCAAAACAAAAAGAGAATTTAGGCGAGATAACAACTAAATA
chr5_denovo_ERVK        GGAACAAGATCAGCCTTCCCCAAAACAAAAAGAGAATTTAGGCGAGATAACAACTAAATA
                        ************************************************************

chr11_APOB              TCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAAC
chr19_denovo_ERVK       TCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAAC
ChrX_ERV_denovo         TCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAAC
chr18_denovo_ERVK       TCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAAC
chr2_denovo_ERVK        TCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAAC
chr5_denovo_ERVK        TCACTCTGATGGAGATTGGCATCTCTTAAACAAAGATGCCCCAAAAGGCCTAAGAGAAAC
                        ************************************************************

chr11_APOB              TTCCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTT
chr19_denovo_ERVK       TTCCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTT
ChrX_ERV_denovo         TTCCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTT
chr18_denovo_ERVK       TTCCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTT
chr2_denovo_ERVK        TTCCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTT
chr5_denovo_ERVK        TTCCCCTGTCAGACCATCCGCTCCTAGGAGCTTGAGGGAAACATCCCCGATCAGACAGTT
                        ************************************************************

chr11_APOB              TGTAAGATTTAGTCACCAAACAGTTAAAGGATCTCCGGAGCAGGAGAAGAGAAATAGGGG
chr19_denovo_ERVK       TGTAAGATTTAATCACCAAACAGTTAAAGGATCTCCGGAGCAGGAGAAGAGAAATACGGG
ChrX_ERV_denovo         TGTAAGATTTAATCACCAAACAGTTAAAGGATCTCCGGAGCAGGAGAAGAGAAATACGGG
chr18_denovo_ERVK       TGTAAGATTTAGTCACCAAACAGTTAAAGGATCTCCGGAGCAGGAGAAGAGAAATACGGG
chr2_denovo_ERVK        TGTAAGATTTAATCACCAAACAGTTAAAGGATCTCCGGAGCAGGAGAAGAGAAATACGGG
chr5_denovo_ERVK        TGTAAGATTTAATCACCAAACAGTTAAAGGATCTCCGGAGCAGGAGAAGAGAAATACGGG
                        ***********.************************************************ ***

chr11_APOB              ACGCTCCCATCCCTCCATGCCTCCTCCTCTGTGTGGGGGTAAAGGGCCTCCACTAGGAGT
chr19_denovo_ERVK       ACGCTCCCTTCCCCCCATGCCTCCTCCTCCGTGTGGGGGTAAAGGGCCTCCTCTAGGAGT
ChrX_ERV_denovo         ACGCTCCCTTCCCCCCATGCCTCCTCCTCCGTGTGGGGGTAAAGGGCCTCCTCTAGGAGT
chr18_denovo_ERVK       ACGCTCCCATCCCTCCATGCCTCCTCCTCTGTGTGGGGGTAAAGGGCCTCCACTAGGAGT
chr2_denovo_ERVK        ACGCTCCCTTCCCCCCATGCCTCCTCCTCCGTGTGGGGGTAAAGGGCCTCCTCTAGGAGT
chr5_denovo_ERVK        ACGCTCCCATCCCTCCATGCCTCCTCCTCTGTGTGGGGGTAAAGGGCCTCCACTAGGAGT
                        ********:**** *************** **********************:********

chr11_APOB              TTCCCCGAGAAGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTC
chr19_denovo_ERVK       TTCCCCGAGAAGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTC
ChrX_ERV_denovo         TTCCCCGAGAAGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTC
chr18_denovo_ERVK       TTCCCCGAGAAGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTC
chr2_denovo_ERVK        TTCCCCGAGAAGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTC
chr5_denovo_ERVK        TTCCCCGAGAAGGGTTTGCTCTAGTCCCAAGGATAAATTTGATCCTCACCTTGAGGCTTC
                        ************************************************************

chr11_APOB              TTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGTACAAAATAAATAAAAATAT
chr19_denovo_ERVK       TTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGTACAAAATAAATAAAAATAT
ChrX_ERV_denovo         TTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGTACAAAATAAATAAAAATAT
chr18_denovo_ERVK       TTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGTACAAAATAAATAAAAATAT
chr2_denovo_ERVK        TTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGTACAAAATAAATAAAAATAT
chr5_denovo_ERVK        TTTTCTAGTCCCCTCTGCAGTTTAGATAGGGTAAAGAAGGTACAAAATAAATAAAAATAT
                        ************************************************************

chr11_APOB              TCTCAAAAAAAAAAAAAATAAATAAAAAAAAAAAATGGGTTTCTCTGCATCTAAGAATATAC
chr19_denovo_ERVK       TCTCAAAAAAAAAA------TTA-AAAAAAAAATGGGTTTCTCTGCATCTAAGAGA----
ChrX_ERV_denovo         TCTCAAAAAAAAAA------TTAAAAAAAAAAAATGGGTTTCTCTGCATCTAAGAGA----
chr18_denovo_ERVK       TCTCAAAAAAAAAA---------AAAAAAAAAATGGGTTTCTCTGCATCTAAGAATATAC
chr2_denovo_ERVK        TCTCAAAAAAAAAA------TTAAAAAAAAAAATGGGTTTCTCTGCATCTAAGAGA----
chr5_denovo_ERVK        TCTCAAAAAAAAAA------TTAAAAAAAAAAATGGGTTTCTCTGCATCTAAGAGA----
                        **************          ******************************.:

chr11_APOB              AAATGTTTATTGGCTCTGTGAAATATGATTAGAAATGTCCTGGACCCCTCTCATGAGGCT
chr19_denovo_ERVK       -AATGTTTATTGGCTCTGTGAAATATGATTAGAAATGTCCTGGACCCCTCTCATGAGGCT
ChrX_ERV_denovo         -AATGTTTATTGGCTCTGTGAAATATGATTAGAAATGTCCTGGACCCCTCTCATGAGGCT
chr18_denovo_ERVK       AAATGTTTATTGGCTCTGTGAAATATGATTAGAAATGTCCTGGACCCCTGTCATGAGGCT
chr2_denovo_ERVK        -AATGTTTATTGGCTCTGTGAAATATGATTAGAAATGTCCTGGACCCCTCTCATGAGGCT
chr5_denovo_ERVK        -AATGTTTATTGGCTCTGTGAAATATGATTAGAAATGTCCTGGACCCCTGTCATGAGGCT
                         ********************************************** **********

chr11_APOB              GTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCAGTTTCTGCA
chr19_denovo_ERVK       GTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCACTTTCTGCA
ChrX_ERV_denovo         GTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCACTTTCTGCA
chr18_denovo_ERVK       GTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCACTTTCTGCA
chr2_denovo_ERVK        GTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCACTTTCTGCA
chr5_denovo_ERVK        GTTAAATAGCCTTTGAGGGAGGCTATAGCGGTAGATGGTGGTGAGTCTCCACTTTCTGCA
```

```
                    ************************************************** ********

chr11_APOB          TAGATCATAATTTCTGCTATTGGAAAAAAA--AAAAAAAAGGAGGGAGAAAGTGCTTTAC
chr19_denovo_ERVK   TAGATCATAATTTCTGCTATTGGAAAAAAAAAAAAAAAAAAGGAGGGAGAAAGTGCTTTAC
ChrX_ERV_denovo     TAGATCATAATTTCTGCTATTGGAAAAAAAA-AAAAAAAAGGAGGGAGAAAGTGCTTTAC
chr18_denovo_ERVK   TAGATCATAATTTCTGCTATTGGAAAAAAA--AAAAAAAAGGAGGGAGAAAGTGCTTTAC
chr2_denovo_ERVK    TAGATCATAATTTCTGCTATTGGAAAAAAAAAAAAAAAAAAGGAGGGAGAAAGTGCTTTAC
chr5_denovo_ERVK    TAGATCATAATTTCTGCTATTGAAAAAAAA--AAAAAAAAGGAGGGAGAAAGTGCTTTAC
                    **********************.******* ****************************

chr11_APOB          CTCCCGAGCAAGGAGAGGGTTTACAGGACGCTGCGGCCGAGCGCCCTGGCGAGCTCCCTC
chr19_denovo_ERVK   CTCCCGAGGAAGGAGAGGGTTTACAGGACGCTGCGGCCGAGCGCCCTGGCGAGCTCCCTC
ChrX_ERV_denovo     CTCCCGAGGAAGGAGAGGGTTTACAGGACGCTGCGGCCGAGCGCCCTGGCGAGCTCCCTC
chr18_denovo_ERVK   CTCCCGAGGAAGGAGAGGGTTTACAGGACGCTGCGGCCGAGCGCCCTGGCGAGCTCCCTC
chr2_denovo_ERVK    CTCCCGAGGAAGGAGAGGGTTTACAGGACGCTGCGGCCGAGCGCCCTGGCGAGCTCCCTC
chr5_denovo_ERVK    CTCCCGAGGAAGGAGAGGGTTTACAGGACGCTGCGGCCGAGCGCCCTGGCGAGCTCCCTC
                    ******** ************************************************

chr11_APOB          CCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCA
chr19_denovo_ERVK   CCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCA
ChrX_ERV_denovo     CCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCA
chr18_denovo_ERVK   CCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCA
chr2_denovo_ERVK    CCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCA
chr5_denovo_ERVK    CCCCTCTGCGCAAACCTTTAAAAAACTTATCCACCACTTTCTGATTTAAGGTGATCGCCA
                    ************************************************************

chr11_APOB          CCGCCTCCGTTTGTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTC
chr19_denovo_ERVK   CCGCCTCCGTTTGTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTC
ChrX_ERV_denovo     CCGCCTCCGTTTGTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTC
chr18_denovo_ERVK   CCGCCTCCGTTTGTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTC
chr2_denovo_ERVK    CCGCCTCCGTTTGTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTC
chr5_denovo_ERVK    CCGCCTCCGTTTGTGCCTCCCAGGACCCAGGAGTTCCCCACTTTGCCCCCAAAGCCTCTC
                    ************************************************************

chr11_APOB          AAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTGAACAGTGGAGCTTTTAAAAC
chr19_denovo_ERVK   AAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTGAACAGTGGAGCTTTTAAAAC
ChrX_ERV_denovo     AAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTGAACAGTGGAGCTTTTAAAAC
chr18_denovo_ERVK   AAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTGAACAGTGGAGCTTTTAAAAC
chr2_denovo_ERVK    AAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTGAACAGTGGAGCTTTTAAAAC
chr5_denovo_ERVK    AAAGCGTTGCCTAAGCCTGAGGAAGATAAAAAGTTTTTTGAACAGTGGAGCTTTTAAAAC
                    ************************************************************

chr11_APOB          AGACTGCGTGCACAATATGCAGAGCATGCTTCTTAGAGAAAACCCCCTCAGGGGGGGTCT
chr19_denovo_ERVK   AGACTGCGTGCACAATATGCAGAGCATGCTTCTTAGAGAAAACCCCCTCAGGGGGGGTCT
ChrX_ERV_denovo     AGACTGCGTGCACAATATGCAGAGCATGCTTCTTAGAGAAAACCCCCTCAGGGGGGGTCT
chr18_denovo_ERVK   AGACTGCGTGCACAATATGCAGAGCATGCTTTTTAGAGAAAACCCCCTCGG-GGGGGTCT
chr2_denovo_ERVK    AGACTGCGTGCACAATATGCAGAGCATGCTTCTTAGAGAAAACCCCCTCAGGGGGGGTCT
chr5_denovo_ERVK    AGACTGCGTGCACAATATGCAGAGCATGCTTCTTAGAGAAAACCCCCTCAGGGGGGGTCT
                    ****************************** ****************.* ********

chr11_APOB          CACCCAGGCTAGGAGAAAAGCGAAACCGCAAGGGGATTTAGCAACAATATTAACCCCTGA
chr19_denovo_ERVK   CACCCAGGCTAGGAGAAAAGCGAAACCGCAAGGGGATTTAGCAACAATATTAACCCCTGA
ChrX_ERV_denovo     CACCCAGGCTAGGAGAAAAGCGAAACCGCAAGGGGATTTAGCAACAATATTAACCCCTGA
chr18_denovo_ERVK   CACCCAGGCTAGGAGAAAAGCGAAACTGCAAGGGGATTTAGCAACAATATTAACCCCTGA
chr2_denovo_ERVK    CACCCAGGCTAGGAGAAAAGCGAAACCGCAAGGGGATTTAGCAACAATATTAACCCCTGA
chr5_denovo_ERVK    CACCCAGGCTAGGAGAAAAGCGAAACCGCAAGGGGATTTAGCAACAATATTAACCCCTGA
                    ************************** *********************************

chr11_APOB          TGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTGGACATTGTAAG
chr19_denovo_ERVK   TGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTGGACATTGTAAG
ChrX_ERV_denovo     TGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTGGACATTGTAAG
chr18_denovo_ERVK   TGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTGGACATTGTAAG
chr2_denovo_ERVK    TGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTGGACATTGTAAG
chr5_denovo_ERVK    TGTGCCGCTTACTCCAATTCCAACGGGAGTGGCTGGCCCCCTACCTTTGGACATTGTAAG
                    ************************************************************

chr11_APOB          ATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAAA---AAAAAAAAAATTCGG
chr19_denovo_ERVK   ATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAA---------AAAAAAATTCGG
ChrX_ERV_denovo     ATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAAAAAAAAAAAAAAAATTCGG
chr18_denovo_ERVK   ATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAAA---AAAAAAAAAATTCGG
chr2_denovo_ERVK    ATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAA-----AAAAAAAAATTCGG
chr5_denovo_ERVK    ATCTGTCCTGGGGCATAGCTCGCTTTCTTTTCAAAAAAAAAA---AAAAAAAAAATTCGG
                    ************************************        ************

chr11_APOB          TGGTGCATGGTGTAGTAGATTCTGATTATATTGAGAAATTAAAGTCTTGATATCACCGCC
chr19_denovo_ERVK   TGGTGCATGGTGTAGTAGATTCTGATTATATTGAGAAATTAAAGTCTTGATATCACCGCC
ChrX_ERV_denovo     TGGTGCATGGTGTAGTAGATTCTGATTATATTGAGAAATTAAAGTCTTGATATCACCGCC
```

```
chr18_denovo_ERVK       TGGTGCATGGTGTAGTAGATTCTGATTATATTGAGAAATTAAAGTCTTGATATCACCGCC
chr2_denovo_ERVK        TGGTGCATGGTGTAGTAGATTCTGATTATATTGAGAAATTAAAGTCTTGATATCACCGCC
chr5_denovo_ERVK        TGGTGCATGGTGTAGTAGATTCTGATTATATTGAGAAATTAAAGTCTTGATATCACCGCC
                        ************************************************************

chr11_APOB              TACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTA
chr19_denovo_ERVK       TACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTA
ChrX_ERV_denovo         TACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTA
chr18_denovo_ERVK       TACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTA
chr2_denovo_ERVK        TACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTA
chr5_denovo_ERVK        TACCAAAACTGTACAAATTAATAAAGGTCAAAGAGTAACACAGCCTTTGCTTTTACCTTA
                        ************************************************************

chr11_APOB              TTATCAGACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACGAAACATTTGGATCTAG
chr19_denovo_ERVK       TTATCAGACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACAAAACATTTGGATCTAG
ChrX_ERV_denovo         TTATCAGACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACAAAACATTTGGATCTAG
chr18_denovo_ERVK       TTATCAGACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACGAAACATTTGGATCTAG
chr2_denovo_ERVK        TTATCAGACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACAAAACATTTGGATCTAG
chr5_denovo_ERVK        TTATCAGACAAGAAAAAACTTGACTTCTCAAGTTAAGAGCCACGAAACATTTGGATCTAG
                        ******************************************.*****************

chr11_APOB              TGATCTAGCCTTTTGGATGCAGGAAATTAC------------------------------
chr19_denovo_ERVK       TGATCTAGCCTTTTGGGTGCAGGAAATTACAGCTCCAAGGCCTTTAAAAGATCTTTTAAT
ChrX_ERV_denovo         TGATCTAGCCTTTTGGGTGCAGGAAATTACAGCTCCAAGGCCTTTAAAAGATCTTTTAAT
chr18_denovo_ERVK       TGATCTAGCCTTTTGGGTGCAGGAAATTACAGCTCCAAGGCCTTTAAAAGATCTTTTAAT
chr2_denovo_ERVK        TGATCTAGCCTTTTGGGTGCAGGAAATTACAGCTCCAAGGCCTTTAAAAGATCTTTTAAT
chr5_denovo_ERVK        TGATCTAGCCTTTTGGGTGCAGGAAATTACAGCTCCAAGGCCTTTAAAAGATCTTTTAAT
                        ****************.*************

chr11_APOB              ------------------------------------------------------------
chr19_denovo_ERVK       TCCAGAGAATAAAATGCCAGGGCTATTGGACACAGGAACAGACGTCTCTTAGCATTGCTG
ChrX_ERV_denovo         TCCAGAGAATAAAATGCCAGGGCTATTGGACACAGGAACAGACGTCTCTTAGCATTGCTG
chr18_denovo_ERVK       TCCAGAGAATAAAATGCCAGGGCTATTGGACACAGGAACAGACGTCTCTTAGCATTGCTG
chr2_denovo_ERVK        TCCAGAGAATAAAATGCCAGGGCTATTGGACACAGGAACAGACGTCTCTTAGCATTGCTG
chr5_denovo_ERVK        TCCAGAGAATAAAATGCCAGGGCTATTGGACACAGGAACAGACGTCTCTTAGCATTGCTG

chr11_APOB              --------------AGCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAG
chr19_denovo_ERVK       GGAAAGACTGGCCCAGCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAG
ChrX_ERV_denovo         GGAAAGACTGGCCCAGCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAG
chr18_denovo_ERVK       GGAAAGACTGGCCCAGCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAG
chr2_denovo_ERVK        GGAAAGACTGGCCCAGCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAG
chr5_denovo_ERVK        GGAAAGACTGGCCCAGCTCCTGGCCAACACATACTACTGAAAATGAGTTGGTGGGATTAG
                                  *********************************************

chr11_APOB              AGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATTTAACAATACTTTATAT
chr19_denovo_ERVK       AGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATTTAACAATACTTTATAT
ChrX_ERV_denovo         AGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATTTAACAATACTTTATAT
chr18_denovo_ERVK       AGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATTTAACAATACTTTATAT
chr2_denovo_ERVK        AGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATTTAACAATACTTTATAT
chr5_denovo_ERVK        AGAAAGTGGTATGTGAGGTGGGGATGCTATACAGGTGGTACATTTAACAATACTTTATAT
                        ************************************************************

chr11_APOB              ATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTAT
chr19_denovo_ERVK       ATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTAT
ChrX_ERV_denovo         ATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTAT
chr18_denovo_ERVK       ATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTAT
chr2_denovo_ERVK        ATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTAT
chr5_denovo_ERVK        ATTGTTATAAATACATAATATAAATATATTTGCCTAATTACAGTTTGCCTAATTAGGTAT
                        ************************************************************

chr11_APOB              GGGTATAAATAAAATATAATAAAGGTATACCTAATTCTACTTATAGATCTATACTTAATT
chr19_denovo_ERVK       GGGTATAAATAAAATATAATAAAGGTATACCTAATTCTACTTATAGATCTATACTTAATT
ChrX_ERV_denovo         GGGTATAAATAAAATATAATAAAGGTATACCTAATTCTACTTATAGATCTATACTTAATT
chr18_denovo_ERVK       GGGTATAAATAAAATATAATAAAGGTATACCTAATTCTACTTATAGATCTATACTTAATT
chr2_denovo_ERVK        GGGTATAAATAAAATATAATAAAGGTATACCTAATTCTACTTATAGATCTATACTTAATT
chr5_denovo_ERVK        GGGTATAAATAAAATATAATAAAGGTATACCTAATTCTACTTATAGATCTATACTTAATT
                        ************************************************************

chr11_APOB              TGTATATAAATTAATATGTATACTATATATACATATATATATACTGTATAATTAAATGTA
chr19_denovo_ERVK       TGTATATAAATTAATATGTATACTATATATACATATATATATACTGTATAATTAAAGGTA
ChrX_ERV_denovo         TGTATATAAATTAATATGTATACTATATATACATAT--ATATACTGTATAATTAAAGGTA
chr18_denovo_ERVK       TGTATATAAATTAATATGTATACTATATATACATAT--ATATACTGTATAATTAAATGTA
chr2_denovo_ERVK        TGTATATAAATTAATATGTATACTATATATACATAT--ATATACTGTATAATTAAAGGTA
chr5_denovo_ERVK        TGTATATAAATTAATATGTATACTATATATACATAT--ATATACTGTATAATTAAATGTA
                        ********************************  ****************** ***
```

```
chr11_APOB            TATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGAT
chr19_denovo_ERVK     TATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGAT
ChrX_ERV_denovo       TATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGAT
chr18_denovo_ERVK     TATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGAT
chr2_denovo_ERVK      TATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGAT
chr5_denovo_ERVK      TATTGCAGTAATATAATGTGTGTGGCTGATATTAATTGGTATGGATTGGTGTGCTGTGAT
                      ************************************************************

chr11_APOB            GATATATGTTCTATATACTGTATAATTGTATATGTGTGACATGCATTATTGCAGTACATT
chr19_denovo_ERVK     GATATATGTTCTATATACTGTATAATTGTATATGTGTGACATGTATTATTACAGTACACT
ChrX_ERV_denovo       GATATATGTTCTATATACTGTATAATTGTATATGTGTGACATGTATTATTACAGTACACT
chr18_denovo_ERVK     GATATATGTTCTATATACTGTATAATTGTATATGTGTGACATGTATTATTACAGTACACT
chr2_denovo_ERVK      GATATATGTTCTATATACTGTATAATTGTATATGTGTGACATGTATTATTACAGTACACT
chr5_denovo_ERVK      GATATATGTTCTATATACTGTATAATTGTATATGTGTGACATGTATTATTACAGTACACT
                      ****************************************** ******.******* *

chr11_APOB            GGTTTGTGTTGGTTGGTATTAGCTGTGTATGTGTTGTATTGCTGTAATATATATTAATTA
chr19_denovo_ERVK     GATTTGTGTTGGTTGGTATTAGCTGTGTACGTGTTGTATTGCTGTAATATATATTAATTA
ChrX_ERV_denovo       GGTTTGTGTTGGTTGGTATTAGCTGTGTATGTGTTGTATTGCTGTAATATATATTAATTA
chr18_denovo_ERVK     GGTTTGTGTTGGTTGGTATTAGCTGTGTACGTGTTGTATTGCTGTAATATATATTAATTA
chr2_denovo_ERVK      GGTTTGTGTTGGTTGGTATTAGCTGTGTATGTGTTGTATTGCTGTAATATATATTAATTA
chr5_denovo_ERVK      GGTTTGTGTTGGTTGGTATTAGCTGTGTACGTGTTGTATTGCTGTAATATATATTAATTA
                      *.************************* ******************************

chr11_APOB            ATATATTAATTATAAAGATTAATTCAAATATATTGATTTATATATATTATATGTCAATAA
chr19_denovo_ERVK     ATATATTAATTATAAAGATTAATTCAAATATATTGATTTATATATATTATATGTCAATAA
ChrX_ERV_denovo       ATATATTAATTATAAAGATTAATTCAAGTATATTGATTTATATATATTATATGTCAATAA
chr18_denovo_ERVK     ATATATTAATTATAAAGATTAATTCAAATATATTGATTTATATATATTATATGTCAATAA
chr2_denovo_ERVK      ATATATTAATTATAAAGATTAATTCAAGTATATTGATTTATATATATTATATGTCAATAA
chr5_denovo_ERVK      ATATATTAATTATAAAGATTAATTCAAATATATTGATTTATATATATTATATGTCAATAA
                      ***************************.**********************************

chr11_APOB            GTTTATACTTGTTGCCATATATAAATACATATATAAATACATTTTACATTTAGGTATGTC
chr19_denovo_ERVK     GTTTATACTTGTTGCCATATATAAATACATATATAAATACATTTTACATTTAGGTATGTC
ChrX_ERV_denovo       GTTTATACTTGTTGCCATATATAAATACATATATAAATACATTTTACATTTAGGTATGTC
chr18_denovo_ERVK     GTTTATACTTGTTGCCATATATAAATACATATATAAATACATTTTACATTTAGGTATGTC
chr2_denovo_ERVK      GTTTATACTTGTTGCCATATATAAATACATATATAAATACATTTTACATTTAGGTATGTC
chr5_denovo_ERVK      GTTTATACTTGTTGCCATATATAAATACATATATAAATACATTTTACATTTAGGTATGTC
                      ************************************************************

chr11_APOB            TGTATACCAAAATGAGATAAGGAGGTTATACATTTATTACTTAAATTTATACAGAGACCT
chr19_denovo_ERVK     TGTATACCAAAATGAGATAAGGAGGTTATACATTTATTACTTAAATTTATACAGAGACCT
ChrX_ERV_denovo       TGTATACCAAAATGAGATAAGGAGGTTATACATTTATTATTTAAATTTATACAGAGACCT
chr18_denovo_ERVK     TGTATACCAAAATGAGATAAGGAGGTTATACATTTATTACTTAAATTTATACAGAGACCT
chr2_denovo_ERVK      TGTATACCAAAATGAGATAAGGAGGTTATACATTTATTATTTAAATTTATACAGAGACCT
chr5_denovo_ERVK      TGTATACCAAAATGAGATAAGGAGGTTATACATTTATTACTTAAATTTATACAGAGACCT
                      ************************************* ********************

chr11_APOB            AAACTAAACATTAGACCTAAATTAACATGTCCCTAGGAAAACAGCTGGACAAATAGTAAA
chr19_denovo_ERVK     AAACTAAACATTAGACCTCAATTAACATGTCCCTAGGAAAACAGCTGGACAAATAGTAAA
ChrX_ERV_denovo       AAACTAAACATTAGACCTAAATTAACATATCCCTAGGAAAACAGCTGGACAAATAGTAAA
chr18_denovo_ERVK     AAACTAAACATTAGACCTAAATTAACATGTCCCTAGGAAAACAGCTGGACAAATAGTAAA
chr2_denovo_ERVK      AAACTAAACATTAGACCTAAATTAACATATCCCTAGGAAAACAGCTGGACAAATAGTAAA
chr5_denovo_ERVK      AAACTAAACATTAGACCTAAATTAACATGTCCCTAGGAAAACAGCTGGACAAATAGTAAA
                      ******************.**********.******************************

chr11_APOB            GTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACACAGGGGCAGACTACAATGA
chr19_denovo_ERVK     GTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACACAGGAGCAGACTACAATGA
ChrX_ERV_denovo       GTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACACAGGGGCAGACTACAATGA
chr18_denovo_ERVK     GTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACACAGGAGCAGACTACAATGA
chr2_denovo_ERVK      GTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACACAGGGGCAGACTACAATGA
chr5_denovo_ERVK      GTTATGTCCAAATTGTAATTAATCACTTAAAGGTAAACACACAGGGGCAGACTACAATGA
                      ********************************************.***************

chr11_APOB            TGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACC
chr19_denovo_ERVK     TGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACC
ChrX_ERV_denovo       TGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACC
chr18_denovo_ERVK     TGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACC
chr2_denovo_ERVK      TGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACC
chr5_denovo_ERVK      TGAGACACTAGATGCCTGAATTACAGGCTACGACTCGGCCCCTGGTCAAGTAAAAAGACC
                      ************************************************************

chr11_APOB            TCCTTACTGGAGAGTGTAAAAGGCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTT
chr19_denovo_ERVK     TCCTTACTAGAGAGTGTAAAAGGCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTT
ChrX_ERV_denovo       TCCTTACTGGAGAGTGTAAAAGGCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTT
chr18_denovo_ERVK     TCCTTACTGGAGAGTGTAAAAGGCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTT
chr2_denovo_ERVK      TCCTTACTGGAGAGTGTAAAAGGCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTT
```

```
chr5_denovo_ERVK        TCCTTACTGGAGAGTGTAAAAGGCCAGATGTGTTGCTAACTTGTGGGAGAGGGTATGCTT
                        ********.**************************************************

chr11_APOB              ATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTCGTCATGTCAC
chr19_denovo_ERVK       ATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTCGTCATGTCAC
ChrX_ERV_denovo         ATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTCGTCATGTCAC
chr18_denovo_ERVK       ATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTCGTCATGTCAC
chr2_denovo_ERVK        ATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTCGTCATGTCAC
chr5_denovo_ERVK        ATATATTTCTACAGAATGCAGATTCTCCGATTTGGATCTCAGAAAAATTCGTCATGTCAC
                        ************************************************************

chr11_APOB              TTTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGC
chr19_denovo_ERVK       ATTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGC
ChrX_ERV_denovo         TTTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGC
chr18_denovo_ERVK       ATTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGC
chr2_denovo_ERVK        TTTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGC
chr5_denovo_ERVK        ATTCCCCAAAGACCAGGTTCGCGGCGCGCGCACTCGCTCGCTCCCCCTCCCCCAGCCAGC
                        :************************************************************

chr11_APOB              TCTCTCC-----GCCGCCAGCGGCGCCCTGCACCCTCCTTGCTGCACCCCGAGACCTAGA
chr19_denovo_ERVK       TCTCTCCTCCGCGCCGCCATCGGCGCCCCACACCCTCCTTGCTGCACCCCGCGACCTAGA
ChrX_ERV_denovo         TCTCTCC-----GCCGCCAGCGGCGCCCTGCACCCTCCTTGCTGCACCCCGAGACCTAGA
chr18_denovo_ERVK       TCTCTCC-----GCCGCCAGCGGCGCCCTGCACCCTCCTTGCTGCACCCCGAGACCTAGA
chr2_denovo_ERVK        TCTCTCC-----GCCGCCAGCGGCGCCCTGCACCCTCCTTGCTGCACCCCGAGACCTAGA
chr5_denovo_ERVK        TCTCTCC-----GCCGCCAGCGGCGCCCTGCACCCTCCTTGCTGCACCCCGAGACCTAGA
                        *******     ******* ******** .**********************.********

chr11_APOB              GCAAAGAAGTCTGTGCGGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGCAGTGCAGT
chr19_denovo_ERVK       GCAAGGAAGTCTGTGCG-CGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGGAGTGCAGT
ChrX_ERV_denovo         GCAAAGAAGTCTGTGCGGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGCAGTGCAGT
chr18_denovo_ERVK       GCAAAGAAGTCTGTGCGGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGCAGTGCAGT
chr2_denovo_ERVK        GCAAAGAAGTCTGTGCGGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGCAGTGCAGT
chr5_denovo_ERVK        GCAAAGAAGTCTGTGCGGCGAGTGAGGGCCAGAGAGGAAAGCGCGCCCGCGCAGTGCAGT
                        ****.************ ********************************** ********

chr11_APOB              CCAGACCAGCGCGCTCCCGCCGCGGGCGACAGGAAAACGCCCCGAGCCCAAGCGGC----
chr19_denovo_ERVK       CCAGACCAGCGCGCTCCCGCC---GGCGACAGGAAAACGCCCGGAGCCCAAGCGGCGGCG
ChrX_ERV_denovo         CCAGACCAGCGCGCTCCCGCCGCGGGCGACAGGAAAACGCCCCGAGCCCAAGCGGC----
chr18_denovo_ERVK       CCAGACCAGCGCGCTCCCGCCGCGGGCGACAGGAAAACGCCCCGAGCCCAAGCGGC----
chr2_denovo_ERVK        CCAGACCAGCGCGCTCCCGCCGCGGGCGACAGGAAAACGCCCCGAGCCCAAGCGGC----
chr5_denovo_ERVK        CCAGACCAGCGCGCTCCCRCCGCGGGCGACAGGAAAACGCCCCGAGCCCAAGCGGC----
                        ****************** **   ****************** *************

chr11_APOB              GCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCCGTAGACGCCTCGGCCCACA
chr19_denovo_ERVK       GCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCCGTAGACGCCGCGGCCCACA
ChrX_ERV_denovo         GCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCCGTAGACGCCTCGGCCCACA
chr18_denovo_ERVK       GCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCCGTAGACGCCTCGGCCCACA
chr2_denovo_ERVK        GCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCCGTAGACGCCTCGGCCCACA
chr5_denovo_ERVK        GCGGCTAGCCCGAGTCCGCGAACCCCGCCCCTCCGCCCGCCGTAGACGCCTCGGCCCACA
                        ************************************************ *********

chr11_APOB              ACCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCT
chr19_denovo_ERVK       GCCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCT
ChrX_ERV_denovo         ACCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCT
chr18_denovo_ERVK       ACCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCT
chr2_denovo_ERVK        ACCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCT
chr5_denovo_ERVK        ACCTGCCGTTCGCGTTCTGCCTCCTGATGCTGCGGACACTGCTCCGCCGAGCCGATGCCT
                        .************************************************************

chr11_APOB              ACAGCTGCTCCTTGTCTCATCTTGTTTAAATCCTCAA-----------------------
chr19_denovo_ERVK       ACAGCTGCTCCTTGTTTCATCTTGTTTAAATCCTCAA-----------------------
ChrX_ERV_denovo         ACAGCTGCTCCTTGTTTCATCTTGTTTAAATCCTCAATATAATGTTTGCTCCTTTGTGGT
chr18_denovo_ERVK       ACAGCTGCTCCTTGTTTCATCTTGTTTAAATCCTCAATATAATGTTTGCTCCTTTGTGGT
chr2_denovo_ERVK        ACAGCTGCTCCTTGTTTCATCTTGTTTAAATCCTCAATATAATGTTTGCTCCTTTGTGGT
chr5_denovo_ERVK        ACAGCTGCTCCTTGTTTCATCTTGTTTAAATCCTCAATATAATGTTTGCTCCTTTGTGGT
                        *************** **********************

chr11_APOB              ---------------------------------------------TATAATGTTTGCTCC
chr19_denovo_ERVK       ---------------------------------------------TATAATGTTTGCTCC
ChrX_ERV_denovo         GGTGTTGTGAACAAATGTATGCTCTCATCTTGTTTAAATCCTCAATATAATGTTTGCTCC
chr18_denovo_ERVK       GGTGTTGTGAACAAATGTATGCTCTCATCTTGTTTAAATCCTCAATATAATGTTTGCTCC
chr2_denovo_ERVK        GGTGTTGTGAACAAATGTATGCTCTCATCTTGTTTAAATCCTCAATATAATGTTTGCTCC
chr5_denovo_ERVK        GGTGTTGTGAACAAATGTATGCTCTCATCTTGTTTAAATCCTCAATATAATGTTTGCTCC
                                                              ***************

chr11_APOB              TTTGTAATGTTACAATGTCCACCTTATCTTATGATGTAACCACTTATTGAATTATGCTCT
chr19_denovo_ERVK       TTTGTAATGTTACAACGTCCACCTTATCTTATGATGTAACCACTTATTGAATTATGCTCT
```

```
ChrX_ERV_denovo         TTTGTAATGTTACAACGTCCACCTTACCTTATGATGTAACCACTTATTGAATTATGCTCT
chr18_denovo_ERVK       TTTGTAATGTTACAACGTCCACCTTATCTTATGATGTAACCACTTATTGAATTATGCTCT
chr2_denovo_ERVK        TTTGTAATGTTACAACGTCCACCTTACCTTATGATGTAACCACTTATTGAATTATGCTCT
chr5_denovo_ERVK        TTTGTAATGTTACAACGTCCACCTTACCTTATGATGTAACCACTTATTGAATTATGCTCT
                        *************** ********** ********************************

chr11_APOB              TGCTGTATTACAACAACTGCAGGATTTAATGCGAACACGACGGTTTGTGGGCTTACTTAT
chr19_denovo_ERVK       TGCTGTATTACAACAACTGCAGGATTTAATGCGAACACGACGGTTTGTGGGCTTACTTAT
ChrX_ERV_denovo         TGCTGTATTACAACAACTGCAGGATTTAATGCGAACACGACGGTTTGTGGGCTTACTTAT
chr18_denovo_ERVK       TGCTGTATTACAACAACTGCAGGATTTAATGCGAACACGACGGTTTGTGGGCTTACTTAT
chr2_denovo_ERVK        TGCTGTATTACAACAACTGCAGGATTTAATGCGAACACGACGGTTTGTGGGCTTACTTAT
chr5_denovo_ERVK        TGCTGTATTACAACAACTGCAGGATTTAATGCGAACACGACGGTTTGTGGGCTTACTTAT
                        ************************************************************

chr11_APOB              CTTGGGAATAGCAGCTTTGATAAGCGCAATTACTTCTGTTACTGTGGCAGCAATATCATT
chr19_denovo_ERVK       CTTGGGAATAGCAGCTTTGATAAGTGCAATTACTTCTGTTACTGTGGCAGCAATATCATT
ChrX_ERV_denovo         CTTGGGAATAGCAGCTTTGATAAGTGCAATTACTTCTGTTACTGTGGCAGCAATATCATT
chr18_denovo_ERVK       CTTGGGAATAGCAGCATTGATAAGTGCAATTACTTCTGTTACTGTGGCAGCAATATCATT
chr2_denovo_ERVK        CTTGGGAATAGCAGCTTTGATAAGTGCAATTACTTCTGTTACTGTGGCAGCAATATCATT
chr5_denovo_ERVK        CTTGGGAATAGCAGCTTTGATAAGTGCAATTACTTCTGTTACTGTGGCAGCAATATCATT
                        ***************:******** **********************************

chr11_APOB              GACTCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGC
chr19_denovo_ERVK       GACTCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGC
ChrX_ERV_denovo         GACTCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGC
chr18_denovo_ERVK       GACTCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGC
chr2_denovo_ERVK        GACTCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGC
chr5_denovo_ERVK        GACTCAACAAGTACATACTGCTCAATATGTTGATTCTATGTCCAAAAATGTTTCTTTAGC
                        ************************************************************

chr11_APOB              ATTGGCAACACAGGAAGCTATAGACAGGAAATTAGAAATGAGGGTAGATGCCCTAGAGGA
chr19_denovo_ERVK       ATTGGCAACACAGGAAGCTATAGACAGAAAATTAGAAATGAGGGTAGATGCCCTAGAGGA
ChrX_ERV_denovo         ATTGGCAACACAGGAAGCTATAGACAGAAAATTAGAGATGAGGGTAGACGCCCTAGAAGA
chr18_denovo_ERVK       ATTGGCAACACAGGAAGCTATAGACAGAAAATTAGAGATGAGGGTAGACGCCCTAGAGGA
chr2_denovo_ERVK        ATTGGCAACACAGGAAGCTATAGACAGAAAATTAGAGATGAGGGTAGACGCCCTAGAAGA
chr5_denovo_ERVK        ATTGGCAACACAGGAAGCTATAGACAGAAAATTAGAGATGAGGGTAGACGCCCTAGAAGA
                        ***************************.********.*********** ********.**

chr11_APOB              AGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAGGTGAAAATGGCATTGTCCTG
chr19_denovo_ERVK       AGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAGGTGAAAATGGCATTGTCCTG
ChrX_ERV_denovo         AGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAGGTGAAAATGGCATTGTCCTG
chr18_denovo_ERVK       AGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAGGTGAAAATGGCATTGTCCTG
chr2_denovo_ERVK        AGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAGGTGAAAATGGCATTGTCCTG
chr5_denovo_ERVK        AGCAGTAATACATATTGGGACTGAATTGCAGGCTTTAAAGGTGAAAATGGCATTGTCCTG
                        ************************************************************

chr11_APOB              TCATGCTGACTACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGAGACAGATTTTGA
chr19_denovo_ERVK       TCATGCTGACTACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGAGACAGATTTTGA
ChrX_ERV_denovo         TCATGCTGACTACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGACACAGATTTTGA
chr18_denovo_ERVK       CCATGCTGACTACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGAGACAGATTTTGA
chr2_denovo_ERVK        TCATGCTGACTACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGACACAGATTTTGA
chr5_denovo_ERVK        TCATGCTGACTACCGGTGGATATGTGTAACACCCCTGAAAGTAAATGACACAGATTTTGA
                         ********************************************** ************

chr11_APOB              ATGGGAAAAGATTAAAAACCATATTTCAGGTATTTGGAACAGCTCTGACATTAGCTTAGA
chr19_denovo_ERVK       ATGGGAAAAGATTAAAAACCATATTTCAGGTATTTGGAACAGCTCTGACATTAGCTTAGA
ChrX_ERV_denovo         ATGGGAAAAGATTAAAAACCATATTTCAGGTATTTGGAACAGCTCTGACATTAGCTTAGA
chr18_denovo_ERVK       ATGGGAAAAGATTAAAAACCATATTTCAGGTATTTGGAACAGCTCTGACATTAGCTTAGA
chr2_denovo_ERVK        ATGGGAAAAGATTAAAAACCATATTTCAGGTATTTGGAACAGCTCTGACATTAGCTTAGA
chr5_denovo_ERVK        ATGGGAAAAGATTAAAAACCATATTTCAGGTATTTGGAACAGCTCTGACATTAGCTTAGA
                        ************************************************************

chr11_APOB              CTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCCGATTAGATTTTACTGC
chr19_denovo_ERVK       CTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCCGATTAGATTTTACTGC
ChrX_ERV_denovo         CTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCCGATTAGATTTTACTGC
chr18_denovo_ERVK       CTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCTGATTAGATTTTACTGC
chr2_denovo_ERVK        CTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCCGATTAGATTTTACTGC
chr5_denovo_ERVK        CTTAGGGAAACTTCACAATCAAATAGCAACCCTGGAACACTCCCGATTAGATTTTACTGC
                        ****************************************** ***************

chr11_APOB              CGCTGGAACAGCAAATGATTTCTTCCATACTTTCTCTAACTACATTTCAGGAAAAAATAT
chr19_denovo_ERVK       CGCTGGAACAGCAAATGATTTCTTCCATACTTTCTCTAACTACATTTCAGGAAAAAATAT
ChrX_ERV_denovo         CGCTGGAACAGCAAATGATTTCTTCCATACTTTCTCTAACTACATTTCAGGAAAAAATAT
chr18_denovo_ERVK       CGCTGGAACAGCAAATGATTTCTTCCATACTTTCTCTAACTACATTTCAGGAAAAAATAT
chr2_denovo_ERVK        CGCTGGAACAGCAAATGATTTCTTCCATACTTTCTCTAATTACATTTCAGGAAAAAATAT
chr5_denovo_ERVK        CGCTGGAACAGCAAATGATTTCTTCCATACTTTCTCTAATTACATTTCAGGAAAAAATAT
                        ************************************ ********************
```

```
chr11_APOB            TCTGTCTACCTTCCTCGGCTACGCTACCTTGGCTGTTTTAATTTTATTTCTAATAATCAT
chr19_denovo_ERVK     TCTGTCTACCTTCCTCGGCTACGCTACCTTGGCTGTTTTAATTTTATTTCTAATAATCAT
ChrX_ERV_denovo       TCTGTCTACCTTCCTCGGCTACGCTACCTTGGCTGTTTTAATTTTATTTCTAATAATCAT
chr18_denovo_ERVK     TCTGTCTACCTTCCTCGGCTACGCTACCTTGGCTGTTTTAATTTTATTTCTAATAATCAT
chr2_denovo_ERVK      TCTGTCTACCTTCCTCGGCTACGCTACCTTGGCTGTTTTAATTTTATTTCTAATAATCAT
chr5_denovo_ERVK      TCTGTCTACCTTCCTCGGCTACGCTACCTTGGCTGTTTTAATTTTATTTCTAATAATCAT
                      ************************************************************

chr11_APOB            TCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGCTACA
chr19_denovo_ERVK     TCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGCTACA
ChrX_ERV_denovo       TCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGCTACA
chr18_denovo_ERVK     TCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGCTACA
chr2_denovo_ERVK      TCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGCTACA
chr5_denovo_ERVK      TCTTCCTTGTATTGTCAGGATTCTTCGGCAGAGCATTCAGAGGCTCGCGACTGAGCTACA
                      ************************************************************

chr11_APOB            TCTGGCTGTTTTAAGAAATAAAAAAGGGGGAGATGCGGGGAGCCGGTGAGGCATTCCACT
chr19_denovo_ERVK     TCTGGCTGTTTTAAGAAATAAAAAAGGGGGAGATGCGGGGAGCCGGTGAGGCATTCCACT
ChrX_ERV_denovo       TCTGGCTGTTTTAAGAAATAAAAAAGGGGGAGATGCGGGGAGCCGGTGAGGCATTCCACT
chr18_denovo_ERVK     TCTGGCTGTTTTAAGAAATAAAAAAGGGGGAGATGCGGGGAGCCGGTGAGGCATTCCACT
chr2_denovo_ERVK      TCTGGCTGTTTTAAGAAATAAAAAAGGGGGAGATGCGGGGAGCCGGTGAGGCATTCCACT
chr5_denovo_ERVK      TCTGGCTGTTTTAAGAAATAAAAAAGGGGGAGATGCGGGGAGCCGGTGAGGCATTCCACT
                      ************************************************************

chr11_APOB            CGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGG
chr19_denovo_ERVK     CGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGG
ChrX_ERV_denovo       CGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGG
chr18_denovo_ERVK     CGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGG
chr2_denovo_ERVK      CGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGG
chr5_denovo_ERVK      CGTGACAAAGGTCATGAGGAAGGAGGCTCGGCATACGCAAAGGCGGGATCGAGCCTCAGG
                      ************************************************************

chr11_APOB            AGTCCCCCCGGATATTCTCGAGCATTTTCCCCCAA--AAAACCAGAGTCTGCCTACTTTA
chr19_denovo_ERVK     AGTCCCCCCGGATATTCTCGAGCATTTTCCCCCAA-AAAAACCAGAGTCTGCCTACTTTA
ChrX_ERV_denovo       AGTCCCCCCGGATATTCTCGAGCATTTTCCCCCA-AAAAAACCAGAGTCTGCCTACTTTA
chr18_denovo_ERVK     AGTCCCCCCGGATATTCTCGAGCATTTTCCCCCCAAAAAAACCAGAGTCTGCCTACTTTA
chr2_denovo_ERVK      AGTCCCCCCGGATATTCTCGAGCATTTTCCCCCA--AAAAACCAGAGTCTGCCTACTTTA
chr5_denovo_ERVK      AGTCCCCCCGGATATTCTCGAGCATTTTCCCCCA--AAAAACCAGAGTCTGCCTACTTTA
                      *********************************.   **********************

chr11_APOB            TTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCCTACCACCGTCTCTCTCT
chr19_denovo_ERVK     TTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCCTACCACCATCTCGCTCT
ChrX_ERV_denovo       TTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCCTACCACCATCTCGCTCT
chr18_denovo_ERVK     TTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCCTACCACCGTCTCTCTCT
chr2_denovo_ERVK      TTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCCTACCACCATCTCGCTCT
chr5_denovo_ERVK      TTGCTTTGTGCTCTCACCTCTGACTTTACTGGGGGCTGTCCCCTACCACCATCTCGCTCT
                      *************************************************.**** ****

chr11_APOB            CTCTGTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGT
chr19_denovo_ERVK     CTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGT
ChrX_ERV_denovo       CTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGT
chr18_denovo_ERVK     CTCTGTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGT
chr2_denovo_ERVK      CTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGT
chr5_denovo_ERVK      CTCTCTGTCAAAGAGTTAACTTACAGCTCCAATTAATAAAGTTCCTGGGCAATTAGGAGT
                      **** ********************************************************

chr11_APOB            GTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCC
chr19_denovo_ERVK     GTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCC
ChrX_ERV_denovo       GTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCC
chr18_denovo_ERVK     GTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCC
chr2_denovo_ERVK      GTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCC
chr5_denovo_ERVK      GTTTAAATCCAAACCCCTCTGATGGCTCTCTAACTCGCCTGACAAGTTTACCCGGACTCC
                      ************************************************************

chr11_APOB            TGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGGAAGCTTAAG
chr19_denovo_ERVK     TGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGGAAGCTTAAG
ChrX_ERV_denovo       TGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGGAAGCTTAAG
chr18_denovo_ERVK     TGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGGAAGCTTAAG
chr2_denovo_ERVK      TGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGGAAGCTTAAG
chr5_denovo_ERVK      TGCAGCTATGCATACGATTGTTTACAGTCTCCCAGCCTCGAGAGGCATGGGAAGCTTAAG
                      ************************************************************

chr11_APOB            ATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAG
chr19_denovo_ERVK     ATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAG
ChrX_ERV_denovo       ATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAG
chr18_denovo_ERVK     ATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAG
```

```
chr2_denovo_ERVK        ATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAG
chr5_denovo_ERVK        ATATTCAAATAGCTTAGAGCCTCTCAGAGAGTTAAAAACTGTCAGAATAAACTAGTAAAG
                        ************************************************************

chr11_APOB              GATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTT
chr19_denovo_ERVK       GATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTT
ChrX_ERV_denovo         GATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTT
chr18_denovo_ERVK       GATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTT
chr2_denovo_ERVK        GATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTT
chr5_denovo_ERVK        GATTTCATTGATGAGTCAATGCTTGTTGCCAAGTTTTCACATCCCCTGAATTGTATCCTT
                        ************************************************************

chr11_APOB              GAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAGT
chr19_denovo_ERVK       GAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAGT
ChrX_ERV_denovo         GAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAGT
chr18_denovo_ERVK       GAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAGT
chr2_denovo_ERVK        GAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAGT
chr5_denovo_ERVK        GAATATGTATCAATTAATAGTGGGTATGTAGAAAAAATAAGTAGTGGCCTTGGTGTTAGT
                        ************************************************************

chr11_APOB              AACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTTGTAAACCCATTACACATCCGCC
chr19_denovo_ERVK       AACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTTGTAAACCCATTACACATCCGCC
ChrX_ERV_denovo         AACTTTAGACCCTTAAGGTAATAAATTCTTTCTTT---GTAAACCCATTACACATCCGCC
chr18_denovo_ERVK       AACTTTAGACCCTTAAGGTAATAAATTCTTTCTTTGTTGTAAACCCATTACACATCAGCC
chr2_denovo_ERVK        AACTTTAGACCCTTAAGGTAATAAATTCTTTCTTT---GTAAACCCATTACACATCCGCC
chr5_denovo_ERVK        AACTTTAGACCCTTAAGGTAATAAATTCTTTCTTT---GTAAACCCATTACACATCCGCC
                        **********************************   ******************.***

chr11_APOB              CTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCAAACCTTGAAATAATTACTCTTA
chr19_denovo_ERVK       CTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCAAACCTTGAAATAATTACTCTTA
ChrX_ERV_denovo         CTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCAAACCTTGAAATAATTACTCTTA
chr18_denovo_ERVK       CTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCAAACCTTGAAATAATTACTCTTA
chr2_denovo_ERVK        CTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCAAACCTTGAAATAATTACTCTTA
chr5_denovo_ERVK        CTATAGGAATGCAATTTTATCTTTGGAAGATGGTGCCAAACCTTGAAATAATTACTCTTA
                        ************************************************************

chr11_APOB              GAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTT
chr19_denovo_ERVK       GAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTT
ChrX_ERV_denovo         GAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTT
chr18_denovo_ERVK       GAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTT
chr2_denovo_ERVK        GAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTT
chr5_denovo_ERVK        GAGAAAGTAAGTCTTTGTTGATAAGTCCTTGTCAAGAGTCATAAAATGTTAGTAGGCCTT
                        ************************************************************

chr11_APOB              CTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAG
chr19_denovo_ERVK       CTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAG
ChrX_ERV_denovo         CTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAG
chr18_denovo_ERVK       CTGGCCAGAAGATGATGTAAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAG
chr2_denovo_ERVK        CTGGCCAGAAGATGATGTGAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAG
chr5_denovo_ERVK        CTGGCCAGAAGATGATGTGAATCACCTAAACCATTTGTATACGATACATTTGCAGGAAAG
                        ******************.*****************************************

chr11_APOB              AAACCTTGGTTTTTGATAAGAATCAAAGACTGCTGACTTTGCATCCCCTATTATCCTCTA
chr19_denovo_ERVK       AAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACTTTGCATCCCCTATTATCCTCTA
ChrX_ERV_denovo         AAACCTTGGTTTTTGATAAGAATCAAAGACTGCTGACTTTGCATCCCCTATTATCCTCTA
chr18_denovo_ERVK       AAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACTTTGCATCCCCTATTATCCTCTA
chr2_denovo_ERVK        AAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACTTTGCATCCCCTATTATCCTCTA
chr5_denovo_ERVK        AAACCTTGGTTTTTGATAAGAACCAAAGACTGCTGACTTTGCATCCCCTATTATCCTCTA
                        ********************* ************************************

chr11_APOB              TGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAA
chr19_denovo_ERVK       TGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAA
ChrX_ERV_denovo         TGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAA
chr18_denovo_ERVK       TGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAA
chr2_denovo_ERVK        TGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAA
chr5_denovo_ERVK        TGTGTAACTTAGGGTATAAAAGCCCCTGTTAAAAATAAAGCTACGGGCCTTGCTCACCAA
                        ************************************************************

chr11_APOB              CGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGA
chr19_denovo_ERVK       CGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGA
ChrX_ERV_denovo         CGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGA
chr18_denovo_ERVK       CGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGA
chr2_denovo_ERVK        CGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGA
chr5_denovo_ERVK        CGCTTGGTCTCCCCATGTCATTCTTTTAACTTCCAGCTGAGTCTCCATCTGGAGCGCGGA
                        ************************************************************

chr11_APOB              ACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAGGTGTCTAGG
```

```
chr19_denovo_ERVK       ACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAGGTGTCTAGG
ChrX_ERV_denovo         ACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAGGTGTCTAGG
chr18_denovo_ERVK       ACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAGGTGTCTAGG
chr2_denovo_ERVK        ACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAGGTGTCTAGG
chr5_denovo_ERVK        ACCCACCACGCTTACTAATCATGCCTGGGCTTCTAAGACCCACTCGAGAAGGTGTCTAGG
                        ************************************************************

chr11_APOB              GTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTC
chr19_denovo_ERVK       GTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTC
ChrX_ERV_denovo         GTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTC
chr18_denovo_ERVK       GTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTC
chr2_denovo_ERVK        GTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTC
chr5_denovo_ERVK        GTGAGACACCTTCCGCTATTCGAGAGGGCGCCTGCGGCCTACGTAAGTGGTGCAAACTTC
                        ************************************************************

chr11_APOB              TTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCAC
chr19_denovo_ERVK       TTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCAC
ChrX_ERV_denovo         TTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCAC
chr18_denovo_ERVK       TTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCAC
chr2_denovo_ERVK        TTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCAC
chr5_denovo_ERVK        TTGTCTTGAAGTTTTATTGGTCTCCCGCGTAAACCAAGCTACTCAGCTTCTTTTCTCCAC
                        ************************************************************

chr11_APOB              TGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCATATA
chr19_denovo_ERVK       TGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCATATA
ChrX_ERV_denovo         TGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCATATA
chr18_denovo_ERVK       TGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCATATA
chr2_denovo_ERVK        TGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCATATA
chr5_denovo_ERVK        TGAAATTTCCTACTGAGCTATCCTCATTCTATTGTTCTCTATATCCCTAATTAGCATATA
                        ************************************************************

chr11_APOB              AATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGG
chr19_denovo_ERVK       AATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGG
ChrX_ERV_denovo         AATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGG
chr18_denovo_ERVK       AATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGG
chr2_denovo_ERVK        AATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGG
chr5_denovo_ERVK        AATAGTCGCCGACGCCGTCTCCCCTTCGAATACCCTGGATCAGCCGGGGCTGGTCCTCGG
                        ************************************************************

chr11_APOB              CATGAAGGCAAGCA---AGTT--CTA-------------------CTTTACCCAGAG---
chr19_denovo_ERVK       CANNNAGGGTCCATTC--------AAATAGAGGTGAA-------AATATTGCCTTAT---
ChrX_ERV_denovo         CAGGactGTTGTTTCTTGGATATCATTTTG-GATGAACTATGAATCTTC-TGCAGCCAGG
chr18_denovo_ERVK       CANNNATTTATGCTA--AG-TTCCACCTAGATACGTACTGTGGGTTTTGTTCCTTTGTAG
chr2_denovo_ERVK        CANNNGCTGCCTGT----G-ATGCA--------GGAGATATGGGTTCAATCCCTGGATCA
chr5_denovo_ERVK        CANNNATTTTCCAT----G-TGT-------------------GGGT----TACCTGAAA--
                        **          :                                 *:

chr11_APOB              ----------------------------------------------AAAGAAGAGCCT
chr19_denovo_ERVK       --------------------------------AGAGCCAATCTTT---------------
ChrX_ERV_denovo         GTGACAGT-------------CGGTAAGTCCTGTCCCCATCTCAGGAAAGAACTTGGCTC
chr18_denovo_ERVK       GAGTCA----------------------AC-----C-AGCA-------------------
chr2_denovo_ERVK        GAAAGATCCCTGGATAAGAAACTGGCGACCCACTCC-AGTATT-----CTTGCCTGGGAA
chr5_denovo_ERVK        --------------AACAAAACTGGTTTTCCTGCCC-TGCCTTTTGAAAATGATGGGCTA


chr11_APOB              AAACA--------CAGCCTCAACA-----------------------TNNNNNNNNNNN-
chr19_denovo_ERVK       ---------------GT----G----------AAAGCAAG--------NNNNNNNNNNNN
ChrX_ERV_denovo         ATCT---TGCTC--------AGAAAAAAGCTGAGAGTAAG--GAGCAGTGTG--------
chr18_denovo_ERVK       ---------------GAAACAGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
chr2_denovo_ERVK        ATCCCATGGACAGAGGAACCTGGCAGA-----------AATAGTCCATGG-GGC-TG---
chr5_denovo_ERVK        ATCAGAAGGCCCTAAGACCATGTATGGAGATGAAAAGATGATTACCAGGTTGGA-AGATT
                                           .

chr11_APOB              -------NNNN---
chr19_denovo_ERVK       NN-NNNNNNNNN-
ChrX_ERV_denovo         --------------
chr18_denovo_ERVK       NNNNNNNNNNNNNN
chr2_denovo_ERVK        --------------
chr5_denovo_ERVK        TG-AAT-TA--AAA
```

Experimental Section

Study 4:

A stop-gain in the *laminin, alpha 3* gene causes recessive junctional epidermolysis bullosa in Belgian Blue cattle.

Sartelet A, Harland C, Tamma N, Karim L, Bayrou C, Li W, Ahariz N, Coppieters W, Georges M, Charlier C.

Abstract

Four newborn purebred Belgian Blue calves presenting a severe form of epidermolysis bullosa were recently referred to our heredo-surveillance platform. SNP array genotyping followed by autozygosity mapping located the causative gene in a 8.3-Mb interval on bovine chromosome 24. Combining information from (i) whole-genome sequencing of an affected calf, (ii) transcriptomic data from a panel of tissues and (iii) a list of functionally ranked positional candidates pinpointed a private *G* to *A* nucleotide substitution in the *LAMA3* gene that creates a premature stop codon (p.Arg2609*) in exon 60, truncating 22% of the corresponding protein. The *LAMA3* gene encodes the alpha 3 subunit of the heterotrimeric laminin-332, a key constituent of the lamina lucida that is part of the skin basement membrane connecting epidermis and dermis layers. Homozygous loss-of-function mutations in this gene are known to cause severe junctional epidermolysis bullosa in human, mice, horse, sheep and dog. Overall, our data strongly support the causality of the identified gene and mutation.

ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics
SHORT COMMUNICATION
doi: 10.1111/age.12342

A stop-gain in the *laminin, alpha 3* gene causes recessive junctional epidermolysis bullosa in Belgian Blue cattle

Arnaud Sartelet*[1], Chad Harland[†1], Nico Tamma[†], Latifa Karim[†‡], Calixte Bayrou[§], Wanbo Li*, Naima Ahariz[†‡], Wouter Coppieters[†‡], Michel Georges[†] and Carole Charlier[†]

*Bovine Clinic, FARAH and Faculty of Veterinary Medicine, University of Liège, Liège, Belgium. [†]Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, Liège, Belgium. [‡]GIGA-Genomic platform, University of Liège, Liège, Belgium. [§]Department of Pathology, FARAH and Faculty of Veterinary Medicine, University of Liège, Liège, Belgium.

Summary

Four newborn purebred Belgian Blue calves presenting a severe form of epidermolysis bullosa were recently referred to our heredo-surveillance platform. SNP array genotyping followed by autozygosity mapping located the causative gene in a 8.3-Mb interval on bovine chromosome 24. Combining information from (i) whole-genome sequencing of an affected calf, (ii) transcriptomic data from a panel of tissues and (iii) a list of functionally ranked positional candidates pinpointed a private *G* to *A* nucleotide substitution in the *LAMA3* gene that creates a premature stop codon (p.Arg2609*) in exon 60, truncating 22% of the corresponding protein. The *LAMA3* gene encodes the alpha 3 subunit of the heterotrimeric laminin-332, a key constituent of the lamina lucida that is part of the skin basement membrane connecting epidermis and dermis layers. Homozygous loss-of-function mutations in this gene are known to cause severe junctional epidermolysis bullosa in human, mice, horse, sheep and dog. Overall, our data strongly support the causality of the identified gene and mutation.

Keywords *LAMA3*, laminin-332, herlitz-type junctional epidermolysis bullosa, autozygosity mapping, whole-genome sequencing, Belgian Blue cattle breed

The Belgian Blue cattle breed (BBCB) is a reputed highly specialized beef breed. Over the last decade, we have implemented a surveillance platform to closely monitor emerging disorders of suspected genetic origin in BBCB. This national program has been very successful, with a current list of seven recessive diseases elucidated at the molecular level and virtually eradicated (Charlier *et al.* 2008; Fasquelle *et al.* 2009; Sartelet *et al.* 2012a,b, 2014, 2015). This study illustrates how modern genomic tools helped to solve the eighth one.

In 2015, four newborn calves presenting a severe form of epidermolysis bullosa (EB), originating from three different farms, were collected. Inherited EB is a heterogeneous skin fragility disorder characterized by skin blistering and/or erosions upon trauma, with or without associated mucosal defects. In human, four main subtypes are described, commonly classified based on the disturbed layer responsible for its skin fragility: EB simplex, junctional EB, dystrophic EB and Kindler syndrome (reviewed by Has & Bruckner-Tuderman 2014). Up to now, several hundred causative mutations in a total of 18 mammalian genes have been molecularly characterized. Altogether, these genes encode a group of membrane-bound and structural proteins required for epidermal and/or dermal–epidermal adhesion. Clinical expression and severity largely depends on both the causative gene and mutation type (reviewed by Has & Bruckner-Tuderman 2014). In the Belgian Blue cases, the main clinical symptoms, present at birth, were the following: extensive skin blistering predominantly located at pressure points and on limb extremities and articulations, hoof exungulation with erythema and multiple mucosal ulcerations in the oral cavity and on the tongue (Fig. 1). These clinical findings were grossly reminiscent of junctional EB disorders reported in Gir, Charolais and Hereford cattle breeds (Medeiros *et al.* 2012; Michot *et al.* 2015; Murgiano *et al.* 2015; Peters *et al.* 2015).

The four BBCB cases were genotyped on a 50-K SNP array (BovineSNP50 Genotyping BeadChip, Illumina) following standard procedures, and genotypes were analyzed

Address for Correspondence

C. Charlier, Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège (B34), 1 Avenue de l'Hôpital, 4000-Liège, Belgium.
E-mail: carole.charlier@ulg.ac.be

[1]Equal contribution.

Accepted for publication 07 July 2015

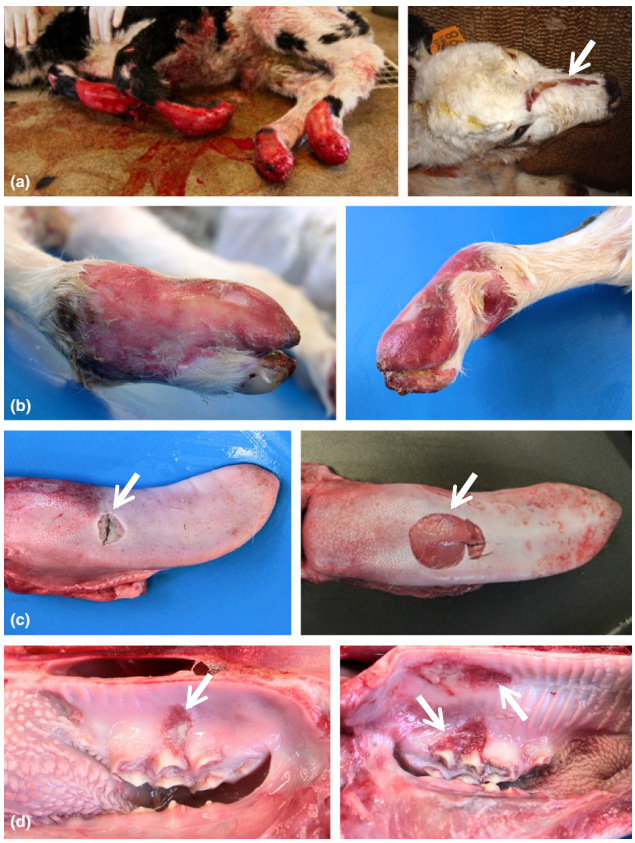


Figure 1 Clinical spectrum of EB cases. (a) Affected calf presenting a severe EB phenotype (left); EB case with a large skin-blistering lesion on its forehead (right, arrow). (b) Hoof exungulation accompanied by pronounced erythema. (c) Large ulceration of the tongue in two cases (arrows). (d) Multiple mucosal ulcerations in the oral cavity (arrows).

together with those of a panel of unrelated controls to search for shared autozygous segments in cases—a hallmark of recessive mode of inheritance—as previously described (e.g. Charlier *et al.* 2008). A unique homozygous 8.3-Mb identical-by-descent (IBD) haplotype on bovine chromosome (BTA) 24 (32 171 355–40 490 628 bp; *Bos taurus* assembly: BosTau6/UMD3.1) was found to be shared by the four cases (Fig. 2a). In cattle, this IBD segment encompassed a list of 36 annotated RefSeq positional candidates. Comparative analysis defined two orthologous regions on human chromosome (HSA) 18 (112 526–7 059 564: 6 947 039 bp and 18 528 880–22 12 730: 3 983 851 bp; *Homo Sapiens assembly*: GRCh37/hg19). A total of 74 RefSeq candidate genes were present in these two human orthologous regions (data not shown).

A PCR-free whole-genome sequencing of one affected case was undertaken and run on one lane of an Illumina HiSeq2000 apparatus to obtain the complete sequence of the shared IBD segment. Paired-sequence reads (2 × 100 bp) were mapped to the UMD3.1 bovine reference genome assembly using BWA, yielding a mean coverage of 12.8-fold (Li & Durbin 2009). Variants were called using the PLATYPUS software package (Rimmer *et al.* 2014). Within the IBD region, a total of 11 207 homozygous variants were identified. The corresponding 8.3-Mb local BAM file has been deposited in the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/data/view/PRJEB9432). After filtering for polymorphisms segregating in other cattle breeds and unrelated Belgian Blue sires, a short list of 39 private variants remained and were annotated with the SNP EFFECT PREDICTOR web-based tool (McLaren *et al.* 2011) (Table S1). Out of these, only two were located in transcribed regions—confirmed by intersecting their positions with available transcriptomic data (RNA-seq) from a panel of bovine tissues including fetal skin (Fig. 2b). It highlighted a single *G* to *A* substitution at position 33 111 473 bp on the forward strand. This substitution created a premature stop codon (CGA>TGA; c.7825C>T;

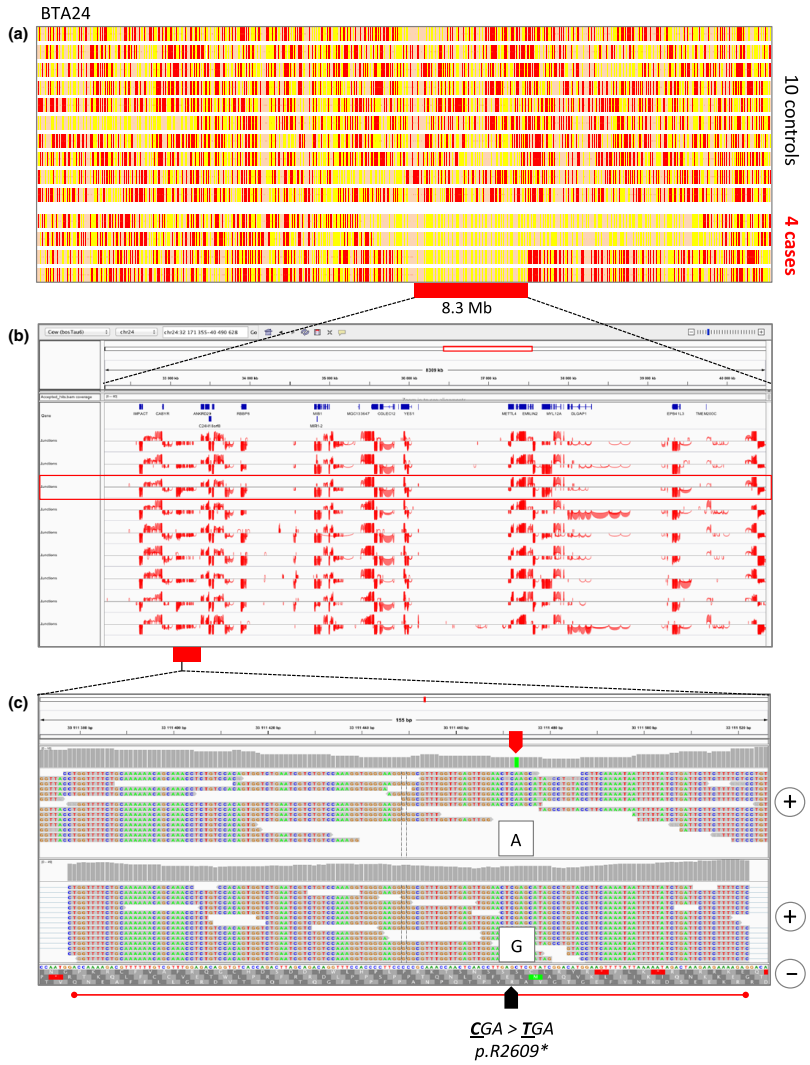


Figure 2 Autozygosity mapping of the locus for epidermolysis bullosa (EB), positional candidate gene delineation by transcriptomic analysis and identification of the causative gene and mutation by WGS of a case. (a) Genotypes of the 1312 BTA24 SNP markers for 10 unrelated controls (top) and four EB cases (bottom). Alternate homozygous genotypes are shown respectively in yellow and orange and heterozygous genotypes in red. The 8.3-Mb homozygous haplotype shared by the four cases is underlined by a red bar. (b) INTEGRATIVE GENOMICS VIEWER (IGV; Robinson *et al.* 2011) screen capture of the 8.3-Mb genomic region with, from top to bottom, the track of RefSeq gene annotation (blue bars) and RNA-seq junctions outputs tracks obtained with TOPHAT (Trapnell *et al.* 2012) corresponding to nine fetal tissues (pituitary, skeletal muscle, skin, cerebral cortex, duodenum, kidney, heart, mammary gland, hypothalamus); the 'skin' track is highlighted by a red rectangle; the position of the *LAMA3* transcript within the candidate region is underlined by a red bar. (c) Screen capture of an IGV output for a 155-bp genomic region encompassing the *LAMA3* 145-bp exon 60 (underlined by a red line) and displaying (i) the genomic sequence reads of an EB homozygous mutant calf (top) and (ii) fetal skin cDNA sequence reads from a wild-type animal (bottom); reads are aligned on the bovine genomic reference sequence and are presented on the forward strand ('+'); the red arrow points toward the homozygous '*A*' mutation in the case of genomic DNA (green) in contrast to the homozygous wild-type '*G*' allele present in the control; the black arrow points toward the position of the mutated arginine (p.Arg2609*) in protein translation on frame 3 displayed on the reverse strand ('-').

p.Arg2609*) in coding exon 60 of the *laminin, alpha 3* (*LAMA3)* gene, which is transcribed from the reverse strand (Fig. 2c and Table S1). The gene spans ~283 kb (BTA24: 33 067 058–33 349 900 bp), encompasses 75 coding exons and is highly expressed in fetal skin. The *LAMA3* gene encodes the alpha 3 subunit of the heterotrimeric laminin-332 (or epiligrin/kalinin/nicein), a structural protein that is a key component of the cytoskeleton in the skin basement membrane. Homozygous loss-of-function (LoF) mutations in this gene cause severe generalized junctional EB (Herlitz-type) in human, mice, horse, sheep and dog (reviewed by Medeiros *et al.* 2012). In cattle, this is the first occurrence of a LoF mutation in the *LAMA3* gene. The p.Arg2609* premature stop codon is predicted to lead to a 22% truncation of the corresponding protein, and the mutant mRNA is very likely targeted by the non-sense-mediated decay pathway.

We developed a TaqMan-based direct diagnostic test to interrogate the c.7825C>T mutation (Appendix S1). The four cases were confirmed homozygous mutant, and a carrier frequency of ~1% was estimated within the BBCB by screening a cohort of 3000 animals. Retrospectively, a haplotype-based analysis was performed on available 50-K SNP array genotypes (for a total of 1202 animals) to identify additional putative carriers in this breed (data not shown). All 11 carriers of the haplotype associated with the disease were subsequently confirmed to be heterozygotes for the causative mutation by direct testing.

Altogether, these data strongly support the causality of the p.Arg2609* mutation in the emerging recessive severe junctional EB disorder in BBCB. Jointly with all previously identified deleterious and functional mutations in BBCB, direct genotyping of this new variant is now routinely applied on a low-density custom array to avoid carrier–carrier matings in breeding schemes. This study confirms the power of state-of-the-art genomic tools to rapidly elucidate the molecular basis of emerging monogenic disorders in livestock.

Acknowledgements

We are grateful to the breeders and practitioners for their collaboration and to the Walloon Breeding Association (AWE) for pedigree data. We also thank all the members of the GIGA-Genomic platform for their technical assistance. CC is senior research associate of the Fonds National de la Recherche Scientifique (FNRS, Belgium). This work was funded by grants from the Walloon Ministry of Agriculture (Rilouke), the Belgian Science Policy Organisation (SSTC Genefunc PAI) and the University of Liège.

Authors' contributions

AS performed case's collection, phenotyping and epidemiological analysis; CH analyzed WG sequencing data (mapping, variant calling, filtering against known variations and variant effect prediction); NT processed case samples, developed and applied the direct genotyping test; LK constructed the PCR-free WGS library; CB was responsible for case's necropsy and pictures; WL produced and analyzed the RNA-seq data; NA genotyped samples on the 50-K SNP array; WC supervised WGS and SNP array genotyping and performed haplotype-based analysis; and MG and CC designed the study, analyzed the data and wrote the manuscript with the help of all co-authors.

References

Charlier C., Coppieters W., Rollin F. *et al.* (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nature Genetics* **40**, 449–54.

Fasquelle C., Sartelet A., Li W. *et al.* (2009) Balancing selection of a frame-shift mutation in the *MRC2* gene accounts for the outbreak of the crooked tail syndrome in Belgian Blue cattle. *PLoS Genetics* **5**, e1000666.

Has C. & Bruckner-Tuderman L. (2014) The genetics of skin fragility. *Annual Review of Genomics and Human Genetics* **15**, 245–68.

Li H. & Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60.

McLaren W., Pritchard B., Rios D., Chen Y., Flicek P. & Cunningham F. (2011) Deriving the consequences of genomic variants with the Ensembl API and SNP EFFECT PREDICTOR. *Bioinformatics* **26**, 2069–70.

Medeiros G.X., Riet-Correa F., Armién A.G., Dantas A.F., de Galiza G.J. & Simões S.V. (2012) Junctional epidermolysis bullosa in a calf. *Journal of Veterinary Diagnostic Investigation* **24**, 231–4.

Medeiros G.X. & Riet-Correa F. (2015) Epidermolysis bullosa in animals: a review. *Veterinary Dermatology* **26**, 3–13.

Michot P., Fantini O., Braque R. *et al.* (2015) Whole-genome sequencing identifies a homozygous deletion encompassing exons 17 to 23 of the *integrin beta 4* gene in a Charolais calf with junctional epidermolysis bullosa. *Genetic Selection and Evolution* **47**, 37.

Murgiano L., Wiedemar N., Jagannathan V., Isling L.K., Drögemüller C. & Agerholm J.S. (2015) Epidermolysis bullosa in Danish Hereford calves is caused by a deletion in *LAMC2* gene. *BMC Veterinary Research* **11**, 23.

Peters M., Reber I., Jagannathan V., Raddatz B., Wohlsein P. & Drögemüller C. (2015) DNA-based diagnosis of rare diseases in veterinary medicine: a 4.4 kb deletion of *ITGB4* is associated with epidermolysis bullosa in Charolais cattle. *BMC Veterinary Research* **11**, 48.

Rimmer A., Phan H., Mathieson I., Iqbal Z., Twigg S.R., WGS500 Consortium, Wilkie A.O., McVean G. & Lunter G. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics* **46**, 912–8.

Robinson J.T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E.S., Getz G. & Mesirov J.P. (2011) Integrative genomics viewer. *Nature Biotechnology* **29**, 24–6.

Sartelet A., Druet T., Michaux C., Fasquelle C., Géron S., Tamma N., Zhang Z., Coppieters W., Georges M. & Charlier C. (2012a) A splice site variant in the bovine *RNF11* gene compromises

growth and regulation of the inflammatory response. *PLoS Genetics* **8**, e1002581.

Sartelet A., Klingbeil P., Franklin C.K., Fasquelle C., Géron S., Isacke C.M., Georges M. & Charlier C. (2012b) Allelic heterogeneity of crooked tail syndrome: result of balancing selection? *Animal Genetics* **43**, 604–7.

Sartelet A., Stauber T., Coppieters W. *et al.* (2014) A missense mutation accelerating the gating of the lysosomal Cl-/H+ -exchanger ClC-7/Ostm1 causes osteopetrosis with gingival hamartomas in cattle. *Disease Models & Mechanisms* **7**, 119–28.

Sartelet A., Li W., Pailhoux E. *et al.* (2015) Genome-wide next-generation DNA and RNA sequencing reveals a mutation that perturbs splicing of the *phosphatidylinositol glycan anchor biosynthesis class H* gene (*PIGH*) and causes arthrogryposis in Belgian Blue cattle. *BMC Genomics* **16**, 316.

Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D.R., Pimentel H., Salzberg S.L., Rinn J.L. & Pachter L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TOPHAT and CUFFLINKS. *Nature Protocols* **7**, 562–78.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 List of private variants located on the 8.3 Mb disease causing haplotype and their predicted effect.

Appendix S1 Materials and methods: direct diagnostic test of the *LAMA3* c.7825C>T mutation.

Table S1: List of private variants located on the 8.3 Mb disease causing haplotype and their predicted effect. The G to A causative variant in *LAMA3* is highlighted in orange.

Position	Non RefSeq allele	Consequence	Symbol	Ensembl Gene
24:32500929-32500930	-	intergenic_variant	-	-
24:32604861-32604861	A	intergenic_variant	-	-
24:32721922-32721922	C	intron_variant	*OSBPL1A*	*ENSBTAG00000023259*
24:32721922-32721922	C	intron_variant	*OSBPL1A*	*ENSBTAG00000023259*
24:33111473-33111473	A	stop_gained	*LAMA3*	*ENSBTAG00000027181*
24:33148767-33148767	G	intron_variant	*LAMA3*	*ENSBTAG00000027181*
24:33347294-33347294	T	intergenic_variant	-	-
24:33465206-33465206	C	intron_variant	*NPC1*	*ENSBTAG00000015195*
24:33688008-33688009	-	3_prime_UTR_variant	*CABLES1*	*ENSBTAG00000008705*
24:33825343-33825343	T	intergenic_variant	-	-
24:33885757-33885757	G	intergenic_variant	-	-
24:34069358-34069358	C	intergenic_variant	-	-
24:34069387-34069388	-	intergenic_variant	-	-
24:34557847-34557848	-	intron_variant	*GATA6*	*ENSBTAG00000005734*
24:34911709-34911710	-	intergenic_variant	-	-
24:35020336-35020336	G	upstream_gene_variant	*ESCO1*	*ENSBTAG00000017597*
24:35081219-35081219	A	intron_variant	*GREB1L*	*ENSBTAG00000008275*
24:35233407-35233408	-	intergenic_variant	-	-
24:35664039-35664039	C	intron_variant	*COLEC12*	*ENSBTAG00000007705*
24:35720088-35720089	-	intron_variant	*COLEC12*	*ENSBTAG00000007705*
24:36039630-36039630	A	upstream_gene_variant	*YES1*	*ENSBTAG00000001523*
24:36138830-36138830	C	intergenic_variant	-	-
24:36149647-36149647	G	intergenic_variant	-	-
24:36465248-36465248	C	intergenic_variant	-	-
24:36595321-36595322	-	intergenic_variant	-	-
24:36894948-36894949	-	intergenic_variant	-	-
24:36915373-36915373	T	intergenic_variant	-	-
24:37124770-37124770	T	intergenic_variant	-	-
24:37125558-37125558	T	intergenic_variant	-	-
24:37140883-37140897	TGACCCAGATAATCC	intergenic_variant	-	-
24:37141206-37141206	T	intergenic_variant	-	-
24:37141566-37141566	G	intergenic_variant	-	-
24:37145935-37145935	G	intergenic_variant	-	-
24:37146355-37146355	C	intergenic_variant	-	-
24:37147853-37147853	A	intergenic_variant	-	-
24:37407960-37407961	-	intron_variant	*SMCHD1*	*ENSBTAG00000003354*
24:37503295-37503295	T	intron_variant	*EMILIN2*	*ENSBTAG00000003880*
24:37612142-37612143	-	intergenic_variant	-	-
24:38602887-38602888	-	intergenic_variant	-	-

Appendix S1 Materials and methods: Direct diagnostic test of the *LAMA3* c.7825C>T mutation.

A 5' exonuclease assay was developed to genotype the *LAMA3* c.7825C>T mutation, using 5'- TGA TTG GTG TTG ACA GGA GAA AAG A -3' and 5'- GTC TGA ATC GTC TGT CCA AAG GT - 3' as PCR primers, and 5'- AGT TGG AAC TC[**G**] AGC ATA G - 3' (wild-type G allele) and 5' - AGT TGG AAC TC[**A**] AGC ATA G - 3' (mutant A allele) as probes (Taqman, Applied Biosystems, Fosters City, CA). Reactions were carried out on an ABI7900HT instrument (Applied Biosystems, Fosters City, CA) using standard procedures.

Experimental Section

Study 5:

Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle

Nature Comms 5: 5861, 2014

Mathew D. Littlejohn, Kristen M. Henty, Kathryn Tiplady, Thomas Johnson, Chad Harland, Thomas Lopdell, Richard G. Sherlock, Wanbo Li, Steven D. Lukefahr, Bruce C. Shanks, Dorian J. Garrick, Russell G. Snell, Richard J. Spelman & Stephen R. Davis

Abstract

Lactation, hair development and homeothermy are characteristic evolutionary features that define mammals from other vertebrate species. Here we describe the discovery of two autosomal dominant mutations with antagonistic, pleiotropic effects on all three of these biological processes, mediated through the prolactin signalling pathway. Most conspicuously, mutations in prolactin (*PRL)* and its receptor *(PRLR)* have an impact on thermoregulation and hair morphology phenotypes, giving prominence to this pathway outside of its classical roles in lactation.

nature COMMUNICATIONS

ARTICLE

Received 13 Aug 2014 | Accepted 13 Nov 2014 | Published 18 Dec 2014

DOI: 10.1038/ncomms6861 OPEN

Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle

Mathew D. Littlejohn[1,2,*], Kristen M. Henty[2,*], Kathryn Tiplady[1], Thomas Johnson[1], Chad Harland[1], Thomas Lopdell[1], Richard G. Sherlock[1], Wanbo Li[3], Steven D. Lukefahr[4], Bruce C. Shanks[5], Dorian J. Garrick[6], Russell G. Snell[2], Richard J. Spelman[1] & Stephen R. Davis[1]

Lactation, hair development and homeothermy are characteristic evolutionary features that define mammals from other vertebrate species. Here we describe the discovery of two autosomal dominant mutations with antagonistic, pleiotropic effects on all three of these biological processes, mediated through the prolactin signalling pathway. Most conspicuously, mutations in prolactin *(PRL)* and its receptor *(PRLR)* have an impact on thermoregulation and hair morphology phenotypes, giving prominence to this pathway outside of its classical roles in lactation.

[1] Livestock Improvement Corporation, Cnr Ruakura and Morrinsville Roads, Newstead, Hamilton 3240, New Zealand. [2] School of Biological Sciences, University of Auckland, 3A Symonds Street, Auckland 1010, New Zealand. [3] Unit of Animal Genomics, GIGA-R, Faculty of Veterinary Medicine, University of Liège (B34), Liège B-4000, Belgium. [4] Department of Animal, Rangeland and Wildlife Sciences, Texas A&M University-Kingsville, MSC 228, Kingsville, Texas 78363-8202, USA. [5] Department of Agriculture and Environmental Science, Lincoln University, 820 Chestnut Street, Jefferson City, Missouri 65101, USA. [6] Department of Animal Science, Iowa State University, 225 Kildee, Ames, Iowa 50011-3250, USA. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to M.D.L. (email: mlittlejohn@lic.co.nz) or to S.R.D. (email: steve.davis@lic.co.nz).

Hallmark characteristics of mammals include the secretion of milk, the development of body hair and the homeothermic regulation of body temperature. These latter two processes have clear physiological links, with one of the key functions of body hair being to insulate the endothermic animal. Lactation also shares some common biology with these processes, where similarities in the structure and function of mammary, sweat and sebaceous glands has led to the hypothesis that mammary glands evolved from a pilosebaceous apocrine structure in the skin[1]. The literature describing the cellular and molecular physiology of each of these phenomena is vast, and in the case of mammary and hair follicle biology, these processes are known to be broadly regulated by a range of hormones including oestrogen[2–4], testosterone[4–6], growth hormone[7,8], prolactin[9,10] and others[11].

In 2011 we identified a spontaneous, dominant genetic syndrome in *Bos taurus*, presenting as a collection of unusual phenotypes including lactation failure, excessively 'hairy' pelage and thermoregulatory dysfunction. We report mapping of the mutation for this syndrome, and further report identification of a novel, phenotypically reciprocal mutation in the same molecular pathway, defining the slick-coated, thermotolerant characteristics of the Senepol breed of cattle.

Results

A novel pleiotropic syndrome in dairy cattle. As part of routine animal screening in a large dairy cattle-breeding programme in New Zealand, we identified a dominant genetic syndrome that had been rapidly propagated through the population through widespread use of semen representing a bull and his son. Animals within the pedigree ($N>6{,}000$) segregated for abnormally long and 'hairy' coats, symptoms of heat stress including increased respiration rates and the tendency to wallow in mud and drinking troughs, and major defects in lactation. Detailed examination of 12 affected and 12 control females showed that hair was more slender (two-sided *t*-test, $P=1.4\times10^{-4}$) and approximately twice as long in affected animals (two-sided *t*-test, $P=1.4\times10^{-7}$; Fig. 1a,b; Supplementary Fig. 1). There was also an increase in hair mass per cm^2 of skin area in affected animals (two-sided *t*-test, $P=0.012$; Fig. 1b), although this effect was not apparent when adjusted for hair diameter and length (two-sided *t*-test, $P=0.226$), suggesting a similar density of hair of increased fibre weight.

At an ambient temperature of 22 °C (thermoneutral for *Bos taurus*), rectal temperatures were elevated in affected animals compared with controls (two-sided *t*-test, $P=1.3\times10^{-8}$; Fig. 1c). Heart rates were not significantly different between groups (two-sided *t*-test, $P=0.149$; Supplementary Fig. 2); however, respiration rates were approximately four times greater in affected individuals (two-sided *t*-test, $P=2.6\times10^{-14}$; Fig. 1d). These effects were reproducible over multiple time points and days (Supplementary Table 1 and Supplementary Fig. 2). Since this heat stress response could have been partly attributable to increased hair length, five of twelve affected

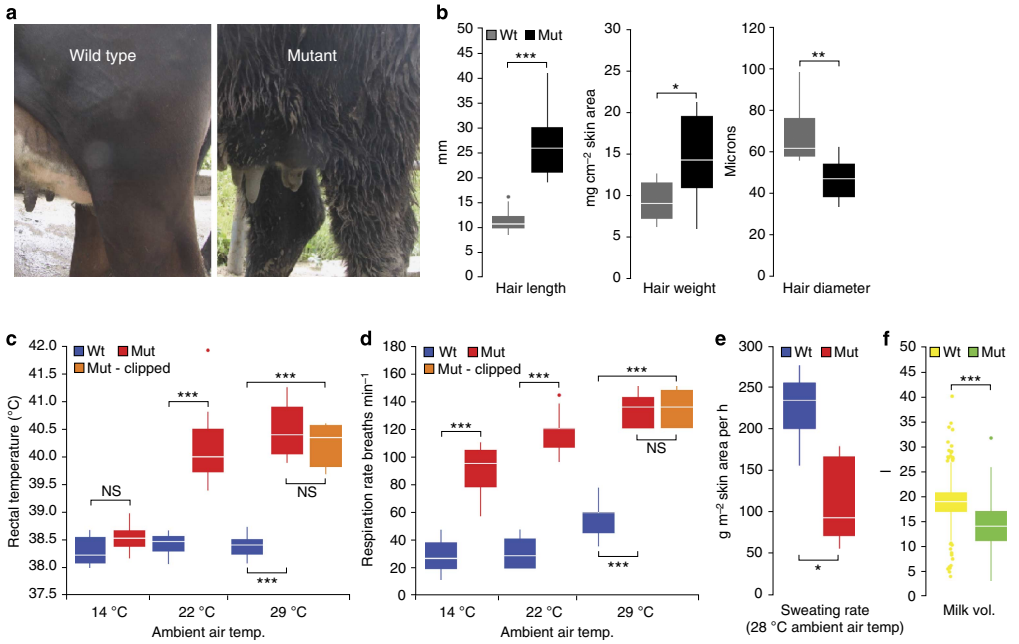


Figure 1 | Phenotypic characteristics of hairy syndrome cattle. (**a**) Photograph showing coat differences between wild-type and mutant half-sibs, with muddy coat due to wallowing behaviour typical of affected animals. (**b**) Hair morphology differences between mutant ($N=12$) and wild-type ($N=12$) half-sibs. (**c**,**d**) Heat stress response phenotypes of mutant ($N=12$) and wild-type ($N=12$) half-sibs measured at different ambient temperatures. Responses of twelve wild-type, seven mutant and five clipped mutants also indicated. (**e**) Sweating rate contrast between mutant ($N=6$) and wild-type ($N=6$) cows. (**f**) Differences in milk volumes between wild-type ($N=740$) and mutant ($N=77$) half-sibs. These differences underestimate the extent of lactation effects since at least 25% of mutant animals failed to initiate lactation. Box plots define the median, upper and lower quartiles for the various phenotypes, with whiskers representing the furthest data points within 1.5 × of the interquartile range, and outlier samples indicated beyond this range. $^{*}P<0.05$, $^{**}P<0.001$, $^{***}P<0.0001$ (two-sided *t*-tests, Bonferroni-adjusted).

animals were clipped to approximate the coat lengths of controls (Supplementary Fig. 1). This had no effect on body temperatures or respiration rates (Fig. 1c,d). Since sweating and panting are the primary modes of active heat loss in cattle, we next assessed the sweating rates of six affected and six control animals. At an ambient indoor temperature of 28 °C, control cattle produced twice the weight of sweat compared with hairy animals (two-sided *t*-test, $P=0.001$; Fig. 1e), implicating sweat gland dysfunction as the likely source of thermoregulatory failure.

Affected females also failed to lactate or produced markedly less milk (two-sided *t*-test, $P=3.7\times10^{-21}$; $N=817$; Fig. 1f). Although heat stress may have contributed to these effects, they appeared to be a primary feature of the syndrome, since >95% of lactation records from affected animals were measured during spring at cool to moderate temperatures (September to November 2013; mean daily temperature <16 °C for all geographic regions). Further, the influence on milk yield was similar when comparing farms between the North Island and South Island (Supplementary Table 1), where the mean daily temperature from September to November was 14.1 and 11.6 °C for each island, respectively.

Mapping the 'hairy' mutation. To identify the '*hairy*' locus and mutation, we undertook genome-wide transmission disequilibrium testing using 628,278 single-nucleotide polymorphisms (SNPs) in 22 nuclear trios and 55 half-sib offspring of the two founder sires. This analysis revealed a single significant locus on chromosome 23 (sib-transmission/disequilibrium test, $P=1.7\times10^{-12}$; Fig. 2a), with the most highly associated SNP (rs110103404) mapping within 0.5 Mbp of the *MIR2284C*, *HDGFL1* and *PRL* genes. Given the key roles of prolactin signalling in mammary differentiation[12], and hair follicle growth and cycling[13], we considered *PRL* as a candidate gene at the *hairy* locus. Sanger sequencing of *PRL* in both sires revealed a single candidate mutation that was not present in the National Center for Biotechnology Information (NCBI) database for short genetic variations (dbSNP), or our own whole-genome sequence database of 554 contemporary animals (ss1067289409; chr23:35105313A>C; Fig. 2b). This nonsynonymous SNP in exon 5 encodes a p.Cys221Gly substitution highly conserved across vertebrates and other structurally related hormones, disrupting one of three disulphide bonds defining the three-dimensional (3D) structure of mature prolactin hormone (Fig. 2c,d). To assess the candidacy of other mutations at this locus, we then conducted genome sequencing of the two founder sires. Filtering all previously unobserved variants assuming a dominant, heterozygous genetic model yielded only seven variants chromosome-wide, only one of which mapped to exonic sequence, being the same *PRL* mutation discovered using our candidate-led approach (Supplementary Table 2). The p.Cys221Gly variant was then genotyped in 2,205 progeny of the

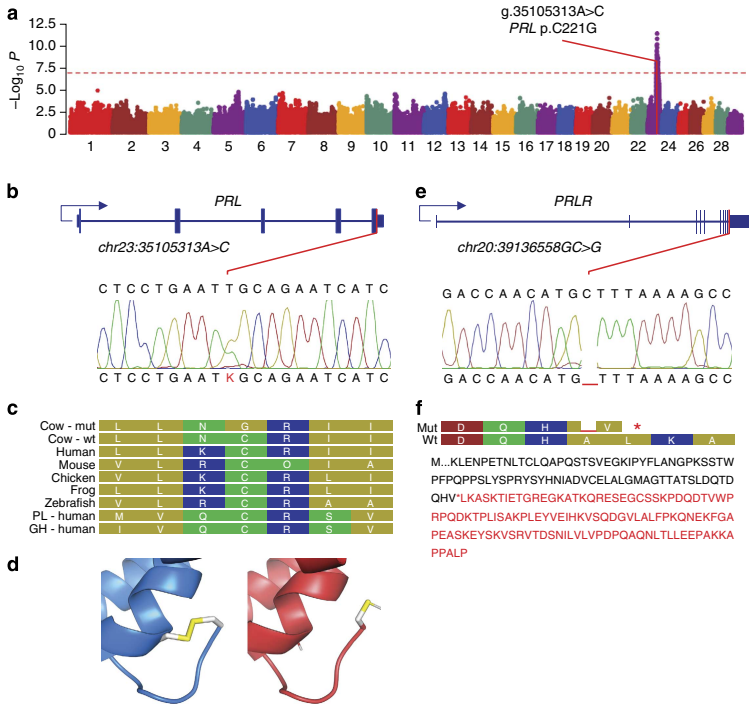


Figure 2 | Mapping and bioinformatic characterization of *PRL* and *PRLR* mutations. (**a**) Manhattan plot showing the *hairy* locus on chromosome 23, with significance plotted on the *y* axis, and chromosome number and position indicated on the *x* axis. (**b**,**e**) Graphics depicting *PRL* and *PRLR* gene structures, showing locations of the respective p.Cys221Gly and *PRLR* p.Leu462* mutations and representative Sanger sequence traces. (**c**) ClustalW alignment showing conservation of the prolactin Cys221 residue in five vertebrates, and in human placental lactogen and growth hormone (residues coloured by polarity). (**d**) Disruption of the C-terminal disulphide bridge because of p.Cys221Gly, modelled on the 3D structure of human prolactin (1RW5.pdb). (**f**) 200 C-terminal amino acids of *PRLR*, with truncated residues because of the p.Leu462* mutation indicated in red.

two sires, demonstrating complete concordance between affected ($N=1{,}045$) and unaffected ($N=1{,}160$) individuals.

A candidate pathway for thermoregulatory mutations in other cattle. With genetic data from the hairy pedigree strongly supporting the causative status of the *PRL* p.Cys221Gly variant, we next contemplated whether coat conformation and heat tolerance in other cattle might be influenced by other mutations in prolactin signalling pathways. The individual coat types of domesticated bovine breeds vary widely, with yak breeds (*Bos grunniens*) selected for hair length and cold tolerance, and short-haired cattle such as zebu (*Bos indicus*) selected for hot, tropical environments. Most *Bos taurus* breeds are temperate-adapted; however, Senepol is one of a small number of breeds that is heat-tolerant, ostensibly due to their unusually short, 'slick' coats (Fig. 3). This trait is thought to be determined by a single, dominant mutation[14], with the '*slick*' locus spanning a region on chromosome 20 (refs 15–17) that includes the prolactin receptor (*PRLR*). We thus considered *PRLR* as a positional candidate gene for the slick coat phenotype, and sequenced *PRLR* in a purebred Senepol sire. We identified a single homozygous frameshift mutation not present in dbSNP or our sequence database, consisting of a single base deletion in exon 10 that introduces a premature stop codon (p.Leu462*) and loss of 120 C-terminal amino acids from the long isoform of the receptor (ss1067289408; chr20:39136558GC>G; Fig. 2e,f).

Association analysis at the *slick* locus. We next typed the *PRLR* p.Leu462* mutation in four purebred Senepol sires whose progenies were known to segregate for slick coat type, with the mutation confirmed as heterozygous in these animals. We then genotyped a collection of 82 highly crossbred cattle containing 0.5–0.0625 Senepol ancestry. Coat length was scored on a quantitative scale (where 1 = slick, 4 = long), since polygenic background effects in crossbreeds can result in slight increases in hair length over that seen in purebred Senepol animals[14]. The mutation was highly associated with coat length in these animals (genotypic test assuming dominance, $P=7.3\times10^{-20}$), and when considered as a binary trait comprising 42 cases and 40 controls (1 = slick, >1 = not slick), the mutation segregated in 79 of 82 individuals (genotypic test assuming dominance, $P=4.7\times10^{-17}$). This proportion of nonsegregating animals was similar to that reported for slick-coat phenotype transmission rates in other crossbreeds[14], and for the two nonslick animals that carried the *PRLR* p.Leu462* mutation, both had quantitative scores of '2' (Supplementary Table 3), supporting a hypothesis of phenotype ambiguity or misassignment in these animals. Haplotype-based analysis was then conducted using 25 Illumina SNP50 BeadChip SNPs in a 1-Mbp consensus *slick* interval reported in independent analyses of Senepol[16] and Senepol crossbreeds[17]. This analysis revealed maximum significance for a 229-kb haplotype block bearing the p.Leu462* mutation (two-sided *t*-test assuming dominance, $P=2.4\times10^{-19}$), with the corresponding ancestral-allele haplotype unassociated with coat length (Supplementary Tables 4 and 5). Notably, haplotypes of the third nonsegregating animal did not share an obvious lineage with the 229-kb contiguous block found in all other slick-coded animals, making the existence of an alternative, hidden causative mutation shared by all slick-coded animals unlikely. These data suggested that the *PRLR* p.Leu462* mutation, or some other, unknown variant carried by the same haplotype was responsible for the slick-coat phenotype.

Exome sequence analysis. To look for alternative mutations at the *slick* locus, we next obtained exome sequence data from 115 animals representing Senepol, Angus, Belgian Blue, Brahman, Charolais, Holstein Friesian, Jersey, Nelore, Simmental and Yak breeds. Restricting analysis to the 1-Mbp *slick* interval used for haplotype testing, and filtering to nonreference variants that were present in all Senepol, but absent in all other breeds yielded only the *PRLR* p.Leu462* variant. Our exome sequence panel included *Bos indicus* breeds that are also short-coated and heat tolerant (Brahman and Nelore). Although the short coat of indicus cattle is not reported as a segregating trait, it is conceivable that Senepol

Figure 3 | 'Slick' coat type. Photographs contrasting slick and nonslick Senepol crossbreeds. The animal pictured on the left (**a**) carries the *PRLR* p.Leu462* mutation and is a three-way cross of Tuli (0.5), Senepol (0.25) and Red Angus (0.25); the animal on the right (**b**) is wild-type and contains Senepol (0.375), Red Angus (0.25), Beefmaster (0.1875) and Simmental (0.1875) ancestry. Pictured animals are representative of the crossbreeds used for genetic analysis of the *slick* locus, representing coat scores of 1 and 4, respectively.

coat type was derived from this species, given the recent proposal that Senepol contains minor proportions of indicus ancestry[16]. As an alternative analysis, we pooled Senepol, Nelore and Brahman animals, and filtered to all nonreference variants that were shared by these breeds, but were absent in all others. This yielded no variants in the 1-Mbp interval of interest, suggesting that Senepol coat type did not arise through introgression of fixed *Bos indicus* alleles, and further supporting the *PRLR* p.Leu462* variant as the only plausible causative mutation.

Histological and molecular characteristics of hairy and slick cattle. Histological analyses of ear tissue biopsies were conducted using 12 wild-type, 11 *PRL* mutant and three *PRLR* mutant animals to further investigate the cutaneous phenotypes of hairy and slick cattle. Although the number of samples representing *PRLR* p.Leu462* carriers precluded formal statistical analysis, there appeared to be no differences in the size, shape and density of sweat glands or hair follicles compared with wild-type animals (Fig. 4). Notably, the sweat glands of *PRL* p.Cys221Gly mutants were indistinguishable from wild-type cows, despite the dysfunctional sweating exhibited by these animals. Likewise, there were no other qualitative or quantitative anatomical differences between wild-type and hairy syndrome skin sections (Fig. 4; Supplementary Table 6). This included hair follicle density, a result consistent with analysis of length and diameter-adjusted hair weight data.

To investigate the molecular mechanism of prolactin dysfunction in hairy syndrome animals, we obtained pituitary samples representing four *PRL* mutant and four unrelated controls. Sequencing of pituitary RNA showed no difference in the expression of *PRL*, with western blotting of pituitary extracts also indicating comparable levels of prolactin peptide between groups (Supplementary Fig. 3). Serum enzyme-linked immunosorbent assay (ELISA) data from mutant ($N=6$) and wild-type ($N=6$) animals also indicated comparable prolactin-secretory responses when infused with thyrotropin-releasing hormone (Fig. 5). Together, these data suggested that mutant prolactin transcripts and hormone are expressed in the pituitary gland, and are actively secreted into circulation.

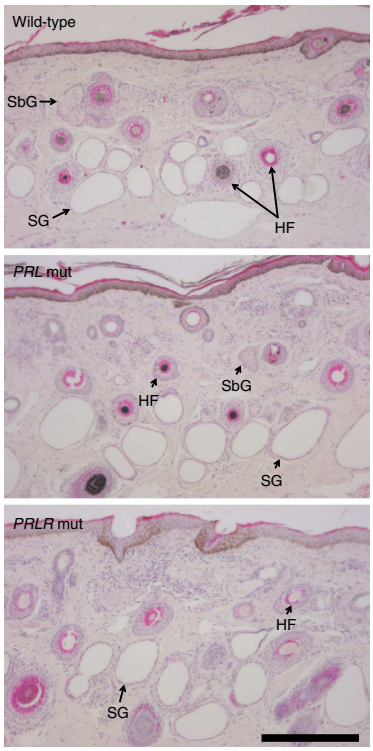


Figure 4 | Skin histology of hairy and slick cattle. Example haemotoxylin/eosin-stained skin sections at 100 × magnification representing wild-type ($N=11$), hairy ($N=12$) and slick ($N=3$) cows. The epidermis is top of field in each panel, sweat glands (SG), sebaceous glands (SbG) and hair follicles with and without fibre cross-sections (HF) are indicated. No qualitative or quantitative differences were observed between the different genotypes. Scale bar: 300 μm.

Discussion

The complementarity of phenotypes, genetic association data and predicted functional impact of the *PRL* and *PRLR* mutations strongly suggests these as causal in determining the characteristics of hairy and slick cattle. Associations between circulating prolactin and thermal stress have been observed in various mammals including humans[18], although a direct modulatory role for prolactin in thermoregulation has remained unproven. Our findings confirm such a role, presenting two bovine models to further explore these responses. Remarkably, these effects appear to occur through control of sweat secretion, with histological similarities between wild-type and *PRL* mutant cows indicating an acute signalling role for prolactin, as opposed to one affecting sweat gland morphogenesis.

These observations suggest that the *PRLR* p.Leu462* mutation may confer additional thermotolerance to cattle beyond its effects on short coat length. Two studies of Senepol–Holstein crossbreeds suggest that slick cattle sweat at higher rates than nonslick controls[19,20]. These studies present conflicting data regarding the mechanism of increased sweating rate, proposing this as a secondary effect related to coat length[19], and alternatively as a consequence of genuinely higher secretory capacity[20]. The precise role of the *PRLR* mutation in sweating rate remains to be resolved; however, observations in hairy syndrome cattle would

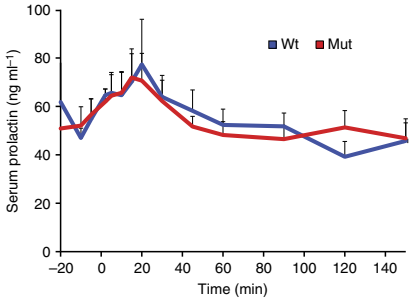


Figure 5 | Prolactin secretory responses to TRH infusion. Serum ELISA results showing mean prolactin secretory responses to TRH challenge in *PRL* p.Cys221Gly mutant ($N=6$) and wild-type ($N=6$) animals. The *x* axis denotes time relative to TRH infusion (time = 0), only positive values for error bars (s.e.m.) are plotted. Peak serum prolactin response was not significantly different between groups (two-sided *t*-test, $P=0.96$).

suggest direct secretory control. The reciprocity of *PRL* and *PRLR* mutations on coat length (and possibly sweating), and the observation of severe lactation dysfunction in hairy syndrome animals also suggests a role for the *PRLR* p.Leu462* mutation on lactation phenotypes. This seems especially likely, given the milk fat and protein yield effects attributed to a p.Ser18Asn substitution in bovine *PRLR*[21]. In studies where the *slick* haplotype has been introgressed into Holstein dairy cattle, slick-haired animals demonstrate higher milk yields than nonslick contemporaries[14,20]. These effects are assumed to be due to enhanced thermotolerance, with one study presenting winter milk yield data for which there was no apparent difference between slick and nonslick cows[20]. It should be noted however that the number of slick animals in that study was small ($N=11$), leaving the role of the *PRLR* p.Leu462* variant in lactation an open-ended question.

The molecular mechanisms by which the *PRL* and *PRLR* mutations could exert their effects remain unclear. Mutant and wild-type *PRL* transcripts are equivalently expressed in the pituitary gland, and the level of prolactin hormone is also similar between groups. Stimulated release of prolactin also appears comparable between hairy syndrome animals and controls. Although the relevance of these data to extrapituitary sites of prolactin synthesis is unknown, these findings suggest a receptor-binding-based mechanism underpinning the hairy syndrome, possibly involving receptor antagonism, given the haplosufficiency exhibited by *Prl* knockout mice[12]. The dominance of the *PRLR* mutation is also curious, since truncation of 120 C-terminal amino acids could be expected to result in a loss of function. Prolactin receptor knockout mice exhibit marginally larger diameter hair, although other phenotypes are reminiscent of hairy syndrome cattle, namely longer hair fibres, and failure to lactate[9,13]. These observations suggest enhanced prolactin pathway signalling as a result of the *PRLR* p.Leu462* mutation. An example of a functionally coupled, C-terminal *PRLR* mutant has recently been described in chickens, where, notably, this variant has been proposed as the causative mutation underlying a dominantly inherited feather-growth retardation phenotype[22].

Irrespective of molecular mechanism, discovery of the *PRL* p.Cys221Gly and *PRLR* p.Leu462* mutations highlights new facets of prolactin biology, expanding the already-extensive repertoire of exocrine functions attributed to this hormone. The impact of the *PRLR* p.Leu462* mutation on thermotolerance carries additional industrial significance, and represents one of few dominant, beneficial alleles reported in livestock. This is of particular interest in dairy farming contexts, where most selection has occurred in heat-intolerant *Bos taurus* breeds. As a frameshift mutation, its amenability to gene editing will allow relatively simple assessment within diverse genetic backgrounds, potentially unlocking hot farming environments to the highest performance genetic lines.

Methods

Primary data. Genotype, phenotype and sequence data sets representing all experimental populations have been deposited in the Dryad digital data repository (http://doi.org/10.5061/dryad.nh6v4[23]), and NCBI SRA (SRP043521). Semen representing *PRL* p.Cys221Gly heterozygous animals may also be available for research purposes on request.

Ethics statement. Ethics approval for all animal experiments was granted by the Ruakura Animal Ethics Committee, Hamilton, New Zealand, under approvals 13134 (heat-stress measurements), 13198 (thyrotropin-releasing hormone (TRH) infusion experiment) and 13346 (sweating analysis).

Animal populations. Individuals used for genetic analyses of the hairy syndrome comprised a two-generation pedigree of 2,274 animals of predominantly Holstein–Friesian ancestry. This pedigree consisted of two large sire families representing the presumed *de novo* sire (67 progenies), and one of his affected male offspring (2,185 progenies). In addition, included were 21 dams forming 22 nuclear trios used for genome-wide analysis. Individuals targeted for *PRLR* genotyping and coat-length analysis consisted of four purebred Senepol sires, three Senepol × Charolais F1 sires, 41 crossbred animals of mixed Senepol, Barzona, Red Angus and Hereford ancestry, and 38 crossbred animals of predominantly Senepol, Red Angus and Tuli ancestry. Three Senepol × Holstein–Friesian animals were assessed for skin histological analysis. Genetic mapping was conducted retrospectively with sample sizes representing all animals for which phenotypic data were available. For prospective analyses (that is, heat stress, TRH infusion and sweating measurements), power calculations were conducted to restrict sample sizes to a minimum based on ethical approvals.

Phenotypic analysis. Phenotypic classification across hairy and slick cohorts was made visually, the former coded as a binary trait and the latter on a quantitative coat-length scale scored 1–4. Slick cohort quantitative scores were also re-classified for binomial analysis into slick (score 1) and not slick (scores 2–4) classes. There were 37 cases and 62 controls representing 22 nuclear trios and 55 half-sibs used for mapping of the *hairy* locus. For *slick* analysis, the distribution of coat lengths is indicated in Supplementary Table 3.

Cows representing the hairy pedigree were distributed across various North Island and South Island commercial farms, with lactation data for affected ($N=111$) and unaffected half-sibs ($N=760$) extracted from a national database of milk yield and composition test results. Milk yield data were absent for 30% of affected animals compared with 3% of controls, largely due to failure of these animals to initiate lactation. Cows were tested at ~60 days of lactation, with >90% of all records measured during spring (September to November) in 2013. The mean daily temperatures reported in text represent the 3-month average for September to November 2013, with source temperature data obtained from the National Institute of Water and Atmospheric Research (https://www.niwa.co.nz/).

For quantitative assessment of hair phenotypes, a 100-cm^2 area of skin was clipped in the left dorsolumbar region of 12 hairy and 12 control cows (Supplementary Fig. 1). Collected hair was weighed with a subsample photographed on a glass microscope slide. Images were analysed using ImageJ (http://imagej.nih.gov/ij/), with randomly selected hairs measured for diameter ($N=20$) and length ($N=10$).

Measurements of rectal temperature, respiration rate and heart rate were made on the same 24 cows used for hair morphological analysis, assessed outdoors without shade in the morning (9 a.m.) and afternoon (3 p.m.) on two consecutive days. Rectal temperature was measured using a clinical thermometer. Respiration and heart rate were assessed over a 30-s period. Five of the twelve hairy cows were subsequently clipped with grooming shears to a coat-length-matching control cows (Supplementary Fig. 1). Respiration rate and rectal temperature were then measured 5 days later to allow for re-acclimation.

Measurement of sweating rates was conducted indoors in a heated room maintained at 28 °C. Six affected and six age-matched control animals (two unaffected half-sibs and four unrelated animals) were used for analysis, assessed in batches of four animals per measurement period. Sweating rates (g m^{-2} skin area per h) were measured by the $CaCl_2$ capsule method[24], using inverted 82-mm diameter Petri dishes filled with 50 g anhydrous $CaCl_2$ (Sigma-Aldrich), separated from the skin by a gauze membrane. Animals were introduced to the hot room 1 h before sweat measurement, with capsule weight change measured over the following hour. The mean sweating rate across two clipped skin areas per animal was quantified, measured at the fore flank posterior the right shoulder and the right dorsolumbar region. Respiration rates and rectal temperatures were also measured at the end of the heat exposure period (Supplementary Fig. 2).

Sanger sequencing and custom genotyping. Semen, hair or ear punch tissue samples were used for DNA extraction following standard protocols, with samples processed by GeneMark (Hamilton, New Zealand) or GeneSeek (Lincoln, NE, USA). For Sanger sequencing of *PRL*, primers were designed to amplify all exons, intron–exon boundaries, and 2 kb of 5′ non-coding sequence according to the RefSeq transcript NM_173953 (Supplementary Table 7). For *PRLR* sequencing, all exons, intron–exon boundaries and 3 kb of 5′ non-coding sequence were amplified according to annotations derived from mammary RNA-sequence data (not shown), targeting an additional 5′ untranslated region (UTR) exon and 9 kb of additional 3′ UTR sequence relative to the RefSeq gene structure NM_001039726 (Supplementary Table 7). Amplicons were sequenced using BigDye version 3.1 chemistry on a 3130xL instrument (Applied Biosystems) at the University of Auckland DNA Sequencing Facility (Auckland, New Zealand). Custom genotyping of the chr23:35105313A>C *PRL* SNP was performed by GeneMark using a TaqMan assay (Applied Biosystems). Genotyping of the chr20:39136558GC>G *PRLR* variant was conducted by GeneSeek using Sequenom iPLEX (Sequenom), targeting alleles in both forward and reverse strand orientations.

High-throughput genotyping and imputation. For genome-wide analysis within the hairy pedigree, 74 animals were genotyped using the Illumina BovineSNP50 BeadChip (Illumina), and 24 using the GeneSeek Genomic Profiler BeadChip (Super GGP; GeneSeek/Illumina). These data were used to impute a total of 712,123 SNPs from the Illumina BovineHD BeadChip using Beagle software[25] (v4),

from a reference population of 3,222 animals. Senepol and Senepol crossbred animals were typed using the BovineSNP50 BeadChip. All variant positions reference the UMD3.1 *Bos taurus* genome assembly.

Association analysis. Milk yield phenotypes were derived from linear models fitted to a wider data set that included all herd contemporaries. Residuals from these models, which included milk yield as the dependent variable, and independent variables for herd, stage of lactation, age at calving, breed and heterosis were used for association testing based on two-sided *t*-tests. Cows were also stratified for analysis based on their geographical location (North or South Island). Two-sided *t*-tests were also used to evaluate associations with heat tolerance traits, hair-related traits and histological and molecular phenotypes.

All genotype data were filtered to exclude markers for minor allele frequency (<1%), and per-individual genotype call rate (<90%). Family-based genome-wide association testing in the hairy pedigree was conducted using the DFAM procedure in PLINK[26], combining both full and incompletely genotyped trios in a single TDT-based analysis. For haplotype analysis of the slick locus, the Beagle software was used to phase the *PRLR* p.Leu462* mutation together with 25 Illumina SNP50 BeadChip SNPs representing the chromosome 20 38.6–39.6 Mbp target interval. Six-marker sliding window haplotypes were used to span the interval, incorporating a three-marker overlap per tile (50% redundancy). Individual haplotypes (minimum $N=5$) were tested for association with coat length using two-sided *t*-tests, assuming a dominance model. For association analysis of the *PRLR* genotype with coat length, dominant genetic models were assessed using genotypic tests in PLINK. An alpha level of 0.05 was used for all tests, incorporating Bonferroni corrections for multiple hypothesis testing within each experiment. Associations were considered significant at $P<0.016$ for lactation, hair morphology, sweating rate and histological analyses (three tests each), and $P<0.002$ for heat stress phenotypes (24 tests). Associations from DFAM analysis were considered significant at $P<7.96\times10^{-8}$ (628,278 tests), haplotype tests were considered significant at $P<7.14\times10^{-4}$ (70 tests). Unadjusted *P* values are reported in text.

Genome and exome sequencing. Whole-genome 100-bp paired-end sequencing was performed by Illumina FastTrack using the Illumina HiSeq 2000 instrument. Sequenced animals included 135 Holstein–Friesians, 102 Jerseys, 316 Holstein–Friesian × Jersey crossbreeds and 1 Ayrshire. The mean mapped read-depth was 11× per animal with a median mapped depth of 7×. The sequence database included the unaffected father (6× mapped depth) of the presumed *de novo* sire, with the two *PRL* mutant sires sequenced to a mapped depth of 8×. Exome sequencing was conducted on 115 animals representing Holstein–Friesian ($N=10$), Jersey ($N=10$), Angus ($N=9$), Belgian Blue ($N=29$), Brahman ($N=10$), Charolais ($N=10$), Nelore ($N=10$), Senepol ($N=9$), Simmental ($N=10$) and Yak ($N=8$) breeds. Custom capture targeting all RefSeq, Ensembl and human paralogous genes was performed using the SureSelect Target Enrichment System (Agilent), with 101-bp paired-end sequencing conducted on the HiSeq 2000. The mean sequencing depth across exome targets was 25–40× per sample.

Sequence informatics and variant filtering. Sequence reads were aligned to the *Bos taurus* UMD3.1 genome assembly using RTG map[27] (v2.7.2) for genome sequence, and BWA aln[28] (v0.6.2) for exome data. Variant calling was performed using RTG Population[27] caller and GATK HaplotypeCaller[29] (v2.8) on genome and exome alignments, respectively. Variant lists were filtered based on affection status criteria and functional predictions according to gene structures from Ensembl gene build 74.

Histological analysis. Skin samples were obtained by biopsy using a 3.5-mm needle ear-punch (Allflex, Brisbane, Australia). Eleven hairy syndrome and twelve wild-type animals were sampled for analysis, with 18 of these overlapping with the animals used for hair morphological analysis, and the remainder representing unrelated wild-type animals. Three Senepol x Holstein–Friesian crossbreeds heterozygous for the *PRLR* p.Leu462* mutation were also sampled for analysis. Tissue samples were fixed for 24 h in 10% neutral-buffered formalin, dehydrated and embedded in paraffin wax. Tissues were then sectioned perpendicular to the skin surface at 7 µm and stained with haemotoxylin and eosin. Slides were visualized using a DMI 3000 B research microscope (Leica) and images captured using a DFC295 camera (Leica). For quantitative assessment of sweat gland and hair follicle density phenotypes, skin surface length was measured using ImageJ software, with features counted and normalized to this length. As a proxy of sweat gland size, sweat gland perimeters were also measured, with the mean perimeter length per individual used for statistical analysis. Since biopsies were double-sided (that is, represented both 'inside' and 'outside' ear surfaces), both surfaces were used for quantification.

Pituitary RNA sequencing and western blotting. Pituitaries were obtained from four hairy syndrome animals and four unrelated age-matched control calves following their killing on commercial slaughter premises. Pituitaries were pulverized in liquid nitrogen with samples divided for protein and RNA analyses. For RNA extraction and sequencing, tissue was homogenized in TRIzol Reagent (Life Technologies) and total RNA recovered using standard protocols by NZ Genomics Limited (NZGL; Auckland, NZ). Illumina sequencing libraries were prepared and sequenced by NZGL (Dunedin, NZ) using 100-bp paired-end reads on the HiSeq 2000 instrument, yielding 24–30 million read pairs per sample. Reads were mapped with Tophat2 (ref. 30) (v2.0.8), and *PRL* expression quantified using the 'variance stabilizing transformation' function in DESeq[31] (v1.14.0). Pituitary protein extracts were resolved on 12% SDS–PAGE gels and blotted on polyvinylidene difluoride (Bio-Rad). Western blotting was performed using antibodies to bovine prolactin (1:1,000 National Hormone and Peptide Program, USA) with beta-tubulin included as a loading control (1:2,000 AbCam ab6046).

TRH challenge. TRH infusion was conducted using the same 12 animals used for analysis of sweating rate. Jugular catheters were inserted on the day before the challenge, with catheter patency maintained with heparinized saline. Lyophilized hormone was obtained from Peptide Sciences (http://www.peptidesciences.com/trh) as 99% pure. Peptide was reconstituted in PBS at 20 µg ml^{-1} and administered at 0.3 µg TRH per kg of body weight. Blood samples (10 ml) were collected at −20, −10, −5, 2, 5, 10, 15, 20, 30, 45, 60, 90, 120 and 150 min relative to the time of the TRH bolus injection. Blood plasma was separated using centrifugation and ELISA assays conducted with AgResearch (Ruakura Research Centre, Hamilton), using bovine prolactin kits (USCN Life Science Inc., Cat. No. CEA846Bo).

References

1. Oftedal, O. T. & Dhouailly, D. *Evo-devo* of the mammary gland. *J. Mammary Gland Biol. Neoplasia* **18,** 105–120 (2013).
2. Oh, H. S. & Smart, R. C. An estrogen receptor pathway regulates the telogen-anagen hair follicle transition and influences epidermal cell proliferation. *Proc. Natl Acad. Sci. USA* **93,** 12525–12530 (1996).
3. Bocchinfuso, W. P. *et al.* Induction of mammary gland development in estrogen receptor-alpha knockout mice. *Endocrinology* **141,** 2982–2994 (2000).
4. Kondo, S., Hozumi, Y. & Aso, K. Organ culture of human scalp hair follicles: effect of testosterone and oestrogen on hair growth. *Arch. Dermatol. Res.* **282,** 442–445 (1990).
5. Chang, C., Lee, S. O., Wang, R.-S., Yeh, S. & Chang, T.-M. Androgen receptor (AR) physiological roles in male and female reproductive systems: lessons learned from AR-knockout mice lacking AR in selective cells. *Biol. Reprod.* **89,** 21 (2013).
6. Kaufman, K. D. Androgen metabolism as it affects hair growth in androgenetic alopecia. *Dermatol. Clin.* **14,** 697–711 (1996).
7. Blok, G. J., de Boer, H., Gooren, L. J. & van der Veen, E. A. Growth hormone substitution in adult growth hormone-deficient men augments androgen effects on the skin. *Clin. Endocrinol. (Oxf)* **47,** 29–36 (1997).
8. Kleinberg, D. L. Early mammary development: growth hormone and IGF-1. *J. Mammary Gland Biol. Neoplasia* **2,** 49–57 (1997).
9. Brisken, C. *et al.* Prolactin controls mammary gland development via direct and indirect mechanisms. *Dev. Biol.* **210,** 96–106 (1999).
10. Foitzik, K. *et al.* Human scalp hair follicles are both a target and a source of prolactin, which serves as an autocrine and/or paracrine promoter of apoptosis-driven hair follicle regression. *Am. J. Pathol.* **168,** 748–756 (2006).
11. Stenn, K. S. & Paus, R. Controls of hair follicle cycling. *Physiol. Rev.* **81,** 449–494 (2001).
12. Horseman, N. D. *et al.* Defective mammopoiesis, but normal hematopoiesis, in mice with a targeted disruption of the prolactin gene. *EMBO J.* **16,** 6926–6935 (1997).
13. Craven, A. J. *et al.* Prolactin signaling influences the timing mechanism of the hair follicle: analysis of hair growth cycles in prolactin receptor knockout mice. *Endocrinology* **142,** 2533–2539 (2001).
14. Olson, T. A., Lucena, C., Chase, C. C. & Hammond, A. C. Evidence of a major gene influencing hair length and heat tolerance in Bos taurus cattle. *J. Anim. Sci.* **81,** 80–90 (2003).
15. Mariasegaram, M. *et al.* The slick hair coat locus maps to chromosome 20 in Senepol-derived cattle. *Anim. Genet.* **38,** 54–59 (2007).
16. Flori, L. *et al.* A quasi-exclusive European ancestry in the Senepol tropical cattle breed highlights the importance of the slick locus in tropical adaptation. *PLoS ONE* **7,** e36133 (2012).
17. Huson, H. J. *et al.* Genome-wide association study and ancestral origins of the slick-hair coat in tropically adapted cattle. *Front. Genet.* **5,** 101 (2014).
18. Mills, D. E. & Robertshaw, D. Response of plasma prolactin to changes in ambient temperature and humidity in man. *J. Clin. Endocrinol. Metab.* **52,** 279–283 (1981).
19. Dikmen, S. *et al.* Differences in thermoregulatory ability between slick-haired and wild-type lactating Holstein cows in response to acute heat stress. *J. Dairy Sci.* **91,** 3395–3402 (2008).
20. Dikmen, S. *et al.* The SLICK hair locus derived from Senepol cattle confers thermotolerance to intensively managed lactating Holstein cows. *J. Dairy Sci.* **97,** 5508–5520 (2014).

21. Viitala, S. *et al.* The role of the bovine growth hormone receptor and prolactin receptor genes in milk, fat and protein production in Finnish Ayrshire dairy cattle. *Genetics* **173,** 2151–2164 (2006).
22. Bu, G. *et al.* Characterization of the novel duplicated PRLR gene at the late-feathering K locus in Lohmann chickens. *J. Mol. Endocrinol.* **51,** 261–276 (2013).
23. Littlejohn, M. D. *et al.* Data from: Functionally Reciprocal Mutations of the Prolactin Signalling Pathway Define Hairy and Slick Cattle. *Dryad Digit. Repos.* http://doi.org/10.5061/dryad.nh6v4 (2014).
24. Ferguson, K. & Dowling, D. The function of cattle sweat glands. *Aust. J. Agric. Res.* **6,** 640 (1955).
25. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84,** 210–223 (2009).
26. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
27. Cleary, J. G. *et al.* Joint variant and *de novo* mutation identification on pedigrees from high-throughput sequencing data. *J. Comput. Biol.* **21,** 405–419 (2014).
28. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
29. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
30. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
31. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

Acknowledgements

We would like to acknowledge the University of Auckland Centre for Genomics, Proteomics, and Metabolomics, and New Zealand Genomics Limited (in particular L. Williams) for RNA preparation and sequencing. We are also grateful to J. Chen from the School of Biological Sciences, the University of Auckland, for histological sample preparation. We also thank T. Hale, C. McMahon, A. Cullum and C. Berry from AgResearch Ltd. (Hamilton, NZ) for animal care and ELISA analysis, and B. Gudex and J. Arias from LIC (Hamilton, NZ), and G. Maynard from 5-Star Senepol (Queensland, Australia) for supplying Senepol DNA samples. We also gratefully acknowledge the assistance of S. Morgan and staff at DairyNZ Ltd. (Hamilton, NZ), D. Bartrum, S. Bandari and A. Bates of VetLife (Temuka, NZ), D. Fairweather of Dairy SolutioNZ Ltd. (Hamilton, NZ), and A. Reid (Winchester, South Canterbury, NZ) for phenotypic analysis of cattle. We also thank C. Charlier and M. Georges from the University of Liege, Belgium for facilitating access to bovine exome data, and M. Walker and G. Corbett of GeneMark (Hamilton, NZ) for conducting Taqman assays.
We are also grateful for financial support from the Ministry for Primary Industries (Wellington, New Zealand), who co-funded the work through the Primary Growth Partnership.

Author contributions

M.D.L., K.M.H., R.J.S. and S.R.D. conceived, designed and interpreted the experiments; K.M.H. conducted functional experiments; T.J., C.H., T.L. and W.L. conducted sequence analysis; M.D.L., K.T. and R.G. Sherlock performed statistical analyses; S.D.L., B.C.S. and D.J.G. provided animal resources; M.D.L., R.G. Snell, R.J.S. and S.R.D. supervised the project; M.D.L. wrote the manuscript.

Additional information

Accession codes: Genotype, phenotype and sequence data sets representing all experimental populations have been deposited in NCBI Short Read Archive under accession code SRP043521 and in Dryad Digital Repository under DOI: 10.5061/dryad.nh6v4.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: M.D.L., K.T., T.J., C.H., T.L., R.G.S., R.J.S. and S.R.D. are employees of Livestock Improvement Corporation, a commercial provider of bovine germplasm. The remaining authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Littlejohn, M. D. *et al.* Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. *Nat. Commun.* 5:5861 doi: 10.1038/ncomms6861 (2014).

Supplementary Information

Functionally Reciprocal Mutations of the Prolactin Signalling Pathway Define Hairy and Slick Cattle

Mathew D Littlejohn[1,2]*, Kristen M Henty[2]*, Kathryn Tiplady[1], Thomas Johnson[1], Chad Harland[1], Thomas Lopdell[1], Richard G Sherlock[1], Wanbo Li[3], Steven D Lukefahr[4], Bruce C Shanks[5], Dorian J Garrick[6], Russell G Snell[2], Richard J Spelman[1] & Stephen R Davis[1]

1. Livestock Improvement Corporation, Cnr Ruakura & Morrinsville Roads, Newstead, Hamilton 3240, New Zealand
2. School of Biological Sciences, University of Auckland, 3A Symonds Street, Auckland 1010, New Zealand
3. Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège (B34), Liège, Belgium
4. Department of Animal, Rangeland & Wildlife Sciences, MSC 228, Texas A&M University-Kingsville, Texas 78363-8202, USA
5. Department of Agriculture and Environmental Science, Lincoln University, 820 Chestnut Street, Jefferson City, Missouri 65101, USA
6. Department of Animal Science, Iowa State University, 225 Kildee, Ames, Iowa 50011-3250, USA

* These authors contributed equally to this work

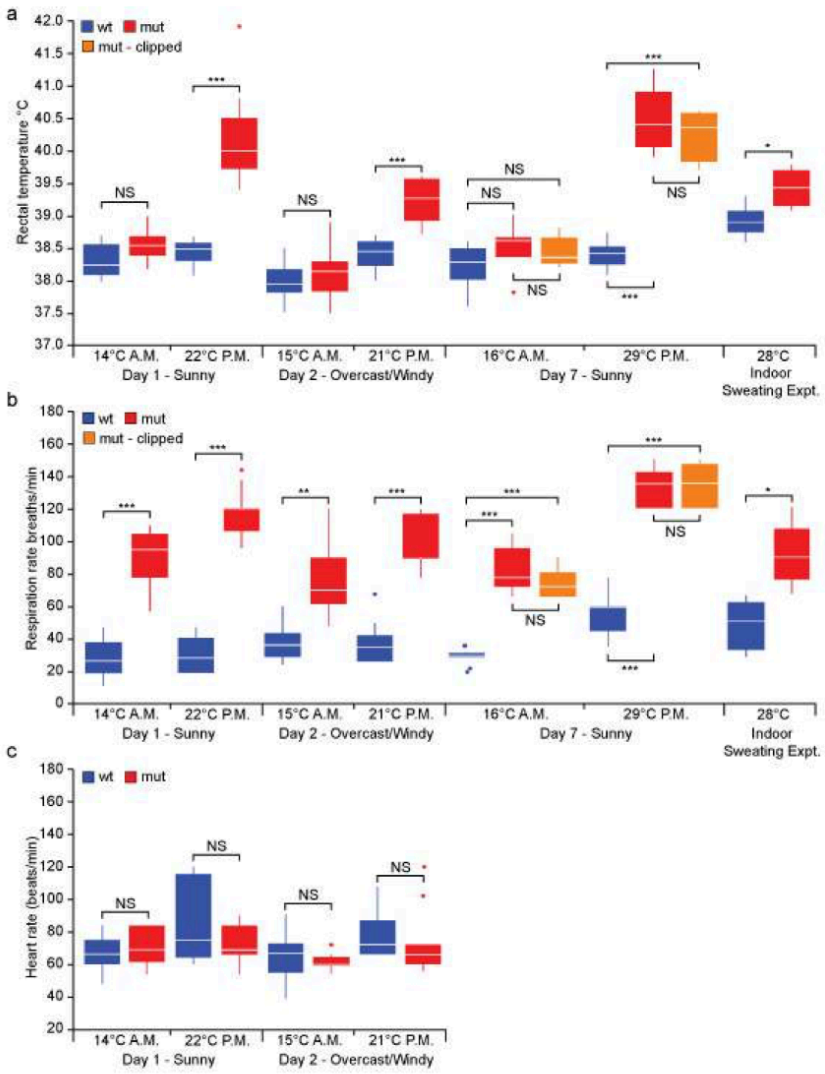
Contents

Supplementary Figure 1 **2**
Supplementary Figure 2 **3**
Supplementary Figure 3 **5**
Supplementary Table 1 **6**
Supplementary Table 2 **8**
Supplementary Table 3 **9**
Supplementary Table 4 **10**
Supplementary Table 5 **13**
Supplementary Table 6 **14**
Supplementray Table 7 **15**

Supplementary Figure 1: Coat analysis of hairy syndrome cattle

Photographs indicating coat differences between wild-type and mutant half-sibs (top), showing shaved skin areas used for collection of hair for morphological analysis (the white appearance is due to application of sunscreen). The bottom photograph shows contrast of coat-length in a clipped mutant animal, where coat-length was reduced to match those of wild-type animals for heat stress response measurements.

2

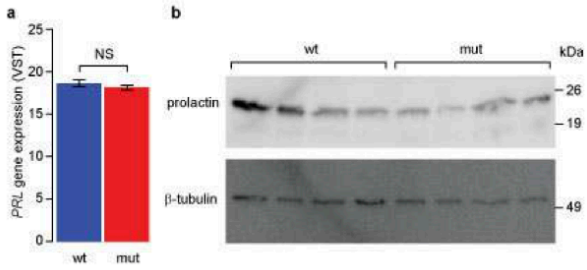


Supplementary Figure 2: Heat stress response measures of hairy pedigree animals at varying ambient temperatures

3

Rectal temperature (**a**), respiration rate (**b**), and heart rate (**c**) responses to different ambient temperatures in mutant and wild-type half-sibs. These data include additional time-points to those displayed in Figure 1, representing morning and afternoon measurements made on four separate days. Day 1 and Day 2 measurements compare 12 wild-type and 12 mutant animals, Day 3 measurements contrast the responses of clipped mutant animals (N=5) to un-clipped mutant (N=7), and control (N=12) animals. Rectal temperature and respiration rate data for the indoor sweating rate experiment are also indicated. Box plots define the median, upper and lower quartiles for the various phenotypes, with whiskers representing the furthest data-points within 1.5x of the interquartile range, and outlier samples indicated beyond this range. **P*<0.05, ***P*<0.001, ****P*<0.0001 (two-sided *t* tests, Bonferroni-adjusted).

4



Supplementary Figure 3: Pituitary prolactin gene and protein expression

(**a**) Mean pituitary prolactin gene expression in *PRL* mutant (N=4) and control (N=4) animals, quantified using RNA sequencing. Data represent variance transformed read counts, with no significant difference observed between groups (two-sided *t* test, *P*=0.32; error bars are ±s.e.m). Allele-tagged read counts within mutant samples was also similar for mutant and reference alleles (data not shown; between 51-54% 'C'-tagged reads for each of the four heterozygous mutant animals). (**b**) Prolactin western blot of pituitary extracts representing the same samples used for RNA sequencing, showing comparable levels of pituitary hormone between mutant and wild-type animals.

5

Phenotype	Control N	Control Mean	Control SE	Mutant N	Mutant Mean	Mutant SE	Clipped Mutant N	Clipped Mutant Mean	Clipped Mutant SE	*t* value	*P*-value	Bonferroni *P*
Milk Yield (L)												
Milk Yield – NZ-wide	740	18.9	0.14	77	14.3	0.57	.	.	.	-9.71	3.73E-21	1.12E-20
Milk Yield – North Island	446	18.7	0.17	39	14.9	0.78	.	.	.	-6.23	1.00E-09	3.00E-09
Milk Yield – South Island	294	19.1	0.24	38	13.8	0.84	.	.	.	-7.38	1.27E-12	3.81E-12
Hair Morphology												
Hair length (mm)	12	11.6	0.65	12	26.8	1.9	.	.	.	-7.58	1.42E-07	4.26E-07
Hair diameter (microns)	12	67.2	3.72	12	46.4	2.62	.	.	.	4.58	1.48E-04	4.44E-04
Hair weight (mg/cm2 skin)	12	9.1	0.68	12	16.1	2.46	.	.	.	-2.73	0.012	0.037
Heat stress												
Rectal Temperature (°C)												
Day 1 A.M. 14°C	12	38.3	0.068	12	38.6	0.073	.	.	.	-2.5	0.021	0.504
Day 1 P.M. 22°C	12	38.5	0.053	12	40.2	0.193	.	.	.	-8.73	1.33E-08	3.19E-07
Day 2 A.M. 15°C	12	38	0.085	12	38.1	0.116	.	.	.	-0.93	0.364	1
Day 2 P.M. 21°C	12	38.4	0.063	12	39.2	0.094	.	.	.	-6.91	6.19E-07	1.49E-05
Day 3 A.M. 16°C	12	38.2	0.083	7	38.5	0.138	.	.	.	-1.73	0.101	1
Day 3 A.M. 16°C	12	38.2	0.083	.	.	.	5	38.4	0.104	-1.32	0.280	1
Day 3 A.M. 16°C	.	.	.	7	38.5	0.138	5	38.4	0.104	0.37	0.716	1
Day 3 P.M. 29°C	12	38.4	0.054	7	40.5	0.187	.	.	.	-13.25	2.17E-10	5.21E-09
Day 3 P.M. 29°C	12	38.4	0.054	.	.	.	5	40.2	0.175	-13.23	1.13E-09	2.71E-08
Day 3 P.M. 29°C	.	.	.	7	40.5	0.187	5	40.2	0.175	1.01	0.337	1
Indoors 28°C (Sweating Expt)	6	38.9	0.098	6	39.5	0.11	.	.	.	-3.59	0.005	0.015
Respiration Rate (breaths/min)												
Day 1 A.M. 14°C	12	29	3.2	12	91	5	.	.	.	-10.49	5.04E-10	1.21E-08
Day 1 P.M. 22°C	12	32	3.1	12	118	3.9	.	.	.	-17.34	2.57E-14	6.17E-13
Day 2 A.M. 15°C	12	37	2.9	12	75	5.8	.	.	.	-5.87	6.62E-06	1.59E-04
Day 2 P.M. 21°C	12	35	3.6	12	99	4.3	.	.	.	-11.45	9.73E-11	2.34E-09
Day 3 A.M. 16°C	12	30	1.4	7	81	5.4	.	.	.	-11.69	1.50E-09	3.60E-08

Day 3 A.M. 16°C	12	30	1.4	.	.	.	5	73	4.4	-12.61	2.20E-09	5.28E-08
Day 3 A.M. 16°C	.	.	.	7	81	5.4	5	73	4.4	1.05	0.316	1
Day 3 P.M. 29°C	12	57	3.3	7	132	4.6	.	.	.	-13.52	1.59E-10	3.82E-09
Day 3 P.M. 29°C	12	57	3.3	.	.	.	5	134	6.1	-11.98	4.40E-09	1.06E-07
Day 3 P.M. 29°C	.	.	.	7	132	4.6	5	134	6.1	-0.26	0.800	1
Indoors 28°C (Sweating Expt)	6	48	6.2	6	91	7.7	.	.	.	-4.33	0.002	0.006
Heart Rate (beats/min)												
Day 1 A.M. 14°C	12	67	3.1	12	71	3.4	.	.	.	-0.98	0.336	1
Day 1 P.M. 22°C	12	85	7	12	74	3.2	.	.	.	1.49	0.149	1
Day 2 A.M. 15°C	11	65	4	12	62	1.3	.	.	.	0.75	0.462	1
Day 2 P.M. 21°C	12	77	3.8	12	73	5.5	.	.	.	0.67	0.507	1
Sweating Rate (grams/m^2 skin area/hr)												
Indoors 28°C	6	225	40	6	111	49	.	.	.	4.45	0.001	0.004

Supplementary Table 1: Association statistics for lactation, hair morphology, and heat stress phenotypes for hairy syndrome mutant and wild-type half-sibs

CHR	POS	HOMREF	HET	HOMALT	EFFECT	Amino_Acid_Change	Gene_Name	Transcript_ID
Chr23	27549182	554	2	0	DOWNSTREAM		NFKBIL1	ENSBTAT00000035817
Chr23	27549182	554	2	0	DOWNSTREAM		NFKBIL1	ENSBTAT00000019268
Chr23	27549182	554	2	0	INTERGENIC			
Chr23	27610230	554	2	0	INTERGENIC			
Chr23	27610230	554	2	0	UPSTREAM		ENSBTAG00000031913	ENSBTAT00000055663
Chr23	30971021	554	2	0	INTRON		ZNF184	ENSBTAT00000015283
Chr23	35105313	554	2	0	NON_SYNONYMOUS_CODING	C221G	PRL	ENSBTAT00000020313
Chr23	35159091	554	2	0	INTERGENIC			
Chr23	39362478	554	2	0	INTERGENIC			
Chr23	47446572	554	2	0	INTRON		BMP6	ENSBTAT00000025614

Supplementary Table 2: Annotation details for the seven chromosome 23 variants identified from whole genome sequence analysis of the two hairy syndrome founder sires

8

	PRLR p.Leu462* carrier	
Hair Length	Yes	No
1	41	1
2	2	21
3	0	11
4	0	6
Slick		
Yes	41	1
No	2	38

Supplementary Table 3: Distribution of quantitative and binary hair-length scores for *PRLR* p.Leu462* carrier and wild-type crossbreeds

9

			Haplo Count by Coat-Length								
Tile	Chr20 Position (bp)	Tile Size (kb)	Haplotype	1	2	3	4	Total	*P*-value	Bonferroni *P*
1.1	38606353-38892683	286	CCAAGT	6	4	2	1	13	0.749	1
1.2			CTAATT	3	7	1	3	14	0.293	1
1.3			TCGCGT	1	3	2	0	6	0.287	1
1.4			TCGCTT	46	6	5	4	61	4.00E-05	0.003
1.5			TTAAGT	7	0	0	2	9	0.735	1
1.6			TTGAGT	8	21	4	0	33	0.484	1
1.7			TTGCGT	10	1	2	2	15	0.875	1
2.1	38794140-38989299	195	AGTAAA	6	0	0	2	8	0.955	1
2.2			AGTAAG	5	2	2	1	10	0.642	1
2.3			AGTAGG	11	18	3	0	32	0.818	1
2.4			AGTGGG	0	5	1	0	6	0.287	1
2.5			ATTGGG	3	10	1	3	17	0.313	1
2.6			CGTAGG	8	3	5	2	18	0.146	1
2.7			CTTAGG	42	2	2	0	46	1.23E-16	8.60E-15
2.8			CTTGAG	3	3	5	4	15	1.95E-04	0.014
3.1	38920878-39071965	151	AAAGCC	6	0	0	2	8	0.955	1
3.2			AAGACT	6	3	3	1	13	0.524	1
3.3			AGGACC	53	7	3	2	65	1.42E-06	9.97E-05
3.4			AGGACT	4	7	3	0	14	0.490	1
3.5			AGGGCC	5	8	3	0	16	0.743	1
3.6			AGGGCT	1	2	2	0	5	0.296	1
3.7			GAGACT	3	3	5	4	15	0.000195	0.014
3.8			GGGACC	1	8	1	0	10	0.412	1
3.9			GGGGCC	2	8	1	3	14	0.163	1
4.1	39017985-39141815	124	ACCG*T	43	2	0	0	45	2.40E-19	1.68E-17
4.2			ACCGCT	10	1	1	2	14	0.589	1
4.3			ACCTCT	4	12	3	0	19	0.822	1

10

4.4			ACTTCC	10	11	9	5	35	0.007	0.491	
4.5			ACTTCT	3	2	2	0	7	0.797	1	
4.6			GCCGCT	5	16	2	1	24	0.351	1	
4.7			GCCTCT	8	0	2	4	14	0.329	1	
4.8			GCTGCT	1	2	2	0	5	0.296	1	
5.1	39100174-39242226	142	G*TGGA	42	2	0	0	44	8.39E-18	5.87E-16	
5.2			GCTGGA	1	2	2	0	5	0.296	1	
5.3			GCTGGG	3	7	0	0	10	0.809	1	
5.4			GCTTAA	11	10	3	3	27	0.294	1	
5.5			TCCTAA	9	8	6	1	24	0.588	1	
5.6			TCCTAG	1	3	3	4	11	5.81E-06	4.07E-04	
5.7			TCTGGA	3	8	5	0	16	0.369	1	
5.8			TCTTAA	7	0	0	2	9	0.735	1	
5.9			TCTTAG	4	6	0	2	12	0.362	1	
6.1	39194261-39350230	156	GGACAC	41	2	0	0	43	1.89E-16	1.32E-14	
6.2			GGATAT	3	1	2	0	6	0.863	1	
6.3			GGATGT	1	8	5	0	14	0.100	1	
6.4			GGGTAC	3	7	0	0	10	0.809	1	
6.5			TAACAC	7	0	2	1	10	0.809	1	
6.6			TAACAT	20	18	7	5	50	0.117	1	
6.7			TAGTAC	4	9	3	4	20	0.001	0.085	
7.1	39275917-39509383	234	CACTCG	21	0	0	0	21	5.28E-06	3.70E-04	
7.2			CACTCT	5	0	2	1	8	0.739	1	
7.3			CACTTT	22	2	0	0	24	3.17E-05	0.002	
7.4			CATCCG	3	7	3	0	13	0.339	1	
7.5			CATCTT	9	1	2	2	14	0.940	1	
7.6			CATTCG	4	9	2	3	18	0.210	1	
7.7			CATTCT	6	1	0	2	9	0.975	1	
7.8			TACCCT	4	7	0	0	11	0.622	1	
7.9			TACCTT	0	4	1	1	6	0.049	1	

11

7.10			TACTCG	2	1	1	3	7	0.005	0.350
7.11			TATTTT	3	1	3	0	7	0.502	1
7.12			TGTCCT	1	6	1	0	8	0.470	1
7.13			TGTTTT	1	2	2	0	5	0.296	1
8.1	39422828-39576637	154	CCGCG	3	7	3	0	13	0.339	1
8.2			CCTCG	6	13	1	0	20	0.829	1
8.3			CTTCA	0	4	1	1	6	0.049	1
8.4			CTTCG	9	1	2	2	14	0.940	1
8.5			TCGCA	24	9	1	3	37	0.009	0.626
8.6			TCGTA	2	0	1	3	6	0.004	0.252
8.7			TCTCG	12	3	3	3	21	0.621	1
8.8			TTTCG	7	1	0	0	8	0.042	1
8.9			TTTTA	20	7	5	0	32	0.212	1

Supplementary Table 4: Association results for 6-marker sliding window haplotypes in 82 Senepol crossbreeds

Eight consecutive tiles (3-marker overlaps) cover the 38.6-39.6Mbp *slick* locus, representing 70 haplotype states. Haplotype sequences and distribution across coat-length scores are indicated, with the *PRLR* p.Leu462* mutation represented by a red asterisk. Peak significance is observed for tile 4.1, carrying the *PRLR* p.Leu462* mutation and nested within a larger, 229kb contiguous haplotype shared by 41 of 42 slick-coded animals (Supplementary Table 5). Notably, the ancestral-allele haplotype 4.2 is un-associated with coat length.

ID	Haplotype	N
1	GACCG*TGG	45
2	**GACTTCCTA**	35
3	**GGCCGCTTA**	14
4	GACCGCTTA	14
5	GGCCGCTGG	10
6	GACCTCTTA	9
7	GACCTCTGG	9
8	AGCCTCTTA	8
9	GACTTCTGG	7
10	GGCTGCTGG	5
11	GGCCTCTTA	4
12	GGCCTCTGA	2
13	GATCGCTTA	1
14	GACCTCTTG	1

Supplementary Table 5: Senepol crossbreed haplotypes spanning a 229kb interval at the *slick* locus

A 229kb interval on chromosome 20 (38989299-39218755bp) demarcating a contiguous haplotype block found in all *PRLR* p.Leu462* carriers, and 41 of 42 slick-coded animals (haplotype 1). The single slick-coded animal that did not carry the p.Leu462* mutation is heterozygous for haplotypes 2 and 3 (bolded).

13

Phenotype	Control N	Control Mean	Control SE	Mutant N	Mutant Mean	Mutant SE	*t* value	*P*-value	Bonferroni *P*
Histology									
Sweat gland density (glands/mm skin surface length)	12	1.454	0.157	11	1.255	0.164	0.88	0.390	1
Hair follicle density (follicles/mm skin surface length)	12	2.532	0.391	11	3.468	0.408	-1.65	0.113	0.339
Sweat gland perimeter length (mm)	12	0.472	0.034	11	0.447	0.035	0.51	0.616	1
Prolactin expression									
PRL gene expression (VST counts)	4	18.7	0.39	4	18.2	0.24	1.08	0.321	.
Peak serum prolactin (ng/mL)	6	157.5	12.93	6	158.7	20.16	-0.05	0.962	.

Supplementary Table 6: Association statistics for histological and prolactin expression phenotypes

	Primer Name	Sequence
PRL		
Promoter	PRL_Promo_For1	TTGGAGAAGGAAATGGCAAC
	PRL_Promo_Rev1	CGTCAACTTAAAGCTGGGTCA
	PRL_Promo_For2	TGGGGCAGCATTAATTTCTT
	PRL_Promo_Rev2	TCAGGAGGGATGTGAAGAGG
	PRL_Promo_For3	GTGTGCCCTTGAAAACCACT
	PRL_Promo_Rev3	CCAGAAATGAACATCTAGGAAGG
Exon 1	PRL_Exon_1_For1	TGCAGAGAAATAAAGGCAAATG
	PRL_Exon_1_Rev1	CAAATGTCTCTGGAAGACAGTCC
Exon 2	PRL_Exon_2_For1	TTTACACAGTGGAAGGTGTTGC
	PRL_Exon_2_Rev1	CAGGTGCTTTAAATTTATTTTGGAA
Exon 3	PRL_Exon_3_For1	GGATGAAATGAAACAAGGGAAA
	PRL_Exon_3_Rev1	CACCTTTCTCTGCATGTCCA
Exon 4	PRL_Exon_4_For1	GGTCAATCACTCTGAGCAAAAA
	PRL_Exon_4_Rev1	GCCATCTGTACCCAGGAAGA
Exon 5	PRL_Exon_5_For1	TGGCTCCAAAATCCAAGTGT
	PRL_Exon_5_Rev1	CCCAGAATAATTTGCTGTGATTC
3' UTR	PRL_3'_For1	CAACAACTGCTAAGCCCACA
	PRL_3'_Rev1	CCACCTGACCATTTCCAAAC
PRLR		
Promoter	PRLR_Promo_For1	GCAAATGGGATTCTCCAGAC
	PRLR_Promo_Rev2	GGAAAGAACCCAGCTTTTTG
	PRLR_Promo_For3	TGAGGTTTAGGGAAGCCAAG
	PRLR_Promo_Rev3	TTCCCTTTTGAATGATTGACAC
	PRLR_Promo_For4	GGACTCAGCAGGTGGCTACT
	PRLR_Promo_Rev5	GGTGTCCCAACCTGGACTC
Exon 1	PRLR_5UTR_For1	CCTCCTCTCGCAAAGAAAGA
	PRLR_5UTR_Rev2	TTAGGGTAAGGTGGGCTGCT
Exon 2	PRLR_Exon2_For2	TCGATACCTGGGTCTGGAAG
	PRLR_Exon2_Rev2	AACAGCAGAATGCAACAACG
Exon 3	PRLR_Exon3_For	ATTTTTCCCAGCGTATGCAC
	PRLR_Exon3_Rev	TCCCAGAATGAGGATGGAAG
Exon 4	PRLR_Exon4_For	TGCTGACATCTTGGCACTTC
	PRLR_Exon4_Rev	AATTAACGCAGGGTCAGTGG
Exon 5	PRLR_Exon5_For	AGCAAGGAAGCTCCATACCA
	PRLR_Exon5_Rev	GGAAGAAGGGTCAAGGGAAG
Exon 6	PRLR_Exon6_For	CACCTCATGTCACCCACTTG
	PRLR_Exon6_Rev	GTCCTGGGAGAGCTCTGATG
Exon 7	PRLR_Exon7_For	GCAGAGAGGGTGAAATGGTG
	PRLR_Exon7_Rev	TGGCCTAGGGAAAGATGCTA
Exon 8	PRLR_Exon8_For2	TATAGAGGGGCAGGGGACTT
	PRLR_Exon8_Rev2	GCCTCCATTTGATGGAAAGA
Exon 9	PRLR_Exon9_For2	GCAGCCATTTGGAAATAAGG
	PRLR_Exon9_Rev2	TAGCAGCAGCTAAGCGACAA
Exon 10	PRLR_3UTR_For1	TGACATCAGCCACTGTGAGG
	PRLR_3UTR_Rev2	CAGCCCAACTGGAGTCTGC
	PRLR_3UTR_For2	CCTATTTTCTGGCCAATGGA

15

PRLR_3UTR_Rev3	TGCAAAGGTTAAGCAACTGG
PRLR_3UTR_For4	GGCCTTCATGGTTTCGTATG
PRLR_3UTR_Rev5	TTTCACCCAGAGAAGTGAAAA
PRLR_3UTR_For5	TGAAATAAACCAGACATAGAAAGACAA
PRLR_3UTR_Rev6	CTGCTAGGGCAATGCTTCTC
PRLR_3UTR_For7	CACTGCTTGGAAATGCAGAA
PRLR_3UTR_Rev8	TTTGCACATCACTTTAGACTATGATTC
PRLR_3UTR_For8	TGGCTGAAGACTCAAACTGAA
PRLR_3UTR_Rev9	AGATTCCCCCTTTCAGTTGG
PRLR_3UTR_For9	ATGAGTCGGCCACAGGTTTA
PRLR_3UTR_Rev10	GCATGCAATTCAAAGCCATT
PRLR_3UTR_For11	CAACCGCTGAGTGGACTTTT
PRLR_3UTR_Rev12	TGAAGGTCTCAGAGCCAAAA
PRLR_3UTR_For12	CAGGAAAACCCTGCTGATGA
PRLR_3UTR_Rev14	CAGTTTTTGAACCACATAAAGCA
PRLR_3UTR_For12	CAGGAAAACCCTGCTGATGA
PRLR_3UTR_Rev15	TATGGAACCTTTGGCTGTCC
PRLR_3UTR_For15	TCTTTCGAAGCTGCTTATTGC
PRLR_3UTR_Rev16	GGCATTGTAACTAGTTCATGTAGAAA
PRLR_3UTR_For17	TTTCCCTAAAGCCCCTAGAAA
PRLR_3UTR_Rev18	AAAGCAGATGGCACCAGTGT
PRLR_3UTR_For18	TAACACATGCCTGGCTGAAA
PRLR_3UTR_Rev19	TCCCATATTGCAGGTGGATT
PRLR_3UTR_For20	ATCCCCTGGAGAAGGGATAG
PRLR_3UTR_Rev21	AGAATGCATGCCTGGAAAAA

Supplementary Table 7: *PRL* and *PRLR* oligonucleotide sequences

Primer sequences used to generate *PRL* and *PRLR* PCR amplicons for Sanger sequencing are indicated. Each 'For' and 'Rev' primer pair for a given amplicon are denoted by blue or pink background shading. Sequencing was conducted using the same primers used for amplification.

16

Experimental Section

Study 6:

NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock

Genome Res. 26:10, 1333-1341, 2016

Carole Charlier, Wanbo Li, Chad Harland, Mathew Littlejohn, Wouter Coppieters, Frances Creagh, Steve Davis, Tom Druet, Pierre Faux, François Guillaume, Latifa Karim, Mike Keehan, Naveen Kumar Kadri, Nico Tamma, Richard Spelman and Michel Georges

Abstract

We herein report the result of a large-scale, next generation sequencing (NGS)-based screen for embryonic lethal (EL) mutations in Belgian beef and New Zealand dairy cattle. We estimated by simulation that cattle might carry, on average, ~0.5 recessive EL mutations. We mined exome sequence data from >600 animals, and identified 1377 stop-gain, 3139 frame-shift, 1341 splice-site, 22,939 disruptive missense, 62,399 benign missense, and 92,163 synonymous variants. We show that cattle have a comparable load of loss-of-function (LoF) variants (defined as stop-gain, frame-shift, or splice-site variants) as humans despite having a more variable exome. We genotyped >40,000 animals for up to 296 LoF and 3483 disruptive missense, breed-specific variants. We identified candidate EL mutations based on the observation of a significant depletion in homozygotes. We estimated the proportion of EL mutations at 15% of tested LoF and 6% of tested disruptive missense variants. We confirmed the EL nature of nine candidate variants by genotyping 200 carrier × carrier trios, and demonstrating the absence of homozygous offspring. The nine identified EL mutations segregate at frequencies ranging from 1.2% to 6.6% in the studied populations and collectively account for the mortality of ~0.6% of conceptuses. We show that EL mutations preferentially affect gene products fulfilling basic cellular functions. The resulting information will be useful to avoid at-risk matings, thereby improving fertility.

Research

NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock

Carole Charlier,[1,5] Wanbo Li,[2,5] Chad Harland,[1,3] Mathew Littlejohn,[3] Wouter Coppieters,[1,4] Frances Creagh,[3] Steve Davis,[3] Tom Druet,[1] Pierre Faux,[1] François Guillaume,[1,6] Latifa Karim,[1,4] Mike Keehan,[3] Naveen Kumar Kadri,[1] Nico Tamma,[1] Richard Spelman,[3] and Michel Georges[1]

[1]*Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège (B34), 4000-Liège, Belgium;* [2]*State Key Laboratory for Pig Genetic Improvement and Production Technology, Jiangxi Agricultural University, Nanchang, 330045, Jiangxi Province, P.R. China;* [3]*Livestock Improvement Corporation, Newstead, Hamilton 3240, New Zealand;* [4]*Genomics Platform, GIGA, University of Liège (B34), 4000-Liège, Belgium*

We herein report the result of a large-scale, next generation sequencing (NGS)-based screen for embryonic lethal (EL) mutations in Belgian beef and New Zealand dairy cattle. We estimated by simulation that cattle might carry, on average, ~0.5 recessive EL mutations. We mined exome sequence data from >600 animals, and identified 1377 stop-gain, 3139 frame-shift, 1341 splice-site, 22,939 disruptive missense, 62,399 benign missense, and 92,163 synonymous variants. We show that cattle have a comparable load of loss-of-function (LoF) variants (defined as stop-gain, frame-shift, or splice-site variants) as humans despite having a more variable exome. We genotyped >40,000 animals for up to 296 LoF and 3483 disruptive missense, breed-specific variants. We identified candidate EL mutations based on the observation of a significant depletion in homozygotes. We estimated the proportion of EL mutations at 15% of tested LoF and 6% of tested disruptive missense variants. We confirmed the EL nature of nine candidate variants by genotyping 200 carrier × carrier trios, and demonstrating the absence of homozygous offspring. The nine identified EL mutations segregate at frequencies ranging from 1.2% to 6.6% in the studied populations and collectively account for the mortality of ~0.6% of conceptuses. We show that EL mutations preferentially affect gene products fulfilling basic cellular functions. The resulting information will be useful to avoid at-risk matings, thereby improving fertility.

[Supplemental material is available for this article.]

Livestock productivity has dramatically increased over the last 50 years. Milk production in Holstein cows has doubled from ~6000 in 1960 to ~12,000 kgs in 2000, and ~75% of this change was genetic (Dekkers and Hospital 2002). However, gains for producers were partially eroded by concomitant decreases in disease resistance and fertility. Pregnancy rate decreased by ~6% in this population over the same period (Norman et al. 2009). It is assumed that the reduced fertility results from the negative energy balance of high-producing cows. A complementary explanation might be an increase in premature pregnancy termination due to homozygosity for embryonic lethal (EL) mutations.

This is supported by several observations. One is the recent positional cloning of a quantitative trait locus (QTL) for fertility in Nordic Red Cattle (Kadri et al. 2014). It was shown to be due to a 660-kb deletion on Chromosome 12 that causes early embryonic lethality in homozygotes. The deletion was shown to segregate at high frequencies in Nordic cattle (up to 16% in Finnish Ayrshire) as a result of its positive effect on milk yield in heterozygotes. Prior to its detection, it caused the death of up to ~0.64% of conceptuses in these breeds. Also, the realization that all or a substantial proportion of embryos homozygous for the DUMPS (deficiency of uridine monophosphate synthase) (Robinson et al. 1983), CVM (complex vertebral malformation) (Thomsen et al. 2006), or BS (brachyspina syndrome) (Charlier et al. 2012) mutations die before birth and are therefore never reported suggests that other fully lethal (i.e., early mortality of all embryos) and hence unsuspected ELs might be segregating at fairly high frequencies. As an example, the BS mutation was shown to segregate at a frequency of 3.7% in Holstein Friesian and hence to cause the mortality of ~0.14% of conceptuses. The 660-kb deletion, as well as the CVM and BS mutations, were identified using standard forward genetics approaches (Georges 2007). In the case of CVM and BS, this was possible because samples from affected individuals could be used for linkage and association analyses. The population frequency of the 660-kb deletion was high enough in Finnish Ayrshire to significantly affect the breeding values for fertility of carrier bulls, hence allowing QTL analysis. It is worth

[5]**These authors contributed equally to this work.**
[6]**Present address: Evolution NT, 35706 Rennes, France**
Corresponding authors: carole.charlier@ulg.ac.be, michel.georges@ulg.ac.be
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.207076.116.

noting that in other Scandinavian breeds, in which the deletion was segregating at frequencies ≤6% (hence still causing mortality of 0.09% of conceptuses), QTL analysis was not possible, as the effect on the breeding values for fertility of this recessive EL was too modest. Thus, phenotype-driven forward genetic approaches are not suitable to identify ELs segregating at frequencies <~10% which is likely to be the case for the majority.

An alternative, genotype-driven approach has recently been devised that takes advantage of the large cattle cohorts that have been genotyped with genome-wide SNP arrays for genomic selection. The signals that are sought are depletions in homozygotes (among live animals) for specific haplotypes assumed to be associated with EL mutations. This approach, combined with follow-up studies of the corresponding haplotypes, has led to the identification of six ELs in cattle (Fritz et al. 2013; Sonstegard et al. 2013; Daetwyler et al. 2014; Pausch et al. 2015). However, at least two conditions need to be met for this strategy to be effective: (1) Very large cohorts (tens to hundreds of thousands of animals) genotyped with medium- to high-density SNP arrays need to be available in the breeds of interest; and (2) linkage disequilibrium (LD) between the EL and the cognate haplotype needs to be very high, if not perfect ($r^2 \sim 1$). The former condition is only met for few very popular breeds, including Holstein-Friesian, in which three of six detected ELs were found. It is likely to remain a considerable bottleneck, as low-density (~10K) SNP arrays (which are not suitable for haplotype-based analyses) are increasingly replacing medium-density (~50K) ones. The latter condition is likely to be met for only part of the ELs, as most of the time LD between the EL and the haplotype will be complete ($D' \sim 1$) but not perfect (i.e., cognate haplotypes without the EL are also segregating in the population). Thus, it is almost certain that other as of yet unknown ELs still segregate in most livestock populations.

To make further progress in the identification of ELs in cattle, we hereby apply a reverse genetics approach that takes advantage of the growing amount of whole-exome and whole-genome NGS data in livestock. The proposed approach consists in (1) mining available sequence data for predicted loss-of-function (LoF) and damaging missense (MS) variants, (2) genotyping large cohorts for the corresponding candidates and identifying putative ELs on the basis of a significant depletion in homozygotes, and (3) confirming the EL nature of the corresponding super-candidates on the basis of a significant depletion in homozygotes in carrier × carrier matings.

Results

Expectations for the number of EL mutations carried per individual

Diploidy has allowed the genome to increase in size while insuring at least one functional copy of each gene in the majority of individuals. Accordingly, most diploid individuals are assumed to carry a number of lethal mutations in the heterozygous state. In *Drosophila melanogaster*, this number has been estimated at ~1.6 (e.g., Simmons and Crow 1977). Humans have been estimated to carry an average of the order of ~0.29 recessive mutations that lead to complete postnatal sterility or death by reproductive age when homozygous (Gao et al. 2015), or ~1.4 postnatal "lethal equivalents" (e.g., Sutter and Tabah 1953; Morton et al. 1956; Bittles and Neel 1994). It remains unknown, however, how many recessive mutations causing prenatal death when homozygous are carried, on average, by humans or any other mammal. The total number of recessive lethals (pre- and postnatal) carried by individuals is a function of the number of recessive lethals that segregate in the population as well as the frequency distribution of their occurrence in the population. The actual values of these parameters are unknown but can be estimated from the knowledge of (1) the genomic target size for recessive lethal mutations, (2) the rate of recessive lethal mutations in this target space, and (3) the present and past effective population size. Systematic knock-out programs conducted in the mouse indicate that ≤25% of mammalian genes are essential, i.e., defined as causing complete or partial preweaning lethality in homozygotes (International Mouse Phenotype Consortium [IMPC] at https://www.mousephenotype.org). This corresponds to a target space of ~2,500,000 codons (or ~7,500,000 nt), and ~90,000 splice-sites (or ~180,000 nt) (Ng et al. 2009). Assuming (1) a single nucleotide substitution rate of $\sim 10^{-8}$ per base pair and per gamete, (2) that 3% of single nucleotide substitutions in codon space cause illegitimate stop-gains (given the mammalian codon usage and a transition/transversion ratio of 2), (3) that all single nucleotide substitutions in splice-sites perturb splicing, and (4) a ~25% proportion of stop-gains and splice-site variants among lethal mutations (deduced from the equivalent proportion among mutations causing known recessive genetic defects; see, for instance, The Human Gene Mutation Database [HGMD at http://www.hgmd.cf.ac.uk]), the rate of recessive lethal mutations can be estimated at ~0.015 per gamete. We performed simulations under these assumptions and estimated that the number of recessive lethals (pre- and postnatal; hereafter collectively termed ELs) carried, on average, per individual increases with population size from ~0.85 for an effective population size (N_e) of 100 to ~7.7 for N_e = 10,000. Interestingly, the frequency of death as a result of homozygosity for EL remains nearly constant, diminishing only very slightly from ~1.73% at $N_e = 100$ to ~1.54 at $N_e = 10{,}000$. However, the proportion of these deaths due to "common" EL mutations (defined as having a minor allele frequency [MAF] ≥ 2%) ranges from ~98% when $N_e = 100$ to ~0% when $N_e = 10{,}000$ (Table 1). Despite an actual population of several tens of millions of animals, the effective population size of Holstein-Friesian dairy cattle has been estimated at ~100, as a result of intense selection and widespread use of artificial insemination (de Roos et al. 2008). Despite an actual population size of billions, the effective population size of humans has been estimated at ~10,000, as a result of past bottlenecks. Thus, our simulations indicate that the number of ELs segregating in dairy cattle populations may be of the order of tens, and that the population frequency of many of these may be of the order of 2% or more. Identifying these common ELs may be an effective first step to reduce the number of embryonic deaths from homozygosity for recessive lethals, thereby improving fertility.

Identification of ~94,000 nonsynonymous variants in domestic cattle

We resequenced the whole genome of 496 animals from the New Zealand dairy cattle (NZDC) population and 50 Belgian Blue Cattle (BBC) at an average depth of 11 (range: 3–148). In addition, we resequenced the exome of 78 animals representing six cattle breeds (*Bos taurus*) at an average depth of 40 (range: 18–100). Sequencing was carried out using reversible terminator chemistry on HiSeq 2000 instruments (Illumina) and SureSelect Target Enrichment reagents (Agilent) for exome sequencing. Sequence reads were mapped to the Bostau6 bovine reference genome using BWA (Li

Table 1. Estimation, by simulation (≥2000 generations), about lethal mutations as a function of the effective population size (N_e; range: 50–10,000) and the rate of recessive lethal mutations per gamete (MU; 0.01 or 0.015)

N_e	MU	NR SEGR SITES[a]	NR MUT/IND[b]	MUT FREQ[c]	% DEATH[d]	% > 0.02[e]
50	0.01	4.84 (2.30)	0.37 (0.22)	3.74 (1.64)	1.05 (1.82)	1
	0.015	7.36 (2.85)	0.58 (0.31)	3.96 (1.44)	1.87 (2.38)	0.98
100	**0.01**	**11.01 (3.34)**	**0.53 (0.21)**	**2.41 (0.69)**	**1.01 (1.18)**	**0.94**
	0.015	**17.19 (4.60)**	**0.85 (0.30)**	**2.49 (0.58)**	**1.73 (1.60)**	**0.98**
500	0.01	68.42 (8.86)	1.14 (0.22)	0.84 (0.11)	1.02 (0.60)	0.7
	0.015	104.77 (10.83)	1.78 (0.28)	0.85 (0.09)	1.69 (0.75)	0.69
1000	0.01	151.58 (13.51)	1.65 (0.21)	0.54 (0.05)	1.07 (0.40)	0.48
	0.015	220.54 (15.18)	2.29 (0.22)	0.52 (0.04)	1.39 (0.45)	0.43
5000	0.01	899.31 (27.71)	3.53 (0.19)	0.2 (0.01)	0.99 (0.18)	0.02
	0.015	1366.06 (41.38)	5.37 (0.22)	0.2 (0.01)	1.5 (0.19)	0.02
10,000	**0.01**	**1925.4 (43.46)**	**4.95 (0.16)**	**0.13 (0.01)**	**0.99 (0.12)**	**0.0006**
	0.015	**2936.68 (55.92)**	**7.7 (0.19)**	**0.13 (0.00)**	**1.54 (0.14)**	**0.0001**

Simulations were conducted assuming complete selection against homozygotes. Numbers in parentheses correspond to standard deviations. Values for $N_e = 100$ and $N_e = 10{,}000$, corresponding to the effective population size of cattle and human, respectively, are in bold.
[a]Number of segregating recessive lethal mutations.
[b]Number of recessive lethals carried, on average, per individual.
[c]Average frequency of the corresponding recessive lethals in the population.
[d]Percentage (total) of conceptuses dying as a result of homozygosity for a recessive lethal mutation.
[e]Percentage of these deaths (cf. footnote d) that are due to homozygosity for common recessive lethal mutations (defined as MAF ≥ 0.02).

and Durbin 2009). Exomic variants were identified using GATK and corresponding best practices (McKenna et al. 2010). Effects on gene function of the identified variants were predicted using Variant Effect Predictor (McLaren et al. 2010). We identified a total of 186,112 exonic variants, including 1377 stop-gain, 112 stop-loss, 3139 frame-shift, 1341 splice-site, 85,338 missense, and 92,163 synonymous variants (Supplemental Table S1). Of the missense variants, 22,939 were predicted by SIFT and/or PolyPhen to be disruptive/damaging (Kumar et al. 2009; Adzhubei et al. 2010). To ensure that the differences in nucleotide diversity observed between the human (BAM files downloaded from the 1000 Genomes Project) and bovine samples (sequenced at the University of Liège [ULg]) would not be merely technical artifacts, we compared the nucleotide diversity obtained with the 1000 Genomes BAM files with those obtained for 10 human samples sequenced at the ULg using the same experimental conditions (Supplemental Material S1).

Domestic cattle have a comparable LoF load as humans despite a more variable exome

It has been shown that humans carry, on average, ~120 loss-of-function variants defined by MacArthur et al. (2012) as frame-shift, splice-site, stop-gains, and large deletions. To rigorously compare the mutational load of humans and domestic cattle, we selected 148,913 conserved coding exons from the human-bovine genome alignment (amounting to ~58% of coding exon space) (see Methods) captured by Agilent's bovine SureSelect Target Enrichment assay. Within this sequence space, we called genetic variants in 59 exome-sequenced cattle and 60 humans using BAM files that were either generated in-house or downloaded from the 1000 Genomes Project (http://www.1000genomes.org/). From these data, we extrapolated (to the entire exome) that Yorubans are, on average, heterozygous at ~9000 (9 K) synonymous (S) and ~5.4 K nonsynonymous (NS) sites, while European and Asians are heterozygous at ~6.3 K S and ~4.0 K NS positions, in agreement with previous estimates (Fig. 1A; e.g., The 1000 Genomes Project Consortium 2010, 2012). In contrast, domestic cattle are, on average, heterozygous at 13.2 K S and 5.9 K NS positions (Fig. 1B). Thus, present-day domestic cattle are genetically more variable than humans, including Africans. The observed S/NS ratios are ~4.6- and ~6.3-fold larger than expected in humans and cattle, respectively, supporting enhanced purifying selection on NS variants as expected (more so in cattle; see hereafter). Humans were estimated to be heterozygous for 58 (range: 31–85) and homozygous for 9 (range: 0–21) LoF variants (excluding large deletions), which is also in agreement with previous studies (Fig. 1C; MacArthur et al. 2012). Domestic cattle were heterozygous for 51 (range: 25–82) and homozygous for 7 (range: 0–21) LoF variants (excluding large deletions), and this was significantly ($P = 0.002$) lower than humans (Fig. 1D). Thus, despite the higher overall genetic variation observed in domestic cattle, their load of LoF variants is equivalent, if not somewhat lower than that of humans.

Estimating the proportion of EL among LoF and missense variants from population data

The observed number of ~120 LoF variants per individual is ~20-fold larger than the ~1–5 recessive lethals estimated to be carried, on average, by individuals (see above). This discrepancy is thought to reflect the importance of molecular redundancy and the high proportion of developmentally nonessential genes. The identification of the minority of EL mutations among the many LoF variants remains a considerable challenge.

To gain insights into the proportion and nature of EL mutations among LoF variants in cattle, we mined the available lists of bovine variants for frame-shift, splice-site, and stop-gain variants. Moreover, we identified missense variants predicted by PolyPhen2 to be damaging and/or by SIFT to be deleterious (Kumar et al. 2009; Adzhubei et al. 2010). The corresponding list of candidate ELs was manually curated for possible sequencing or alignment artifacts using IGV (Robinson et al. 2011), including confirmation of the gene models using fetal RNA-seq data. We further selected variants for which none of the well-covered sequenced individuals were homozygous and which were breed-specific (see Methods). We selected 3779 candidate EL variants in the NZDC population (including 296 LoF and 3483 missense), and 1050 in the BBC population (108 LoF, 942 missense), and added them as custom variants to new designs of the Illumina bovine LD SNP arrays. Moreover, we added 200 breed-specific

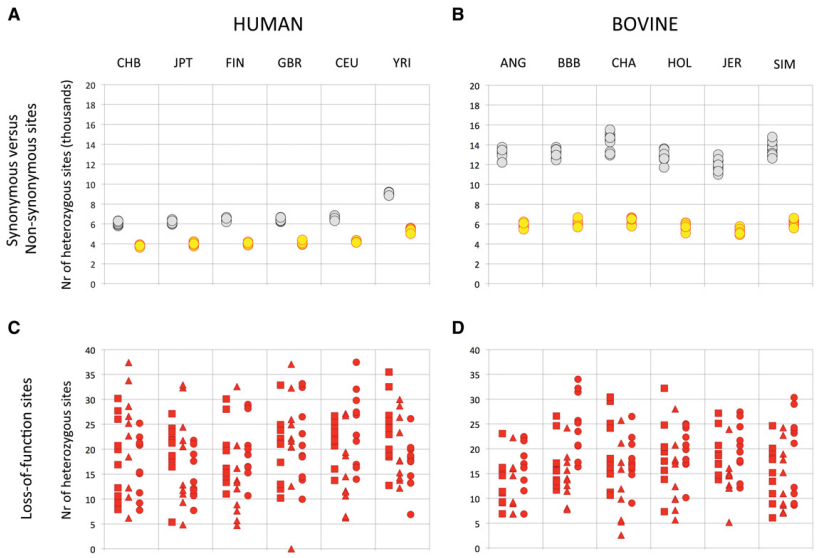


Figure 1. (*A*,*B*) Number of heterozygous synonymous (gray) and nonsynonymous (yellow) sites per individual (*A*: humans; *B*: bovine). (*C*,*D*) Number of heterozygous stop-gain (squares), splice-site (triangles), and frame-shift (circles) sites per individual. CHB: Chinese; JPT: Japanese; FIN: Finns; GBR: Britons; CEU: Northern Europeans; YRI: Yorubans; ANG: Angus; BBB: Belgian Blue; CHA: Charolais; HOL: Holstein-Friesian; JER: Jerseys; SIM: Simmentals.

synonymous variants as "matched controls" to one of the BBC designs (Supplemental Table S2). We genotyped ~35,000 NZDC and ≥6300 BBC healthy animals. For all variants on the array, we computed the statistical significance (log[1/*p*]) of the depletion in homozygosity for the minor allele (versus within-breed Hardy-Weinberg expectation) (Fig. 2; Supplemental Fig. S1; see Methods). We were struck by the occurrence, in both populations, of candidate EL variants without homozygote mutant animals despite population frequencies ≥1.3% (NZDC) and 1.8% (BBCB), while this was never observed for any one of the thousands of neutral (N) variants on the arrays. This suggested that the interrogated LoF and missense variants might indeed harbor EL mutations. Alternatively, the observed difference between candidate EL and N variants might reflect their distinct ascertainment scheme. As an example, interrogated LoF and missense variants were selected to be breed-specific and hence probably younger on average than the N variants shared by multiple breeds. To account for this possible discrepancy, we compared the behavior of candidate EL variants with a set of breed-specific synonymous variants, selected using the same criteria as the LoF and missense variants in BBC. Contrary to LoF and missense variants, there was not a single synonymous variant with population frequency ≥2.2% without homozygote individuals, again suggesting the occurrence of ELs among interrogated LoF and missense variants. The proportion of LoF variants without homozygotes was 0.348 (±0.050), while it was 0.228 (±0.038) for equally sized (50) sets of frequency-matched synonymous variants. The same numbers were 0.233 (±0.056) and 0.185 (±0.044) for frequency-matched sets of missense and synonymous variants. From this, we estimated the proportion of ELs at 15.5% of tested LoF variants and 5.9% of tested missense variants (see Methods).

Confirming the embryonic lethality of nine common LoF variants in carrier-carrier matings

To provide direct evidence of their embryonic lethality, we retrospectively genotyped 25 trios (carrier sire, carrier dam, healthy offspring), on average (range: 8–50), for the (at the time) most significant four LoF and single missense variants in BBC, and for the (at the time) most significant three LoF and single missense variants in NZDC, all with MAF ≥1.2%, and without observed homozygotes. Using information from the matched S variants in BBC, we estimated the proportion of ELs among LoF and missense variants without homozygotes at 0.44 and 0.25, respectively (see Methods). Genotyping was done directly for 141/200 trios and by combining direct genotyping in the parents with linkage analysis for 59/200 trios (see Methods). No homozygous offspring were observed in the 200 offspring, supporting the embryonic lethality of the nine tested variants (four in NZDC and five in BBC). Ratios between homozygote wild-type and carrier animals did not depart significantly from the expected 1:2 in these crosses ($P \geq 0.13$). Eight of these genes are broadly expressed and code for proteins fulfilling essential housekeeping processes, such as DNA replication, transcription, and RNA processing. Expected *cis*-eQTL effects were observed in mammary gland for the three LoF variants predicted to cause nonsense mediated RNA decay (*OBFC1* frame-shift, *TTF1* stop-gain, and *RNF20* stop-gain) (Supplemental Material S2). Frequencies of the identified ELs averaged 3.2% and ranged from 1.2% to 6.6% (Table 2).

Identifying nonlethal coding variants with phenotypic effects

Some variants were characterized by a pronounced depletion in homozygotes in the general population despite the occurrence

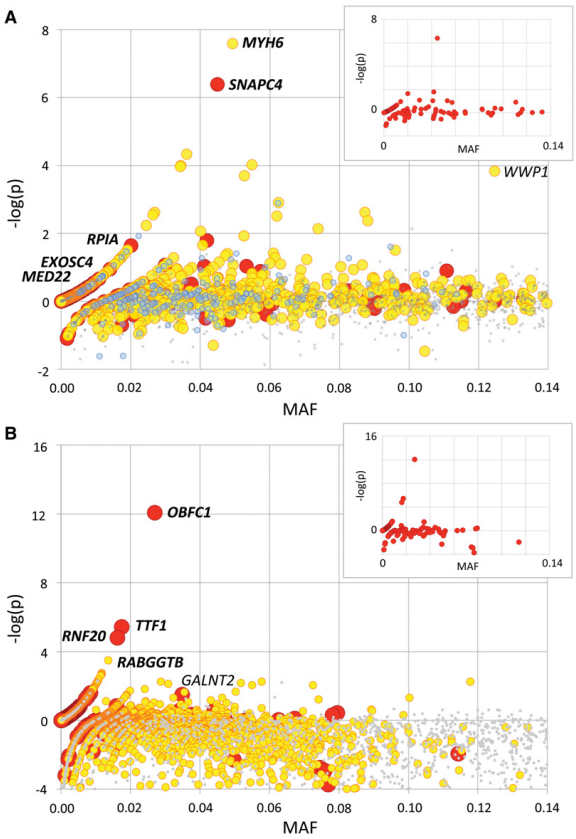


Figure 2. Statistical significance [−log(*p*): *y*-axis] of the depletion (positive values) or excess (negative values) in homozygotes for loss-of-function (red; defined as frame-shift, splice-site, and stop-gain variants), missense (yellow), matched synonymous (blue), and random neutral (small gray) variants ordered by minor allele frequency (MAF: *x*-axis), based on the genotyping of 6385 healthy BBC (*A*) and 35,219 healthy NZDC (*B*) animals. Variants that have been subsequently tested in carrier × carrier matings and proven to be embryonic lethals (EL) are labeled in italics and bold. *WWP1*, shown to affect muscularity, and *GALNT2*, shown to cause growth retardation, are labeled in italics. For NZDC (*B*), MAFs were computed across breeds (NZ Holstein-Friesian, NZ Jersey, and NZ cross-bred), explaining the differences with the within-breed MAF reported in Table 2, and the high proportion of variants with negative −log(*p*) values. *Insets*: loss-of-function-variants-alone graphs for the corresponding BBC (A) and NZDC (B) populations.

of presumably healthy homozygous individuals. This suggests that selection acts against homozygotes, albeit without causing early death. Indeed, one of these variants proved to be a splice-site mutation in the *GALNT2* gene, encoding polypeptide N-acetylgalactosaminyltransferase 2. It was recently identified by a standard forward genetic approach as the mutation causing "Small Calf Syndrome" in NZDC (M Littlejohn, pers. comm.). Another is a common (13% frequency in BBC) missense variant in the *WWP1* gene, encoding the WW domain containing E3 ubiquitin protein ligase 1. The ≥6300 genotyped BBC animals included 581 bulls with an average of 331 (range: 1–4706) offspring records for more than eight traits pertaining to muscularity and stature, allowing computation of breeding values. A genome wide association study (GWAS) using these breeding values indicated that the R844Q *WWP1* variant very significantly increased muscularity, while decreasing stature (Supplemental Fig. S2). This strongly suggests that its observed high frequency in BBC results from yet another example of balanced polymorphism operating in intensely selected livestock populations (Hedrick 2015).

Lack of evidence for synergistic epistasis

It has been suggested that deleterious variants are more effectively purged from populations as a result of synergistic epistasis, i.e., that multiple deleterious genetic variants have a larger cost on fitness than expected from their multiplicative effects. This hypothesis predicts that individuals carrying multiple deleterious variants will be fewer than expected assuming random assortment. Recent analyses of the GoNL sequence data suggest that synergistic epistasis might be operating in humans (Sohail et al. 2016). We tested the hypothesis using the large genotype database generated as part of this study. Analyses were conducted in the BBC population, separately for LoF and missense variants. We observed no evidence for an underrepresentation of animals carrying multiple LoF or missense variants in either of these populations (Supplemental Material S3).

Discussion

Making reasonable assumptions about the genomic target size for recessive lethal mutations (~9×10^6 bp), the proportion of lethal mutations among all mutations in this space (~15%), and a mutation rate per base pair of 10^{-8}, we herein estimate by simulation that the number of ELs carried, on average, per individual increases with effective population size (N_e) from ~0.5 for $N_e = 100$ to ~5 for $N_e = 10{,}000$, corresponding to estimates of the effective population size for domestic cattle and humans, respectively. We show that the percentage of conceptuses that will die from homozygosity for EL mutations is independent of effective population size and on the order of ~1% under our model. We show that the majority of these deaths involve on the order of tens of ELs segregating at frequencies >2% in domestic cattle, while likely involving a very large number of rare EL variants in human.

We then show that the exome of domestic cattle is more variable than that of humans, when considering both synonymous and nonsynonymous variants. These findings are in agreement

Table 2. Main features of nine confirmed embryonic lethal (EL) mutations in cattle

Gene			Mutation				
Symbol	Name	Function	Type[a]	Symbol	Population[b] (MAF %)	Offspring genotypes ++/+M/MM	*P*-values[c] MM/T (++/+M)
OBFC1	oligonucleotide/oligosaccharide-binding fold containing 1	Initiation of DNA replication and telomere protection	FS	p.Lys127Valfs*28	JER (6.59)	12/18/0	$1.8 \times 10^{-4***}$ (4.4×10^{-1})
TTF1	transcription termination factor	Ribosomal gene transcription regulation	SG	p.Arg527*	HF (3.52)	11/18/0	$2.4 \times 10^{-4***}$ (6.0×10^{-1})
RABGGTB	Rab geranylgeranyltransferase beta subunit	Post-translational addition of geranylgeranyl groups to Rab GTPases.	MS	p.Tyr195Cys	HF (2.13)	4/4/0	$1.00 \times 10^{-1(*)}$ (3.2×10^{-1})
RNF20	ring finger protein 20, E3 ubiquitin protein ligase	Regulation of chromosome structure by monoubiquitinating histone H2B	SG	p.Lys693*	HF (1.82)	4/9/0	$3.4 \times 10^{-2**}$ (8.4×10^{-1})
MYH6[d]	myosin, heavy chain 6	Myofibril formation and contraction, cardiac development	Del 1aa	p.Lys1730del	BBC (4.99)	18/28/0	$1.8 \times 10^{-6***}$ (4.0×10^{-1})
SNAPC4	small nuclear RNA activating complex polypeptide 4	Transcription of RNA pol II and III small-nuclear RNA genes	FS	p.Leu1227Alafs*134	BBC (5.13)	6/20/0	$5.6 \times 10^{-4***}$ (2.7×10^{-1})
RPIA	ribose 5-phosphate isomerase A	Conversion between ribose-5-phosphate and ribulose-5-phosphate in the pentose-phosphate pathway	SS	c.826+1G>A	BBC (1.89)	3/15/0	$5.6 \times 10^{-3***}$ (1.3×10^{-1})
EXOSC4	exosome component 4	Participation in RNA processing and degradation	SG	p.Arg64*	BBC (1.33)	6/12/0	$5.6 \times 10^{-3***}$ (1.0×10^{0})
MED22	mediator complex subunit 22	Transcription regulation by bridging interactions between regulatory factors, RNA pol II, and transcription factors	FS	p.Leu38Argfs*25	BBC (1.15)	5/7/0	$3.2 \times 10^{-2*}$ (5.4×10^{-1})

OMIA 002042-9913 Abortion (embryonic lethality), EXOSC4 in Bos taurus (cattle) Gene: EXOSC4.
OMIA 002043-9913 Abortion (embryonic lethality), MED22 in Bos taurus (cattle) Gene: MED22.
OMIA 002039-9913 Abortion (embryonic lethality), MYH6 in Bos taurus (cattle) Gene: MYH6.
OMIA 002035-9913 Abortion (embryonic lethality), OBFC1 in Bos taurus (cattle) Gene: OBFC1.
OMIA 002037-9913 Abortion (embryonic lethality), RABGGTB in Bos taurus (cattle) Gene: RABGGTB.
OMIA 002038-9913 Abortion (embryonic lethality), RNF20 in Bos taurus (cattle) Gene: RNF20.
OMIA 002041-9913 Abortion (embryonic lethality), RPIA in Bos taurus (cattle) Gene: RPIA.
OMIA 002040-9913 Abortion (embryonic lethality), SNAPC4 in Bos taurus (cattle) Gene: SNAPC4.
OMIA 002036-9913 Abortion (embryonic lethality), TTF1 in Bos taurus (cattle) Gene: TTF1.
[a]FS: frame-shift; SG: stop-gain; MS: missense; Del: deletion; SS: splice-site.
[b]BBC: Belgian Blue Cattle Breed; JER: New Zealand Jerseys; HF: New Zealand Holstein-Friesian.
[c]*MM/T*: *P*-value of MM/total ratio assuming viable MM genotype; ++/+M: *P*-value of ++/+M ratio assuming lethality of MM genotype. (*) $P < 0.05$; (**) $P < 0.01$; (***) $P < 0.001$.
[d]The *MYH6* p.Lys1730del mutation is in high linkage disequilibrium with a p.Thr202Met missense mutation in the *API G2* gene for which healthy homozygous individuals have been observed and which can therefore be excluded as being responsible for the observed EL effect.

with recent estimates of nucleotide diversity (based on whole-genome sequence data) shown to be higher in domestic cattle (1.44/kb) than in humans (Yoruba: 1.03/kb; European: 0.68/kb) (Daetwyler et al. 2014). The mutation rate in the cattle germ-line has recently been estimated at $\sim 1.1 \times 10^{-8}$ per base pair per gamete (M Georges, unpubl.), hence near identical to human. This strongly suggests that the past effective population size of domestic cattle was larger than that of humans (e.g., MacEachern et al. 2009). Thus, against expectations, the bottlenecks undergone by cattle as the result of the domestication process appeared to have been less severe than the bottlenecks undergone by humans, including Africans. One explanation for this is that domestication of cattle has been a long-lasting process with a sustained flow of genes from the wild (with large effective population size) into the domestic populations. Another possible cause of the observed higher nucleotide diversity in domestic cattle when compared to humans is that domestication involved subspecies of wild bovids carrying highly divergent haplotypes. Thus, present-day domestic taurine cattle might in fact have a mosaic genome tracing back to distinct wild subspecies. This phenomenon is certainly well documented in European domestic pig breeds, in which alleles tracing back to Asian wild boars segregate in a genome with originates primarily from European wild boars (e.g., Van Laere et al. 2003; Groenen et al. 2012; Bosse et al. 2014).

When focusing on LoF variants, however, it appears that humans carry, on average, more such variants than cattle. We attribute this apparent conundrum to the fact that deleterious recessive alleles are being purged more effectively and more rapidly from the genome of present-day domestic cattle than from that of humans as a result of the rapid increase in inbreeding following breed creation and initiation of intense selection programs particularly in the nineteenth and twentieth centuries (including the widespread use of artificial selection) (Goddard et al. 2010). In agreement with this hypothesis, we observe that the S/NS ratios are larger in domestic cattle (~6.3) than in humans (~4.6), testifying to stronger purifying selection in cattle than in humans.

We have mined exome sequence data from >500 animals and have identified >400 candidate LoF and >4400 deleterious missense variants which we have genotyped in large cohorts of 35,000 and 6300 animals in NZDC and BBC cattle, respectively. From the observed proportion of variants without homozygotes among healthy individuals, we have estimated that ~15% of tested LoF variants and ~6% of tested missense variants might be ELs. These percentages increase to 44% (LoF) and 25% (missense) when restricting the analysis to variants without homozygotes among healthy individuals. We have tested the ELs of nine of the most common of these candidate EL variants in carrier × carrier matings, indeed confirming their lethality. Not unexpectedly, the corresponding genes are broadly expressed and code for proteins fulfilling essential housekeeping functions, including DNA replication, transcription, and RNA processing. We estimated the proportion of affected conceptuses (i.e., homozygous for at least one of the nine reported ELs) to be ~0.64% in the NZDC and ~0.61% in the BBC populations, corresponding to ~7600 and ~3000 embryos, and an estimated cost of 13.8 million NZ$ and 2.7 million €, respectively. In offspring of sires that are carrying the most common ELs, these proportions reach ~3.3% in the NZDC and ~2.7% in the BBC populations, respectively. Knowing the genotypes of sires and dams for the corresponding EL variants will assist in avoiding at-risk matings, thereby improving fertility.

There remain two frame-shift and eight missense variants with population frequency >1% in the BBC population, of which the EL status has not yet been confirmed in carrier × carrier matings. At least four of these affect genes fulfilling essential functions (Supplemental Table S2). Our prediction is that these are likely ELs as well and work to test this is in progress.

Thus far, a number of ELs have been identified in livestock by taking advantage of large cohorts that were SNP-genotyped for genomic-selection purposes and identifying haplotypes never observed in homozygous form. The corresponding haplotypes are then sequenced to identify the putative EL mutations (e.g., Pausch et al. 2015). This approach is only effective if (1) the ELs are in complete linkage disequilibrium ($r^2 \sim 1$) with the corresponding haplotypes, and (2) large enough SNP-genotyped cohorts are available (which is the case for only very few breeds). Retrospective analyses indicate that only the single most common of the four EL mutations in NZDC (in *OBFC1*) would have been detected using this standard approach (Supplemental Material S4). For the remaining ones, the ELs are only in perfect LD ($D' \sim 1$; $r^2 < 1$) with flanking haplotypes, indicating that equivalent wild-type haplotypes still segregate in the population, hence obscuring the signal. Thus, more ELs are likely to segregate in the studied populations than might be suspected from haplotype-based analyses alone. The absence of large SNP-genotyped cohorts in BBC (as in most other smaller breeds) precluded the use of the haplotype-based approach. Our results demonstrate the efficacy of an NGS-based reverse genetic screen even in smaller populations.

We observe a significant departure from Hardy-Weinberg equilibrium for some of the tested variants showing a depletion in (yet existence of) homozygotes. This could be due to the contamination of the supposedly healthy cohort with affected individuals, particularly for variants causing mild phenotypes. This was likely the case for the splice-site variant in the *GALNT2* gene causing a form of dwarfism in the NZDC population. Alternatively, the expressivity may be variable up to the point of incomplete penetrance, such that a proportion of homozygotes may appear healthy and be included in the cohort. Selection should nevertheless act against such variants and drive them toward low frequencies. We observed a missense variant in the *WWP1* gene showing a striking depletion in homozygotes yet having an allelic frequency as high as 13% in the BBC population. We provide evidence that this is likely due to the fact that it positively affects desirable phenotypes in the heterozygotes while being deleterious in the homozygote state. Thus, these variants, especially the most frequent ones, possibly encompass additional examples of balancing selection, which increasingly appear to be commonplace in domestic animals (e.g., Hedrick 2015).

In addition, this study yields a cohort of animals that appear normal at first glance despite being homozygous for obvious LoF variants in genes deemed essential. The list included homozygous mutants for *NME7 (NDK7)*, *SYNE2*, *SLC9A9*, and *FREM1* (Supplemental Table S2). Such animals will be deeply phenotyped to possibly uncover physiological perturbations that might be medically pertinent as illustrated by *PCSK9*, *CCR5*, *ACTN3*, *CASP12*, and *SCN9A* knockouts in humans (Kaiser 2014; Alkuraya 2015).

Methods

Simulations

To estimate the number of ELs carried on average per individual, we simulated the reproduction of panmictic populations with constant effective population size ranging from 50 to 10,000 for

10,000 generations. At every generation, gametes had a probability of 0.01 to be affected by a novel recessive lethal mutation, which was always considered to be distinct and affecting another gene compared to all other mutations already present in the population. All mutations were assumed to segregate independently of each other (no linkage). Individuals that were homozygous for any of the segregating mutations were removed from the population with compensatory reproduction.

Next generation sequencing

Genomic DNA was extracted from sperm or whole blood using standard procedures. For whole-genome resequencing, PCR-free libraries were generated and sequenced (100-bp paired ends) on HiSeq 2000 instruments by Illumina's FastTrack services for the NZDC samples, and at the CNAG (Barcelona) for the BBC samples. For exome sequencing, enrichment was conducted using the Sure Select Target Enrichment Reagents (Agilent), and sequencing conducted on HiSeq 2000 instruments at the GIGA Genomics platform at the University of Liège.

Variant calling

Sequence reads were aligned to the bosTau6 reference genome using BWA (Li and Durbin 2009). PCR duplicates were identified using Picard (http://picard.sourceforge.net/). Local indel realignment and base quality score recalibration was conducted with GATK (McKenna et al. 2010). Variants were called using UnifiedGenotyper for the NZDC and exome samples, and using the GATK Haplotype caller (McKenna et al. 2010) for the BBC samples. Variant quality score recalibration was conducted using GATK VariantRecalibrator (McKenna et al. 2010) using the Illumina BovineHD Genotyping BeadChip variants and a subset of newly detected sequence variants showing correct Mendelian segregation within a large sequenced nuclear pedigree as reference sets.

Comparing the mutational load of human and bovine exomes

Bovine exome sequencing was generated as described above. Data from 60 unrelated human exomes were downloaded from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/ and down-sampled to match the distribution of sequence depth of the bovine samples using GATK. Variants were called using GATK's UnifiedGenotyper (McKenna et al. 2010) as described above. The comparison of the mutational load was restricted to 148,913 coding exons that were nonredundant, 1:1 alignable, and of equal size in human and bovine, flanked by canonical splice-sites, and autosomal. Olfactory receptor genes were ignored. Variant sites were only considered if coverage ≥20 and mapping quality ≥30. Additional filters for qualifying SNPs were: QD < 2.0, MQ < 40.0, FS > 60.0, ReadPosRankSum <−8.0, MQRankSum <−12.5, and for qualifying indels: QD < 2.0, FS > 200.0, ReadPosRankSum <−20.0. Heterozygosity was calculated for each individual as the number of heterozygous sites divided by the total number of qualifying sites (coverage ≥ 20 and MQ ≥ 30). Variants were annotated as S, MS, SS, FS, and SG mutation based on the human RefSeq gene model.

Testing for depletion in homozygosity

The significance of the depletion in homozygosity was computed using a standard likelihood ratio test corresponding to $LRT = 2\ln(\langle L|H1\rangle/\langle L|H0\rangle$ in which

$$\langle L|H1\rangle = \left(\frac{n_{mm}}{n_{mm}+n_{m+}+n_{++}}\right)^{n_{mm}} \times \left(\frac{n_{m+}+n_{++}}{n_{mm}+n_{m+}+n_{++}}\right)^{n_{m+}+n_{++}}$$

and

$$\langle L|H0\rangle = \left(\frac{2\times n_{mm}+n_{m+}}{2\times(n_{mm}+n_{m+}+n_{++})}\right)^{2\times n_{mm}} \times \left(1-\left(\frac{2\times n_{mm}+n_{m+}}{2\times(n_{mm}+n_{m+}+n_{++})}\right)^{2}\right)^{(n_{m+}+n_{++})}.$$

In these, n_{xx} corresponds to the number of animals with corresponding genotype (*m*: mutant, +: wild-type allele). *LRT* was assumed to have a χ^2 distribution with one degree of freedom under the null.

Estimating the proportion of ELs among LoF and missense variants

The proportion of ELs among LoF (respectively, missense) variants, p, was estimated as $p = b - a/1 - a$, where b is the proportion of interrogated LoF (respectively, missense) variants without homozygotes and a is the average proportion of variants without homozygotes among size- and frequency-matched sets of control "S" variants (cf. main text). This is derived from the assumption that $b = p + (1-p)a$.

The proportion of ELs among LoF (respectively, missense) variants without homozygous animals, q, was estimated as

$$q = \frac{p}{b} = \frac{b-a}{b(1-a)},$$

where b and a are defined as above.

Testing for synergistic epistasis

To test for synergistic epistasis, we permuted (1000×) genotypes for LoF and/or missense variants among genotyped individuals (separately for each variant). We then examined the distribution of the number of individuals carrying 0, 1, 2, … n LoF/missense variants, looking for a depletion of individuals carrying multiple mutations when compared to the simulated data.

Data access

VCF files (GATK) and individual BAM files (BWA) from this study, corresponding to the annotated exonic sequences of the full bovine data set, have been submitted to the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/) under accession number PRJEB14827.

Acknowledgments

This work was supported by grants from the European Research Council (Damona), the Walloon Region (DGARNE Rilouke), and the EU Framework 7 program (GplusE). We used the supercomputing facilities of the Consortium des Equipements de Calcul Intensitf en Fédération Wallonie Bruxelles (CECI) funded by the F.R.S.-FNRS. We thank the Ministry for Primary Industries (Wellington, New Zealand) for financial support, who cofunded the work through the Primary Growth Partnership. We also thank the Asssociation Wallonne de l'Elevage and Herdbook Blanc-Bleu Belge for support. We thank Arnaud Sartelet, Xavier Hubin, and Kristof De Fauw for their assistance in sample collection.

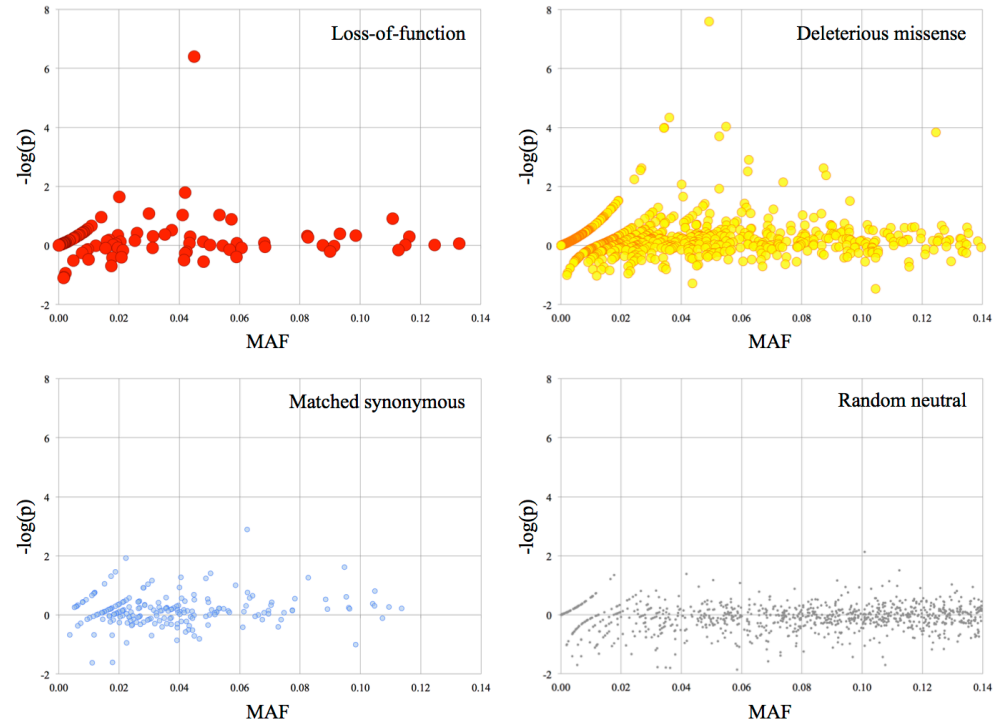
References

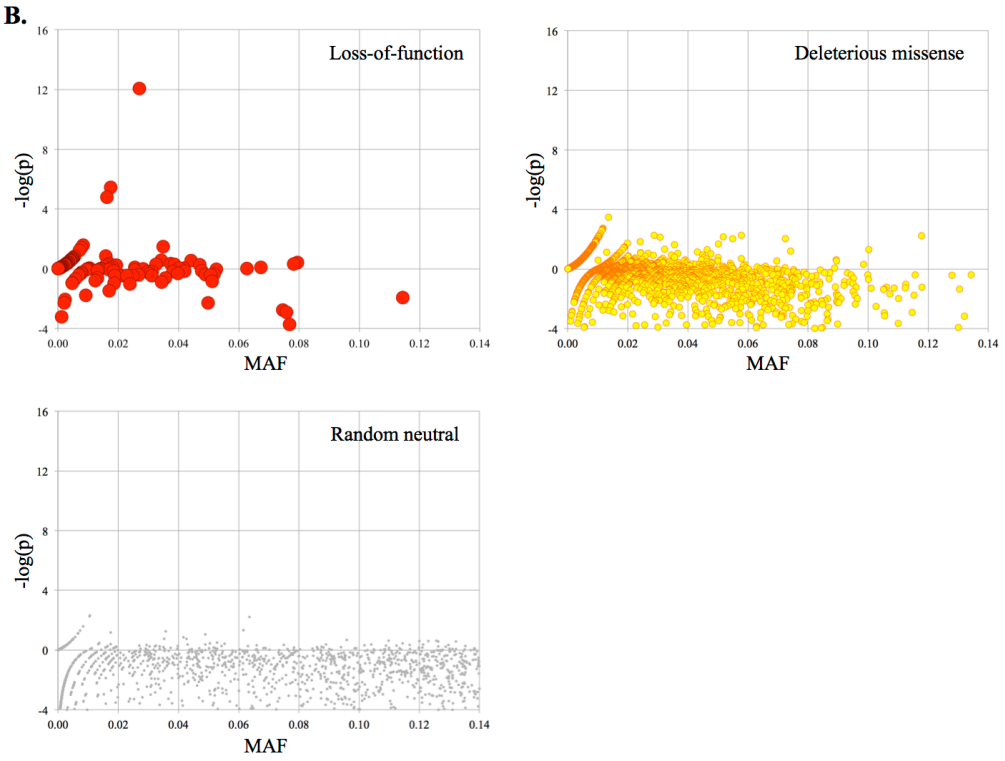
The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.
The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky V, Gerasimova A, Bork P, Kondrashov A, Sunyaev S. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7:** 248–249.
Alkuraya FS. 2015. Human knockout research: new horizons and opportunities. *Trends Genet* **31:** 108–115.
Bittles AH, Neel JV. 1994. The costs of human inbreeding and their implications for variations at the DNA level. *Nat Genet* **8:** 117–121.
Bosse M, Megens HJ, Frantz LA, Madsen O, Larson G, Paudel Y, Duijvesteijn N, Harlizius B, Hagemeijer Y, Crooijmans RP, et al. 2014. Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nat Commun* **5:** 4392.
Charlier C, Agerholm JS, Coppieters W, Karlskov-Mortensen P, Li W, de Jong G, Fasquelle C, Karim L, Cirera S, Cambisano N, et al. 2012. A deletion in the bovine *FANCI* gene compromises fertility by causing fetal death and brachyspina. *PLoS One* **7:** e43085.
Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, Liao X, Djari A, Rodriguez SC, Grohs C, et al. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46:** 858–865.
de Roos APW, Hayes BJ, Spelman RJ, Goddard ME. 2008. Linkage disequilibrium and persistence of phase in Holstein Friesian, Jersey and Angus cattle. *Genetics* **179:** 1503–1512.
Dekkers JCM, Hospital F. 2002. Multifactorial genetics: the use of molecular genetics in the improvement of agricultural populations. *Nat Rev Genet* **3:** 22–32.
Fritz S, Capitan A, Djari A, Rodriguez SC, Barbat A, Baur A, Grohs C, Weiss B, Boussaha M, Esquerré D, et al. 2013. Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in GART, SHBG and SLC37A2. *PLoS One* **8:** e65550.
Gao Z, Waggoner D, Stephens M, Ober C, Przeworski M. 2015. An estimate of the average number of recessive lethal mutations carried by humans. *Genetics* **199:** 1243–1254.
Georges M. 2007. Mapping, fine mapping, and molecular dissection of quantitative trait loci in domestic animals. *Annu Rev Genomics Hum Genet* **8:** 131–162.
Goddard ME, Hayes BJ, Meuwissen TH. 2010. Genomic selection in livestock populations. *Genet Res* **92:** 413–421.
Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491:** 393–398.
Hedrick PW. 2015. Heterozygote advantage: the effect of artificial selection in livestock and pets. *J Hered* **106:** 141–154.
Kadri NK, Sahana G, Charlier C, Iso-Touru T, Guldbrandtsen B, Karim L, Nielsen US, Panitz F, Aamand GP, Schulman N, et al. 2014. A 660-kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. *PLoS Genet* **10:** e1004049.
Kaiser J. 2014. The hunt for missing genes. *Science* **344:** 687.
Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4:** 1073–1081.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760.
MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335:** 823–828.
MacEachern S, Hayes B, McEwan J, Goddard M. 2009. An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in domestic cattle. *BMC Genomics* **10:** 181.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303.
McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26:** 2069–2070.
Morton NE, Crow JF, Muller HJ. 1956. An estimate of mutational damage in man from data on consanguineous marriages. *Proc Natl Acad Sci* **42:** 855–863.
Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461:** 272–276.
Norman HD, Wright JR, Hubbard SM, Miller RH, Hutchison JL. 2009. Reproductive status of Holstein and Jersey cows in the United States. *J Dairy Sci* **92:** 3517–3528.
Pausch H, Schwarzenbacher H, Burgstaller J, Flisikowski K, Wurmser C, Jansen S, Jung S, Schnieke A, Wittek T, Fries R. 2015. Homozygous haplotype deficiency reveals deleterious mutations compromising reproductive and rearing success in cattle. *BMC Genomics* **16:** 312–325.
Robinson JL, Drabik MR, Dombrowski DB, Clark JH. 1983. Consequences of UMP synthase deficiency in cattle. *Proc Natl Acad Sci* **80:** 321–323.
Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26.
Simmons MJ, Crow JF. 1977. Mutations affecting fitness in Drosophila populations. *Annu Rev Genet* **11:** 49–78.
Sohail M, Vakhrusheva OA, Sul JH, Pulit S, Francioli L, GoNL Consortium, Alzheimers Disease Neuroimaging Initiative, van den Berg LH, Veldink JH, de Bakker P, et al. 2016. Negative selection in humans and fruit flies involves synergistic epistasis. *bioRxiv 066407*. doi: http://dx.doi.org/10.1101/066407.
Sonstegard TS, Cole JB, Vanraden PM, Van Tassell CP, Null DJ, Schroeder SG, Bickhart D, McClure MC. 2013. Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. *PLoS One* **8:** e54872.
Sutter J, Tabah L. 1953. Structure de la mortalité dans les familles consanguines. *Population* **8:** 511–526.
Thomsen B, Horn P, Panitz F, Bendixen E, Petersen AH, Holm LE, Nielsen VH, Agerholm JS, Arnbjerg J, Bendixen CA. 2006. Missense mutation in the bovine *SLC35A3* gene, encoding a UDP-*N*-acetylglucosamine transporter, causes complex vertebral malformation. *Genome Res* **16:** 97–105.
Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M, et al. 2003. A regulatory mutation in *IGF2* causes a major QTL effect on muscle growth in the pig. *Nature* **425:** 832–836.

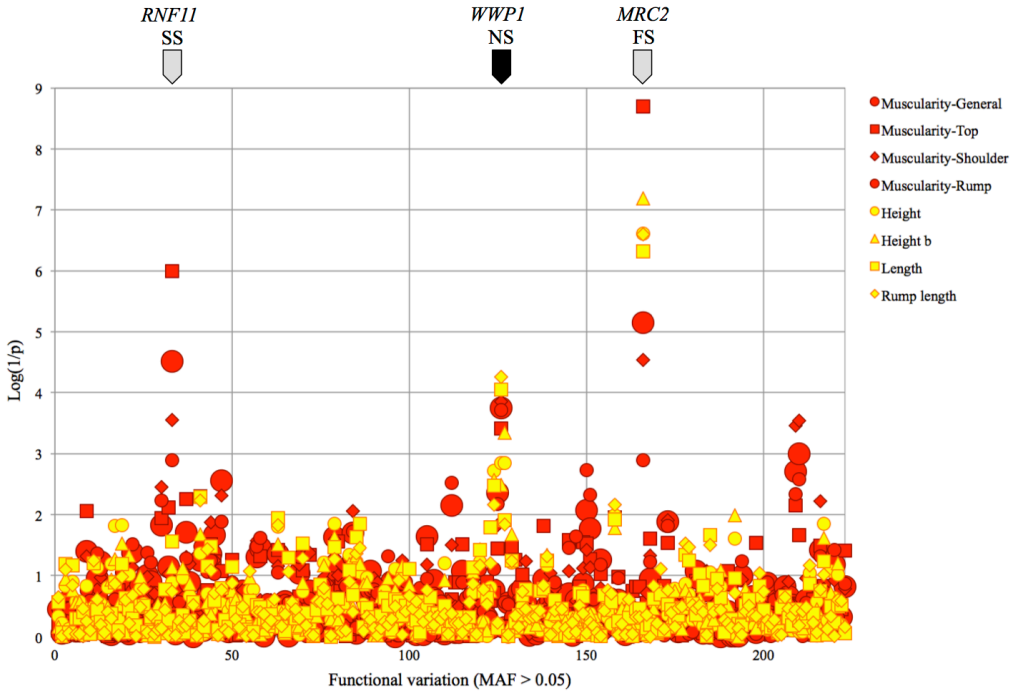
Received March 16, 2016; accepted in revised form August 19, 2016.

A.


Loss-of-function
Deleterious missense
Matched synonymous
Random neutral
-log(p)
MAF
8
6
4
2
0
-2
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14



Supplemental Figure S1 : Statistical significance (-log p: Y-axis) of the depletion (positive values) or excess (negative values) in homozygotes shown separately for loss-of-function (red; defined as frame-shift, splice-site and stop-gain variants), missense (yellow), matched synonymous (blue), random neutral (small grey) variants ordered by minor allele frequency (MAF: X-axis), based on the genotyping of 6,385 healthy BBC (**A**) and 35,219 healthy NZDC (**B**) animals. For NZDC (**B**), MAF were computed across breeds (NZ Holstein-Friesian, NZ Jersey and NZ cross-bred), explaining the differences with the within breed MAF reported in Table 2, and the high proportion of variants with negative -log(p) values.



Supplemental Figure S2: GWAS conducted in 580 BBC bulls using breeding values for eight trait pertaining to muscularity and stature and genotypes for 223 LoF and missense variants (MAF>5%) sorted according to their genomic position. The two strongest signals correspond to the previously identified frame-shift (FS) *MRC2*[1] and splice-site (SS) *RNF11*[2] mutations, causing Crooked tail Syndrome and Growth Stunting, respectively. The third signal corresponds to the newly identified R844Q non-synonymous (NS) mutation in the *WWP1* gene. The statistical model used was as in Druet et al.[3].

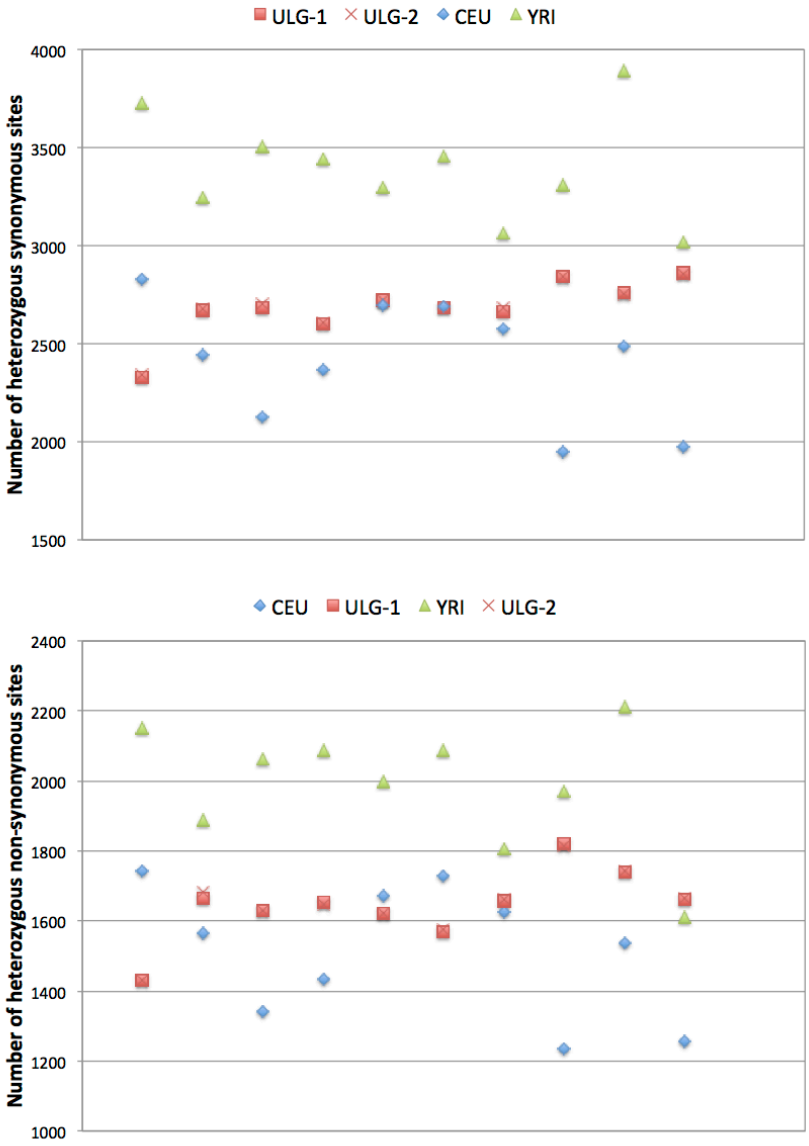
1. Fasquelle et al. Balancing selection of a frame-shift mutation in the MRC2 gene accounts for the outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. *PLoSGenetics* 5: e1000666 (2009).
2. Sartelet et al. A splice site variant in the bovine RNF11 gene compromises growth and regulation of the inflammatory response. *PLoSGenetics* 8: e1002581 (2012).
3. Druet et al. Selection in action: dissecting the molecular underpinnings of the increasing muscle mass of Belgian Blue Cattle. *BMC Genomics* 15: 796 (2014).

Supplemental Material S1:

Effect of dataset/sequencing center on nucleotide diversity.

To ensure that the differences in nucleotide diversity observed between the human (BAM files downloaded from the 1000 Genomes project) and bovine samples (sequenced at the University of Liège - ULg) would not be merely technical artifacts, we compared the nucleotide diversity obtained with the 1000 Genomes BAM files, with those obtained for 10 human samples sequenced at the ULg using virtually the exact same experimental conditions, i.e. chemistry, sequencers and analysis pipeline, as for the bovine samples. The ULg individuals comprised 10 unrelated individuals corresponding to members of families sampled to study a rare form of neurological cancer. They originated from Europe and South America. The sequence coverage for the ULg samples averaged 53.2 fold (range: 43.0 – 67.1). The corresponding human exomes were captured using the SureSelect Human All Exon kit (Agilent). The 1000 Genomes samples were down-sampled to 45.0 fold using GATK "downsample_to_fraction" function.

For each individual, we identified heterozygous positions using GATK and corresponding best practices. Variants were annotated using custom-made scripts and sorted into synonymous and non-synonymous variants. To compare the nucleotide diversity between populations (say A and B) while ensuring that the same exome compartment would be taken into account in the two populations, we only considered variants detected in population A if at least one individual from population B would have a genome coverage ≥ 20 at the corresponding position. Figure 1 shows the number of heterozygous synonymous (A) and non-synonymous (B) positions detected using this procedure when comparing respectively 10 CEU and 10 YRI samples with 10 ULg samples. It can be seen that very similar variant numbers were compiled for the ULg population when confronting it to either the CEU or YRI population, indicating that very similar exome compartments were explored by the CEU and YRI populations. The number of variants that were ignored in the comparisons (because not properly covered by the other population) were 4005, 5690 and 3142/3124 for the CEU, YRI and Ulg (vs CEU/YRI) population respectively. As expected, the number of synonymous and non-synonymous variants detected in individuals from the ULg population overlapped with the corresponding numbers detected in the CEU population, while being inferior to those obtained in the YRI population. Taken together, these results indicate that the observed differences between the bovine and human samples can not be explained by technical artifacts alone.



Legend: Number of synonymous (A) and non-synonymous (B) variant positions detected by exome sequencing in European CEPH samples (CEU: 10) and Yoruban samples (YRI: 10) using BAM files down-loaded from the 1000 Genomes Project (The 1000 Genomes Project; http://www.1000genomes.org/), and in European-ancestry samples sequenced at the University of Liège (ULG-1: comparison CEU vs ULg; ULG-2: comparison YRI vs ULg: 10).

Supplemental Material S2: **Cis-eQTL effects for the three LoF variants predicted to cause nonsense mediated RNA decay in the NZDC population.**

Chr	Position	Gene – Mut.	Ref. allele	BETA	STAT	P value
8	92930920	*RNF20* - SG	T	-0.4591	-15.71	3.05E-42
11	102498942	*TTF1* - SG	A	-0.2396	-8.549	4.19E-16
26	24720154	*OBFC1* - FS	CT	-0.2734	-7.152	5.22E-12

Expression QTL analysis of candidate LoF variants from the New Zealand population was conducted using mammary RNA sequence data and genotypes called directly from the RNAseq alignments. These data represented 406 mostly Holstein-Friesian dairy cows in their second or third lactation, comprising an expanded dataset to that described previously[1]. Briefly, total RNA libraries were prepared and sequenced by NZ Genomics Limited (NZGL; Auckland, New Zealand) or the Australian Genome Research Facility (AGRF; Melbourne, Australia), using 100bp paired end sequencing on the Illumina HiSeq 2000 instrument. Read data were mapped to the UMD3.1 genome using Tophat2[2] (version 2.0.12), yielding a mean mapped depth of 88.9 million read-pairs per individual. Gene expression for the *OBFC1, TTF1 and RNF20* genes was quantified using DESeq[3] (v1.14.0), outputting variance stabilisation-transformed read counts in conjunction with transcript structures defined by the Ensembl genebuild v81. Genome-wide expression outlier individuals were identified using principle component analysis in accordance with published guidelines[4], with 374 quality-filtered animals retained for association analysis. Genotypes were called using Samtools[5] (v1.2), and association testing was performed using PLINK[6] (v1.90). Association models incorporated fixed effects for animal cohort, and covariates to account for population structure using Illumina BovineHD BeadChip genotypes in conjunction with the identity by state and multidimensional scaling procedure implemented in PLINK.

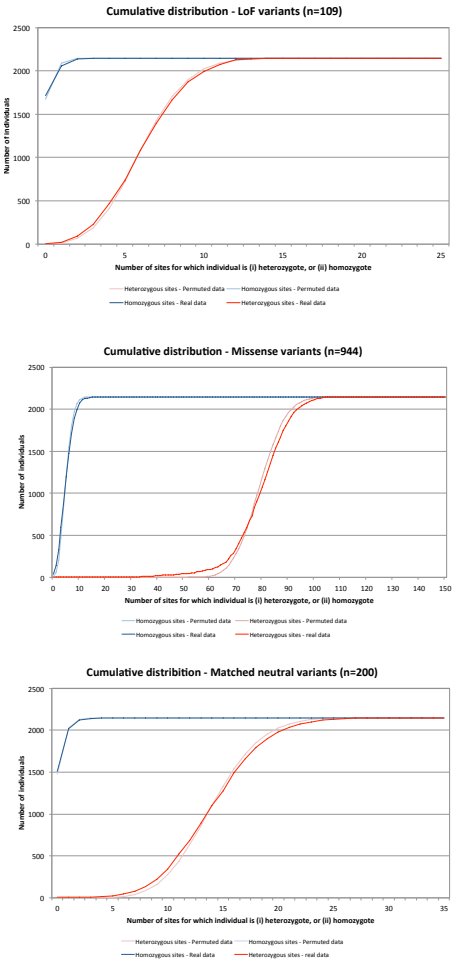
1. Littlejohn, M. D. et al. Expression variants of the lipogenic AGPAT6 gene affect diverse milk composition phenotypes in Bos taurus. *PLoS One* 9, e85757 (2014).
2. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14, R36 (2013).
3. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* 11, R106 (2010).
4. Ellis, S. E. et al. RNA-Seq optimization with eQTL gold standards. BMC Genomics 14, 892 (2013).
5. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–9 (2009).
6. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–75 (2007).

Supplemental Material S3 – Lack of evidence for synergistic epistasis

It has been hypothesized that deleterious variants might be purged from the population by synergistic epistasis, i.e. the fact that multiple deleterious variants have a larger cost on fitness than predicted from their multiplicative effect[1]. This hypothesis predicts that (healthy) individuals carrying multiple deleterious variants will be fewer than expected assuming random assortment.

To test this hypothesis we look at the distribution of the number of individuals that were (i) heterozygote, and (ii) homozygote for *i* genetic variants, where *i* ranged from 0 to the *n* (i.e. the number of genotyped variants). The analysis was conducted separately for "loss-of-function" variants (stop gains, splice site and frame shift variants), missense variants considered by SIFT/POLYPHENE2 to be deleterious/damaging, and matched neutral variants. We used the genotypes of 2,147 BBCB animals generated as part of this study. The distribution for the real genotypes was compared with that obtained by permuting the labels of the individuals (separately for each variant), i.e. by randomizing the genotypes 100 times. For the permuted genotypes, the graphs show the average number of individuals that are heterozygote/homozygote for *i* variants across the 100 permutations.

There was no evidence for a reduction in the number of individuals with the higher number of heterozygous/homozygous sites, whether considering LoF or missense variants, on the contrary (i.e. there were more individuals with large number of heterozygous/homozygous sites with the real than with the permuted data). We observed a slight but significant ($p<0.01$) increase in the variance of the distribution for the real when compared to the permuted data (i.e. there were more individuals on both tails of the distribution with the real versus permuted data). This was observed for the three types of tested variants, including the matched neutral variants. The reason for this systematic increase in variance remains unknown.



We conclude that our data do not provide evidence for synergistic epistasis in this population.

[1] For instance : Keightley PD (2012) Rates and fitness consequences of new mutations in humans. Genetics 190 :295-304.

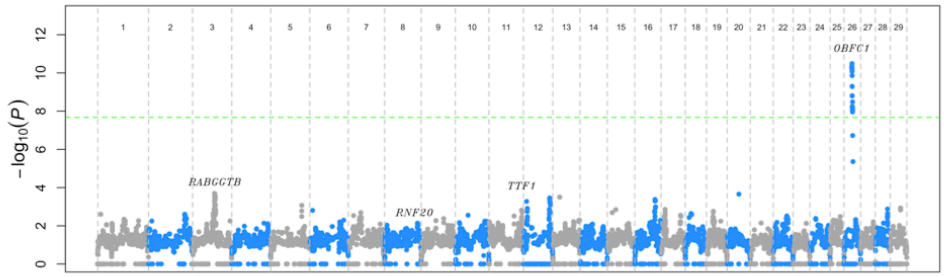
Supplemental Material S4: Haplotype-based genome scan for EL mutations.

We performed a haplotype-based scan to identify regions with haplotypes with significant depletion in homozygotes. We used the data set described previously[1]. It consists in a dairy cattle population from New-Zealand (NZ; mainly Holstein, Jersey and crossbred individuals) including 58,369 individuals genotyped on either Illumina Bovine 50K (v1 and v2) or Illumina BovineHD arrays. We kept markers common to the three arrays and mapping to bovine autosomes (using UMD 3.1 Bovine Reference genome assembly). After checking for parentage errors, we removed markers with a call rate < 95%, generating more than 10 Mendelian inconsistencies, which were monomorphic or strongly deviating from Hardy-Weinberg proportions (p < 1e-8). In addition, we removed 35 small segments that are associated with errors in the genome build[1]. The final data set contained 37,769 SNPs. Remaining Mendelian inconsistencies were erased (removing genotypes in either the offspring, the parent or both).

Haplotypes were first reconstructed based on familial information using LINKPHASE3[1]. The partial haplotypes were further phased (some markers remain unphased) using LD information with DAGPHASE[2] and Beagle[3]. Beagle automatically clusters haplotypes at each marker position based on local similarity using variable length Markov chains as previously described[4].

Regions with putative EL mutations were identified by testing for deviation from Hardy–Weinberg (HW) equilibrium separately for each haplotype cluster (individuals can either carry 0, 1 or 2 copies of a given haplotype cluster). We only considered significant p-values when reflecting a depletion in homozygotes, and when the number of homozygotes for the haplotype numbered < 10.

The results are illustrated in the accompanying Manhattan plot. The genome-wide significance (indicated by green dashed line) was set at p=2.13 e-08 corresponding to Bonferroni corrected p of 0.05 for 2,349,367 tests (performed at 37,769 marker positions each with on average 62 haplotype clusters). Only one region on BTA26 (corresponding to the *OBFC1* EL) showed a genome-wide significant depletion in homozygosity (p = 3.3 e-11). At the most significant position, the haplotype cluster driving the signal had a frequency of 2.7% in the population, while no homozygote individuals were observed. The haplotype-based approach gave no significant signal for the three other ELs detected in the NZ population (*RABGGTB, RNF20* and *TTF1*).



1. Druet T, Georges M. LINKPHASE3: an improved pedigree-based phasing algorithm robust to

genotyping and map errors. *Bioinformatics* 31, 1677-9 (2015).

2. Druet T, Georges M. A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* 184, 789-98 (2010).

3. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84, 210-23 (2009).

4. Browning SR. Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* 78, 903–913 (2006).

Discussion - Perspectives

The focus of this thesis has been the use of next generation, whole genome sequence datasets to investigate the process and characteristics of dnm in dairy cattle. We have also looked at the consequences of dnms in the cattle population by identifying a range of causative de novo and rare variations that are present in the studied populations. Dairy cattle were selected due to their unique population structure, which results from the wide scale use of artificial selection and insemination. This results in a population in which virtually all females in the population contribute to the next generation but only a small number of elite sires contribute via artificial insemination. The population contains relatively few males, with many mates and offspring, providing a large pool of half-sibling that can be drawn from, while females generally bear offspring annually and thus many have four or more living offspring. This combination, allows for the selection of both males and females with multiple offspring and numerous half-sibling grand-offspring, allowing the exploitation of linkage between the half-siblings. Also helpful is the tendency of cattle breeding companies to store biological samples from key individuals in the population long after their passing, especially in the case of semen from key sires. In conjunction with the short (2-6 year) generation time of cattle compared to humans, this improves the availability of three and four generation pedigrees. These consist of at least sire, dam, proband (child) and grand-offspring, with the addition of the grandparents for four generation pedigrees.

Taking this into consideration we created the Damona dataset, of 743 whole genome sequenced Dutch Holstein-Friesian dairy cattle which form 131 trios, for use in the study of the fundamental aspects of genome biology such as dnm and recombination. To study dnms we must first identify them, and the key advantage of WGS data compared to previous methods, is that it allows us to directly detect a large proportion of the dnms in any one individual via the use of trios. When utilising sequence data to directly identify dnms in trios, we identify candidate variants by their presence in the proband's (or child's) DNA and absence in the parental DNA, as well as unrelated individuals or known variant datasets. Most WGS based studies have utilised this method on simple two generation pedigrees which consist of both parents and the child (proband). This allows for easy identification of candidate dnms but does not clearly differentiate between somatic or germline dnms. This is potentially problematic in studies that utilise DNA from established cell-lines (as some human studies have), where many generations of replication and selection can lead to the formation of somatic variants, which by chance gain allelic dosages similar to those from true dnms. Further the use of two generation pedigrees, limits the number of dnms that can be assigned to the germ-line of origin, to those that share a physical DNA fragment (sequence read) with an informative marker that is present in only one of the two parents.

Also due to the cost of whole genome sequencing, there are limits the amount of sequence that will be assigned to a single genome. A consequence of this and the limitations of the current sequencing technologies, is that the depth of coverage for each site in genome can differ markedly and thus our sensitivity for detecting heterozygous variants is not uniform across the genome. Thus, a small

percentage of the sites in a genome which contain a heterozygous variant may be missed by chance, due to insufficient sequence depth. When this occurs in a proband we may have a false negative and miss the dnm, however the small number of dnms expected means this has little impact on our ability to identify dnms. However, when this happens in a parent and a variant which is heterozygous in one parent is misidentified as homozygous reference it can result in the creation of a false positive dnm if the second parent is also homozygous reference. When applied to the millions of variants present in each parent's genome, this results in the creation of numerous false positive dnms.

Several of these issues can be address via the use of additional generations in a pedigree. The first and simplest change is the addition of offspring of the proband to the trio to create a three-generation pedigree. The addition of such individuals allows us to restrict our analysis to germ-line dnms, by requiring the inheritance of any dnm by the offspring, this confirms the dnms presence in the probands germ-line. This will prevent over estimation of the dnm rate by removing somatic dnms. This also improves our ability to identify the germ-line in which the dnm had occurred. If we first utilise the parents to phase the probands variants, then by comparing the offspring's genomes with the phased proband's genome, we can drop the phase information down and determine if the dnm is in linkage with variants inherited from the probands sire or dam. If the dnm in an offspring's genome is flanked by variants inherited from the sire then the dnm occurred on the chromosome inherited from the sire. One downside of requiring inheritance of dnms, is that each offspring only receives half the probands genome, and thus we will on average be discarding half the dnms present in the proband. This can easily be resolved by utilising multiple offspring of the proband, with the power to detect dnms being described by the function $p = 1 - 0.5^n$ where p is the power to detect and n the number of offspring of the proband. Thus, if we use five or more offspring our power to detect a dnm will be greater than 95%. These additional offspring also provide a second advantage, in that they allow us to split our dnms into two different classes, those inherited from the parents (sire or dam) and those that occur during the development of the proband. As previously discussed, dnms can occur at any time during an organism's development, with those that occur in the early stages of embryo development being difficult to differentiate from those that occurred in the parental germline. At least when limited to simple two generation trios. However, with multiple offspring it is possible to differentiate parental inherited dnms which show complete and perfect linkage. From those mosaic dnms that occur during the embryonic development of the proband, which show incomplete but perfect linkage. With multiple offspring the power to detect mosaic dnms is $p = 1 - (1 - AD)^n$ where AD is the true allelic dosage of the mosaic dnm and n is the number of offspring. It is worth noting however, that as a consequence of this our power to detect dnms occurring in the probands genome will be substantially lower than that of the dnms that occurred in the parent's germ-lines, as less than half the children will inherit the dnms.

These two advantages, confirming the germ-line nature and identifying the germ-line of origin for a dnm are substantial and key to our studies, thus all 131 trios in the Damona dataset have grand-offspring

available with 116 having five or more. This combination provides a high degree of confidence in the dnms we identify, as well as allowing us to determine the germ-line of origin for virtually every dnm. Furthermore, utilising this we can classify dnms as early or late occurring and investigate differing mutation patterns between the early and late stages of embryo development and gametogenesis.

The addition of a fourth generation to a pedigree also offers some additional advantages, for the Damona dataset these additional individuals are the grand-parents of the proband. The first benefit of adding the fourth generation is that they further improve our ability to detect and remove ancestral rare variants that might otherwise be mistake as dnms, if by chance they have not been identified in the parents of the proband. As a large proportion of the variants in any one individual are rare variants (minor allele frequency < 1%), they may only be present within the sample population in one parent, the proband and the probands offspring. Thus, they are at high risk of being mistaken for dnms, if by chance insufficient alternative alleles in are observed in the parent to identify them. With the grandparents available there is a second opportunity to observe such variants, drastically lowering the chance that they will be missed twice by chance. For example, assuming a site that is heterozygous on the paternal haplotype in the proband is covered at 20x in the sire, the chance of missing the alternative allele and mistaking the variant as a dnm (assuming a binomial distribution) is 9.54E-7, however if the site was at 10x in the sire the probability would be 9.8E-4. If approximately 5% of the genome was covered at 10x then this could correspond to numerous rare variants being mistaken as dnms. With the addition of the grandparental generation the probability becomes 9.54E-7 for 10x or 9.09E-13 for 20x resulting in several orders of magnitude reduction in the number of false positives. Secondly, the grandparental generations can allow for the detection of early mosaic dnms in the parental generation. Typically if a variant is heterozygous in the proband and present in the associated parent at a very low allelic dosage (< 10%), it is most likely that the variant is an inherited heterozygote, that by chance we only sampled 2 of the expected 10 alternative alleles (Binomial p = 2E-4), rather than a dnm (i.e. p = 15/3E7 = 5E-6) that had occurred within the first 4-5 cell divisions of the sires embryo. However, when the grandparents are available we have a second chance to check to see if the variant is present in the inherited haplotype. Thus, if we observe a variant that is heterozygous in the proband and grand-offspring, has a low allelic dosage in the associated parent and is absent in grandparents it is more likely to be a dnm in the parent. Rather than to have been inherited from the grandparent and by chance unobserved in grandparent and at exceedingly low allelic dosage in the parent. Based on this, in four generation families we are able to identify likely parental dnms that occurred relatively early in the parent's development. Thus, for four generation pedigrees we will be able to obtain a more accurate estimate of the number of dnms inherited from each parent, by being able to identify dnms from both the early and late stages of parental development. This, in addition to the benefits gained from utilising three generation pedigrees, allows us to identify a greater proportion of true dnms from both early and late in the development of the parents. To identify dnms that occurred during the development of the proband rather than its parents,

and accurately determine the germ-line of origin for all dnms, while reducing the dnm false positive rate, and removing all somatic dnms.

Aside from the use of relatives to identify dnms, unrelated individuals are also utilised to reduce the number of false positives. Unrelated individuals help account for the relatively high rate of systematic errors in Illumina sequence data as well as biases resulting from the Illumina sequencing chemistry and bioinformatics. These can combine to generate sporadic low-quality variants within the population of which a proportion will be mistaken dnms due to issues in generating reliable genotypes from them. As these result from systemic biases of the sequencing technology the use of unrelated individuals or lists of known variants as a filter will substantially reduce the false positive rate of dnms by removing variants that are common within the population and are thus likely to have been inherited from the parents even if they are not observed by chance in the parental DNA or are the result of systematic sequencing errors and thus found sporadically throughout the population.

One further source of error when identifying dnms, is variants called from collapsed repeats in the reference or other highly repetitive regions. Such regions of the genome can cause issues for short read alignment software resulting in the misplacement of reads or the alignment of a read to multiple locations. With current Illumina sequence the combination of paired reads and read lengths of greater than 100bp allow the unique mapping of the vast majority of reads. For reads with uncertain mapping positions, alignment software such as BWA MEM provide mapping quality scores which aim to quantify some of the uncertainty around the quality of the read mapping. These mapping quality scores can be utilised to filter candidate dnms in regions of the genome where the mapping is likely to be inaccurate and thus variants within this region are unlikely to reliable.

After taking these factors into consideration we developed denovoPedFilter a VCF based dnm identification pipeline (https://github.com/aeonsim/denovoPedFilter). Our software performs an exhaustive search of the supplied pedigree identifying all two, three and four generation pedigrees. For each trio we then divide the supplied individuals into six classes, these are grandparents, parents, direct offspring (of the proband), descendants (all individuals with the proband as an ancestor), extended family (other descendants of the sire or dam) and unrelated individuals. All trio members are then phased following mendelian rules of inheritance on fully phase informative markers and phase information is then dropped down to the grand-offspring based on homozygous variants in the grand-offspring. We then evaluate every variant in the VCF with a QUAL score of greater than 100 (based on GATK best practises recommendations), depth of coverage between 10x (autosome, 5x male X chromosome) and 2x the individuals average depth. Candidate dnms are identified by looking for variants that have been genotyped as homozygous reference in both parents but heterozygous (in autosomes) or homozygous alternate (in male proband's X chromosome) for each trio. If any trio shows this pattern we evaluate each of the probands offspring and record if they carry an alternative allele at the position, regardless of their called genotype. The genotype of every other individual is then evaluated relative to the trio and the number of heterozygous or homozygous alternative individuals is recorded against the appropriate

relationship category for the specified trio (grandparents, descendants, extended family or unrelated). Variants that show the expected pattern of inheritance, have evidence for their existence in the probands offspring and are absent in unrelated individuals are then considered as candidate dnms. For each candidate dnm the probands offspring are then analysed to determine which parental haplotype they received from the proband, and the germ-line of origin of the candidate dnm is determined based its linkage to either the maternal or paternal haplotype. Each candidate dnm is then reported along with the supporting data, the dnms germ-line of origin, it's trinucleotide context and additional statistics.

For our pilot study, this process was applied to the five families and then all candidate dnms were manually evaluated in IGV to confirm phasing, quality of the mappings and the accuracy of the genotypes within the pedigree and population. The resulting confidant set of candidate dnms were then validated via amplicon sequencing, showing a high level of specificity. For our second study utilising 131 trios, the increased size of the dataset made manual evaluation of all candidates impractical, leading to additional bioinformatics filtering to select high candidate variants.

Three primary filters were utilised for the complete dataset, firstly all candidate dnms were recalled via an alternative variant caller (freebayes) using 200bp windows centred on the dnm. After normalisation of variants and their position, any candidate that was not shared by the two variant call sets (GATK and Freebayes) was discarded. The second filter was based on the average allelic dosage of the proband's heterozygous descendants. Assuming a variant is a true dnm in the proband then the descendants of the proband carrying the variant should be heterozygous, with an average allelic dosage of 50% for the variant. Thus, variants with an average allelic dosage significant different from 50% (binomial test) in the grand-offspring were filtered as being enriched for systematic errors. Thirdly, variants with a map quality score of less than 50 were removed from the dataset. The combination of these three filters when applied to the five families from the initial study reduced the number of false positives to ~0 while retaining 95% of the validated dnms.

The combination of these three filters with the pedigree, unrelated individual, quality score and read depth filtering selected a dataset that retained ~95% of the validated variants, with per trio dnm rates similar to that observed for the validated pilot study. Candidate dnms derived from this filtering and selection process are restricted to germ-line dnms that have showed clear transmission to at least one additional generation. Further, for all variants, excepting those few falling directly within a recombination window, the germ-line or origin for the variant has been identified and the variant can be classified as late (non-mosaic) or early (mosaic). Such a dataset provides powerful resource for investigating the process of dnm in the germline and early stages of embryo development.

However, this combination of selection and filtering does have some downsides. By requiring transmission of dnms to the next generation we do lose a proportion of dnms. With five offspring the probability of transmission for a heterozygous variant in the proband is approximately 97%, assuming the sensitivity of the offspring is approximately 100% for detecting a heterozygous variant. However, for our dataset many offspring have an average depth of between 5-10x which gives an actual

heterozygous sensitivity of 81-99% (estimated from a binomial distribution). For a five-offspring family, each with an average sequence depth of 5x the overall power to detect variant that is heterozygous in the proband is approximately 92%. This reduction in power must be accounted for in estimating the dnm rate, by applying a per family adjustment based on the number of offspring and their average depth of coverage. A similar correction is needed to correct for the percentage of the genome that is considered, as a result of the depth of coverage and mapping quality filters. This is estimated during the dnm identification process by tracking the percentage of SNPs in each trio that meet the filtering requirements compared to those that fail them and is reported by the software. For the 131 trio dataset the use of Freebayes as a second pipeline to confirm the dnms, has little effect on SNPs or small indels, due to the previous filtering removing complex or difficult to call regions of the genome. Rather larger or more complex variants are lost, however these are not the focus of this study and thus can be ignored for now. The final issue with this approach for identifying dnms is that the requirement that the variant be absent in unrelated individuals does not allow for the detection of recurrent dnms. Naively it would be assumed by chance that in a 3GB haploid genome the chance of dnms occurring at the same site in differing individuals would be incredibly low. However, when we consider that the rate of dnm is not even across the genome but may vary by one to two orders of magnitude, depending on local base context, combined with a dataset that consists of over one hundred families it becomes increasingly likely that some recurrent dnms will exist and be lost via this approach. To allow detection of recurrent dnms our software allows disabling the requirement that the dnm be absent in unrelated individuals, instead relying on the pedigree structure and other filters to identify likely dnms. Compared to two generation pedigrees our three and four generation pedigrees provide considerable improvement in specificity by requiring the transmission of the variant to the next generation. While the four generation pedigrees provide further improvements to specificity by requiring two ancestral generations to lack the candidate dnm identified in the proband. The combination of these additional generations with the filters described should allow for the reliable detection of recurrent dnms, at a reasonable level of specificity. In addition, it may be possible to utilising linkage and analysis of the phased genomes to identify the specific haplotypes surrounding each dnm and thus identify recurrent events by their occurrence on different haplotypes. Further work is needed to adjust the pipeline and filtering to identify recurrent dnms with a decent level of specificity.

Taking this into consideration we utilised the Damona dataset to characterise the rates and properties of dnm in dairy cattle. Our initial study focused on a subset of the population consisting of five pedigrees (two three and three four generation pedigrees). We estimate that the rate of dnm in cattle is approximately 0.9×10^{-8} per bp per generation or 0.18×10^{-8} per bp per year, when identifying dnms in the same manner as the human studies. This rate is slightly lower than the average of 1.2×10^{-8} from human studies (Campbell and Eichler 2013; Ségurel et al. 2014) but is within the range reported. We also observed the expected 2.5:1excess of paternal dnms compared to the dnms originating from a maternal

germ-line. This is substantially lower than the 3.9:1 rate observed in humans or the 5.5:1 rate in chimpanzees. The paternal bias in dnm is thought to be a consequence of the additional 23 (human) cell divisions that occur yearly post-puberty during spermatogenesis in mammals, under the assumption that most dnms occur during DNA replication. Considering that the average generation time is 4-5 years in the cattle we studied (compared to 30 and 25 years respectively in humans and chimpanzees), it is reasonable to expect a lower paternal to maternal ratio.

We then looked at the additional generations available in our pedigrees to identify gonosomal and germ-line mosaic dnms, that would have occurred during the development of the sires, dams and probands. We utilised complete but imperfect linkage among the grand-offspring and probands half-siblings, as well as allelic dosages with significant departures from the expectation of 0.5 (alternative allele / total reads) for a heterozygous variant to identify mosaic mutations in both the probands and parents. We observed that ~30% of the dnms in a sperm cell and 50% of those in an egg are detectably mosaic in the parental DNA. The actual number of mosaic dnms in either a sperm cell or egg are similar (14 to 12). With the large number of late, non-mosaic mutations occurring during spermatogenesis, lowering the proportion of dnms the mosaics account for in a sperm cell compared to an egg. Moreover, approximately 17% of the detected dnms in these families occurred within the germ-line of the proband rather than that of its parents. The differences in the proportions of mosaic dnms detected in the proband compared to the gametes of the parents is due to two factors. The first is that we detect different sets of dnms in each case for the PM dnms we only detect dnms that have a high enough allelic dosage in the proband (> 10-25% depending on depth of coverage) to be called as possible heterozygous variants by GATK. While for the parental mosaics (DM + SM) we only identify dnms where the allelic dosage is below the threshold to be called as heterozygous (< 10-25% depending on the coverage). Secondly for PM dnms we are identifying all the mosaic dnms that have been inherited by at least one of the five children. While for parental mosaics (DM + SM) we are identifying the number of mosaics present in a single egg or sperm. With the addition of the parental mosaic dnms, and the removal of the proband mosaic dnms, the dnm rate in cattle is 1.2×10^{-8} per bp per generation, with a 2.4:1 paternal to maternal ratio. This represents a ~30% increase in the cattle dnm rate and should this apply to humans the dnm rate would increase to $\sim 1.6 \times 10^{-8}$ per bp per generation. This higher estimate of the human mutation rate is remarkably similar to that reported by the Genome of the Netherlands project of 1.66×10^{-8}, which utilised a multigenerational approach that is inclusive of mosaic dnms (Palamara et al. 2015). Though it is still substantially lower than the rate of 2.5×10^{-8} estimated by phylogenetic methods (Nachman and Crowell 2000). The average transition-transversion ratio (Ts/Tv) for the five families was 1.33 for the dnms, substantially below the expected 2.1 from SNPs segregating in the same population. This was primarily due to an excess of C>A dnms present in two of the pedigrees. For the complete dataset of dnms from the 131 trios, the overall transition-transversion ratio is 1.96, close to expectation. Additionally, the number of dnms per family varied substantially suggesting possible interindividual

variation in the rate and signatures of dnm. Further, the degree of mosaicism observed was incompatible with the standard model of embryo development, when assuming a constant mutation rate. Simulation of embryo development and gametogenesis suggests that the observed level of mosaicism in the five families is most compatible with a 20-fold increased rate for the first four cell divisions after fertilisation. Thus, suggesting that the early stages of embryo development are particularly error prone.

In our second study, we expanded our analysis to the complete 131 pedigrees in the Damona dataset identifying 7,498 dnms. We estimated the average rate of dnm to be $1.21x10^{-8}$ per bp per generation (95% CI, $1.15\text{-}1.34x10^{-8}$) when accounting for mosaicism, and $\sim0.8x10^{-8}$ (95% CI, $0.73\text{-}0.88x10^{-8}$) when not. The degree of mosaicism was comparable to our pilot study with 37% and 55% of the dnms in a sperm or egg being mosaic. However, there was substantial variation in the dnm rate between families with an ~9x difference between the minimum and maximum rates. The majority of the variation between families was due to mosaic variants. With a coefficient of variation of greater than one for mosaic mutations, compared to ~0.5 for non-mosaic dnms. We identified a possible environmental effect on mutation rate, as we observe a significant difference in the average number of proband mosaic dnms associated with different reproductive technologies. For artificial insemination, an average of 1.9 proband mosaic dnms were observed compared to 2.7 for MOET and 4.5 for in vitro fertilisation and maturation. We identified four outlier animals with 5-17x the average number of mosaic dnm and differing signatures of mutation compared to the population. The first outlier, with a 17x increase in the number of mosaic mutations, presented an eight-times increase in C>T mutations outside of the NpCpG trinucleotide context. Assuming that this phenotype is genetically determined, we identified two candidate causative mutations. The first is a unique homozygous predicted deleterious missense mutation in the proband's REV1 polymerase that is involved in translesion repair. The second is a maternal heterozygous 140Kb deletion of the *TFB1M* gene, (mitochondrial transcription factor B). The *TFB1M* gene is critical for mitochondrial function and has been reported to show a possible haploinsufficency effect (Koeck et al. 2011). Due to the critical importance of mitochondria in early development and the reliance on maternal genome products before the zygote genome activates, this variant could potentially effect mitochondrial efficacy. Such a reduction could in turn reduce the energy budget for the zygote reducing the efficiency of DNA repair or increasing the oxidative stress for the zygote. The second outlier had a 5x increase in mosaic dnms, with a mutational signature dominated by a 4.5x excess in C>A/G>T mutations, compared to the general population. It is noteworthy that cytosine to adenosine mutations are associated with DNA damage in the form of 8-oxo-guanine (van Loon et al. 2010). No candidate causative events could be identified for this family. This outlier was present in the pilot study as the sire of probands three and four, and was primarily responsible for the lower than expected Ts/Tv ratio of 1.33. The third outlier with a 5x increase in mosaic dnms, presented a third distinct mutational signature, with an excess of tandem mutations CC>TT, CT>TC, CC>GT and CG>TA along with an excess of C>T mutations. This mutational signature resembles that observed from UV damage (Sinha and Häder 2002). For the fourth, a ~5x increase in DM mutations was observed,

with a mutational signature that showed a 2x increase in C>T mutations outside of the CpG context. Together the per pedigree variation in the rates of dnm, and the four outlier pedigrees with their distinct mutational signatures, suggest considerable interindividual variation in the rate of dnm is present in the cattle population. The distinct mutational signatures for the four outliers suggests that possible mutator alleles are present and active in the cattle population. In addition, the larger number of dnms in this study allowed us to begin the initial characterisation of dnms in the cattle germ-line. We were able take the categories of dnms identified in our pilot study and show that there are substantial differences in the mutational spectrum, and the rate at which they occur in the male and female germ-lines between the early occurring PM, SM and DM dnms and the late occurring SNM and DNM dnms. This suggests that the rates, properties and possibly mechanisms of dnm differ depending on the stage of development at which they occur.

We then sought to identify additional classes of dnm, beyond the typical SNPs and small INDELs, by developing a pipeline to find polymorphic transpositions of endogenous retroviruses (ERVs) and long interspersed nuclear elements (LINEs). Utilising this pipeline we discovered ~1,600 polymorphic ERVs and ~8,000 polymorphic LINEs in the Dutch dairy and Belgian Blue cattle populations. We identified the de novo transposition of five ERVs belonging to the ERVK family and one LINE within the Damona dataset, for estimated transposition rates of 1 per 50 gametes (ERVs) and 1 per 260 gametes (LINE). Three of the five ERVs occurred within the germ-line of one individual, with two of the five being present in the same sperm, suggesting considerable interindividual variation. Interestingly, after assembling the complete ERV sequence for the five cases, we observe numerous mutations present within each ERV. The mutations were sufficient to compromise the function of all the ERV proteins, and the pattern of mutations differed between the de novo ERVs, including those that occurred in the same germline. As such it seems likely that the de novo ERVs were not in of themselves fully functional, requiring external assistance in the form of functional machinery provided by a different ERV. Once the functional enzymes were present they did not copy a single ERV multiple times, but instead allowed the transposition of several different full length ERVs in the genome. The presence of de novo transpositions and breed specific polymorphic ERVs suggest that ERVs are still active in the bovine genome, and are likely to have phenotypic impacts.

Looking at ERVs present in genes we identified a transposition of a full length ERV-K element into exon 5 of *Apolipoprotein B* (*APOB*) gene causes premature transcriptional termination. The transcriptional shutdown of the gene was determined to be responsible for the Cholesterol Deficiency Syndrome (CD) in several Holstein-Friesian populations sharing a common ancestor. The causative variant is estimated to have a MAF of ~2% and was traced back eight generations to the Canadian bull *Maughlin Storm* (born in 1991). The mutation is unique to its descendants. Thus, the de novo transposition was likely to have occurred in one of its ancestors or during its own embryonic development. Assembly of the full length ERV revealed that like the de novo events, the genes present

in the ERV were likely to be inactive, while the pattern of variants present indicated it was unique compared to the other de novo events.

We also observed evidence of selection against ERVs, with a lower MAF for ERVs that were inserted in genes in the same (concordant) orientation as the gene, compared to ERVs that inserted into the gene in a discordant orientation, as well as intronic and intergenic ERVs. With a notable exception, a polymorphic ERV in the *AGBL4* gene had a MAF of 35%. This gene has been identified as being under positive selection in cattle (Flori et al. 2009; Aliloo et al. 2015). These data suggest that ERVs of at least the ERV-K family are currently active in the bovine genome and are a source of functional variation. The non-functional nature of the genes in the five de novo ERVs and the CD ERV suggest that the state of a specific ERVs genes does not limit its ability to spread, providing there is an alternative source of the enzymes available. The complex pattern of SNPs and INDELs present in the de novo ERVs show that they do not originate from a single function ERVK element, but instead are derived from multiple different copies of ERVK that are present in the bovine genome. The presence of three de novo ERVs in one germ-line, with two events being present in a single sperm cell, suggest that ERV transposition can happen in bursts. When the suppression of ERVs is temporarily inactivated, functional copies of the ERV enzymes will transpose multiple different full-length elements.

From the Cholesterol Deficiency study, we observed a rare or de novo ERV transposition with a deleterious phenotype that nevertheless reached a MAF of ~2% in 8-9 generations. This shows that under the current dairy cattle population structure and breeding model it is possible for dnms to rapidly increase in frequency, even in cases where they are deleterious. A study in the New Zealand cattle (*Bos taurus*) population provides a second example of this. In 2011, incidents of unusually hairy, heat intolerant, dairy cattle that failed to milk were reported in New Zealand. Genome wide transmission disequilibrium testing identified a non-synonymous variant in exon 5 of the prolactin gene. The p.Cys221Gly substitution results in the loss of one of the three disulphide bridges in the prolactin hormone, altering its three dimensional structure. Genotyping of 2,205 progeny of two affected sires showed complete concordance between the variant and affected individuals. This deleterious dominant variant was determined to be a dnm in the sire of the elite bull Matrix, with semen from the two bulls having produced greater than 6,000 offspring. The phenotypes resulting from this mutation are the direct opposite of those reported in ‘slick’ cattle such as Senepol, who show short hair length, increased heat tolerance and increased milk production. Identification of the variants responsible for this phenotype has been of considerable economic interest, with previous studies having identified a single dominant locus as being responsible, but not a causative mutation. This ‘slick’ locus on chromosome 20 includes the prolactin receptor gene. Sequencing of the *PRLR* gene identified a premature stop codon (p.Leu462*) in exon 10 of the gene. Genotyping and phenotype analysis of additional Senepol animals confirmed a significant association between the variant and the slick phenotype, while additional screens of slick cattle identified no other causative variant in the region. The prolactin study demonstrates two aspects of dnm in cattle. First due to the population structure of cattle, where a few elite bulls act as sires for a

large proportion of the population, it is possible for dnms to have a large immediate impact. In this case, two bulls gave rise to over 6,000 offspring in a couple of generations, of which approximately 2,000 were affected. Due to the dominant nature of the mutation and the noticeable phenotype, the variant was identified relatively quickly, allowing for its control and elimination from the population. In the case of a recessive event, the phenotype would not have been observed at such an early stage, allowing numerous additional generations for the variant to have increased its MAF within the population. The rapid increase in the MAF of a dnm can be greatly accelerated by the continued selection of elite sires from the progeny of the initial bull, due to their own 'elite' nature. While deleterious dnms are commercially undesirable, they can act as a source of natural 'knockout' models when they disrupt genes resulting in new phenotypes. These natural 'knockouts' can in turn provide new insights into the biology and importance of specific genes and genetic pathways and phenotypes. These new insights into the underlying biology can in turn be utilised to drive genetic improvements within the cattle population.

In our next study, we investigated four cases of newborn Belgian Blue calves presenting a severe form of epidermolysis bullosa, a skin fragility disorder. Genotyping of the four cases, with an Illumina 50K SNP chip, and comparison with unrelated controls, identified a unique identical by descent (IBD) 8.3Mb homozygous region. Whole genome sequencing of one individual, and intersection with RNA-Seq transcriptome data from fetal skin, identified an exonic G>A substitution in the laminin gene, alpha 3 (*LAMA3*). This substitution creates a premature stop codon in exon 60 of *LAMA3*, resulting in the truncation of 22% of the corresponding protein and the likely non-sense-mediated decay of the associated mRNAs. Homozygous LOF mutations of this gene in human, mice, horse, sheep and dog have been reported to cause severe junctional epidermolysis bullosa, supporting the causative nature of the mutation. Custom genotyping of 3000 Belgian Blue cattle estimated the carrier frequency to be ~1%. Although the *LAMA3* variant is present in ~1% of the Belgian Blue population, incidences of epidermolysis bullosa in other cattle have been associated with differing genes (Peters et al. 2015; Pausch et al. 2016) suggesting that the *LAMA3* mutation is restricted to Belgian blue. Further, the relatively large size (8.3Mb) of the homozygous IBD region, suggests the *LAMA3* variant originated, or was present within a shared ancestor relatively recently. Taking these factors into account and its presence in 1% of the population, this may be an additional case where a dnm has rapidly spread. With the spread being driven by the widespread use of artificial insemination with a small number of elite sires. Further, the recessive nature of the resulting phenotype allowed the variant to initially avoid purifying selection, allowing its spread to ~1% of the population. This is unlike the *PRL* mutation where the dominant phenotype lead to its rapid discovery and elimination. While the dominant *PRL* event may have had a greater initial impact on the population, the long-term effect of the *LAMA3* may be greater.
In our final study, we focused on the rare recessive variants present in cattle populations and their effect on fertility. The *APOB*, *PRL*, and *LAMA3* variants were all identified as the result of forward genetic screens, in which detectable phenotypes (Cholesterol deficiency, Hairy Syndrome, and Epidermolysis Bullosa) were utilised to determine the causative variants and genes. For this study, we utilised a reverse

genetic screen in the New Zealand Dairy cattle (NZDC) and Belgian Blue Breed (BBB) populations, with the aim of identifying Embryonic Lethal variants present in the population. This was achieved by identifying likely causative variants from sequence data and then determining the associated phenotype. This approach utilised exome sequence for ~500 NZDC and 80 BBB to identify 1377 stop-gain, 3139 frame-shift, 1341 splice-site, 22,939 disruptive missense, 62,399 benign missense and 92,163 synonymous variants. For cattle, the average number of heterozygous synonymous and nonsynonymous variants was ~13,200 and ~5,900 respectively. This is higher than the ~9,000 synonymous or ~5,400 nonsynonymous heterozygous variants observed in Yorubans, or the ~6,300 and ~4,000 observed in Asian and European individuals. The average number of LOF variants observed in cattle was 51 heterozygous and 7 homozygous compared to 58 and 9 for humans. We then selected frame-shift, splice-site, stop gain, predicted deleterious missense mutations and breed-specific variants with no homozygous individuals in the sequenced population. The resulting 3,779 NZDC and 1,050 BBB variants were then genotyped in ~35,000 NZDC and >6,300 BBB animals utilising Illumina custom SNP arrays, alongside 200 BBB specific synonymous variants as matched controls. Genotyped variants were then tested for a depletion in homozygotes for the minor allele compared to Hardy-Weinberg expectation. In both populations, we observed LOF and missense variants with no homozygous mutant animals despite population frequencies of >1.3% for NZDC and 1.8% for BBB. In comparison, there were no synonymous variants in the same population at greater than 2.2% without homozygous mutants. The proportion of LOF variants without homozygotes was 0.348 compared to 0.228 for matched synonymous variants. For missense variants, the proportion without homozygous mutants was 0.233 and 0.185. Thus, the estimated proportion of embryonic lethal variants was 15.5% of the tested LOF variants and 5.9% of the tested missense variants. For nine candidate embryonic lethal variants with a MAF 1.2-6.6%, 200 carrier x carrier matings where examined. No homozygous mutant offspring were observed, supporting embryonic lethal nature of the nine variants. The nine embryonic lethal variants are estimated to effect ~0.64% and ~0.61% of conceptions in the NZDC and BBB for an estimated cost of ~9 and 2.7 million Euros respectively. These embryonic lethal mutations are examples of strongly deleterious variants that are unique to specific populations and have successfully reached MAFs of >1.15%. Three of the variants are effectively common variants within their population having reached MAF's of ~5% or higher. All nine are also specific to individual breeds. This, combined with their homozygous lethal nature, suggests they are either relatively recent dnms or exceedingly rare ancestral variants. In either case they provide examples of rare variants that have accidentally been selected for by the use of elite bulls as artificial insemination sires, resulting in their rapid increase in frequency within the population. They demonstrate the cost that rare deleterious variants can have when accidentally selected for in domesticated species where a small number of elite individuals contribute to the majority of conceptions in each generation. In addition, a key conclusion that can be taken from this study, is that it is possible to take complex phenotypes, such as fertility, and use a reverse genetic screen to unravel parts of the contributing biology.

This thesis has focused on the properties and characteristics of dnms in cattle, for SNPs, INDELs and ERVs. We have made advances in understanding the interaction between mutation rate, embryo development and gametogenesis. The use of three and four generation pedigrees has allowed us to show that the rate and characteristics of dnms varies depending on the stage of embryo development. With a substantially higher mutation rate for the initial stages of embryo development resulting in the formation of mosaic dnms. This agrees with the findings of several recent human studies that suggest the dnm rate may vary over three stages of development, early development (pre-PGC formation), late development (post-PGC formation) and post-puberty. Further, there appear to be significant differences in the mutational signatures of early dnms compared to late dnms and that early occurring dnms show greater similarity between genders than those occurring later. This suggests that the early dnms occur before sexual differentiation and thus have similar rates and characteristics in both the male and female germline. Simulations suggest the numbers of dnms observed in the early stages of development are incompatible with a model of embryo develop that assumes a constant rate of dnm through development and that rather the mutation rate in the first 4-6 cell divisions is likely greater than 10x that of the later stages of development. While we observe a higher initial dnm rate in cattle than has been observed in humans, the increased power for detecting early dnms from our use of three and four generation pedigrees, compared to the two generation pedigrees in human studies, may partly explain it. Also we would note there is considerable inter-individual variation in the early mutation rate in both our cattle dataset and the small number of human studies (Dal et al. 2014; Rahbari et al. 2016). This inter-individual variation may partly be explained by the use of reproductive technologies which appear to increase the early dnm rate in cattle. If this is the case then the substantially higher use of such technologies in the cattle population (>50% of the damona dataset) compared to the human population may explain part of the difference. An additional contributing factor is the presence of genetic effects from mutator loci in the cattle population with four outliers showing a 5-17x increase in the number of early dnms. For one of these outliers we were able to identify two candidate causative variants that are currently being investigated. This suggests that the rate and patterns of dnm are not static but continually evolving.

Our work suggests that both the environmental and potential genetic effects are primarily occuring during the early stages of embryo development and significantly increases the number of early dnms. Such dnms occurring early in embryo development will tend to be mosaic in nature, being present in a subset of the soma and germ-line (1-30%), and thus are likely to be inherited by multiple offspring. Consequently, such mosaic dnms are likely to have little or no effect even when dominant in the originating individual, due to their restriction to only a small proportion of the cells forming the soma. However, their offspring that inherit the variant will receive the full phenotypic impact of the variant (if dominant), and due to the variant being mosaic, upto 30% of the individuals offspring may be affected due to being fully heterozygous for the variant. For dairy cattle were a single artificial insemination bull

can have thousands of offspring, this can produce large numbers of affected offspring within one generation. For recessive variants, mosaicism also substantially increases risk as allows the variant to rapidly spread into the population compared to a non-mosaic dnm. For example a deleterious mosaic dnm in a bull with 10,000 offspring could result in 3,000 carriers, well a non-mosaic dnm occurring in that same bull would only have one carrier. Thus, the non-mosaic dnm will be highly susceptible to genetic drift, while the mosaic dnm is like to be far more resistant to genetic drift due to the many carriers.

When looking at the overall dnm rate, the cattle dnm rate is similar to that reported for humans and chimpanzees of approximately $1.2x10^{-8}$ bp/gen. As this is the third large mammalian species to have a WGS based estimate of the dnms rate, with a similar genome size, but differing generation time and having diverged from humans and chimpanzees around fifty million years ago it may provide a useful outgroup for investigating the evolution of dnm.

We have also demonstrated the impact dnms or rare variant can have on cattle populations, due to their presence in a small number of highly reproductively successful individuals. The combination of artificial selection and artificial insemination can rapidly propagate strongly deleterious variants into such populations, at considerable economic cost. With the current level of accuracy in whole genome sequencing and its lowering costs, we would argue that it is potentially time to consider developing reverse genetic screening programs for dnm. Especially for domesticated species such as cattle, where a small number of individuals contribute to the majority of the progeny. Whole genome sequencing of the candidate sires and their parents would allow the identification of the majority of any new deleterious mutations in the sires' germ-lines. This is especially critical with the recent widespread use of genomic selection, which has reduced the generation time for cattle breeding from four - six years to two - three years. This reduced generation time has significantly increased the risk of dominant mutations reaching the population, due to the lack of phenotypic proofing of the sires' offspring before it's widespread use as an artificial insemination sire. Especially in the case of dominant deleterious mosaic dnms for which the sire may be asymptomatic, or dnms where the primary deleterious phenotype is specific to females, such as the case for the PRL mutation and its' associated loss of milk production. Identifying sires carrying dnms in key genes would allow breeding organisations to decide if the potential risk of a deleterious phenotype resulting from a dnm out weights the potential benefits of the sires' genetics. For potentially exceptional sires, small scale breeding experiments could be carried out to rapidly determine the phenotypic effect of the mutation, while the sire was used in the population. Should the dnm turn out to be deleterious, then carefully controlled breeding of the carriers would allow much of the sires' genetic merit to be retained in the population while the effect of the deleterious variant is minimised. Finally, the constant identification of new dnms could provide insight into the role of key genes and their effect on phenotypes of interest.

When considering dnm in humans in light of our work, there are three areas that are of considerable importance. Firstly, there is considerable inter-individual variation in the patterns and rates of dnm in the early stages of development. This can lead to a substantial proportion of the dnms in any one individual being gonosomal or germline mosaic. In humans this is particularly relevant with regards to ultra-rare genetic disorders, and the probability of multiple offspring being affected. The degree of variability in the rate of dnm would suggest that considerable care needs to be taken in predicting the chance of additional affected offspring being born, when dealing with de novo disorders where the causative variant has not been identified or has not been tested for in the parents germ-line. The high degree of variability makes it risky to apply average estimates to individual cases and should be considered when providing genetic counselling around the risk of additional affected children to couples with an affected child. Secondly our work suggests that reproductive technologies can have an impact on the rate of dnm during the early stages of development. It is possible that this also applies to human embryos resulting from IVF/IVM, if so there may be an increase in the numbers of mosaic dnms present in individuals resulting from such procedures. Due to the mosaic nature of the early occurring dnms, there may be little or no impact on the individual themselves but instead may show up as a slight increase in the occurrence of genetic disorders in the individuals offspring. While the increase in the number of mosaic dnms is substantial in our work the absolute increase in numbers of dnms is substantially less. Still this would be worth investigating in humans and improving IVF/IVM techniques to reduce the increased rate of dnm formation may improve the viability of the resulting embryos. One final observation with regards to humans, is that while the overall dnm rate appears to be similar between humans and cattle, it is possible that the rates of mutation differ at the same stages of development for the two species. If this is the case then it may provide an interesting starting point for investigating both differences in embryo development along with the evolution of mutation rate and how substantially differences in the patterns and characteristics of dnm can occur while maintaining the same overall dnm rate. Or due to the increased power we have to mosaic detect dnms in our three and four generation pedigrees it may suggest that the early rates of dnm in humans is being substantially underestimated.

In summary, the dnm rate in cattle is $1.2x10^{-8}$ bp/gen, similar to the rate reported in humans and chimpanzees. This rate varies depending on the stage of development, with the initial rate in the first 4-6 cell divisions after fertilisation, being potentially 10-20x higher than the rate later in development and exhibiting different mutational signatures. This suggests mutational processes may differ between stages of development and that a steady state mutation rate cannot be assumed when estimating the risk of recurrence for de novo genetic disorders within a family. Thus, dnm is an actively dynamic and evolving process with both environmental factors such as the use of reproductive technologies or genetic factors such as mutator loci affecting the rate and signatures of dnm. As such further work looking at how

reproductive technologies influence mutation rate maybe be beneficial for both the human and dairy cattle populations as the use of these technologies continues to increase. Finally in dairy cattle it is worth considering the sequencing of elite sires to proactively reduce the risk and impact to the industry of deleterious dominant and recessive variants.

Bibliography

Adelson DL, Raison JM, Edgar RC. 2009. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proc Natl Acad Sci U S A* **106**: 12855–12860.

Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR. 2015. Clock-like mutational processes in human somatic cells. *Nat Genet* **47**: 1402–1407.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, et al. 2013a. Signatures of mutational processes in human cancer. *Nature* **500**: 415–421.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013b. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep* **3**: 246–259.

Aliloo H, Pryce JE, González-Recio O, Cocks BG, Hayes BJ. 2015. Validation of markers with non-additive effects on milk yield and fertility in Holstein and Jersey cows. *BMC Genet* **16**: 89.

Andreassen M. 1943. Haemofili i Danmark. Op. ex domo biol. hered. Hum. Univ. Hafn. VI.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Avery OT, Macleod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. *J Exp Med* **79**: 137–158.

Awadalla P, Gauthier J, Myers RA, Casals F, Hamdan FF, Griffing AR, Cote M, Henrion E, Spiegelman D, Tarabeux J, et al. 2010. Direct Measure of the De Novo Mutation Rate in Autism and Schizophrenia Cohorts. *Am J Hum Genet* **87**: 316–324.

Bensaude O, Babinet C, Morange M, Jacob F. 1983. Heat shock proteins, first major products of zygotic gene activity in mouse embryo. *Nature* **305**: 331–333.

Besenbacher S, Liu S, Izarzugaza JMG, Grove J, Belling K, Bork-Jensen J, Huang S, Als TD, Li S, Yadav R, et al. 2015. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* **6**. http://www.nature.com/ncomms/2015/150119/ncomms6969/full/ncomms6969.html (Accessed January 26, 2015).

Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, Jonasdottir A, Magnusson OT, Thorsteinsdottir U, Masson G, et al. 2016. Multi-nucleotide de novo Mutations in Humans. *PLOS Genet* **12**: e1006315.

Boeke JD, Stoye JP. 1997. Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements. In *Retroviruses* (eds. J.M. Coffin, S.H. Hughes, and H.E. Varmus), Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY) http://www.ncbi.nlm.nih.gov/books/NBK19468/ (Accessed August 15, 2017).

Brandler WM, Antaki D, Gujral M, Noor A, Rosanio G, Chapman TR, Barrera DJ, Lin GN, Malhotra D, Watts AC, et al. 2016. Frequency and Complexity of De Novo Structural Mutation in Autism. *Am J Hum Genet* **98**: 667–679.

Braude P, Bolton V, Moore S. 1988. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature* **332**: 459–461.

Brown T. 2002. *Genomes*. 2nd ed. Oxford: Wiley-Liss https://www.ncbi.nlm.nih.gov/books/NBK21114/.

Buis JM, Cheek J, Kalliri E, Broderick JB. 2006. Characterization of an Active Spore Photoproduct Lyase, a DNA Repair Enzyme in the Radical S-Adenosylmethionine Superfamily. *J Biol Chem* **281**: 25994–26003.

Cadenas E, Davies KJ. 2000. Mitochondrial free radical generation, oxidative stress, and aging. *Free Radic Biol Med* **29**: 222–230.

Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, Han L, Vives L, O'Roak BJ, Sudmant PH, Shendure J, et al. 2012. Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* **44**: 1277–1281.

Campbell CD, Eichler EE. 2013. Properties and rates of germline mutations in humans. *Trends Genet*. http://www.cell.com/trends/genetics/abstract/S0168-9525(13)00070-X (Accessed September 5, 2013).

Campbell IM, Yuan B, Robberecht C, Pfundt R, Szafranski P, McEntagart ME, Nagamani SCS, Erez A, Bartnik M, Wiśniowiecka-Kowalnik B, et al. 2014. Parental Somatic Mosaicism Is Underrecognized and Influences Recurrence Risk of Genomic Disorders. *Am J Hum Genet* **95**: 173–182.

Charlier C, Li W, Harland C, Littlejohn M, Coppieters W, Creagh F, Davis S, Druet T, Faux P, Guillaume F. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome Res* **26**: 1333–1341.

Christmann M, Verbeek B, Roos WP, Kaina B. 2011. O(6)-Methylguanine-DNA methyltransferase (MGMT) in normal tissues and tumors: enzyme activity, promoter methylation and immunohistochemistry. *Biochim Biophys Acta* **1816**: 179–190.

Church DN, Briggs SEW, Palles C, Domingo E, Kearsey SJ, Grimes JM, Gorman M, Martin L, Howarth KM, Hodgson SV, et al. 2013. DNA polymerase ε and δ exonuclease domain mutations in endometrial cancer. *Hum Mol Genet* **22**: 2820–2828.

Coffin JM, Hughes SH, Varmus HE, eds. 1997. *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY) http://www.ncbi.nlm.nih.gov/books/NBK19376/ (Accessed August 15, 2017).

Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Cassals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, et al. 2011. Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**: 712–714.

Cooke MS, Evans MD, Dizdaroglu M, Lunec J. 2003. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J* **17**: 1195–1214.

Cooper DN, Youssoufian H. 1988. The CpG dinucleotide and human genetic disease. *Hum Genet* **78**: 151–155.

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. 1978. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274**: 775–780.

Crosby IM, Gandolfi F, Moor RM. 1988. Control of protein synthesis during early cleavage of sheep embryos. *J Reprod Fertil* **82**: 769–775.

Crow JF. 2000. The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* **1**: 40–47.

Dal GM, Ergüner B, Sağıroğlu MS, Yüksel B, Onat OE, Alkan C, Özçelik T. 2014. Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *J Med Genet* **51**: 455–459.

De Sousa PA, Watson AJ, Schultz RM. 1998. Transient expression of a translation initiation factor is conservatively associated with embryonic gene activation in murine and bovine embryos. *Biol Reprod* **59**: 969–977.

Degtyareva NP, Heyburn L, Sterling J, Resnick MA, Gordenin DA, Doetsch PW. 2013. Oxidative stress-induced mutagenesis in single-strand DNA occurs primarily at cytosines and is DNA polymerase zeta-dependent only for adenines and guanines. *Nucleic Acids Res* **41**: 8995–9005.

DePristo MA, Banks E, Poplin RE, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas M., Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Drake JW. 1993. Rates of spontaneous mutation among RNA viruses. *Proc Natl Acad Sci* **90**: 4171–4175.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of Spontaneous Mutation. *Genetics* **148**: 1667–1686.

Drost JB, Lee WR. 1995. Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. *Environ Mol Mutagen* **25 Suppl 26**: 48–64.

Essen LO, Klar T. 2006. Light-driven DNA repair by photolyases. *Cell Mol Life Sci CMLS* **63**: 1266–1277.

Feng C, Pettersson M, Lamichhaney S, Rubin C-J, Rafati N, Casini M, Folkvord A, Andersson L. 2017. Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. *eLife* **6**: e23907.

Flori L, Fritz S, Jaffrézic F, Boussaha M, Gut I, Heath S, Foulley J-L, Gautier M. 2009. The Genome Response to Artificial Selection: A Case Study in Dairy Cattle. *PLoS ONE* **4**. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2722727/ (Accessed June 16, 2017).

Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, Genome of the Netherlands Consortium, van Duijn CM, Swertz M, Wijmenga C, et al. 2015. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **advance online publication**. http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3292.html (Accessed May 19, 2015).

Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, Georgieva L, Rees E, Palta P, Ruderfer DM, et al. 2014. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**: 179–184.

Gao Z, Wyman MJ, Sella G, Przeworski M. 2016. Interpreting the Dependence of Mutation Rates on Age and Time. *PLoS Biol* **14**: e1002355.

Garcia-Etxebarria K, Jugo BM. 2010. Genome-Wide Detection and Characterization of Endogenous Retroviruses in Bos taurus. *J Virol* **84**: 10852–10862.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio*. http://arxiv.org/abs/1207.3907 (Accessed August 28, 2017).

Gilbert. SF, ed. 2000. *Developmental Biology*. 6th ed. Sunderland (MA): Sinauer Associates;

Girard SL, Bourassa CV, Perreault L-PL, Legault M-A, Barhdadi A, Ambalavanan A, Brendgen M, Vitaro F, Noreau A, Dionne G, et al. 2016. Paternal Age Explains a Major Portion of De Novo Germline Mutation Rate Variability in Healthy Individuals. *PLOS ONE* **11**: e0164212.

Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, Dionne-Laporte A, Spiegelman D, Henrion E, Diallo O, et al. 2011. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet* **43**: 860–863.

Goldmann JM, Seplyarskiy VB, Wong WSW, Vilboux T, Bodian DL, Solomon BD, Veltman JA, Deeken JF, Gilissen C, Niederhuber JE. 2017. Germline De Novo Mutation Clusters Arise During Oocyte Aging In Genomic Regions With Increased Double-Strand Break Incidence. *bioRxiv* 140111.

Goldmann JM, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, Glusman G, Vissers LELM, Hoischen A, Roach JC, et al. 2016. Parent-of-origin-specific signatures of de novo mutations. *Nat Genet* **48**: 935–939.

Goodier JL. 2016. Restricting retrotransposons: a review. *Mob DNA* **7**: 16.

Goodier JL, Kazazian HH. 2008. Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. *Cell* **135**: 23–35.

Haldane JB. 1947. The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Ann Eugen* **13**: 262–271.

Haldane JBS. 1935. The rate of spontaneous mutation of a human gene. *J Genet* **31**: 317–326.

Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mullaart E, Coppieters W, Georges M. 2017a. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions. *bioRxiv* 079863.

Harland C, Karim L, Durkin K, Artesi M, Sartelet A, Knapp E, Tamma N, Mullaart E, Coppieters W, Georges M, et al. 2017b. A polymorphic ERV element that is mobilized in the germline of specific individuals causes abetalipoproteinemia and hypolipidemia in cattle by disrupting the APOB gene. *Preperation*.

Harris K. 2015. Evidence for recent, population-specific evolution of the human mutation rate. *Proc Natl Acad Sci U S A* **112**: 3439–3444.

Harris K, Nielsen R. 2014. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res* **24**: 1445–1454.

Harris K, Pritchard JK. 2017. Rapid evolution of the human mutation spectrum. *eLife* **6**: e24284.

Hershey AD, Chase M. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**: 39–56.

Hodgkins PS, O'Neill P, Stevens D, Fairman MP. 1996. The Severity of Alpha-Particle-Induced DNA Damage Is Revealed by Exposure to Cell-Free Extracts. *Radiat Res* **146**: 660.

Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, et al. 2014. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**: 216–221.

Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE. 2010. De novo rates and selection of large copy number variation. *Genome Res* **20**: 1469–1481.

Josse J, Kaiser AD, Kornberg A. 1961. Enzymatic Synthesis of Deoxyribonucleic Acid VIII. FREQUENCIES OF NEAREST NEIGHBOR BASE SEQUENCES IN DEOXYRIBONUCLEIC ACID. *J Biol Chem* **236**: 864–875.

Ju YS, Martincorena I, Gerstung M, Petljak M, Alexandrov LB, Rahbari R, Wedge DC, Davies HR, Ramakrishna M, Fullam A, et al. 2017. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**: 714–718.

Julie G, Hamdan FF, Rouleau GA. 2011. A Strategy to Identify de Novo Mutations in Common Disorders such as Autism and Schizophrenia. *J Vis Exp JoVE*. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3197027/ (Accessed June 18, 2014).

Keightley PD, Ness RW, Halligan DL, Haddrill PR. 2014. Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a Drosophila melanogaster Full-Sib Family. *Genetics* **196**: 313–320.

Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, Davey JW, Jiggins CD. 2015. Estimation of the Spontaneous Mutation Rate in Heliconius melpomene. *Mol Biol Evol* **32**: 239–243.

Kimura M. 1967. On the evolutionary adjustment of spontaneous mutation rates. *Genet Res* **9**: 23–34.

Kipp S, Segelke D, Reinhardt F, Reents R, Schierenbeck S, Wurmser C, Pausch H, Fries R, Thaller G, Tetens J, et al. 2015. A new Holstein haplotype affecting calf survival. *Interbull Bull*. https://journal.interbull.org/index.php/ib/article/view/1375 (Accessed August 17, 2017).

Kloosterman WP, Francioli LC, Hormozdiari F, Marschall T, Hehir-Kwa JY, Abdellaoui A, Lameijer E-W, Moed MH, Koval V, Renkens I, et al. 2015. Characteristics of de novo structural changes in the human genome. *Genome Res* gr.185041.114.

Koeck T, Olsson AH, Nitert MD, Sharoyko VV, Ladenvall C, Kotova O, Reiling E, Rönn T, Parikh H, Taneera J, et al. 2011. A Common Variant in TFB1M Is Associated with Reduced Insulin Secretion and Increased Future Risk of Type 2 Diabetes. *Cell Metab* **13**: 80–91.

Kondrashov AS, Crow JF. 1993. A molecular approach to estimating the human deleterious mutation rate. *Hum Mutat* **2**: 229–234.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature* **488**: 471–475.

Kunkel TA. 2004. DNA Replication Fidelity. *J Biol Chem* **279**: 16895–16898.

Lachance J, Tishkoff SA. 2014. Biased Gene Conversion Skews Allele Frequencies in Human Populations, Increasing the Disease Burden of Recessive Alleles. *Am J Hum Genet* **95**: 408–420.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio*. http://arxiv.org/abs/1303.3997 (Accessed August 20, 2017).

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493–496.

Littlejohn MD, Henty KM, Tiplady K, Johnson T, Harland C, Lopdell T, Sherlock RG, Li W, Lukefahr SD, Shanks BC, et al. 2014. Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. *Nat Commun* **5**: 5861.

Löytynoja A, Goldman N. 2017. Short template switch events explain mutation clusters in the human genome. *Genome Res* **27**: 1039–1049.

Lynch M. 2010. Evolution of the mutation rate. *Trends Genet TIG* **26**: 345–352.

Lynch M. 2011. The Lower Bound to the Evolution of Mutation Rates. *Genome Biol Evol* **3**: 1107–1118.

Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* **17**: 704–714.

Mamsen LS, Lutterodt MC, Andersen EW, Byskov AG, Andersen CY. 2011. Germ cell numbers in human embryonic and fetal gonads during the first two trimesters of pregnancy: analysis of six published studies. *Hum Reprod* **26**: 2140–2145.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet* **19**: 330–338.

Marchi E, Kanapin A, Magiorkinis G, Belshaw R. 2014. Unfixed Endogenous Retroviral Insertions in the Human Population. *J Virol* **88**: 9529–9537.

Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, Skov L, Belling K, Theil Have C, Izarzugaza JMG, et al. 2017. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**: 87–91.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

McLaren A, Lawson KA. 2005. How is the mouse germ-cell lineage established? *Differentiation* **73**: 435–437.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 122.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**: 2069–2070.

McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Ménézo Y, Dale B, Cohen M. 2010. DNA damage and repair in human oocytes and embryos: a review. *Zygote* **18**: 357–365.

Menzi F, Besuchet-Schmutz N, Fragnière M, Hofstetter S, Jagannathan V, Mock T, Raemy A, Studer E, Mehinagic K, Regenscheit N, et al. 2016. A transposable element insertion in APOB causes cholesterol deficiency in Holstein cattle. *Anim Genet* **47**: 253–257.

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al. 2012. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell* **151**: 1431–1442.

Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T. 1987a. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb Symp Quant Biol* **52**: 863–867.

Miyata T, Hayashida H, Kuma K, Yasunaga T. 1987b. Male-driven molecular evolution demonstrated by different rates of silent substitutions between autosome- and sex chromosome-linked genes. *Proc Jpn Acad Ser B Phys Biol Sci* **63**: 327–331.

Mohrenweiser HW, Wilson III DM, Jones IM. 2003. Challenges and complexities in estimating both the functional impact and the disease risk associated with the extensive genetic variation in human DNA repair genes. *Mutat Res Mol Mech Mutagen* **526**: 93–125.

Moore GP. 1975. The RNA polymerase activity of the preimplantation mouse embryo. *J Embryol Exp Morphol* **34**: 291–298.

Moorjani P, Amorim CEG, Arndt PF, Przeworski M. 2016. Variation in the molecular clock of primates. *Proc Natl Acad Sci U S A* **113**: 10607–10612.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**: 297–304.

Palamara PF, Francioli LC, Wilton PR, Genovese G, Gusev A, Finucane HK, Sankararaman S, Sunyaev SR, de Bakker PIW, Wakeley J, et al. 2015. Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. *Am J Hum Genet* **97**: 775–789.

Pang D, Berman BL, Chasovskikh S, Rodgers JE, Dritschilo A. 1998. Investigation of Neutron-Induced Damage in DNA by Atomic Force Microscopy: Experimental Evidence of Clustered DNA Lesions. *Radiat Res* **150**: 612.

Pausch H, Ammermüller S, Wurmser C, Hamann H, Tetens J, Drögemüller C, Fries R. 2016. A nonsense mutation in the COL7A1 gene causes epidermolysis bullosa in Vorderwald cattle. *BMC Genet* **17**: 149.

Peters M, Reber I, Jagannathan V, Raddatz B, Wohlsein P, Drögemüller C. 2015. DNA-based diagnosis of rare diseases in veterinary medicine: a 4.4 kb deletion of ITGB4 is associated with epidermolysis bullosa in Charolais cattle. *BMC Vet Res* **11**: 48.

Petljak M, Alexandrov LB. 2016. Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis* **37**: 531–540.

Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. 2002. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21**: 7435.

Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, Dominiczak A, Morris A, Porteous D, Smith B, et al. 2016. Timing, rates and spectra of human germline mutation. *Nat Genet* **48**: 126–133.

Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Pérez JL, Moran JV. 2015. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4498412/ (Accessed August 22, 2017).

Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. 2010. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* **328**: 636–639.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Rosenkranz D. 2016. piRNA cluster database: a web resource for piRNA producing loci. *Nucleic Acids Res* **44**: D223–D230.

Samuels ME, Friedman JM. 2015. Genetic Mosaics and the Germ Line Lineage. *Genes* **6**: 216–237.

Sancar A, Lindsey-Blotz LA, Unsal-Kacmaz K, Linn S. 2004. Molecular Mechanisms of Mammalian DNA Repair and the DNA Damage Checkpoints. *Annu Rev Biochem* **73**: 39–85.

Sartelet A, Harland C, Tamma N, Karim L, Bayrou C, Li W, Ahariz N, Coppieters W, Georges M, Charlier C. 2015. A stop-gain in the laminin, alpha 3 gene causes recessive junctional epidermolysis bullosa in Belgian Blue cattle. *Anim Genet* **46**: 566–570.

Scally A. 2016. Mutation rates and the evolution of germline structure. *Phil Trans R Soc B* **371**: 20150137.

Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* **13**: 745–753.

Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**: 919–925.

Schütz E, Wehrhahn C, Wanjek M, Bortfeld R, Wemheuer WE, Beck J, Brenig B. 2016. The Holstein Friesian Lethal Haplotype 5 (HH5) Results from a Complete Deletion of TBF1M and Cholesterol Deficiency (CDH) from an ERV-(LTR) Insertion into the Coding Region of APOB. *PLOS ONE* **11**: e0154602.

Ségurel L, Wyman MJ, Przeworski M. 2014. Determinants of Mutation Rate Variation in the Human Germline. *Annu Rev Genomics Hum Genet* **15**: null.

Seoighe C, Scally A. 2017. Inference of Candidate Germline Mutator Loci in Humans from Genome-Wide Haplotype Data. *PLOS Genet* **13**: e1006549.

Sinha RP, Häder DP. 2002. UV-induced DNA damage and repair: a review. *Photochem Photobiol Sci Off J Eur Photochem Assoc Eur Soc Photobiol* **1**: 225–236.

Smeds L, Qvarnström A, Ellegren H. 2016. Direct estimate of the rate of germline mutation in a bird. *Genome Res* **26**: 1211–1218.

Stone JE, Lujan SA, Kunkel TA. 2012. DNA Polymerase zeta Generates Clustered Mutations During Bypass of Endogenous DNA Lesions in Saccharomyces cerevisiae. *Environ Mol Mutagen* **53**: 777–786.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci U S A* **109**: 18488–18492.

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**: 1113–1120.

Tomita-Mitchell A, Kat AG, Marcelino LA, Li-Sucholeiki XC, Goodluck-Griffith J, Thilly WG. 2000. Mismatch repair deficient human cells: spontaneous and MNNG-induced mutational spectra in the HPRT gene. *Mutat Res* **450**: 125–138.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**: 562–578.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al* **11**: 11.10.1-11.10.33.

van Loon B, Markkanen E, Hübscher U. 2010. Oxygen as a friend and enemy: How to combat the mutational potential of 8-oxo-guanine. *DNA Repair* **9**: 604–616.

Vassena R, Boué S, González-Roca E, Aran B, Auer H, Veiga A, Belmonte JCI. 2011. Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Dev Camb Engl* **138**: 3699–3709.

Venn O, Turner I, Mathieson I, Groot N de, Bontrop R, McVean G. 2014. Strong male bias drives germline mutation in chimpanzees. *Science* **344**: 1272–1275.

Volkman HE, Stetson DB. 2014. The enemy within: endogenous retroelements and autoimmune disease. *Nat Immunol* **15**: 415–422.

Waters LS, Minesinger BK, Wiltrout ME, D'Souza S, Woodruff RV, Walker GC. 2009. Eukaryotic Translesion Polymerases and Their Roles and Regulation in DNA Damage Tolerance. *Microbiol Mol Biol Rev MMBR* **73**: 134–154.

Waters LS, Walker GC. 2006. The critical mutagenic translesion DNA polymerase Rev1 is highly expressed during G2/M phase rather than S phase. *Proc Natl Acad Sci U S A* **103**: 8971–8976.

Wiseman H, Halliwell B. 1996. Damage to DNA by reactive oxygen and nitrogen species: role in inflammatory disease and progression to cancer. *Biochem J* **313**: 17.

Wong WSW, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, Baker R, Thach DC, Iyer RK, Vockley JG, et al. 2016. New observations on maternal age effect on germline de novo mutations. *Nat Commun* **7**. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4735694/ (Accessed January 31, 2017).

Woodruff RC, Hual H, Thompson JN. 1996. Clusters of identical new mutation in the evolutionary landscape. *Genetica* **98**: 149–160.

Wyman C, Kanaar R. 2006. DNA Double-Strand Break Repair: All's Well that Ends Well. *Annu Rev Genet* **40**: 363–383.

Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen J-Q, Hurst LD, Tian D. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* **523**: 463–467.

Zheng C-J, Luebeck EG, Byers B, Moolgavkar SH. 2005. On the number of founding germ cells in humans. *Theor Biol Med Model* **2**: 32.

Glossary of terms

Abasic site – A location in DNA or RNA where the purine or pyrimidine base has been lost.

Amniotic ectoderm – An epithelial layer of cells that lines the amniotic cavity.

Artificial selection – The deliberate selection of specific phenotypes resulting in improved survival and reproductive success for the selected individuals.

Base pair – A pair of complementary nucleotides in DNA.

Blastocyst – An early stage of embryo development, where the zygote has formed an inner fluid filled cavity which contains the inner cell to one side and is surrounded by the trophoblast.

CNV – Copy number variation, the duplication or deletion of a large (50bp to megabases) region of DNA in the genome, often covering one or more genes.

De novo mutation – A new change in the sequence of an individual's DNA that was absent from the DNA of it's parents.

Dinucleotide – A pair of linked nucleotides

Effective population size – The number of individuals an ideal population would need to show the same level of diversity present in an actual population. An ideal population being one that exhibits a fixed size, random mating, simultaneous formation of the next generation and equal number of children per individual.

Embryonic epiblast – The proportion of the epiblast that contributes to the Embryo

Endogenous processes – The natural internal processes of an organism.

Epiblast – A key structure in the early stage of embryo development that contributes to the formation of the fetus.

ERV – Endogenous retroviruses, are viruses that have colonized the host genome and generally lost the ability to take on their original viral form. They consist of a set of viral genes flanked by a pair of

identical long terminal repeats and spread by their transcription to RNA. Which is in turn reverse transcribed to DNA and integrated at random into the host genome by the viral integrase they code for. Most ERVs carry mutations that deactivate their viral genes, they are form approximately 5-8% of the human genome.

Exogenous processes – External factors or processes affecting the organism.

Gametogenesis – The biological processes responsible for the development and formation of gametes in an organism.

Germ-line – The cells and DNA in an individual that contribute to the formation of the gametes and thus can be passed on to the next generation.

Haemophilia – genetic disorders that prevent the clotting of blood.

Hypoblast – a supporting structure formed from the inner cell mass which does not contribute to the formation of the fetus.
Hypermutability – Increased tendency to mutate.

Inner cell mass – The inner collection of cells in the blastocyst which proceeds to form the embryo.

Inversion – regions of dna in a chromosome that have been flipped around and rejoined to the chromosome so that the 5’ end is now the 3’ end where the DNA fragment rejoins the chromosome.

INDEL – Insertion or deletion, a modification to the sequence of DNA which reduces or increase the total number of nucleotides.

Junctional epidermolysis bullosa – A genetic disorder that results in exceedingly fragile skin which blisters easily

LINE – Long interspersed nuclear elements, are large (kilobase sized) genetic elements consisting of both a 5’ and 3’ untranslated region containing two open reading frames, one of which encodes for an endonuclease and a reverse transcriptase. They replicate via transcription to RNA followed by reverse transcription to DNA and insertion into the host genome. They compose of around 21% of the human genome.

Maternal zygote transition – The stage in early embryo development where the embryo switches from relying on RNA produced from the maternal genome, to that produced by their own genome. It is the stage at which the embryos genome activates and begins transcription.

Meiosis – A type of cell divisions that results in four daughter cells each with half the number of chromosomes compared to the original parent cell, critical for the formation of haploid gametes.

Mitosis – A type of cell division that results in two daughter cells, each with a complete set of the parental chromosomes, typical of normal tissue growth.

Mosaicism – Is when some genetic variants are not present in all the cells in an organism. This results from mutations that occur within a cell after the cell division post fertilization. Such mutations will only be present within that cell and its' descendant cells and is absent in other cells that exist at the same time and there descendant cells.

Natural selection – The interaction between phenotype and the environment that results in differential survival and reproduction of individuals, depending on their phenotype.

Oocyte – The female gamete or egg cell.

Pedigree – The record of the ancestry for an individual or family.

Phenotype - the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment.

Phylogenetics – The study of evolutionary history of individuals or species via the analysis of genetic data.

Polar body – Small haploid daughter cells resulting from the uneven division of a cell and it's cytoplasm when forming an oocyte

Polymerase – Enzymes responsible for the replication of DNA.

Primordial germ-cells – the initial subset of differentiated cells that are ancestral to all germ-line cells

Pseudogenes – Segments of DNA that are derived from copies of genes which have been modified by mutation, resulting in a change or reduction in function.

Retrotransposon – A genetic element that reproduces via transcription to RNA followed by reverse transcription to cDNA.

Reverse transcriptase – An enzyme than can create cDNA from an RNA strand.

SINE - Short interspersed nuclear elements, are small (50-500bp) genetic elements that replicate via transcription to RNA and then reverse transcription to DNA and insertion into the host genome. However unlike the larger elements they do not contain their own genes but instead rely on enzymes produced by LINEs.

SNP – Single nucleotide polymorphism, a single nucleotide change in the DNA of organism.

Somatic – Cells and tissues in the body that do not contribute to the next generation.

Spermatogonia – the primary gametic cells that are responsible for the formation of sperm.

SV – Structural variant, a largescale change in the structure of the genome inclusive of CNVs; transpositions of DNA between chromosomes; inversions, which changing the orientation of a region of DNA in a chromosome; segmental duplications or deletions of a region consisting of multiple copies of the same DNA.

Translocations – A region of DNA that has moved to a different chromosome or different position on the same chromosome.

Trophoblast – The outer layer of cells in a blastocyst which gives rise to the placenta.

Presses de la Faculté de Médecine vétérinaire de l'Université de Liège

4000 Liège (Belgique)

D/2018/0480/1
ISBN 978-2-87543-121-9