UNIVERSITE DE LIEGE
FACULTE DE MEDECINE VETERINAIRE
DEPARTMENT DE GESTION VETERINAIRE DES RESSOURCES ANIMALES
GIGA - UNITE DE GENOMIQUE ANIMALE

Activité des Rétrovirus Endogènes dans la Lignée Germinale des Bovins

Activity of Endogenous Retroviral Elements in the Bovine Germline

Lijing TANG

THESE PRESENTEE EN VUE DE L'OBTENTION DU GRADE DE DOCTORAT EN SCIENCES VETERINAIRES

ANNEE ACADEMIQUE 2024-2025

Cover image: A journey with jumping transposons in Liège

Promotor

Carole CHARLIER (University of Liège, Belgium)

Co-promotor

Haruko TAKEDA (University of Liège, Belgium)

Thesis committee

Michel GEORGES (University of Liège, Belgium)

Laurent GILLET (University of Liège, Belgium)

Jury members

Leif ANDERSSON (Uppsala University, Sweden)

Alex GREENWOOD (Freie Universität Berlin, Germany)

Luc GROBET (University of Liège, Belgium)

Gilles DARCIS (University of Liège, Belgium)

Bernard PEERS (University of Liège, Belgium)

Christophe DESMET (University of Liège, Belgium)

Chair

Anne-Sophie VAN LAERE (University of Liège, Belgium)

Acknowledgements

This PhD journey has in many ways been a rewarding and challenging experience that would not have been possible without the support, guidance, and encouragement of many individuals. I am deeply grateful to everyone who has contributed to this journey.

I would like to express my gratitude to my advisor Carole Charlier, for her invaluable guidance, patience, and encouragement throughout the years. Your mentorship and insights have not only shaped this thesis but have also inspired professional life. Thank you for believing in my potential and for pushing me to reach beyond my limits and for showing me that doing science can and should be for fun. Thank you taking so great care of me and Xinyi. I will remember all of these things in my heart.

I am also grateful to the members of my thesis committee. In particular, I would like to thank Michel Georges for his expertise and involvement, which have been crucial in refining my research. You have always been willing to share your ideas and thoughts, inspiring me in many ways. I will always remember your advice: “don’t be faster, just be better.” Thank you Haruko Takeda for teaching me lab work and for demonstrating high standards in bench practices. I will hold myself to those same high standards. Laurent Gillet, your suggestions have strengthened this work.

I would also like to extend my gratitude to my thesis jury members - Leif Andersson, Alex Greenwood, Luc Grobet, Gilles Darcis, Bernard Peers, and Christophe Desmet as well as to the president, Anne-Sophie Van Laere, for dedicating their time to my thesis and for helping me improve it.

I am especially grateful to Ben for his invaluable scientific contributions to the published paper and to Sebastien for his beautiful bench work. Special thanks go to my colleagues and friends in the Unit of Animal Genomics. Friends like Blanche, Gabriel, Lim, Lati, Wouter, Keith, Maria, Caroline, Yefang, Tom, Can, Jose-Luis, Natalia, Vincent, Anne and Helene have made this journey more collaborative and enjoyable. I would like to thank Miyako for your kind help, which made my stay much easier. I would also like to extend my gratitude to colleagues in sequencing and genotyping platform for their technical support, which made it possible for me to conduct this research.

To my wife, Xinyi, thank you for your love, patience, and encouragement, especially during the more challenging moments of this journey. To my little girl, I am so looking forward to meeting you. To my parents and parents-in-law, thank you for your faith in me and for supporting my aspirations. To my sister, brother-in-law, nephew and niece, your love has meant a lot to me. Lastly, to all members of my extended family who cheered me-thank you for being there. Your constant encouragement has been a source of strength and motivation.

Thank you all for being an essential part of this accomplishment.

Abbreviations

APOB	Apolipoprotein B
CD	Cholesterol deficiency
cDNA	Complementary deoxyribonucleic acid
CRISPR	Clustered regularly interspaced short palindromic repeats
DNA	Deoxyribonucleic acid
eccDNA	Extrachromosomal circular DNA
ENV	Envelop
ERV	Endogenous retrovirus
ESC	Embryonic stem cell
ETn	Early transposon
GAG	Group specific antigen
GWAS	Genome wide association study
HERVK	Human endogenous retrovirus K subfamily
IAP	Intracisternal A-type particle
IN	Integrase
KoRV	Koala retrovirus
KRAB-ZFP	Kruppel-associated box zinc-finger protein
LD	Linkage disequilibrium
LINE	Long interspersed nuclear element
LTR	Long terminal repeat
MusD	Mus musculus type D
NGS	Next generation sequencing
NPC	Nuclear core complex
ORF	Open reading frame
PBS	Primer binding site
PCIP	Pooled CRISPR inverse PCR
PCR	Polymerase chain reaction
PGC	Primordial germ cell
PIC	Preintegration complex
piRNA	P-element induced Wimpy testis interacting RNA
PRO	Protease
RNA	Ribonucleic acid
RT	Reverse transcriptase
SINE	Short interspersed nuclear element
SSC	Spermatogonia stem cell
SU	Surface unit
SVs	Structural Variations
SVA	Sine-VNTR-Alu
TE	Transposable elements

TM	Transmembrane
tRNA	Transfer RNA
TSD	Target site duplication
TSS	Transcription start site
VLP	Viral like particle
WGS	Whole genome sequencing
XRV	Exogenous retrovirus

Résumé - Abstract

Résumé

L'Equipe Moléculaire de l'Unité de Génomique Animale s'est toujours attachée à **disséquer les mécanismes moléculaires** qui sous-tendent les phénotypes d'intérêt agronomique, tels les caractères de production ou les maladies mono- ou oligo-géniques, mais aussi les caractères d'intérêt plus fondamental, tels la recombinaison méiotique, la conversion génique et les **mutations germinales *de novo,*** de tout type. Le présent travail s'est intéressé à caractériser les processus biologiques capables d'influencer la création d'un type particulier de mutation *de novo*, à savoir les nouvelles insertions créées par la **mobilisation d'éléments transposables** (TE) dans la lignée germinale des bovins.

Le point de départ du travail correspond à la dissection moléculaire d'une **pathologie récessive** létale récemment apparue en race bovine Holstein Frisonne (HF) et dénommée '**déficience en cholestérol**'. Grâce à un clonage positionnel, le gène et la mutation causale ont été identifiés et caractérisés : il s'agit de l'insertion d'un TE, et plus particulièrement un TE de la catégorie des **rétrovirus endogènes** (ERVs), dans l'exon 5 d'un gène codant pour une protéine essentielle au transport du cholestérol, le gène *APOB* (***Apolipoprotein B***). Cette insertion, qui génère une perte de fonction totale du gène, est récente et spécifique de la race HF, ce qui laissait à penser que cette famille d'ERVs (**ERV2-1-LTR-BT**) pouvait encore être capable de se mobiliser pour créer de nouvelles insertions germinales chez les bovins.

Dans la première partie du travail, nous avons tiré profit des **séquences 'génome entier'** (WGS) d'un large pédigrée (***Damona,*** 753 WGS au total), constitué de 131 familles nucléaires étendues de bovins HF, pour tenter d'identifier des insertions *de novo* d'ERV répondant donc aux critères suivants : absentes chez les parents, présentes chez le descendant et transmises à la génération suivante. Cinq mutations ont ainsi été détectées et validées après analyse des 2x131 gamètes, elles appartiennent à la même famille que celle de l'insertion d'*APOB*. Quatre d'entre elles étaient d'origine paternelle et une d'origine maternelle. De manière frappante, trois provenaient du même père. Dans l'ensemble, cela prouve la présente activité de cette famille d'ERVs dans les lignées germinales mâle et femelle, et cela nous permet d'émettre l'hypothèse d'une variation interindividuelle du **taux de transposition *de novo* (dnTR)** au sein de cette population bovine.

Cependant, pour évaluer cette variation interindividuelle du dnTR, il fallait développer une **méthode robuste et quantitative**, applicable à large échelle et permettant de mesurer précisément ce phénotype moléculaire (dnTR) pour une cohorte d'individus. **Ce fût la partie la plus longue et techniquement délicate de cette thèse**. La méthode, finalement mise au point, et ensuite validée, a été appliquée à une large cohorte d'échantillons de sperme (gamètes mâles) de 430 taureaux Blanc-Bleu belges. Il s'est avéré que le phénotype dnTR ne variait pas au cours de la vie reproductive d'un individu, mais que, par

contre, il variait de plus de dix fois entre individus, avec une moyenne d'une nouvelle insertion par ~150 gamètes.

Ce **phénotype moléculaire**, maintenant rendu quantitatif, a donc été utilisé pour tenter d'identifier d'éventuelles régions génomiques influençant ce caractère, et ce grâce à des études d'association type 'génome entier' (**GWAS**). Une série de loci significatifs a été mise en évidence et la dissection moléculaire fine de ceux-ci a révélé que, pour quatre sur sept d'entre eux, le variant candidat causatif le plus probable était l'insertion polymorphe (non fixée) d'un ERV de la même famille. Au sein de la cohorte des 430 taureaux, nous avons alors établi et caractérisé leur catalogue personnel d'ERVs, pour un total d'environ 300 ERVs polymorphes capturés dans cette cohorte. Après séquençage complet des éléments du catalogue, il s'est avéré que ces ERVs pouvaient être rangés en deux catégories, une catégorie définie comme '**compétente**' (C), c-à-d encodant toutes les protéines nécessaires à la mobilisation autonome (15 %) et une autre définie comme '**défective'** (D, 85 %), ayant perdu cette capacité. Les quatre loci ERV identifiés par GWAS appartenaient à la catégorie C. Une analyse plus approfondie a montré que plus d'un quart de la variance du dnTR est expliquée par le nombre d'éléments de type C dans le génome, avec une corrélation positive avec les taux de mobilisation.

Enfin, l'analyse d'environ 3.700 insertions *de novo* d'ERVs a révélé qu'elles étaient dominées par des éléments de type D, suggérant que ces derniers prendraient le dessus en détournant la machinerie des éléments de type C. Ces résultats suggèrent donc un mécanisme d'autorégulation, où les éléments de type D agissent comme des '**parasites de parasites**', pouvant potentiellement conduire à l'extinction spontanée de cette famille d'ERV.

Dans la seconde partie de ce travail, notre équipe a été sollicitée par l'équipe d'Etienne Bucher (http://plantepigenetics.ch/) pour tenter de mettre en évidence des mutations *de novo* (nouvelles insertions) d'une famille de TE particulière, faisant partie des **transposons à ADN**, et plus particulièrement de la classe des ***Hélitrons***, une classe spéciale en leur sein. En effet, les *Hélitrons* sont supposés générer de nouvelles insertions via un mécanisme très personnel, qualifié de **'détacher puis coller'** qui implique la génération d'intermédiaires de réplication existant sous forme de **cercles d'ADN**. Cependant, et jusqu'à notre étude, aucune évidence directe, démontrant que des *Hélitrons* avaient gardé la possibilité de bouger de manière autonome, n'avait été démontrée. Nous avons alors adapté la méthode quantitative, développée en bovin, aux *Hélitrons* du blé, et ce avec succès. Amenant ainsi la preuve définitive que certaines lignées de blé renferment au moins une copie pouvant être qualifiée de compétente (ou autonome) au sein de leur génome. Et de manière similaire à la paire d'éléments 'C' et 'D' des ERVs du bovin, il existe aussi une paire semblable 'C' (***Feng8***) et 'D' (***Xuan1***) cohabitant dans certaines lignées de blé.

En conclusion, nous espérons avoir apporté des preuves solides supportant l'activité autonome de certains TE chez le bovin, avoir développé une méthode robuste et quantitative pouvant être adaptée à d'autres espèces d'eucaryotes et à d'autres catégories de TE encore actifs. Cela ouvre la porte à une large série d'études, tant dans les lignées germinales des mammifères, autant mâle que femelle, que dans des tissus somatiques d'intérêt.

Summary

The Molecular Team of the Animal Genomics Unit has consistently focused on dissecting the **molecular mechanisms** underlying agronomically **important phenotypes**, such as production traits or mono- or oligo-genic diseases, as well as traits of more basic interest, such as meiotic recombination, gene conversion, and ***de novo* germline mutations** of all types. The present study aimed to characterize biological processes capable of influencing the creation of a particular type of *de novo* mutation, specifically, new insertions created by the mobilization of **transposable elements** (TEs) in the germline of cattle.

The starting point of our work involved the molecular dissection of a recently identified lethal recessive disease in the Holstein Friesian (HF) cattle breed, known as **'cholesterol deficiency**.' Through positional cloning, the gene and causal mutation were identified and characterized: an insertion of a TE, specifically an **endogenous retrovirus (ERV)**, in exon 5 of a gene encoding a protein essential for cholesterol transport, the ***APOB* gene (*Apolipoprotein B***). This insertion, which results in a full loss of function, is recent and specific to the HF breed, suggesting that this ERV family (**ERV2-1-LTR-BT**) might still be capable of mobilizing to create new germline insertions in cattle.

In the first part of the study, we leveraged **whole genome sequencing** (WGS) data from a large pedigree (***Damona***, totaling 753 WGS), consisting of 131 extended nuclear families of HF cattle, to try to identify *de novo* ERV insertions meeting the following criteria: absent in the parents, present in the offspring, and transmitted to the next generation. Five such mutations were detected and validated after analyzing the 2x131 gametes, all belonging to the same family as the *APOB* insertion. Four and one were of paternal and maternal origin, respectively. Strikingly, three originated from the same sire. All together this demonstrated the current activity of this ERV family in both male and female germlines, and it allows us to hypothesize an inter-individual variation in ***de novo* transposition rate (dnTR**) within this cattle population.

But, to assess this inter-individual variation in dnTR, a robust and quantitative method was needed, one that could be exploited on a large scale and precisely measure this molecular phenotype (dnTR) in a cohort of individuals. **This was the most time-consuming and technically challenging part of this thesis**. The method, which was eventually developed and validated, was applied to a large cohort of sperm samples (male gametes) from 430 Belgian Blue bulls. It turned out that the dnTR phenotype did not vary throughout an individual's reproductive life; nevertheless, it varied more than ten-fold between animals, with an average of 1 new insertion per ~150 sperm cells.

This now-quantified molecular phenotype was subsequently used to attempt to pinpoint genomic regions potentially influencing this trait, through genome-wide association studies (**GWAS**). A series of

significant loci were identified, and detailed molecular analysis of these loci revealed that, in four out of seven cases, the most likely causative candidate variant was a polymorphic (non-fixed) insertion of an ERV from the same family. Within the cohort of 430 bulls, we then established and characterized their personal ERV catalog, capturing a total of around 300 polymorphic ERVs in this cohort. After sequencing the entire catalog, these ERVs were classified into two categories: a '**competent**' category (C), encoding all proteins necessary for autonomous mobilization (15%), and a '**defective**' category (D, 85%), which had lost this capacity. The four ERV GWAS loci belonged to the C category. Further analysis showed that over a quarter of the dnTR variance is explained by the number of C-type elements in the genome, which correlated positively with mobilization rates.

Finally, analysis of approximately 3,700 *de novo* ERV insertions revealed that they were dominated by D-type elements, suggesting that D-type elements may be taking over by hijacking the mobilization machinery of C-type elements. These findings suggest a self-regulation mechanism where D-type elements act as 'parasites of parasites', potentially leading to the spontaneous collapse of this ERV clade.

In the second part of this thesis, our team was invited by Etienne Bucher's team (http://plantepigenetics.ch/) to try to detect *de novo* mutations (new insertions) from a specific family of TEs, namely **DNA transposons,** and particularly the ***Helitron*** class, a unique subgroup. *Helitrons* are thought to create new insertions via a distinctive "**peel and paste**" mechanism that involves generating replication intermediates in the form of **DNA circles**. However, until our study, no direct evidence had demonstrated that *Helitrons* retained the ability to move autonomously. We adapted the quantitative method developed in cattle to ***Helitrons* in wheat**, with success, thereby providing definitive proof that a subset of wheat lines contains at least one competent (or autonomous) copy within their genome. Similarly to the C- and D-element pair of bovine ERVs, there is an equivalent 'C' (***Feng8***) and 'D' (***Xuan1***) pair coexisting in some wheat lines.

In conclusion, we hope to have provided solid evidence supporting the autonomous activity of a subset of TEs in cattle, and to have developed a robust and quantitative method that can be adapted to other eukaryotic species and other classes of active TEs. This opens the door to a wide range of studies in both male and female germlines of mammals, as well as in somatic tissues of interest.

General Preamble

Human and animal populations are characterized by substantial levels of genetic polymorphisms, which are the substrate of Darwinian selection but also the cause of disease. The observed level of genetic polymorphisms reflects a balance between the loss of genetic variation by the process of random drift and the gain of genetic variation by the process of *de novo* mutation (DNM) in the germline. DNMs includes single nucleotide variations (SNVs), small indels (INDELs) and structural variations (SVs), ranging from one base pair to hundreds bases pairs changes, decreasing in frequency in that order yet increasing with regards to the total bases of the genome affected. Although the mutation rate is a very fundamental parameter, measuring it accurately is not trivial, let alone measuring its inter-individual variation. In the past few years, whole-genome sequencing (WGS) of parent-offspring trios, uncovering mutations present in the offspring yet absent in its parents, has facilitated the genome-wide and large-scale detection and study of germline *de novo* mutations. Several cohorts of human and other species have been established to study DNMs using this approach (Bergeron et al., 2023; Goldmann et al., 2016; Harland et al., 2017; Jónsson et al., 2017; Sasani et al., 2019). These studies found that offspring carry 70 single nucleotide DNMs on average and that this number varies between families, ranging from 30 to 100. So far, two factors that may contribute to these variable numbers of DNMs have been identified. The father's age at conception has the largest influence, adding about two single nucleotide DNMs per year of age (Jónsson et al., 2017; Kong et al., 2012). This effect has been attributed to the increased number of mitotic divisions separating zygote from sperm cells with age. The second factor is the mother's age at conception, contributing roughly one extra single nucleotide DNM every four years (Goldmann et al., 2016). Time-dependent but cell division-independent mechanisms are thought to be at play in this effect (Gao et al., 2019). These two factors together explain only part of the variation in the number of single nucleotide DNMs. This raises the question whether the rest of variation in DNM numbers is solely stochastic or whether it can be explained by other factors. One appealing hypothesis is that the genetic make-up (i.e. mutators or mutational modifiers) of parents could influence the number of DNMs in their gametes. This would imply that the single nucleotide mutation rate is a heritable trait. However, evidence supporting this notion is still very sparse.

Most studies mentioned above focus on SNVs and INDELs. However, half of a typical mammalian genome, including that of cattle, is composed of repetitive sequences of which most are in fact selfish transposable elements (TEs) that can be mobile in the genome. Most TEs families are not capable of moving anymore because of raised host silencing mechanisms and accumulation of mutations. It is known however that mobilization of still active TEs generates substantial genetic polymorphism and can cause disease. In addition, mobilization of TEs, is generally included in studies of SVs, although they arise through mutational mechanisms that are largely distinct from those responsible for SNVs, INDELs and other SVs. As they profoundly affect genome evolution, it is reasonable to consider them separately. The rate of TE mobilization is a fundamental parameter. Nonetheless, direct estimates of TE mobilization rate are also scarce. The first insights of TE dynamics, including the rate of TE

mobilization, came from studies of mutation accumulation in *Drosophila* lines (*Drosophila melanogaster* in particular)(Harada et al., 1990; Suh et al., 1995). A more recent study by Adrion and colleagues provided the first genome-wide estimate of TE movement rate in *D. melanogaster* (Adrion et al., 2017). These authors used next generation sequencing data to compare TE contents across laboratory lines before and after ~150 generations of mutation accumulation. They found that the TE movement rate is slightly lower than the point mutation rate: 2.45×10^{-9} per site per generation against 2.8×10^{-9} per site per generation, respectively. The rate of TE insertions is higher than the rate of deletions: 2.11×10^{-9} per site per generation against 1.37×10^{-10} per site per generation, respectively. Considering that there are 270 million sites in the genome, these numbers correspond to approximately 0.57 insertions and 0.037 deletions per generation. Germline *de novo* mobilization of TEs has also been studied in mammalian species by whole genome sequencing of pedigrees. Long interspersed elements (LINE1) mobilized at a rate of one every eight births in C57BL/6J mice, while no mobilization of known active endogenous retrovirus (ERV) elements (IAP and ETn elements) were detected in these mice studies (Richardson et al., 2017a). Using different mice strains, Wolf et. al. observed a high degree of variability in the number of new ETn insertions between mouse strains, suggesting that genetic and epigenetic effects may play an important role (Wolf et al., 2020). This is in line with the observation that 46 IAP insertional mutations with known phenotypic consequences have been described in mice, and the strain of origin has been documented for 43 of them (Gagnier et al., 2019). Among these 43 cases, 36 occurred in C3H/HeJ mice, other C3H substrains or in a hybrid involving C3H, while only seven were identified in other mouse strains (Gagnier et al., 2019; Rebollo et al., 2020). In humans, 26 *de novo* mobilizations of TEs have been reported in 437 births, including human LINE1, *Alu* and SVA elements. The mobilization rate estimates for LINE1, *Alu* and SVA are one in 63 births, one in 40 births, and one in 63 births, respectively (Feusier et al., 2019). Interestingly, the authors observed some potential variation in mobilization activities between pedigrees as 6/26 events occurred in the same pedigree (Feusier et al., 2019). However, pedigree-based approach is not feasible to reveal the extent of inter-individual variation of mobilization rates since TE mobilization is in general rare. Consequently, our goal was to develop a method that would allow to quantitatively estimate the individual rate of TE mobilization, without relying on observations in the offspring, to be able to study the inter-individual variation and genetic and non-genetic factors underlying this variation.

Introduction

1 General introduction about transposable elements (TEs)

1.1 Discovery of TEs

Genes were considered to be relatively stable arranged in an orderly linear fashion along chromosomes until Barbara McClintock challenged this concept by discovering that some elements in maize were capable of mobilization, that is, to change position within the genome (Mcclintock, 1950). These elements became later known as transposons or transposable elements (TEs). The significance of her work was not immediately recognized. Similar mobile genetic elements were later found in bacteriophages, the viruses that infect bacteria (Taylor, 1963). Also, P elements were later detected in *Drosophila* and other eukaryotic organisms in which they cause hybrid dysgenesis (Engels and Preston, 1981; Sharmistha et al., 2015). The scientific community progressively realized that transposons were not just peculiarities of maize but were widespread across species. A landmark study by Kazazian et al. (1988) reported an insertion in the factor VIII gene of patients with hemophilia A, resulting from transposition of an active transposable element, which laid the foundation of the study of long interspersed nuclear element 1 (LINE1) in the human genome (Kazazian et al., 1988). All eukaryotic genomes examined so far are known to harbor TEs, which can account for as much as 85% of the genome in some species.

1.2 Classification of eukaryotic TEs

The most wide accepted classification of TEs was proposed by Finnegan in 1989 (Finnegan, 1989). TEs can be divided into two major classes according to their intermediate form of transposition. Class I TEs, also called retrotransposons, account for the majority of TEs in mammalian genomes and mobilize through a “copy-and-paste” mechanism, in which an RNA intermediate is reverse transcribed into a cDNA that integrates elsewhere in the genome. Class II TEs, also known as DNA transposons, largely move by a “cut-and-paste” process, yet increase copy numbers through indirect mechanisms that rely on the host machinery. DNA transposon can replicate during DNA replication by excising from a newly replicated chromatid to an unreplicated site, resulting in one copy on one replicated chromosome and two on the other (Feschotte and Pritham, 2007). A particular family DNA transposons, called *Helitrons,* is mobilized through a “peel-and-paste” replicative mechanism involving circular DNA intermediates (see chapter 1.6). Class II TEs only account for a small fraction of mammalian genomes.

These two main classes can be further divided into subclasses according to the precise mechanism of transposition, and then into superfamilies, families and subfamilies. Retrotransposons are classified into two main subclasses: long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons. The latter include both long and short interspersed nuclear elements (LINEs and SINEs). In mammals, the majority of LTR retrotransposons are known as endogenous retroviruses (ERVs). Superfamilies and

families are further divided into subfamilies by phylogenetic analyses. Each subfamily generally counts thousands of dispersed copies, attesting of waves of past expansions.

1.3 Genome abundance of TEs

Analyzing TE content at genome wide scale was not feasible until reference genome assemblies became available for a variety of species in the beginning of twenty-first century, following the development of high throughput DNA sequencing technologies. Bioinformatic tools also needed to be developed to analyze and annotate TEs in these reference genomes.

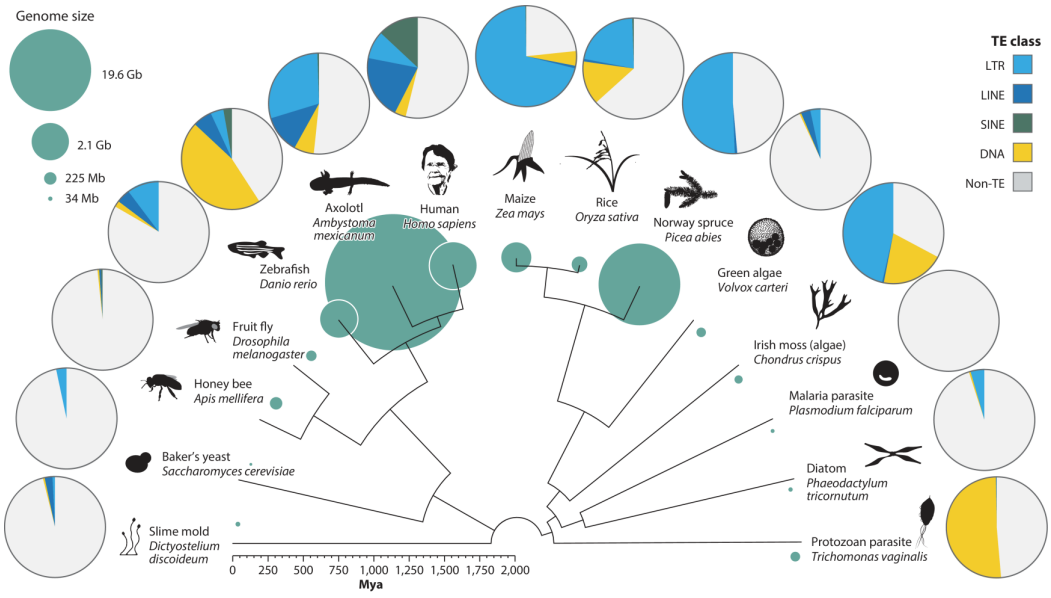


Figure 1 : Distribution of TEs across the eukaryote phylogeny.
Reference genome size (*sea green circles*) varies dramatically across eukaryotes and is loosely correlated with TE content. Here, the honey bee TE content is likely an underestimate, as approximately 3% of the genome derives from unusual large retrotransposon derivatives. YR retroelements have been included with LTRs and all class II elements are included under DNA. Data were acquired from genome RepeatMasker output files. Figure was adapted from (Wells and Feschotte, 2020). Abbreviations: LINE, long interspersed nuclear element; LTR, long terminal repeat; SINE, short interspersed nuclear element; TE, transposable element; YR, tyrosine recombinase. Figure was adapted from Wells and Feschotte (2020).

Systematic in silico annotation of assembled reference genomes with regards to TE content has so far shown that the genome of eukaryotes species contains TEs (Figure 1) (Wells and Feschotte, 2020). Even the most compact genomes, such as that of the pufferfish (*Fugu rubripes*), contain more than 2.7% interspersed repeats (Aparicio et al., 2002). Their widespread distribution among the natural kingdom suggests that TEs existed in the deep phylogenetic root and coevolved with their hosts during evolution. Another striking finding of these analyses is that the abundance of TEs varies between genomes. For

instance, about 50% of the human genome is composed of TEs, in dramatic contrast with the less than 2% of the genome that is protein coding (Lander et al., 2001). Overall, TE content in each of the 248 examined mammalian species ranges from 28% in the star-nosed mole (*Condylura cristata*) to 75% in the aardvark (*Orycteropus afer*), with most species clustering in the middle of that range (around 47%) (Osmanski et al., 2023). The abundance of TEs seems to be positively correlated with genome size, suggesting that the expansion of TE is a major driver of genome size variability (Kapusta et al., 2017). In addition to major differences in TE abundance in general, different TE classes contribute variably to the host genome. In general, retrotransposons are more abundant than transposons due in part to their distinct mechanisms of increasing copy numbers (Bourque et al., 2018). Certain TE families seem to be particularly successful in specific lineages. In mammals, SINEs and LTR retrotransposons are more prevalent in Euarchontoglires, while LINEs dominate in most other lineages, especially the bovids. (Osmanski et al., 2023).

1.4 Activities of TEs

The most notable feature of TEs is that their mobility threatens genome stability and may cause insertional mutagenesis in the germline and somatic tissues. It is noteworthy that, although large fractions of the genome derive from TEs, most TE represent “fossil records” of past activity and only few families of TEs are still capable of transposition in the germline. These may, however, represent an important source of genetic variation. Such events were often recognized because of the phenotypic variation and diseases they caused. Notably, some TE classes are autonomous, meaning that their genome encodes the machinery required for transposition (f.i. LINEs and ERVs), whereas others, such as SINE elements, are non-autonomous and move by hijacking the machinery of other elements (Dewannieux et al., 2003). Phenotypes arising spontaneously in laboratory mice strains often result from insertional mutagenesis by ERVs, typically Early Transposon (ETn) and Intracisternal A Particle (IAP) (Gagnier et al., 2019). The discovery of disease caused by insertions of LINE1 and SINE suggests that these are still active in human (Kazazian et al., 1988). Ongoing and recent activity would produce insertion polymorphisms, shared yet unfixed among individuals in a population. Thus, observing TE insertion polymorphisms in populations allows us to infer recent transposition. Kojima et. al characterized TE insertion polymorphisms in the global human population using high-coverage whole genome sequencing (WGS) data for more than three thousand individuals (Kojima et al., 2023). These polymorphisms predominantly involve young elements known to be active germline mutagens (LINE1, SINE and SVA) (Kojima et al., 2023). In contrast, virtually no polymorphic ERV insertions exist in humans, excluding recent ERV transposition in the human germline (Marchi et al., 2014; Wildschutte et al., 2016). Unlike human ERVs, high levels of ERV insertional polymorphisms have been detected in other mammalians including pig (Chen et al., 2022), horse (Xuexue Liu et al., 2023), sheep (Chessa

et al., 2009) and mule deer (Elleder et al., 2012), suggesting that ERV elements may still transpose actively in the germline of these species.

Nevertheless, we can't be absolutely sure that a family of TEs is currently still active in a species until we observe a *de novo* transposition event in the germline. Sequencing large number of trios (father, mother and offspring) or extended pedigrees in a number of species has revealed *de novo* insertions hence directly demonstrating the activity of the corresponding TEs. Analyzing around 300 human trios demonstrated that LINE1, SINE (*Alu*, in primates) and SVA (SINE-VNTR-*Alu*) are currently active in human in both the female and male germline with the estimated rates ranging from one in 63 births to one in 40 births (Feusier et al., 2019). Richardson et al. applied retrotransposon capture sequencing and WGS to pedigrees of C57BL/6J mice and confirmed that LINE1 is currently active in the mice germline and provided an estimate of one new insertion per eight mice (Richardson et al., 2017b).

TEs not only generate copies of themselves in the germline, but also transpose in somatic tissues in both normal or non-physiological contexts. Somatic retrotransposition of human LINE1, the only autonomous human TE family, has been most extensively studied during normal tissue development and cancerogenesis. Two studies have identified spontaneous somatic LINE1 retrotransposition events in isolated single neurons from normal individuals by applying high-coverage WGS or LINE1 captured target sequencing, respectively, revealing human LINE1 activity during neurogenesis (Baillie et al., 2011; Upton et al., 2015), although the mobilization rate is still controversial (Evrony et al., 2016). Somatic LINE1 retrotransposition has also been observed in human normal colorectal epithelium via WGS of single cell clones which were derived *in vitro*, and the rate of mobilization has been shown to increase in this tissue during tumorigenesis (Nam et al., 2023). A pan-cancer survey has also documented widespread LINE1 somatic retrotransposition in cancerous tissue (Rodriguez-Martin et al., 2020). Moreover, it appears that human ERVs (HERVs) can be reactivated during aging (Xiaoqian Liu et al., 2023) and following cancer chemotherapy using epigenome-modifying drugs (Jones et al., 2019), although no active retrotransposition of HERV has been reported in the human germline.

1.5 Bovinae TEs

The first complete bovine genome assembly (Bos *taurus*) was released in 2009, and systematic detection and characterization of TEs has been conducted along with this release (Mehta et al., 2009). It revealed that the bovine genome contains all the classes of TEs including LINEs, SINEs and LTRs, which together amount to 45% of genome space. Like TE profiles of other eutherian mammals, LINEs and SINEs (~40%) are more abundant than LTRs (~5%) (Adelson et al., 2009). Around one quarter of TEs are lineage-specific and the majority of cattle specific TEs are non-LTR LINE RTE (BovB) and BovB-derived SINEs. Interestingly, BovB repeats are believed to have been horizontally inherited from reptiles around 50 million years ago (Ivancevic et al., 2018). The analysis of BovB's open reading

frames (ORF) suggests absence of active potential. Conversely, more than 800 full-length LINE1 (L1-BT) exist in the bovine genome, of which more than 70 appear to be potentially active based on their ORF content (Adelson et al., 2009).

1.6 Helitrons

DNA transposons usually have inverted terminal repeats at both ends, which are recognized by the transposase. Inside the transposon, there is often a gene coding for the transposase itself. For instance, the Ac/Ds (Activation/Dissociation) is a class of DNA transposons that was first discovered by Barbara McClintock in maize. *Helitron* is a class of DNA transposons that was firstly identified and characterised in the genomes of *Arabidopsis thaliana*, *Oryza sativa*, and *Caenorhabditis elegans* through *in silico* searching (Kapitonov and Jurka, 2001). Further genome-wide scans across hundreds of high-quality reference genomes showed that it is widely present in plant and animal species, and particularly enriched in plant genomes (Li et al., 2023). *Helitrons* are distinct from other DNA transposons with regards to their structure and transposition mechanism. Unlike typical DNA transposons that are flanked by terminal inverted repeats (TIRs, a pair of inverted and identical DNA sequences) which are essential for transposition, *Helitrons* lack TIRs and do not generate target site duplications during transposition (Figure 2). Instead, they often start with TC and end with CTRR (where R stands for A or G). Additionally, *Helitrons* often harbour a short palindromic sequence of 16 to 20 nucleotides that forms a hairpin close to the 3' end. Specific *Helitron* family-defining sequences of 30 bases are present in both termini and are referred to as the left (LTS) and right terminal sequences (RTS) (Figure 2). Another notable distinction is that *Helitrons* use a 'peel-and-paste' transposition process involving extrachromosomal circular DNA (eccDNA) that keeps the original template element unaltered, in contrast with the majority of DNA transposons that move using a 'cut-and-paste' strategy that excises the original template. Furthermore, *Helitrons* demonstrate a strong AT dinucleotide preference at the integration sites (Barro-Trastoy and Köhler, 2024; Grabundzija et al., 2018, 2016).

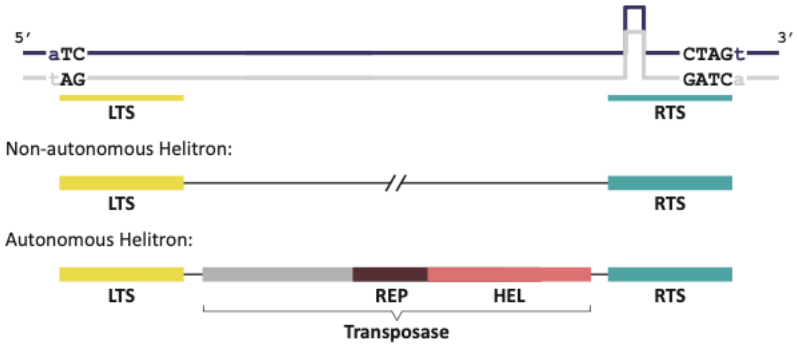


Figure 2. Model of *Helitron*.
Helitrons are characterized by 5'-TC and 3'-CTAG terminal motifs that are part of approximately 30-base pair long left and right terminal sequences, respectively. Non-

autonomous *Helitrons* lack the capability to encode the essential proteins required for self-transposition, whereas autonomous *Helitrons* encode the RepHel transposase. Broken lines indicate that non-autonomous *Helitrons* can vary in length. Abbreviations: HEL, Hel domain of the RepHel transposase; LTS, left terminal sequence; REP, Rep domain of the RepHel transposase; RTS, right terminal sequence. Figure was adapted from Barro-Trastoy and Köhler (2024).

The *Helitrons* discovered so far are all non-autonomous, meaning that they lack the capacity to encode the transposase required for self-transposition. Thus, their mobility relies on the machinery of autonomous counterparts. *In silico* reconstruction of a potentially active *Helitron* showed that it contains a large protein referred to as 'RepHel' (Grabundzija et al., 2016). The RepHel protein includes a replication initiation domain, which shared similarity with HUH endonucleases. Hel represents a DNA helicase domain. Though the activity of reconstructed *Helitrons* has been demonstrated *in vitro*, to date, no *in vivo* active *Helitron* has been isolated, nor has the mobility of *Helitrons* been captured in real time. Thus "*One important open question in the field of Helitron research is whether there are active Helitrons in current genomes and by which trigger they may become activated. While the observation of recent insertional mutations caused by non-autonomous Helitrons in maize suggests that these elements have moved in recent times, direct evidence of Helitron activity in vivo is still lacking.*"(Barro-Trastoy and Köhler, 2024).

2 Endogenous retroviruses (ERVs)

2.1 Discovery of ERV

Research on retroviruses (enveloped single-stranded RNA viruses) started with the identification of the agent causing sarcoma (Rous sarcoma virus, RSV) in chickens in 1911 by Peyton Rous (Rous, 1911; Rubin, 2011). Similar onco-viruses have later been identified in mammals, including murine leukemia virus (MLV), murine sarcoma virus (MSV), and mouse mammary tumor virus (MMTV), as well as sarcoma and leukemia viruses in human (HTLV). One of the major outcomes of these early animal studies was the realization that the genomes of these viruses consisted of RNA, in contrast to the many known DNA viruses (Coffin and Fan, 2016). In the 1970, Howard Temin and David Baltimore, working independently, revealed the existence, in RSV and MLV virions, of an enzyme (now known as reverse transcriptase) with RNA-dependent DNA polymerase activity (Baltimore, 1970; Temin and Satoshi, 1970). This result strongly supported Temin's provirus model predicting that the single-stranded viral RNA was somehow copied into DNA early after infection, and that this DNA was subsequently covalently joined (integrated) into the cell's genomic DNA, where it served as the template for viral RNA synthesis (Weiss, 2006).

Half a century after the initial description of retroviruses, an assay to detect antibodies to viral group specific antigen (where the "Gag" abbreviation comes from) was developed to screen for avian leukosis

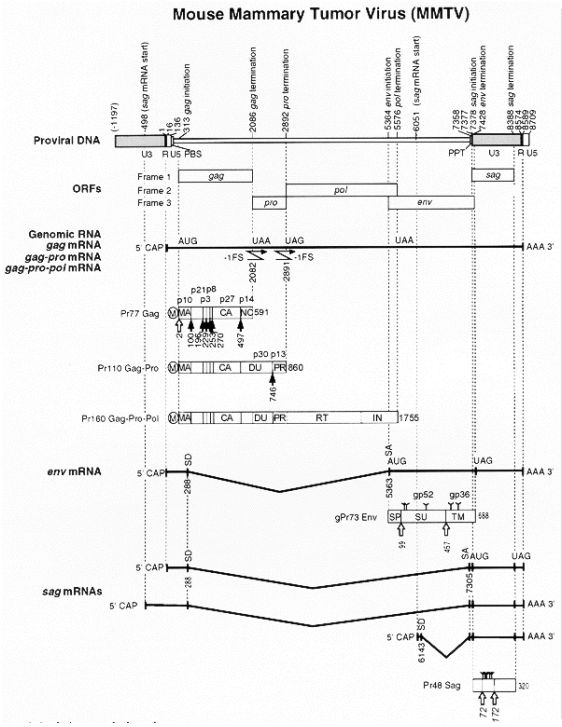
in the chicken (Mager and Stoye, 2015). Intriguingly, the antigen was detected in some chickens free of retroviral infection. Payne et al. subsequently demonstrated that this antigen was inherited as a dominant gene in crosses between Gag-positive and Gag-negative inbred lines of chicken. Similar observations were made for the envelop (Env) protein, and expression of the viral Gag and Env proteins cosegregated in theses crosses. Those findings pointed towards the fact that the viral proteins were encoded in the chicken genome. Mendelian inheritance of retroviral genomes by their hosts was described soon thereafter in the mouse (MLV and MMTV) (Cohen and Varmus, 1979; Jaenisch, 1977, 1976; Stoye and Coffin, 1988). However, the concept of an endogenous retrovirus (ERV), inherited in the germ line as DNA was not widely accepted until the discovery of RT. The chromosomal location and analysis of viral gene expression of endogenous MLV was studied in great detail in the 1970s and 1980s. As with endogenous avian leukemia virus (ALV) many of the MLV genomes are defective, while others maintain complete viral ORFs and are thus potentially infectious (Coffin and Fan, 2016; Weiss, 2006).

These findings raised a huge interest to search for ERVs in other species, particularly in humans, first using probe hybridization and then by the polymerase chain reaction (PCR). Using non-stringent hybridization conditions and a cloned segment of African green monkey DNA that specifically hybridizes to the proviruses of MLV and a baboon endogenous virus as a probe, Martin et al. detected related sequences in three different preparations of human brain DNA, and cloned the first human ERV (HERV) (Martin et al., 1981). Similar approaches or subsequent PCR-based amplification identified more than thirty groups of HERVs, and ERVs in other species (Mager and Freeman, 1995). Ultimately, it was the analysis of the whole human genome sequence that revealed that HERVs constitute 8% of the human genome with 98,000 elements and fragments. Phylogenetic analyses of conserved regions within their Pol and Env genes indicated that they form only a small number of clades closely related to known non-human exogenous retroviruses. Thus, the high copy numbers of those elements probably derived from a small number of germ-line endogenization events followed by an expansion phase (Johnson, 2015).

2.2 ERV structure

The structure of newly inserted ERVs is much like the integrated provirus of XRVs (Figure 3). However, ERVs in mammalian genomes come in distinct forms, reflecting their age and evolutionary history. Firstly, there are intact ERVs that are in essence indistinguishable from the exogenous proviral structure. They contain two identical long terminal repeats (LTRs) ranging in size from several hundred base pairs (bp) to more than one thousand bp, depending on the family, with the encoded open reading frames (ORFs) arrayed between them. LTRs encode the promoter elements, polyadenylation (polyA) signal as well as the *cis*-acting motifs required for transcription. They are made-up of three parts: a repeat region (R), a segment unique to the 5' end (U5) of the LTR and a segment unique to the 3' end (U3) of the

LTR (Figure 3). There is a primer binding site (PBS) immediately downstream of the 5'LTR and upstream of the start of the first ORF. The ORFs in between LTRs include: (i) Gag, which encodes the structural protein that make up the viral particle, (ii) Pro, which encodes the viral protease that cleave Gag, (iii) Pol, which encodes the viral replicative enzymes reverse transcriptase (RT), ribonuclease H (RNase H) and integrase (IN). Some ERVs contain a fourth Env ORF that encodes the polyprotein precursor of the retroviral envelope protein. This protein is proteolytically cleaved to form the surface (SU) and transmembrane (TM) subunits (Coffin et al., 1997) (Figure 3). ERVs arising from different genres of retroviruses may have slightly distinct organization of ORFs. For instance, Pro and Pol are integrated as one ORF in MLV. Additional ORFs other than those four are observed in some XRVs and ERVs, such as the Rec accessory protein of HERVK (Mager and Stoye, 2015). Secondly, a group of commonly observed forms of ERVs lacks one or more ORFs but are otherwise complete. Intriguingly, it is the Env gene, essential for autonomous extracellular replication, that is most often missing. This type of ERV is often also referred to as LTR retrotransposon and is assumed to replicate via intracellular retrotransposition (see hereafter) (Coffin et al., 1997). A third group contains two LTR elements that flank degraded ORFs but no other recognizable homology with retroviral proteins. Such type of elements probably derived from illegitimate recombination with other sequences. Those elements are assumed to be non-autonomous (defective) yet able to mobilize with the help of other complete ERVs as long as they contain RNA packaging sequences. Fourthly, the most abundant form of ERVs in the genome are so-called "solo-LTR" (as opposed to full-length ERV), which are thought to arise by non-allelic homologous recombination between the two LTRs of complete elements. This results in the deletion of the internal sequence, leaving only a single LTR element in the genome. Apparently, this type of ERVs cannot move anymore. Finally, there exist many ERV copies containing partial proviral sequences, with complete or partial LTRs, and having accumulated substitutions, deletions, insertions or more complicated structural rearrangements.



R:	(1–15) The R region of MMTV is the smallest (15 nucleotides) reported for known retroviruses.
U5:	(16–135)
PBS:	(136–153) The primer for minus-strand DNA synthesis is $tRNA^{Lys3}$.
gag:	(313–2086) Proteolytic processing of Pr77 Gag produces MA, CA, and NC. Several cleavage products (pp21, p3, p8) of unknown function are produced from the region between MA and CA. Pr77 Gag is myristylated at amino acid 2Gly. The NC domain provides the amino-terminal region of DU (see below).
pro:	(2082–2892) The *pro* ORF is expressed by using a –1 frameshift (2082), producing a Gag-Pro polyprotein (Pr110 Gag-Pro). The NC region of *gag* and the amino-terminal region of the *pro* ORF encode a p30 protein that exhibits dUTPase activity . The viral protease is encoded in the carboxy-terminal region of the *pro* ORF.
pol:	(2891–5576) The *pol* ORF is expressed by using two –1 frameshifts (2082, 2891) producing a Gag-Pro-Pol polyprotein (Pr160 Gag-Pro-Pol); the locations of the amino-terminal cleavage sites producing the mature RT and IN proteins have not been reported. Integration produces a 6-bp duplication of cellular DNA flanking the provirus.
env:	(5364–7428) Both the *env* subunits are glycosylated. The SU subunit contains three N-linked glycosylation sites at amino acids 127, 143, and 297, and the TM amino acid sequence contains two N-linked glycosylation motifs at amino acid positions 498 and 557. The *env* region is believed to contain an internal transcriptional promoter and start site (6051) that produces a spliced (6143/7305) subgenomic *sag* mRNA.
PPT:	(7358–7376)
sag:	(7378–8338) The *sag* ORF is located within the U3 region and encodes superantigen proteins (Chapters 8 and 10). Sag proteins contain five sites for N-linked glycosylation. The *sag*-encoded protein undergoes proteolytic processing by a cellular protease; Sag contains two furin cleavage sites. Sag has five glycosylation sites: 79, 89, 93, 131, and 146.
U3:	(7377–8573) The U3 region of MMTV is relatively long (1197 nucleotides) compared to other retroviruses (except spumaviruses). The *sag* ORF (7378–8338) is located entirely within U3. The carboxy-terminal region of the *env* ORF also lies within U3; in the provirus structure, the U3 of the 5'LTR region may contain a transcription initiation site for a spliced subgenomic *sag* mRNA .
R:	(8574–8588)

Figure 3 : Genomic structure, transcriptional and translational products of MMTV.
Left panel : The start and end points of U3, R, U5, PBS and PPT are depicted on the proviral DNA of Mouse Mammary Tumor virus (MMTV). The unspliced (genomic RNA, *gag* mRNA, *gag-pro* and *gag-pro-pol*) and spliced (*env* and *sag*) mRNAs are shown in the relation to proviral DNA. The translational products of Gag, Gag-Pro and Gag-Pro-Pol polyproteins and their subunits are illustrated in left panel. All featured domains are elaborated upon in the table in the right panel. The figure and table were adapted from (Coffin et al., 1997).

2.3 ERV replication process

A full round of ERV mobilization is a well delineated and regulated process. Most of the knowledge about this process comes from studies on retroviruses and LTR retrotransposons (e.g. Ty element in yeast). Retroviruses of distinct genres share most of their replication process in common, although some specific features exist. Moreover, XRVs infect target cells extracellularly, while some ERVs can mobilize intracellularly. Here, firstly the full extracellular replication cycle of an exogenous retrovirus (i.e. MMTV) will be briefly described, as the ERV element that we will focus on in this thesis originates from a betaretrovirus. Then the steps of intracellular retrotransposition will be compared with the extracellular reinfection process, and the differences highlighted when necessary.

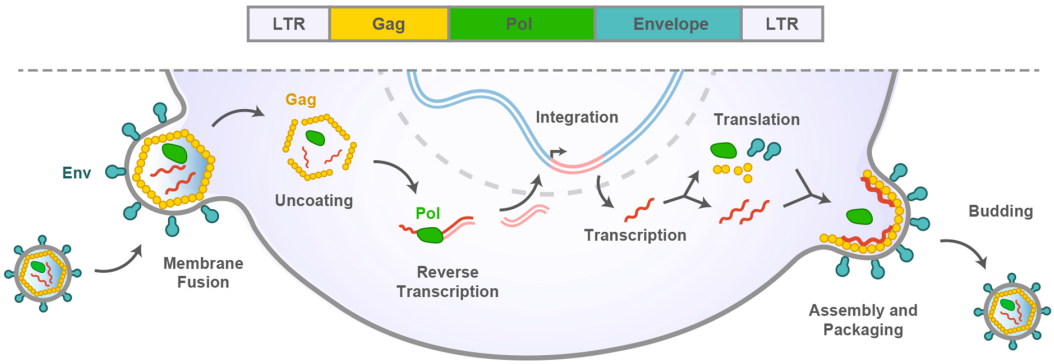


Figure 4: Schematic representation of different stages of retrovirus replication. Infection of cells starts with attachment and fusion of the viral glycoprotein to the cell surface receptor, followed by the release of viral core into the cytoplasm. Reverse transcription of viral RNA begins in the cytoplasm. It then forms DNA, enters nucleus and leading to integration of the viral genome into the host chromosome. Transcription of viral RNA occurs from the integrated "provirus" with the help of regulatory proteins and their *cis*-acting sequences. The unspliced RNAs are then exported from nucleus by the regulatory proteins with the help of host proteins. The mRNA undergo translation. Gag binds to packaging competent RNAs with packaging signal. The Gag-RNA complexes then localize to the plasma membrane where the RNA dimerization occurs during the initial stages of assembly. Finally, the completely assembled virus particles are released. The figure was adapted from Modzelewski et al. (2022).

The whole retroviral replication process is typically divided in eight separate phases: membrane fusion, uncoating, reverse transcription, integration, transcription, translation, particle assembly and budding (Figure 4). An integrated provirus (like ERV) starts the replication process with transcription. **Transcription** is initiated from the promoter within the 5'LTR of the integrated proviral DNA by cellular RNA polymerase II (Pol II). The U3 region that is found upstream of the transcription start site (TSS), contains *cis*-regulatory elements, and the timing and extent of transcription is regulated by the interaction of cellular transcription factors with these elements *in trans* as well as by *cis*-effects of the flanking chromatin states. Like eukaryotic mRNAs, the termination of transcription of retroviral mRNAs involves polyadenylation at the 3' end. The signal for polyadenylation is a highly conserved *cis*-acting sequence AAUAAA located 10–30 bp upstream of polyadenylation site. A copy of full-length genomic RNA (gRNA) that results from transcription contains a unique copy of all of the information encoded in the proviral DNA, plus a short direct repeat at each end termed R (Figure 3). The R region is defined by the TSS in the 5'LTR and the location of 3'end processing event in the 3'LTR. Although, most retroviruses harbor only a single transcription unit, in the case of MMTV, two additional TSSs have been reported. One is present in the U3 region upstream of the U3/R border, and the other TSS is in the Env gene (Figure 3). The cellular machinery caps the 5' end of the gRNA and the 3' end is polyadenylated following splicing. Thus, all viral transcripts are capped and polyadenylated. The gRNAs serve two functions: (i) they encode functional retroviral proteins from both spliced and unspliced RNA isoforms, and (ii) they are packaged into viral particles as genomic RNA. To fulfil the first function, a fraction of retroviral transcripts is spliced, generating subgenomic-sized mRNAs. All

spliced transcripts contain the same 5' end, which spans the U5 region of the 5'LTR, and retain the U3 and R regions encoded by the 3'LTR. While Gag, Pro and Pol are encoded using unspliced viral mRNAs, MMTV, as for all simple retroviruses, contains signals for the removal of at least one intron for the generation of the spliced *env* mRNA using one 5' splice donor (SD) ~300 bp after 5' LTR and one 3' splice acceptor (SA) site at beginning of Env ORF. Alternative splicing can occur to generate the superantigen (*sag*) mRNA in the case of MMTV (Figure 3). Both spliced and unspliced RNAs in retrovirus must be exported out of the nucleus in the cytoplasm, unlike cellular RNAs where only the spliced RNAs are transported out of the nucleus.

The process of **translation** is initiated at the AUG start codon opening the Gag ORF of the unspliced mRNA. Interestingly, Gag-Pro and Gag-Pro-Pol fusion polypeptide precursors are commonly translated from the same unspliced RNAs. Those polypeptides are processed by virally encoded protease to release the enzymatic proteins (Pro and Pol) that are fused to Gag. In the case of MMTV, Gag, Pro and Pol are in three different frames, so Gag-Pro and Gag-Pro-Pol polypeptides were realized by using one and two times ribosomal frameshifting, respectively (Figure 3). Here, occasional ribosomes slip backward one nucleotide (-1 frameshift) at the end of translation of Gag, thus the ribosomes leave the Gag reading frame and shift into an overlapping portion of the Pro reading frame, resulting in the Gag-Pro fusion peptide. The -1 frameshift can happen again at the end of the translation of Pro to generate the Gag-Pro-Pol fusion polypeptide. By using the same initiation codon in the same mRNA to express the Gag, Pro and Pol genes, while occasionally to synthesis the fusion polypeptides via bypassing the Gag termination codon and Pro termination codon, retroviruses ensure the synthesis of appropriate ratios of Gag, Pro and Pol proteins. Env is synthesized from a spliced version of the genomic RNA from which the gag, pro, pol sequences have been removed, and is processed in the same manner as the cellular membrane and secretory proteins. It is glycosylated in the endoplasmic reticulum (ER), followed by proteolytic removal of the leader peptide. In the ER, it is folded, oligomerized, and the transported to the Golgi, where the cellular protease furin mediated cleavage of the polyprotein, resulting in the generation of SU and TM domains (Coffin et al., 1997).

The retroviral particle **assembly** is initiated with the Gag and requires viral RNAs for infectious virus production. The Gag polyprotein provides the basic building blocks of viral particles. Gag alone is sufficient to make non-infectious viral like particles (VLPs). There are many dynamic interactions between the cellular machinery and proteins during this step, including the fusion peptides packaging, cellular membrane budding, and the dimerization and encapsidation of viral RNAs for packaging. Viral RNA only represents a small fraction of the RNA in the cytoplasm, yet it is specifically packaged into virions. There are sequence signals (packaging sequence) in the viral RNA genome that form specific secondary structures. Notably, viral RNAs are present in virions made with Gag as the only constituent, which implies that Gag is the only protein required for selective packaging of viral RNAs. In general,

the packaging sequences are located at the 5'end of the viral RNA, although the exact sequences for packaging vary between retroviruses (Mustafa et al., 2012; Chameettachal et al., 2021). Moreover, transcripts without packaging sequence can be packaged into virions albeit with very low efficacy, suggesting that packaging sequences enhance selective packaging. The Gag polyprotein is sufficient not only for VLP formation, but also for targeting the retroviral assembly to the plasma membrane. To this end, the MA domain of Gag plays a crucial role in determining the subcellular location of the viral assembly. The type-C retroviruses (alpha-, gamma-, and lentiviruses) assemble on the inner leaflet of the plasma membrane, facilitated by a highly basic region in the MA domain and a hydrophobic myristic acid residue while in case of B- and D-type retroviruses, virion assembly occurs in cytoplasm and pre-immature particles target to the plasma membrane for **budding and release** (Zábranský et al., 2009). Core particles assembled in the cytoplasm or at the plasma membrane have an immature morphology. The Env polyprotein precursor is proteolytically cleaved into an external glycosylated peptide (SU) and a membrane spanning protein (TM) during its transport to the surface of the cell, and together these form an oligomeric spike on the surface of the virion. Following release, the immature virions undergo proteolytic processing and extensive structural rearrangements. After this obligatory maturation step, infectious viral particles formed.

To infect target cells, virions must enter these cells. This step is initiated when the SU domain of the virion binds to a specific receptor molecule on the target cell. This binding activates the **membrane fusion**-inducing potential of the TM protein that allows the viral and cell membranes to fuse (Figure 4). The specificity of the SU/receptor interaction defines the tropism of a retrovirus. For instance, human immunodeficiency viruses (HIV) interacts with the CD4 receptor found on the surface of T cells to trigger fusion of viral and cellular membranes (Douek et al., 2002).

At some point after fusion and before integration into host DNA, the viral capsid needs to be shed in a process termed **uncoating** and the RNA genome is reverse transcribed into DNA. However, the timing and mechanism of uncoating and the role of the capsid in early replication remains debate (Muller et al., 2022). **Reverse transcription** starts at the 3'end of a cellular tRNA that primes reverse transcription by hybridizing to the primer binding site (PBS). A short, first-strand cDNA containing R and U5 is synthesized by the RT, followed by the degradation of the template RNA by its RNase H activity. The first strand transfer step moves the short cDNA segment to the 3'end of the viral RNA by hybridizing to R. First strand cDNA is then synthesized up to the PBS. At the same time, the remaining RNA is degraded by RNase H, generating primers for second-strand cDNA synthesis. A synthesized short second strand cDNA is then transferred (second strand transfer) to the 5'end of the first-strand cDNA to complete first and second strand cDNA synthesis, yielding a full-length, double-stranded DNA molecule with two identical LTRs. Notably, two viral RNAs are incorporated into a capsid and template

switching can occur during DNA synthesis, thus allowing for recombination between the two (pseudodiploid) genomes.

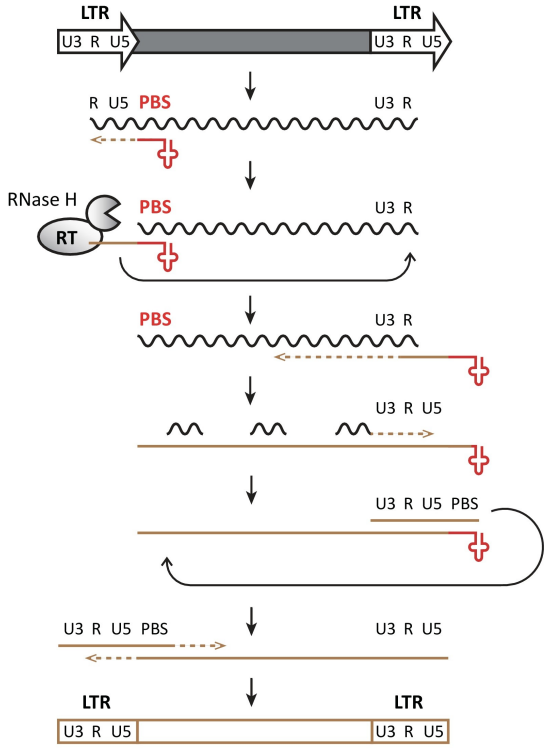


Figure 5: Reverse transcription of retrovirus.
LTRs encode promoter elements and termination signals. The RNA transcript contains a repeat region at either end (R), a segment unique to the 5'-end of the RNA (U5), and a segment only included at the 3'-end of the RNA (U3). The 3'-end of cellular tRNAs (red cloverleaf) primes reverse transcription by hybridizing to the primer binding site (PBS). While this segment is being copied into first-strand cDNA (brown line), also called (–)ssDNA (negative strand strong stop DNA), the RNase H activity of reverse transcriptase (RT) degrades the template RNA. The elongating cDNA is transferred to the 3'-end of the retrotransposon transcript by hybridizing to the R region. The remaining RNA is partially degraded by RNase H, leaving behind primers for second-strand cDNA synthesis. After another transfer event, first- and second-strand synthesis are completed, generating a full-length, double-stranded retroviral DNA that will be integrated into the host genome. The figure and legend were adapted from Schorn and Martienssen (2018).

The preintegration complex (PIC) that is primarily composed of dsDNA, integrase (IN) and host factors must gain access to chromosomes before IN can catalyze its **integration** in the host genome. Different viruses use distinct mechanisms to ensure this outcome. Lentiviruses like HIV are actively transposed to nuclei presumably through nuclear core complexes (NPCs) that reside within the nuclear membrane. Other retroviruses like MLV in contrast require nuclear membrane break down during the M phase of the cell cycle, explaining the inability of these viruses to infect non-dividing cells (Yamashita and Emerman, 2006). Upon entry or release of PIC into the nucleus, host cofactors guide the complex to specific chromatin contexts. In the case of lentiviruses, Lens Epithelium- Derived Growth Factor/p75 (LEDGF/p75) guide PICs to actively transcribe chromatin (Cherepanov et al., 2003; Shinn et al., 2002),

while MLV PICs adopt Bromodomain and Extraterminal domain (BET) proteins to approaching transcription starting sites of active transcribed genes (DeRijck et al., 2013). Subsequently, the integrase (IN) selects the final site, recognizes a target DNA strand, and catalyzes the cutting and joining reactions that fuse viral and host genomes. In a reaction called strand transfer, the viral DNA ends are inserted 4 to 6 bp apart (depending on the retroviral species) into opposing strands of the target DNA (Figure 6). The remaining single-strand gaps are repaired by the host's cell machinery, yielding 4 to 6 bp duplications (called target site duplication, TSD) flanking the provirus, a hallmark of retrotransposition events.

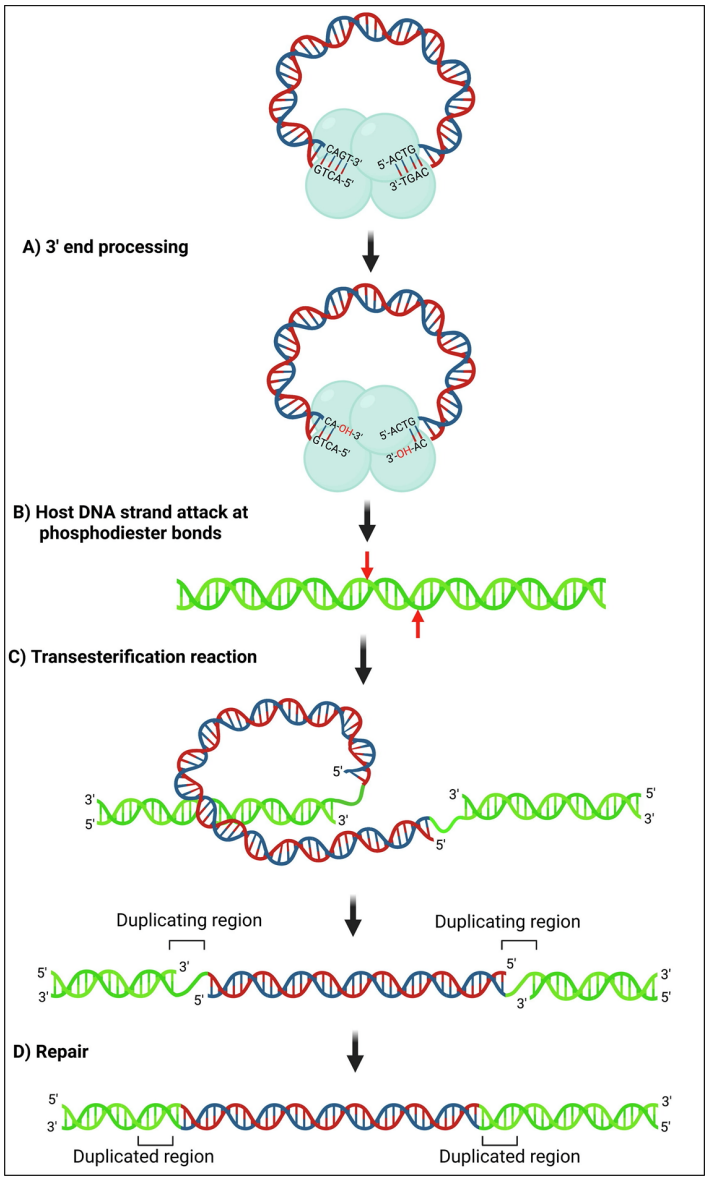


Figure 6: Different stages of the retroviral integration process.
A) The integrase enzyme cleaves at the 3' ends of the viral DNA by its endonuclease activity, releasing 'GT' dinucleotides from the 3' end, resulting in the generation of 5' overhangs. **B & C)** The 3'-OH groups of the ends of the viral DNA attack the phosphodiester bonds on target DNA and the viral DNA 3' ends are joined to the target DNA by a transesterification reaction. The distance between the two phosphodiester bonds at which the transesterification occurs is

dependent upon the virus. **D)** The nucleotide gaps are then repaired by the host cell machinery, resulting in duplication of the site of integration which now flanks the integrated provirus. Figure was adapted from Chameettachal, Mustafa, & Rizvi (2023).

The process of intracellular retrotransposition shares transcription, translation, reverse transcription, target site selection and integration with extracellular infection. It differs in the cellular location of VLPs. VLPs of IAPs are addressed to the cisternae of the endoplasmic reticulum instead of the cell membrane (Heldmann and Heidmann, 1991). Apparently, intracellular retrotransposition doesn't require the function of the Env protein (Dewannieux et al., 2004).

2.4 Endogenization, expansion and demise

ERVs are thought to have evolved from ancient retroviral infections. When an XRV infects its host, the proviral DNA is integrated into the host genome, generally in the targeted somatic cell lineage. Infectious viral particles are then generated, released and transmitted from one host animal to the other by contact or proximity. Infection can also be accomplished by experimental or iatrogenic injection of the retrovirus. Alternatively, transmission may occur from parent to offspring. These two routes of transmission are referred to as horizontal and vertical transmission, respectively. Vertical transmission of a virus from parents to offspring can occur by two different modes: congenital/perinatal infection or genetic transmission. The congenital and perinatal infection as well as horizontal transmission require infection, penetration of the cell membrane, synthesis of provirus, and integration of the virus into the cellular genome. Thus, this type of infection always results in the acquisition of new genetic information by the infected cell. On rare occasions, retroviruses may infect the germ line, allowing the integrated proviral sequences to be transmitted by the gametes (i.e. vertically) to the next generation (Dewannieux and Heidmann, 2013; Jaenisch, 1976). This process is called retroviral endogenization. The fate of such endogenized ERV is determined, as for any kind of *de novo* mutation, by random drift and natural selection. If will often be lost in subsequent generations by random drift and leave no trace in the species' genome. Most ERVs attest of past endogenization events, except for a few known examples including one Koala retrovirus (KoRV) that is currently infecting the Koala germline and is the well-studied retrovirus known to be currently transitioning from exogenous to endogenous form in a natural setting, providing us with the opportunity to study such genome colonization as it takes place (Tarlinton et al., 2006). Notably, it has been shown that KoRV is present, in variable copy numbers, in the germline of all Koalas found in Queensland, but absent in the Koalas from the Kangaroo island off the coast of South Australia. This island was stocked with koalas in the early part of the twentieth century and has remained essentially isolated since then. It appears most likely that the small founding population was entirely free of KoRV (Tarlinton et al., 2006).

Unlike endogenizing KoRV, most ERVs in mammalian genomes are not directly related to an extant XRVs, suggesting that the XRVs responsible for seeding ERVs became extinct or diverged. However,

several ERVs with XRVs homologs have been documented in the literature (Johnson, 2019). One well-studied example is jaagsiekte sheep retrovirus (JSRV) and endogenous retroviral sequences related to JSRV (enJSRV). JSRV is the etiological agent of ovine pulmonary adenocarcinoma (OPA) (Palmarini et al., 1999). The distribution of enJSRV within sheep populations, among small ruminants, and other vertebrate species has been extensively studied by Southern blot hybridization (Hecht et al., 1996). These studies confirmed the presence of enJSRVs in the sheep and goat genomes, suggesting that retroviral endogenization predates the *Ovis* and *Capra* genera divergence (Armezzani et al., 2014). There are other XRV-ERV pairs similar to JSRV/enJSRV, including Avian sarcoma leucosis virus (ASLV)/enASLV, MMTV/*Mtv*, Feline leukemia retrovirus (FeLV)/enFeLV, and Murine leukemia retrovirus (MuLV)/*Emv* (Chiu and Vandewoude, 2021). The interaction between ERVs and XRVs during the brief stage shortly after endogenization provides a unique opportunity to witness host evolution in action.

Following endogenization ERVs will experience an expansion phase in which the copy numbers of ERVs increase. The size of each group of ERVs can vary significantly, presumably reflecting the ease and extent with which viral amplification took place following the initial germ-cell infection. Differential rates of ERV amplification may reflect the properties of the initial provirus. In the mouse genome, ERV family size ranges from a few elements to hundreds. There are many factors including host population dynamics, the ERV itself, and the interaction of those two that may impact the amplification process. One factor - the mechanism of amplification - is believed to play an important role (Figure 7). It is not very hard to imagine that - in the early phases of this process - the provirus maintains the ability to produce infectious viral like particles which produce new integrations by reinfection. This is evidenced by the fact that some Env-containing ERVs can produce VLPs ex-vivo and that those VLPs are infectious (Dewannieux and Heidmann, 2013). This is also in agreement with an *in silico* study of HERVK elements which showed that the Env gene has been under strong purifying selection whose integrity is required for the reinfection process (Belshaw et al., 2004). Conversely, it seems that the more successful families (regarding copy numbers) correspond mainly to elements devoid of Env (Magiorkinis et al., 2012). The two active autonomous ERV families in mouse ETn/MusD and IAP are env-less elements (Dewannieux et al., 2004; Ribet et al., 2007). *In silico* analysis for other ERVs in other genomes showed also that env-less ERVs amplify to higher copy numbers than those with an Env gene. *In vitro* experiments have shown that the transition from extracellular reinfection to intracellular retrotransposition involves sequence changes. Studies of IAP (Intracisternal A-type Particles) and IAPE (Intracisternal A-type Particles elements with an Envelope), two closely related ERVs in the mouse, have shown that in addition to the loss of Env, changes in the N-terminal end of the structural Gag protein cause IAP particles to be addressed to the cisterns of the endoplasmic reticulum whereas particles of IAPE are targeted to the cell membrane (Ribet et al., 2008). It is very hard to know if this transition is a general process during ERV expansion or is a feature of a subset of ERVs. Of note,

it is not uncommon that defective elements of ERV families arose during the stage. They are not able to encode complete proteins required for mobility due to accumulation of deleterious mutations or recombination with degraded retroelements (see hereafter) but are capable of coping themselves via complementing *in trans* by hijacking machinery from other competent elements (Figure 7). For instance, the mobility of mouse ETn elements fully rely on competent MusD copies (Ribet et al., 2004).

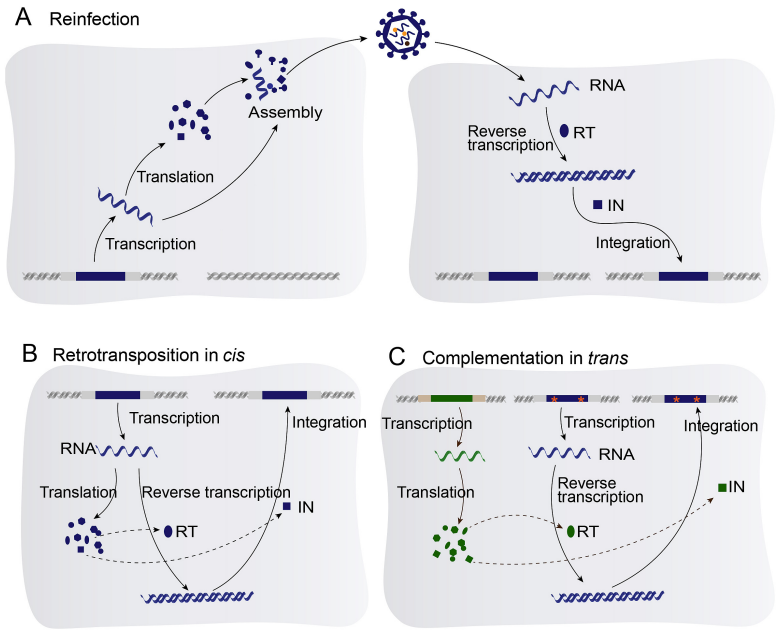


Figure 7: Proliferation mechanisms of ERVs.
A) Reinfection. ERVs in germline cells or somatic cells produce intact virus particles that can be reintegrated into the chromosomes of other germline cells. **B**) Retrotransposition in *cis*. Viruses use their own proteins to proliferate in germ cells, which requires ERVs to have functional *gag* and *pol* genes. **C**) Complementation in *trans*. ERVs increase their copy numbers using proteins encoded by other ERVs (shown in green). ERVs that proliferate in this way do not require functional *gag*, *pol* and *env* genes. The asterisks represent disruptive mutations. Abbreviation: RT, reverse transcriptase; IN, integrase. The figure and legend were adapted from Zheng et al. (2022).

The above-mentioned mechanisms require that an ERV can be transcribed and encode a minimal set of replicative enzymes collectively. The probability of copy-number expansion by those mechanisms may diminish over time as ERV sequences are subject to germline mutations accumulation, preventing the generation of functional infectious particles or they become replication defective (neither reinfection nor retrotransposition). Moreover, the host develops defense systems against the whole process of mobilization of ERVs (see next section). The ERV family eventually dies out in terms of mobilization. Individual ERV loci may become fixed in the host population (present in all genomes in a population) by drift or can be cleared because of negative effects on host fitness. For each ERV family, it went through an endogenization-expansion-demise process, and kinetics of process varies among distinct ERV families.

2.5 ERV classification and nomenclature

There is no uniform way to classify nor to name ERVs, although some proposals have been made (Mager and Stoye, 2015). As the ERVs originate from continuous waves of invasion by exogenous retroviruses and share similarities with exogenous retroviruses, one convenient way to classify ERV was to use the rules used by virologists to classify XRV. According to the International Committee for virus taxonomy, the Retroviridae family is grouped into the order Ortervirales, and comprises two subfamilies, Orthoretroviruses and Spumaretroviruses. Orthoretroviruses comprise six exogenous genera including Alpharetroviruses, Betaretroviruses, Gammaretroviruses, Deltaretroviruses, Epsilonretroviruses and Lentiviruses (https://ictv.global). Large scale phylogenetic analysis integrating XRVs with ERVs sequences demonstrate that ERVs representing all six genera have been detected with ERVs from gamma- and betaretroviruses being the most abundant groups (Hayward et al., 2015). The phylogenetic Pol tree made by ERV sequences with exogenous retroviral sequences shows the seven retroviral genera with ERV sequences placed into three classes (I, II and III) (f.i. Pol). These three classes are primarily related to the exogenous Gammaretrovirus, Betaretroviruses and Spumaretroviruses, respectively (Jern et al., 2005).

The current ERV nomenclature system was developed somewhat arbitrarily reflecting the history of ERV discovery (Mager and Stoye, 2015). Originally, endogenous proviruses were named after the most closely related XRVs (f.i. murine leukemia virus, MLV). It is common to add one or two letters before the ERV to indicate the species in which the ERV was initially identified. For example, “HERV” indicates an ERV first seen in human, MuERV stands for an ERV originally found in mice, and BoERV implies the ERV of Bovinae. HERV can be further classified based on the tRNA that binds to the viral primer binding site (PBS) to prime reverse transcription. For instance, HERVK stands for an ERV using lysine (=K) tRNA. There is no uniform manner to name individual ERV loci at discrete chromosomal locations. The most common way for this purpose is to simply add a number at the end. On some occasions, ERV loci have been denoted based on cytogenetic designation such as “HERVK 11q22” located on the q-arm, chromosomal band 22, of human chromosome 11 (Gifford and Tristem, 2003). The publicly downloaded TE annotations of most assembled genomes were generated by RepeatMasker so the naming of ERVs in repeats tracks follows RepBase conventions (Bao et al., 2015). As explained before, this annotation is based on sequence homology to a set of consensus sequences. Thus, the naming not necessarily reflects phylogenetic relationships among ERVs. Moreover, RepBase annotation treats LTRs and internal sequences of ERVs separately, and the features of internal proviruses need to be further annotated by specific tools such as LTRdigest. The lack of a uniform strategy to classify and name ERVs is worrisome as there is currently no way to link the ERV annotation from one study to another, even when working on the very same species with the same genome assembly.

2.6 ERV annotation

Systematic discovery and annotation of ERVs has usually been conducted together with other TE classes, although specific bioinformatic tools have been developed for ERVs discovery and annotation (Goerner-Potvin and Bourque, 2018). TE discovery can be performed with and without genome assembly. There are mainly two strategies when a genome assembly is available. The first is repository-based annotation, in which genome sequences are queried against a curated collection of TE consensus sequences. The second is *de novo* annotation. It is noteworthy that those two strategies are often implemented in combination. The most widely used tool for this "with genome assembly" strategy is RepeatMasker that queries against the RepBase and Dfam databases and can provide annotated lists of repetitive sequences for a wide range of eukaryotic organisms. Such annotation is crucial for the study of TEs and is deployed as tracks on the UCSC or other genome browsers. However, the quality of annotation using this approach heavily depends on the representatively and completeness of the used database (Ou et al., 2019). Instead of using existing databases, one can build a library of consensus sequences *de novo*. The most common tool for *de novo* annotation of TE for assembled genomes is RepeatModeler, which identifies TEs using two distinct algorithms, RepeatScout and RECON, followed by consensus building and classification steps. The newest version of RepeatModeler (RepeatModeler2) includes a module (LTRharvest) to improve the assembly of ERVs, which were previously too often identified by RepeatScout/RECON alone as fragments with disassociated LTRs and internal regions or even completely missing the internal segment (Flynn et al., 2020). The output of consensus sequences can then be used directly or merged with Repbase's library, to obtain a more comprehensive consensus sequence library. The new library can then be fed to RepeatMasker to identify individual copies in the genome.

2.7 Detection of polymorphic ERV loci

There is huge interest in detecting ERVs (polymorphic ERVs), that are not present in the reference genome, by mining whole genome resequencing (WGS) datasets. Similar to other types of genetic variation (f.i. single-nucleotide polymorphisms), such polymorphic insertions could be associated with specific phenotypes. Numerous tools have been developed for this purpose (see review in (Goerner-Potvin and Bourque, 2018) and benchmarking in (Vendrell-Mir et al., 2019). Most of these tools take alignment files of pair-end next generation sequence (NGS) reads (typically 100-150 bps) from WGS as input. Non-reference ERV insertions are found using combined information from split-reads and discordant read pairs. The ERV insertion generates split reads at the insertion site, with one segment of the read aligning to the reference genome and the remainder to the beginning or end of an ERV sequence. Discordant read pairs result from one read flanking (and pointing towards) the insertion site and its mate mapping to ERV sequences. The primary identification is often followed by a series of refinements and filtrations that exploit various features of ERV insertions.

3 ERV activities in the germline and host responses

During the expansion phase, ERVs are active in germline cells, which produce eggs and sperm. This can result in the creation of new heritable copies of ERVs through processes of extracellular reinfection or intracellular retrotransposition. The activity of ERVs in the germline leaves detectable marks in the germline, including (i) chromatin alterations in the ERV vicinity, (ii) ERV transcripts, (iii) ERV encoded proteins, (iv) *de novo* ERV insertions. Since ERV activity poses a threat to the integrity of the host's genome, it has developed defense strategies.

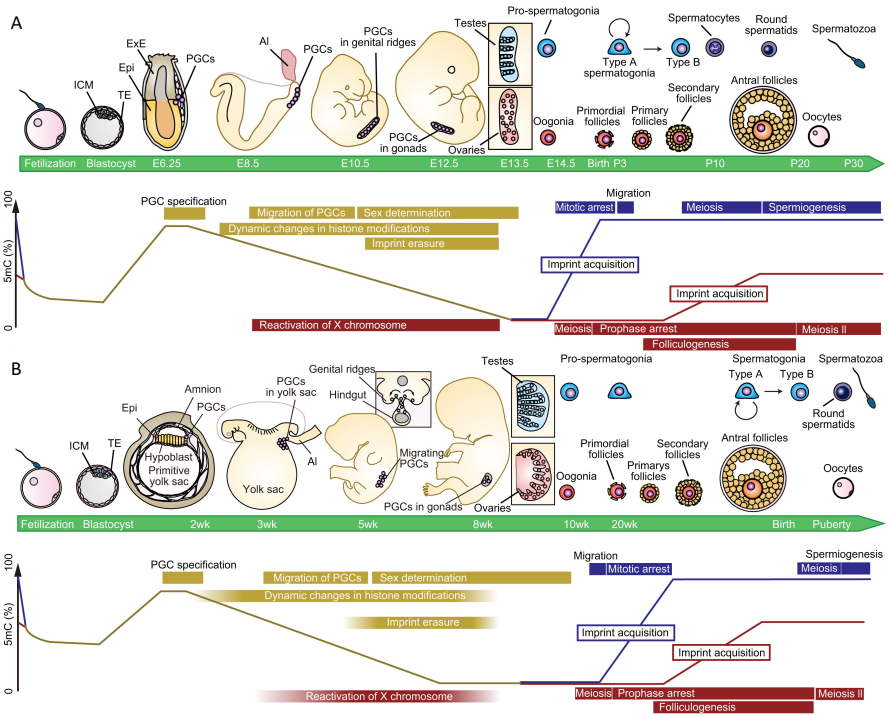


Figure 8: Mouse and human germline development.
Top: schematic of the development of the mouse (**A**) and human (**B**) germ cell lineage. Bottom: key developmental events associated with germ cell development with the dynamics of the 5mC levels. The figure was adapted from Saitou and Miyauchi (2016).

3.1 Germline development

Germline development refers to the differentiation of cells that will give rise to gametes (sperm and eggs). The process of germline development involves several stages and is essential for the transmission of genetic information from one generation to the next. Mouse and human are the most well-studied mammalian species with regard to germline development (Figure 8). Germline development of mice and human begins during embryogenesis when a small group of cells, called primordial germ cells (PGCs), originating from the epiblast, are specified. This step occurs around embryonic day (E) 6.0 and around 2 weeks after fertilization in mice and humans, respectively. Following specification, PGCs migrate to the hindgut, dorsal mesentery, and ultimately into the genital ridges at around E10 in mice

and 4-5 weeks in human (Saitou and Miyauchi, 2016). Once PGCs reach the developing gonads, they undergo a phase of rapid proliferation. This proliferation increases their numbers and ensures that enough germ cells are available for subsequent stages of development. In male fetuses, following colonization and proliferation in the gonads, PGCs enter into mitotic arrest and differentiate into pro-spermatogonia. In male mice, pro-spermatogonia located in seminiferous tubules remain arrested until around birth, a subset of pro-spermatogonia translocate to the basal compartment of seminiferous tubules and differentiate into spermatogonia. At around postnatal day 10, the vast majority of these spermatogonia enter in a first wave of spermatogenesis, resulting in a first wave of spermatozoa at around postnatal day 30. A small population of spermatogonia generates spermatogonial stem cells (SSCs), which will sustain subsequent waves of spermatogenesis throughout reproductive life. Human males lack the equivalent of this first wave of spermatogenesis of male mice. Instead spermatogonia are believed to remain undifferentiated until prior to the puberty (Guo et al., 2021, 2020; Saitou and Miyauchi, 2016). In adult human males, SSCs undergo self-renewal, meaning that they can divide and give rise to SSCs, ensuring a constant population of stem cells throughout adulthood. This self-renewal capacity allows SSCs to produce sperm cells throughout a male's reproductive life. Spermatogenesis begins in the seminiferous tubules, which are coiled structures located in the testes. The process of spermatogenesis can be divided into three main phases: the proliferative phase, the meiotic phase, and the maturation phase. The process starts with the proliferation of germ cells through mitosis. Spermatogonia divide and differentiate into primary spermatocytes. The primary spermatocytes then enter meiosis. During the first meiotic division (meiosis I), the replicated chromosomes pair up and exchange genetic material through a process called genetic recombination or crossing over. This creates genetic diversity among the resulting sperm cells. Each primary spermatocyte divides into two secondary spermatocytes, each containing a set of replicated chromosomes. The second meiotic division (meiosis II) follows, where each secondary spermatocyte divides into two spermatids. The spermatids undergo a process called spermiogenesis, during which they transform into mature spermatozoa. Sperm production is a highly synchronized and continuous process, with millions of sperm cells being produced daily. The entire process of spermatogenesis takes approximately 64 to 72 days in humans, 35 days in mice and around 60 days in cattle from the initial division of spermatogonia to the release of mature spermatozoa into the lumen of the seminiferous tubules (Staub and Johnson, 2018).

In addition to germ cells, Sertoli cells are the most important somatic cells in the testes. Sertoli cells derive from supporting cell precursors in the genital ridge. After sex determination, some supporting cell precursors differentiate into Sertoli cells, which then form the seminiferous tubules' supportive structure (Yang and Oatley, 2015). They are large, columnar cells that extend from the base to the lumen of the seminiferous tubules, creating a supportive framework for germ cell development. Sertoli cells have several functions, including providing essential nutrients and factors required for the growth and maturation of germ cells (Smith et al., 2015; Yao et al., 2015). Leydig cells are located in the interstitial

spaces between the seminiferous tubules and originate from mesenchymal cells in the interstitial tissue of the testes. Under the influence of luteinizing hormone (LH) from the pituitary gland, these cells differentiate into Leydig cells and start producing testosterone during fetal development. The number of Leydig cells increases significantly during puberty. These cells are responsible for the production and secretion of testosterone, a critical male sex hormone that is essential for the development and maintenance of male reproductive structures and secondary sexual characteristics (Sang et al., 2023; Zirkin and Papadopoulos, 2018).

In females, PGCs likewise undergo substantial proliferation in embryonic ovaries but then enter the first meiotic prophase to differentiate into oocytes. The oocytes are arrested at the diplotene stage of the first meiotic prophase. Around birth, a layer of granulosa cells surrounds the oocytes to form primordial follicles. The follicles develop into primary, secondary and then antral follicles. In mice, around 6 weeks after birth, when estrus commences, the development of oocytes resumes with the completion of meiosis I that results in the formation of two cells: a secondary oocyte and the first polar body. The secondary oocyte immediately enters meiosis II but is then arrested in the metaphase stage. This stage is only completed if fertilization occurs. Alternatively, the secondary oocyte degenerates if fertilization does not occur. In women, PGCs continue to proliferate until week 10 of embryonic development and then enter into meiosis to differentiate into oocytes. Moreover, proliferation appears to continue until week 20. Notably, folliculogenesis progresses during the embryonic period, and mature follicles are occasionally formed before birth.

3.2 Two waves of epigenome reprogramming

Epigenetic modifications of the genome that regulate crucial aspects of its function are generally stable and heritable in somatic differentiated cells. There are, however two phases of drastic epigenome reprogramming that occur at distinct stages of mammalian development. The first wave occurs during early embryonic development, and the second wave occurs during germ cell development. These processes are characterized by dynamic modifications of epigenetic marks, including DNA methylation and histone modifications that regulate gene expression (Surani et al., 2008). Epigenome reprogramming plays a crucial role in establishing and maintaining cell identity and function.

During the early stages of embryonic development, epigenome reprogramming erases the epigenetic marks inherited from the gametes (sperm and egg) and establishes a totipotent state, procuring cells the potential to differentiate into any cell type of the body. This first wave of reprogramming involves global demethylation of the genome, removing most of the DNA methylation marks that were inherited from the gametes. This process allows for the activation of key developmental genes and the establishment of a pluripotent cell population. The global demethylation is mediated by enzymes called DNA demethylases, such as the ten-eleven translocation (TET) proteins, which actively remove methyl groups

from the DNA, and by passive loss due to the lack of maintenance methylase during DNA replication for both paternal and maternal genome, with exception of imprinted genes and certain repeats (Guo et al., 2015; Shen et al., 2014) Additionally, histone modifications, such as H3K4me3, H3K9me2/3, and H3K27me3, also undergo changes during this wave of reprogramming (Xu and Xie, 2018). These modifications help establish an open and permissive chromatin state that promotes expression of genes necessary for early embryonic development (Guo et al., 2014; Smith et al., 2014; Xu and Xie, 2018; Zhu et al., 2018). After reaching the lowest methylation level at the blastocyst stage, the embryo proper then undergoes genome-wide re-methylation, establishing a canonical methylation landscape typically found in somatic cells. This process occurs in a short time window (E4.5-6.5 in mice) (Figure 8), and is achieved along with the increase level of *de novo* DNA methyltransferases (DNMT3) at implantation (Chen et al., 2020).

The second wave of epigenome reprogramming occurs during germ cell development. It is crucial for the erasure of epigenetic marks acquired during somatic cell development and the establishment of a new epigenetic state specific to germ cells. During this process, the DNA methylation patterns are largely erased once again, returning the genome to a relatively unmarked state. This ensures that the germ cells have the potential to differentiate into either male or female gametes. Following the near complete erasure of genome-wide DNA methylation in PGCs at E13.5 for mice, male and female germ cells re-methylate their genome in a sex-specific manner. *De novo* DNA methylation in male germ cell is fully established at birth, whereas female germ cells acquire DNA methylation after birth. Besides a different re-methylation timing, mature sperm and oocytes also have very different methylomes (Figure 8). The sperm DNA is globally methylated while oocyte genomes consist of consecutive hyper- and hypomethylated domains (Figure 8) (Gkountela et al., 2015). DNMT3A and DNMT3L are responsible for *de novo* DNA methylation during oogenesis using a transcription-guided mechanism (Veselovska et al., 2015), while the *de novo* DNA methylation in male germ line occurs by default at most genomic elements except the regions marked with H3K4me2 (Singh et al., 2013). The second wave of reprogramming also involves histone modifications, including changes in histone methylation and histone variant incorporation, which contributes to the establishment of the germ cell-specific epigenome (Gkountela et al., 2015; Gruhn et al., 2023; Guo et al., 2015; Hill et al., 2018; Huang et al., 2021; Seisenberger et al., 2012).

3.3 ERV activities during the early embryo and germline development

ERV expression profiles have been extensively studied along with other epigenome profiles in pre-implantation embryos of mammals. By tracking RNA-seq reads aligning to ERVs, Modzelewski et al. established comprehensive retrotransposon expression profiles during early embryonic development for multiple eutherian mammals (Modzelewski et al., 2021). This revealed that pre-implantation embryos from different species exhibit a similar global ERV expression profile with a major switch at zygotic

genome activation. Moreover, expression patterns appear to be specific for distinct ERV subfamilies. For instance, murine MERVL exhibit specific activation at the two-cell stage, and depletion of the MERVL transcripts results in embryonic lethality due to defects in early lineage specification and genome stability. This suggests that MERVL may be essential for mouse preimplantation development (Sakashita et al., 2023). Similarly, ERVs families are transcribed during human early embryogenesis in a stage-specific manner (Göke et al., 2015). Mouse IAP shows stage-specific expression in the germline. In 1996, Dupressoir and Heidmann generated IAP LTR-driven LacZ reporter mice to test transcriptional activity of IAP and found that IAP's transcriptional activity is limited to prespermatogonia and undifferentiated spermatogonia (Dupressoir and Heidmann, 1996). Another study demonstrated that expression of TEs, including ERVs, is dynamically altered when germ cells enter meiosis during spermatogenesis, suggesting that ERV activity is a precisely controlled developmental processes in spermatogenesis (Sakashita et al., 2020). Of note, most studies on ERV activity in early embryo track RNA expression and chromatin changes at the family or subfamily level. They fail to monitor the activity of individual ERV integrants. The repetitive nature of ERV coupled with the use of short-read sequencing methods, precludes reliable mapping to unique genomic sites. Furthermore, knowledge about ERV presence/absence polymorphisms, which may relate to the most dynamic/younger group of ERVs, is not exploited in these analyses.

There are a few studies focusing on the detection of protein products from specific ERV subfamilies in early embryos or germ cell development. Gag protein expression of IAP has been identified in mice in PGCs (Huang et al., 2021) and primordial follicles using immunofluorescence staining (Malki et al., 2014), while ERVL Gag has been detected at the 2 and 4-cell stages (Sakashita et al., 2023). Similarly, Gag of HERVK was detected in human blastocyst (Grow et al., 2015). Strikingly, VLPs of HERVK have been visualized in both human blastocysts (Grow et al., 2015) and placenta (Lyden et al., 1994), although no germline *de novo* integrations have been detected. This indicates that detecting gene expression or the presence of VLPs alone is not sufficient to conclude that there is an active mobilization of ERVs.

3.4 Regulation of ERV activities during the early embryo and germline development

Global DNA demethylation leads to a hypomethylation state of ERVs, thereby it may threaten genome integrity. The host has evolved multiple strategies to control the activity of ERVs at both transcriptional and post-transcriptional levels. Transcriptional silencing of ERVs during early development includes the establishment of repressive epigenetic marks and the action of transcriptional repressors. In the male germline, DNA methylation, chromatin modifications, and the piRNA pathway are crucial for retrotransposon silencing.

Kruppel-associated box zinc-finger proteins (KRAB-ZFP) are the largest subgroup of the tandem C2H2-type ZFP family, which is the largest transcription factor family in the mouse and human genomes (Wolf and Goff, 2009). KRAB-ZFPs are characterized by two domains: the N-terminal KRAB domain and the C-terminal tandem array of C2H2 zinc fingers. The C2H2 zinc fingers mediate binding at specific DNA sequences, whereas the KRAB domain mediates the recruitment of TRIM28 (tripartite motif-containing protein 28, also known as KAP1) which functions as a scaffold for other repressive histone-modifying factors, including the histone methyltransferase (SETDB1; SET domain bifurcated 1) that generates H3K9me3 (histone H3 trimethylated at lysine 9) (Wolf and Goff, 2007; Yang et al., 2017). These ZFPs recognize motifs characteristic of ERV elements and recruit cofactors that will methylate proximal DNA and histones hence blocking local transcription (Wolf et al., 2008). *De novo* H3K9me3 marks in mouse and human embryos are established and progressively strengthened, especially around most ERV elements (Xu et al., 2022). Interestingly, *in silico* analysis indicates that the number of KRAB-ZFPs strongly positively correlates with the number of ERVs in the genome, suggesting their coevolution. KRAB-ZFPs emerge in response to germline invasion of new retroviral families and are eventually lost from the genome as their target ERVs decay, unless they are co-opted for important functions beneficial to the host (Imbeault et al., 2017; Wolf et al., 2020).

P-element induced Wimpy testis (Piwi)-interacting RNAs (piRNAs) are small non-coding RNAs that, with their partner Piwi proteins, constitute another layer of defense against ERVs during germ line development. piRNA and Piwi proteins are highly expressed in the germline and gonadal tissues. Defects of gametogenesis and infertility has been observed in piRNA pathway mutants, correlating with derepression of TEs, including ERVs (Carmell et al., 2007; Guan and Wang, 2021). Mechanically, the piRNA machinery silences TE through two known mechanisms. The first acts post-transcriptionally in the cytoplasm. TE-derived piRNAs are loaded onto Piwi proteins with slicer activity, and they target TE transcripts for degradation. The second acts transcriptionally in the nucleus. piRNAs target nascent ERVs transcripts by sequence complementarity, mediate *de novo* DNA methylation by interacting with components of the *de novo* methylation machinery (DNMT3A and DNMT3L) causing transcriptional silencing of ERVs (Zoch et al., 2020), and recruit the chromatin remodeler SETDB1 to establish heterochromatin (Aravin et al., 2008; Ernst et al., 2017; Kuramochi-Miyagawa et al., 2008; Ozata et al., 2019).

Other means of ERV silencing have been revealed recently. Tien et al. demonstrated that Polycomb complex mediated H3K27me3 regulates ERV repression in the newly hypomethylated germline genome. Genetic loss of Ezh2 (a unit of Polycomb complex 2) leads to upregulation of ERVs expression level (Huang et al., 2021). The human silencing hub (HUSH) complex (conserved from fish to humans) can recruit and deposit H3K9me3 marks, silencing retrotransposons (Tchasovnikarova et al., 2015). Intriguingly, HUSH recognizes and initiates silencing of intronless (a feature of ERV transcripts)

transgenes, thus distinguishing self from non-self DNA without previous exposure to the host (Seczynska et al., 2022). Abundant tRNA-derived small RNA (tRF) can inhibit ERV activity by targeting the PBS essential for ERV reverse transcription (Schorn et al., 2017). Recently, the RNA modification N6-methyladenosine (m6A) was also implicated in ERV regulation (Chelmicki et al., 2021). However, it remains unresolved whether and how these pathways participate in ERV silencing during early embryo or germline development as they were studied *in vitro*.

4 ERV evolution, forces and consequences

The fate of *de novo* ERV insertions in populations is affected by random drift and selective forces like any other *de novo* mutation in the genome. Furthermore, they are themselves subject to all kinds of *de novo* germline mutations and recombination. These processes jointly determine ERV evolution.

4.1 Population genetics of ERV locus

Initially, like for any form of germline *de novo* mutation, the “derived" integrated ERV locus corresponds to the minor allele, whereas the un-integrated allele in the chromosome represents the major allele. It thereafter is subject to the evolutionary processes of selection and random drift, with time it will eventually either be fixed or lost in the host population. The most likely fate for a novel integrated allele is to be lost by chance. In addition to this effect of random drift, ERV loci may be selected against because they reduce the reproductive fitness of the host. There are very little data regarding the rate at which ERVs form and are lost or fixed. These processes are thought to depend on various factors including *de novo* transposition/insertion rate, effective population size, and impact of insertion on host fitness. ERVs can influence the fitness of the host organism in various ways. Few ERVs may be beneficial, providing new functional genes or regulatory elements (see hereinafter), while others can be detrimental, disrupting vital genes or causing disease. ERVs loci (including fixed and unfixed) annotated in the genome are depleted in genic regions, suggesting that ERVs integrated in these regions are more likely to be deleterious, and thus removed by negative, “purifying” selection (Zhang et al., 2008). However, this claim assumes random integration of ERVs in terms of genic/intergenic space, which needs to be supported by data.

At the same time, some integrated ERV loci generate new copies of themselves. The transposition rate in the germline refers to the frequency at which ERVs generate new copies. The transposition rates of active ERVs can be challenging to determine and depend on various factors, including the host species, the specific ERV family, and the genomic environment. For instance, hosts develop defense mechanisms against retroviruses, which affect the integration and activity of ERVs. In turn, ERVs may evolve mechanisms to evade the host's immune response or regulatory systems.

4.2 Evolution of ERV sequences

ERV sequences themselves accumulate mutations at different stages of the ERV's life cycle. The reverse transcription step is known to be error prone. Various attempts to estimate the mutation rate of HIV-1 resulted in a large range of 10^{-5}–10^{-3} errors/bp/cycle due to the use of different types of investigation methods (Yeo et al., 2020). APOBEC3 proteins of the host cell deaminate single-stranded DNA (ssDNA) intermediates leading to C-to-T base changes and mutations within the ERV genome (Ito et al., 2020). This edited ERV genome can be degraded (if heavily edited) or integrated into the genome as a new mutated copy. Finally, viral-like particles generally co-pack two single stranded ERV RNAs, which can produce extrachromosomal recombinant offspring molecules after replication (Figure 9). Frequent recombination has been demonstrated to occur in koala and mule deer during the early phases of endogenization and expansion (Löber et al., 2018; Yang et al., 2021). Once integrated in the host genome, ERV sequences accumulate all types of mutations and are subject to intra- and inter-chromosomal recombination. Recombination can occur in various ways. Recombination between LTRs of the same provirus results in a solo-LTR by eliminating most of the proviral sequences (Mager and Goodchild, 1989). Intra-chromosomal recombination between LTRs from distinct elements may cause deletions not only of the proviral sequences but also of deletion or inversion of the intervening genome sequences. The same type of recombination can also result in tandem proviruses (i.e. two proviruses flanked by LTRs while sharing one LTR). These recombinations underlie the diversification of the ERV sequences and are also an import source of rearrangements of host genome sequences.

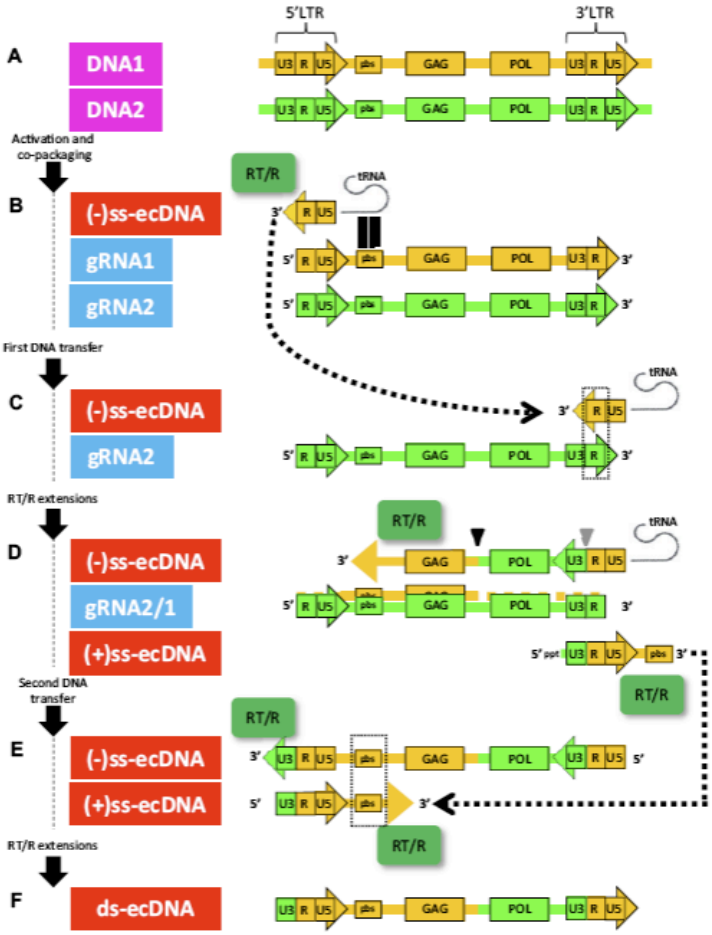


Figure 9: Schematic representation of discontinuous reverse-transcription and recombination steps.
(A) Two members of a LTR retrotransposon clan are transcriptionally activated and the corresponding genomic RNA (gRNA) progenitors (starting and ending in the R regions of 5′- and 3′-LTR, respectively) are copackaged. **(B)** Reverse-transcriptase/RNaseH activity (RT/R, green) is primed by tRNA annealing the priming-binding-site (pbs) of gRNA1 (orange), and minus-single-strand extrachromosomal cDNA ((-)ss-ecDNA) is synthesized. **(C)** First strand transfer: strong-stop (-)ss-ecDNA transfers to a second gRNA2 (green) (first recombinogenic step; the hypothesized recombination point is marked with an inverted grey triangle), using sequence homologies in the R region (marked with a dotted box). **(D)** At the same time that RNaseH activity proceeds (not shown), the (-)ss-ecDNA is extended using gRNAs as alternate templates (second recombinogenic step; a single hypothesized recombination point is marked with an inverted black triangle, but note that more than one event is possible). The color-coded newly synthesized (-)ss-ecDNA molecule is represented as a mosaic of the two progenitor gRNAs. Although RNaseH activity degrades portions of gRNA2 template (not shown), priming of poly-purine-track (ppt) allows nascent plus-single-strand extrachromosomal cDNA ((+)ss-ecDNA) synthesis until the end of the (-)ss-ecDNA molecule used as template. **(E)** Second strand transfer: strong-stop (+)ss-ecDNA swaps toward the 5′ area of the (-)ss-ecDNA. In addition, RT/R final extensions take place. **(F)** After final extensions, a mosaic blunt-ended linear extrachromosomal DNA (ds-ecDNA) molecule with two identical LTRs is generated. U3/R/U5, domains within LTRs; GAG, structural protein; POL, polyprotein. The figure and legend were adapted from Drost et al. (2019).

4.3 ERV co-option and domestication

The mobilization of endogenized ERV elements is not always deleterious for the host. Indeed, ERVs derived sequences and genes can be an important source of genomic novelty. They can be "co-opted" and provide novel advantageous functions to the host.

ERVs sequences contain regulatory elements, such as promoters and enhancers needed for their own expression. Those sequences can affect the expression of nearby host genes by serving as alternative promoters. A study of the FANTOM4 dataset reported that 6 to 30% of cap-selected mouse and human RNA transcripts initiate within TEs, including ERVs (Faulkner et al., 2009). Analysis of approximately 250,000 retrotransposon-derived transcription start sites shows that the associated transcripts are generally tissue-specific thus diversifying RNA repertoires. A mouse-specific MT2B2 LTR-derived promoter generates an N-terminally truncated Cdk2ap1DN with highest, developmentally essential expression in preimplantation embryos. Intriguingly, Cdk2ap1DN has an evolutionarily conserved function yet is driven by different LTR-derived promoters across mammals, representing a convergent co-option event. Hence, species-specific LTR-derived promoters can yield evolutionarily conserved, alternative protein isoforms with new functions and species-specific expression to govern essential biological divergence (Modzelewski et al., 2021). ERV sequences can also create novel long non-coding transcripts by providing internal parts of an exon, transcription start sites (TSS), polyadenylation sites (polyA), splice sites, or combinations of these. Kapusta et. al reported that 22.5% and 29.9% of TSS and polyA sites, respectively, of lncRNA transcripts in the human (Gencode v13 set) are provided by TEs, of which half are co-opted from ERV sequences (Kapusta et al., 2013).

ERV-derived enhancers have being increasingly reported ranging from interferon response enhancers in innate immunity (Chuong et al., 2016), placental enhancers (Chuong et al., 2013; Frost et al., 2023), species-specific germline regulatory elements (Sakashita et al., 2020) to pluripotency regulatory program in mammals (Sexton et al., 2021). Epigenomic profiling of the type II interferon response in bovine cells found thousands of ruminant-specific TEs including MER41_BT (a family of ERV) elements predicted to act as interferon inducible enhancer elements (Kelly et al., 2022). Altogether this demonstrates that lineage-specific ERVs have been independently co-opted to regulate IFN-inducible gene expression in multiple species, supporting ERV co-option as a recurrent mechanism driving the evolution of IFN-inducible transcriptional networks. In embryonic stem cells (ESCs), ERVs bound by pluripotency TFs, including OCT4 and NANOG, have been shown to possess enhancer activities and being able to initiate transcription (Wang et al., 2014).

In some cases, ERV insertions can lead to the formation of repressive epigenetic marks in the vicinity of the insertion site, leading to gene silencing *in cis*. This can occur through various mechanisms, including the recruitment of proteins that promote heterochromatin formation, changes in DNA

methylation patterns, or alterations in histone modifications. Once heterochromatin is established, it can spread along the chromosome, further affecting neighboring genes. Brind'Amour et al. reported that DNA methylation at 4 of 6 mouse-specific and 17 of 110 human-specific imprinted genes are attributed to lineage-specific upstream ERV insertions which induce deposition of DNA methylation marks during oogenesis resulting in gametic differentially methylated regions (Brind'Amour et al., 2018). This provides a genetic explanation of how biallelically expressed genes can be converted to maternally silenced in a lineage specific manner.

ERVs were also shown to regulate genome architecture. Using Hi-C data from 2-cell-like cells (a rare subpopulation of cells in murine embryonic stem cell cultures, possessing features of 2-cell embryo blastomeres) (Genet and Torres-Padilla, 2020), Kruse et al. showed that MERLV elements promote the formation of insulating domain boundaries throughout the genome *in vivo* and *in vitro*. The formation of these boundaries is coupled to the upregulation of directional transcription from MERVL, which results in the activation of a subset of the gene expression program of the 2-cell stage embryo (Kruse et al., 2019). Zhang et al. discovered a role for the HERV-H in creating topologically associated domains (TADs) in human pluripotent stem cells. Interestingly, random integration of an HERV-H element into a new locus was shown to be sufficient to form a *de novo* TAD boundary in a transcription-dependent manner (Zhang et al., 2019).

In addition to playing an important role in gene regulation, coding genes of ERVs were repurposed for new functions as well. One remarkable example of domestication is the creation of the Syncytin gene from a viral envelope gene over 100 million years ago, this gene is considered as essential for the emergence of the placenta (Mi et al., 2000). The first Syncytin gene was likely domesticated in the last common ancestor of placental mammals from an ERV that used the encoded envelope protein for fusion and entry into host cells. Subsequently, independent "convergent" domestication events have been identified in rodents (Dupressoir et al., 2011), lagomorphs (Heidmann et al., 2009), ruminants (Cornelis et al., 2013) and carnivores (Cornelis et al., 2012). However, the use of a cellular receptor for the env derived Syncytin renders the host vulnerable to exogenous retrovirus infection that use the same cell entry. Intriguingly, it has been shown that to avoid the potential danger from exogenous retrovirus infection another retroviral env gene called Suppressyn has been domesticated in human to regulate fusion activity of human Syncytin by occupying its receptor (Frank et al., 2022). Notably, ERV env proteins have been shown to act as restriction factors against related exogenous retroviruses in various species like chickens, sheep, mice and cats (Chiu and Vandewoude, 2021). Many more env-derived restriction factors are expected to be discovered. Two independent studies reported that the neuronal gene Arc encodes a domesticated gag protein and forms virus-like capsid structures that can transfer mRNA between cells in the nervous system (Ashley et al., 2018; Pastuzyn et al., 2018). This concept

was repurposed to engineer a RNA cargo system to achieve cell monitoring and cell-to-cell delivery of mRNA (Horns et al., 2023).

In addition to the direct repurposing of ERV sequences and gene products, the host has also collaterally co-opted adaptively evolved ERV silencing systems, which are now involved in essential physiological processes. The co-option and domestication of ERVs represent a fascinating aspect of evolutionary biology, showcasing how organisms can adapt and repurpose elements from their viral ancestors to their advantage over time.

Objectives

Our objectives are as follows:

Firstly, we seek to establish a comprehensive catalog of polymorphic (unfixed) endogenous retroviral elements (ERVs) within the cattle genome. Utilizing datasets of Whole Genome Sequencing (WGS) from the Holstein Friesian (HF), we aim to compile this catalog, laying the foundation for further investigation.

Building upon this foundation, we intend to utilize WGS-pedigree-based large-scale data within the Holstein Friesian breed, specifically the Damona pedigree. Our goal here is to ascertain whether any ERV family remains active within the cattle male and/or female germline, thus shedding light on the dynamics of ERV activity in these crucial reproductive cells.

Next, we endeavor to develop a high-throughput, sensitive, and quantitative method to measure the *de novo* transposition rate (dnTR) of active ERVs in bovine sperm. This innovative approach will provide valuable insights into the transpositional activity of ERVs within this specific cellular context.

Subsequently, we plan to leverage the normalized dnTR in sperm as a quantitative molecular phenotype. By employing this phenotype, we aim to conduct genome-wide association studies (GWAS), thereby identifying loci that influence this crucial aspect of ERV dynamics within the cattle genome.

Upon identifying GWAS peaks, if any, our focus will shift towards dissecting these peaks. Through rigorous analysis, we endeavor to identify the molecular actors that modulate dnTR within the cattle germline, thus uncovering key regulatory mechanisms governing ERV activity in this context.

Lastly, we aim to validate the broader applicability of our molecular method. By demonstrating its efficacy in analyzing not only ERVs but also other active transposable elements, across diverse species, we seek to establish a versatile tool that can contribute to genomic research beyond the realm of cattle genomics.

Experimental section

Experimental Section

Study 1

GWAS reveals determinants of mobilization rate and dynamics of an active endogenous retrovirus of cattle

Experimental Section

Study 1

GWAS reveals determinants of mobilization rate and dynamics of an active endogenous retrovirus of cattle

Nat Commun. 2024,15:2154.

Lijing Tang[1*], Benjamin Swedlund[1,2], Sébastien Dupont[1], Chad Harland[1,3], Gabriel Costa Monteiro Moreira[1], Keith Durkin[1,4], Maria Artesi[1,4], Eric Mullaart[5], Arnaud Sartelet[1,6], Latifa Karim[1,7], Wouter Coppieters[1,7], Michel Georges[1*], Carole Charlier[1*]

[1]Unit of Animal Genomics, GIGA & Faculty of Veterinary Medicine, University of Liège, Belgium; [2]Keck School of Medicine, University of Southern California, USA; [3]Livestock Improvement Corporation, Hamilton, New Zealand; [4]Laboratory of Human Genetics, GIGA & Faculty of Medicine, University of Liège, Belgium; [5]CRV, Arnhem, The Netherlands; [6]Comparative Veterinary Medicine, FARAH & Faculty of Veterinary Medicine, University of Liège, Belgium; [7]Genomics core facility, GIGA, University of Liège, Belgium.

[*]corresponding authors

Abstract

Five to ten percent of mammalian genomes is occupied by multiple clades of endogenous retroviruses (ERVs), that may count thousands of members. New ERV clades arise by retroviral infection of the germline followed by expansion by reinfection and/or retrotransposition. ERV mobilization is a source of deleterious variation, driving the emergence of ERV silencing mechanisms, leaving "DNA fossils". Here we show that the ERVK[2-1-LTR] clade is still active in the bovine and a source of disease-causing alleles. We develop a method to measure the rate of ERVK[2-1-LTR] mobilization, finding an average of 1 per ~150 sperm cells, with > 10-fold difference between animals. We perform a genome-wide association study and identify eight loci affecting ERVK[2-1-LTR] mobilization. We provide evidence that polymorphic ERVK[2-1-LTR] elements in four of these loci cause the association. We generate a catalogue of full length ERVK[2-1-LTR] elements, and show that it comprises 15% of *C*-type autonomous elements, and 85% of *D*-type non-autonomous elements lacking functional genes. We show that > 25% of the variance of mobilization rate is determined by the number of *C*-type elements, yet that *de novo* insertions are dominated by *D*-type elements. We propose that *D*-type elements act as parasite-of-parasite gene drives that may contribute to the observed demise of ERV elements.

Introduction

Half of mammalian genomes is composed of interspersed repeats, including endogenous retroviruses (ERVs) that occupy ~5-10% of genome space [1]. ERVs derive from retroviral infection of the germline enabling vertical viral transmission. Such retroviral endogenization may lead to the progressive expansion of a clade of ERV elements that may count tens to thousands of members, by an ERV-encoded reverse transcriptase (RT)-dependent copy-paste mechanism. At first, this process entails ERV-encoded envelope (ENV)-associated budding of viral particles followed by germline reinfection. Subsequently, the ERV may sometimes at least in part forgo the extracellular phase yet continue to multiply by more efficient intracellular retro-transposition. In mice, such a transition from IAPE to IAP elements has been shown to result from the combined acquisition of a novel GAG addressing signal and loss of ENV. Expansion of endogenized ERV clades may perdure for hundreds of thousands to millions of years, well after extinction of the initiating exogenous retrovirus. Further expansion of ERV clades is eventually curtailed by the accumulation of ERV-disrupting mutations and the acquisition - by the host - of specific ERV-silencing mechanisms including targeted DNA and chromatin epigenetic modifications, piRNAs, and so-called restriction factors targeting various steps of the ERV life cycle, some of which are co-opted ERV genes. ERV endogenization has been caught in the act at least six times, including in poultry, sheep/goat, mice, cat and koala, indicating that it is a common phenomenon [2-6].

As for other transposable elements, *de novo* ERV transposition events in the germline increase the mutational load of populations including by causing disease. Deleterious ERVs are hence subject to purifying selection. Conversely, ERV elements provide a substrate for the emergence of new functionalities such as novel *cis*-acting regulatory elements [4,7] and even new genes. As an example, the *syncytin* genes, which are essential for placentation, derive from ancient ERVs. Also, ERV elements may condition the host's susceptibility to exogenous retroviruses, including by providing protection. Hence, ERVs are important drivers of genomic innovation [2-6].

As ERVs settle, expand and then regress in a species' genome, the rate of ERV mobilization is bound to be an evolving phenotype. Presently, ERVs are largely silent in the human germline [8], while several clades are still active in the mouse in which they account for a sizeable fraction of deleterious mutations [9]. To what extent the ERV mobilization rate varies between individuals within species during this process, and what the underlying determinants of this variation may be, remains largely unexplored. Modifiers of epigenetic modification and expression of specific ERV elements have been mapped in mice and polymorphisms in KRAB zinc-finger proteins nominated as plausible candidates [10], yet whether these also affect transposition rate is not known.

We herein take advantage of unique features of domestic cattle, and of advances enabled by combining next generation sequencing (NGS) with CRISPR/Cas9, to perform genetic analyses of interindividual variation in ERV mobilization rate in this species.

Results

The insertion of an ERVK[2-1-LTR] element in exon 5 of the APOB gene causes cholesterol deficiency in cattle

A new autosomal recessive defect, referred to as cholesterol deficiency (CD), emerged in Holstein-Friesian dairy cattle around 2015 (OMIA:001965-9913) [11-13]. Autozygosity mapping positioned the corresponding locus on chromosome 11 [11]. We and others sequenced the whole genome of acknowledged carriers and showed that CD is caused by the insertion of an ERV element in exon 5 of the *apolipoprotein B* (*APOB*) gene [12-14]. We PCR-amplified and sequenced the corresponding ERV element and showed that (i) it is ~6.8 Kb long, (ii) belongs to the [2-1-LTR] subgroup of ERVK elements [15,16], and (iii) does not contain full-length open reading frames (ORFs) for either the *GAG*, *PRO*, *POL* or *ENV* genes and is therefore non-autonomous. RNA-Seq analysis of liver of an affected animal showed complete transcriptional shutdown in the ERV's 5'LTR, truncating ~ 97% of the *APOB* ORF (Fig. 1a-b; Supplementary Fig. 1). All carrier animals trace back to *Maughlin Storm*, a sire that was popular in the 1990-ies, suggesting that the birth of the CD mutation might be a recent event and, hence, that ERVK[2-1-LTR] might still be mobile in cattle.

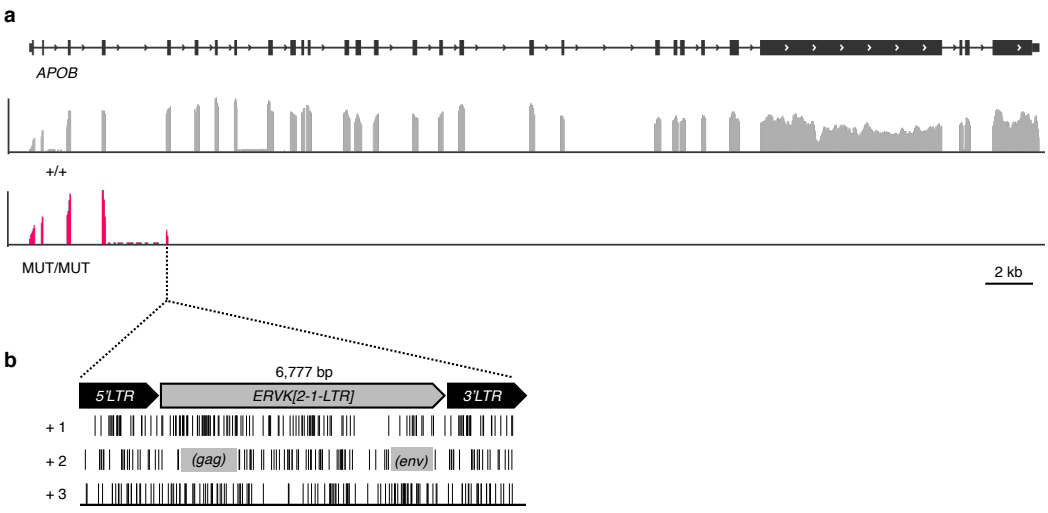


Figure 1: The insertion of an ERVK[2-1-LTR] element in the *APOB* gene causes cholesterol deficiency in Holstein-Friesian cattle.
(a) Genomic structure of the bovine *APOB* gene (chr11: 77,885,988-77,927,967 bp)(upper lane), and liver cDNA sequence coverage tracks for a homozygous wild-type animal (+/+, grey) and an affected homozygous mutant animal (MUT/MUT, red), highlighting the transcriptional shutdown in exon 5, with partial retention of intron 4 (see also Supplementary Fig. 1). **(b)** Schematic representation of the ~6.8 Kb ERVK[2-1-LTR] element sense insertion with the translation of its full-length sequence in the three frames (+ 1, + 2, + 3), revealing the absence of intact open reading frames (ORFs) in the proviral sequence flanked by two identical long terminal repeats (LTRs). Stop codons are represented by black vertical bars. Partial *GAG* and *ENV* ORFs are shown as grey rectangles (frame +2).

ERVK[2-1-LTR] elements are active in the bovine germline

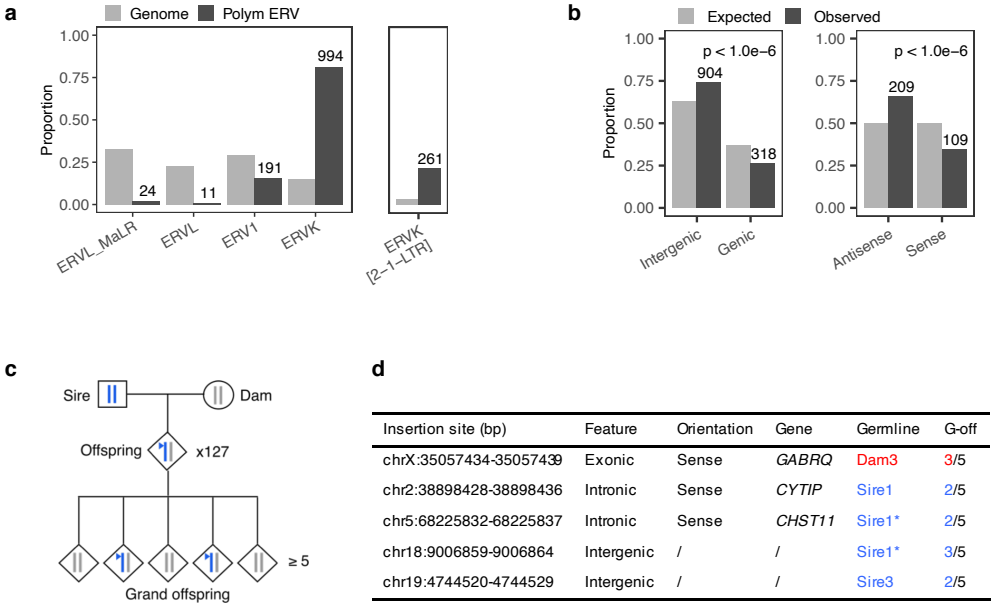
We previously generated the Holstein-Friesian *Damona* pedigree to study *de novo* mutations in the bovine germline. It comprises 743 animals constituting 127 (overlapping) three generation pedigrees including at least two parents (sire and dam), one offspring and ≥ 5 grand-offspring. The whole genome of parents and offspring was sequenced at average 26-fold depth, while that of the grand-offspring was sequenced at average 10-fold depth (Methods). We developed *LocaTER* (*Localization of Transposable Elements and Repeats*), a bioinformatic pipeline for the detection of transposable element insertions that are not present in the reference genome (Methods). *LocaTER* detected 1,222 ERV insertions that are polymorphic in Holstein-Friesian. Repbase [15,16] reports four main groups of ERVs: ERVL-MaLR, ERVL, ERV1 and ERVK, jointly accounting for ~4.7% of genome space. While ERVL (ERVL-MaLR + ERVL) are the most abundant group in the reference genome (55.5% of genome space), followed by ERV1 (29.3%) and ERVK (15.2%), ERVK includes the most abundant amongst polymorphic ERV elements (81. 5% of polymorphic elements), followed by ERV1 (15.6%) and ERVL (2.9%). This is compatible (assuming approximately equal element size) with ERVK being younger and still active (Fig. 2a). Repbase reports 33 subgroups of ERVK, of which four are overrepresented amongst polymorphic ERVK elements, including ERVK[2-1-LTR] (20.2% of ERVK space, 26.6% of polymorphic ERVK elements) and BTLTR1B (10.2% of ERVK space, 27.1% of polymorphic ERVK elements) (Supplementary Fig. 2a).

Nine hundred and four ERV elements mapped to intergenic regions (expected: 611), 109 within genes in sense orientation (expected: 159), and 209 within genes in antisense orientation (expected: 159) ($p < 10^{-6}$), supporting purifying selection against genic sense insertions (Fig. 2b). These trends did not differ significantly between the ERVK and non ERVK groups ($p_{genic\ vs\ intergenic} = 0.61$; $p_{sense\ vs\ antisense} = 0.42$) (Supplementary Fig. 2b-c). Allelic frequency distribution was mildly shifted towards lower values for genic versus intergenic ($p = 0.06$), but not for genic sense versus antisense insertions ($p = 0.99$) (Supplementary Fig. 2d-e; Supplementary Data 1).

Most interestingly, we detected five ERV insertions that were present in an offspring, transmitted to grand-offspring, but absent in both sire and dam, which we considered to be *de novo* mobilization events (Fig. 2c). Linkage analysis in the grand-offspring indicated that four of these occurred in the germline of a sire, and one in the germline of a dam. Intriguingly, three of the four male mobilizations occurred in the germline of the same bull, of which two in the same sperm cell. We PCR-amplified and sequenced the five *de novo* insertions. All five measured ~6.8 Kb and their sequence was very similar to the non-autonomous ERVK[2-1-LTR] insertion in the *APOB* gene (Fig. 2d; Supplementary Fig. 3).

These results confirm that ERVK[2-1-LTR] are presently active in bovine, at a rate of ~ one event per 51 gametes (or ~ one event per 126 gametes when ignoring the exceptional sire), and suggest that the

mobilization rate may differ between individuals. Mining of the whole genome sequence of the bull transmitting three ERVK[2-1-LTR] *de novo* insertions did not reveal striking anomalies in 38 genes that have been connected with control of ERV mobilization [17] (Supplementary Data 2).



Insertion site (bp)	Feature	Orientation	Gene	Germline	G-off
chrX:35057434-35057439	Exonic	Sense	*GABRQ*	Dam3	3/5
chr2:38898428-38898436	Intronic	Sense	*CYTIP*	Sire1	2/5
chr5:68225832-68225837	Intronic	Sense	*CHST11*	Sire1*	2/5
chr18:9006859-9006864	Intergenic	/	/	Sire1*	3/5
chr19:4744520-4744529	Intergenic	/	/	Sire3	2/5

Figure 2: Detection of polymorphic ERVs and *de novo* ERV mobilization events in the *Damona* pedigree.
(a) ERVK elements - including the ERVK[2-1-LTR] clade - are overrepresented amongst polymorphic ERVs detected with *LocaTER* when compared to their abundance (relative to other ERV elements) in the bovine genome, supporting their youth. **(b)** ERVs are overrepresented in intergenic regions, or – when genic – in antisense (as opposed to sense) orientation, when compared to the corresponding genome space, supporting purifying selection. Two-sided p-values ($< 10^{-6}$) were determined from the outcome of $>10^6$ random samples with probabilities of success corresponding to the genomic expectations (sample() function in R). **(c)** The *Damona* pedigree comprises 743 whole genome sequenced animals constituting 127 pedigrees with at least three generations: parents, offspring and (≥ 5) grand-offspring. *De novo* mutations (including ERV mobilization events; blue triangle) are detected by their absence in the parents, presence in the offspring, and transmission to some grand-offspring. Linkage with a grand-parental chromosome (sire's blue chromosome in the example) allows assignment of the *de novo* event to the sire's or the dam's germline. **(d)** Description of the five *de novo* ERVK[2-1-LTR] insertions, with, from left to right: (i) chromosomal position (ARS-UCD1.2 genome assembly), (ii) genome compartment, (iii) when genic, orientation with respect to the affected gene, (iv) when genic, gene symbol of the affected gene, (v) parental germline in which the insertion occurred: Dam (blue) or Sire (red) with numbers identifying individuals according to Supplementary Fig. 3 (three *de novo* insertions occurred in Sire 1; the * identifies the two insertions that were co-transmitted in the same gamete, (vi) Mendelian transmission to "x" (2 or 3) out of five grand-offspring (G-off). Source data are provided as a Source Data file.

Developing a method to measure ERVK[2-1-LTR] mobilization rate in germline and soma

We adapted Pooled CRISPR Inverse PCR (PCIP) [18] to measure the rate of ERVK[2-1-LTR] mobilization in a sample of genomic DNA. As for PCIP, we used CRISPR/Cas9 to augment the efficacy of inverse PCR, yet used Illumina short read sequencing instead of Oxford Nanopore long read sequencing to increase coverage and accuracy (Fig. 3a). Noteworthy features of the method are: (i) the realization of technical replicates by targeting both 5' and 3' LTR, (ii) the ability to recognize PCR duplicates on the basis of shared shearing sites (SS)(if two sperm cells carry the same *de novo* ERVK[2-1-LTR] insertion, the probability that the two corresponding DNA molecules are broken at the exact same position is assumed to be very low; shared SS are therefore assumed to correspond to PCR duplicates), and (iii) the ability to express the number of detected *de novo* mobilization events as a function of the effective number of explored genomes (using inherited ERVK[2-1-LTR] elements as internal controls, see Methods). Of note, ERVK[2-1-LTR] with polymorphisms in the CRISPR target sites may escape detection by PCIP, an issue which is at least partially mitigated by targeting both 5' and 3' proviral ends. The correlation between 5' and 3' rate estimates was 0.68, and 0.85 between averages of 5' and 3' estimates for biological replicates, testifying for the accuracy of the approach (Supplementary Fig. 4a-b; Supplementary Data 3&4).

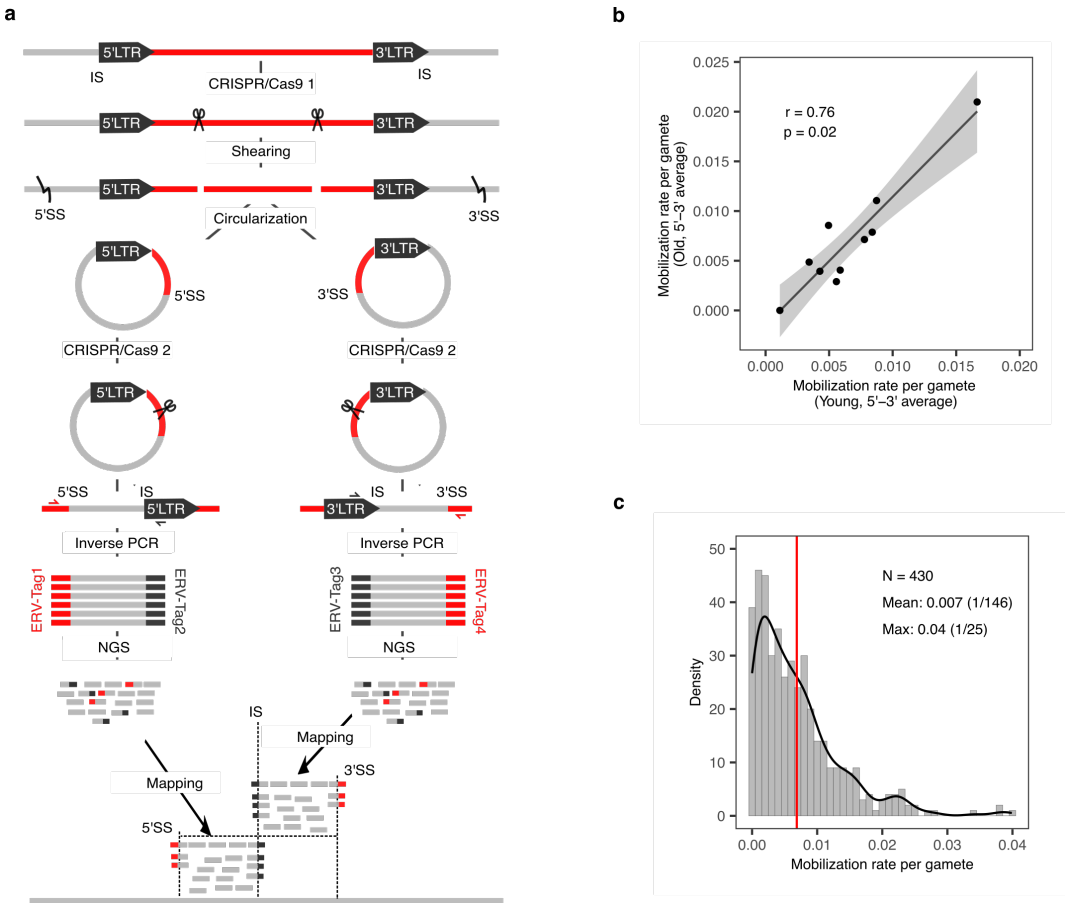


Figure 3: Measuring the *de novo* rate of ERVK[2-1-LTR] mobilization using an adaptation of Pooled CRISPR inverse PCR sequencing (PCIP-Seq).
(a) Method: Genomic DNA is attacked with a pair of CRISPRs targeting the ERVK[2-1-LTR] provirus at 429 and 374 bp from the 5' and 3' LTR sequences, respectively. The DNA is then mechanically sheared (generating 5' and 3' shearing sites (SS)), end-repaired, circularized, and attacked separately by a second pair of CRISPRs targeting proviral sequences adjacent to the LTRs to specifically reopen circles encompassing ERVK[2-1-LTR] proviral insertion sites (IS). Fragments encompassing the IS and SS are then amplified by long-range inverse PCR using divergent ERVK[2-1-LTR] targeting primer pairs. The resulting PCR products are subject to NGS and the obtained sequence reads mapped on the reference genome, revealing ERVK[2-1-LTR] element insertion sites (IS), shearing sites (SS) used as molecular identifiers, and short sequence tags that inform about the inserted ERVK[2-1-LTR] element. PCIP provides information about the genotype of the individual for polymorphic ERVK[2-1-LTR] elements that segregate in the population, the number of genomes that were effectively captured for that individual (based on the number of detected shearing sites for the polymorphic ERVs), and the number of *de novo* mobilization events detected in the individual's DNA sample. **(b)** Correlation between estimates of the mobilization rate of ERVK[2-1-LTR] elements in sperm samples collected at >7 years interval (young vs old) for ten Belgian Blue bulls. Estimates correspond to the average of the 5'LTR and 3'LTR measures. Spearman's correlation was 0.76, with two-sided p-value of 0.02. The 95% confidence region for the regression fit was added using the stat_smooth() ggplot function. **(c)** Frequency distribution of estimated ERVK[2-1-LTR] *de novo* mobilization rate in sperm of 430 Belgian Blue bulls (red vertical bar = mean). Source data are provided as a Source Data file.

The ERVK[2-1-LTR] mobilization rate in the male germline varies between individuals

We first obtained pairs of semen samples from 10 Belgian Blue (BB) bulls, collected at ≥ 7 years interval. We applied our method to these samples and computed the correlation between ERVK[2-1-LTR] mobilization rate (5' and 3' average) estimated at young and old age. The correlation, corresponding to the so-called repeatability of the trait (upper bound of the heritability), was 0.76 (Fig. 3b). This implies that the ERVK[2-1-LTR] mobilization rate (i) varies between animals, and (ii) is stable over long periods of time within animal. There was no evidence of an effect of age on ERVK[2-1-LTR] mobilization rate. We detected 260 *de novo* ERVK[2-1-LTR] mobilization events in these experiments. Their median "dosage" in sperm DNA was 0.00023, or one copy per 4,318 sperm cells (range: 0.00018 – 0.00265). This indicates that the level of mosaicism for these *de novo* insertions is low, and hence that they mostly occur at late stages of spermatogenesis. Yet, we observed 38 cases where the same insertion was captured both in the young and old samples. This suggests that the corresponding insertions are present in spermatogonial stem cells that maintain the capacity to give rise to spermatogenic waves during the entire life of the animal (Supplementary Data 5).

We then applied the method to semen samples from 430 Belgian Blue (BB) bulls. We captured a total of three fixed ERVK[2-1-LTR] elements (i.e. detected in all studied animals), 306 polymorphic ones (i.e. detected in some but not all animals), and 3,669 *de novo* ERVK[2-1-LTR] insertions in this experiment (Supplementary Data 4&6). *De novo* ERVK[2-1-LTR] elements tended to preferentially insert near telomeric ends and in GC-rich regions (even after correcting for distance from chromosome

end). Insertion sites were characterized by a 6-bp duplication and an 8-bp pseudo-palindromic motif [19] (Supplementary Fig. 5). They resided in genic regions more often than expected by chance ($p = 0.0004$), and – in those – equally often in sense as in antisense orientation ($p = 0.40$). We observed a paucity of *de novo* insertions on chromosome X given its size, which is likely due to hemizygosity in males (Supplementary Fig. 6). The estimated ERVK[2-1-LTR] mobilization rate averaged one per 146 sperm cells, yet ranging from zero to one in 25 (Fig. 3c, Supplementary Data 4). There was no evidence of an effect of the bulls' inbreeding coefficient on ERVK[2-1-LTR] mobilization rate (Supplementary Fig. 4c).

GWAS identifies eight loci affecting ERVK[2-1-LTR] mobilization rate

In order to investigate whether the variance in inter-individual transposition rate has a genetic basis, we determined the whole genome sequence (average 40-fold depth) of 40 of the 430 BB bulls and genotyped them at ~14 million variant positions using GATK [20]. The remaining 390 bulls were genotyped using a 50K medium density Illumina array and genotypes augmented to whole genome by imputation. We kept ~10 million variants with minor allele frequency (MAF) ≥ 0.02, together with the PCIP-deduced genotypes (+/+, +/ERV, ERV/ERV) for 87 polymorphic ERVK[2-1-LTR] elements (MAF ≥ 0.02), for further analyses on the complete dataset (Supplementary Data 7). Of note, 80 of 306 polymorphic ERVK[2-1-LTR] elements were singletons, 163 had a population frequency < 0.05 (rare non-singletons), and 63 a frequency ≥ 0.05 (common). We conducted a GWAS for ERVK[2-1-LTR] mobilization rate using a linear model including a fixed variant dosage effect and a random polygenic effect to correct for stratification. We obtained a major, genome-wide significant signal ($-\log(p) = 29.7$) on chromosome 19 (Fig. 4a). We fitted this effect as covariate in the model and repeated the genome scan. This revealed seven additional genome-wide significant (or near) ($-\log(p) = 7$) signals (Fig. 4b).

Strikingly, a polymorphic ERVK[2-1-LTR] element was either the top variant (chromosome 19) or in high linkage disequilibrium (LD) with it ($r^2 \geq 0.62$; chromosomes 2, 4 and 21) for four of the eight loci (Fig. 4c, Supplementary Fig. 7), a coincidence which is very unlikely to have occurred by chance alone ($p < 10^{-8}$; see Methods).

While it may have been expected that inherited ERVK[2-1-LTR] elements promote ERV mobilization rate, this finding raised the question of what differentiates the ERVK[2-1-LTR] elements in the four associated loci from other polymorphic ERVK[2-1-LTR] elements.

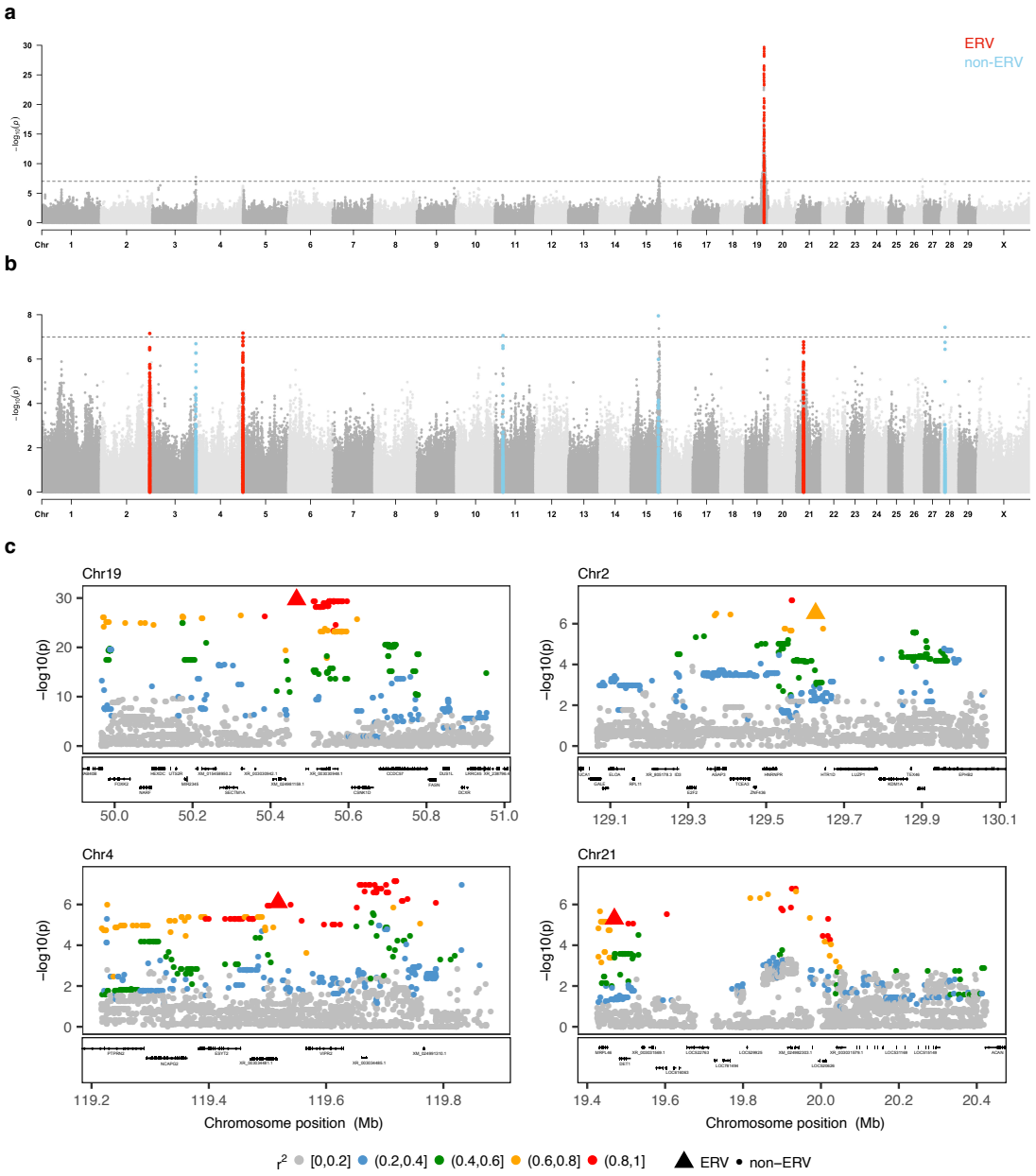


Figure 4: GWAS for the rate of *de novo* ERVK[2-1-LTR] mobilization in the male germline of cattle.
(a) GWAS conducted using PCIP-determined mobilization rate in sperm samples of 430 Belgian Blue bulls and genotypes at ~10 million SNPs, revealing a very strong signal on chromosome 19. **(b)** GWAS conducted using the same population after correcting ERVK[2-1-LTR] mobilization rate for the effect of the chromosome 19 QTL. Seven additional (near) genome-wide significant effects were detected. Loci encompassing an ERVK[2-1-LTR] element are highlighted in red, others in light blue. **(c)** Zoom into the four loci encompassing an ERVK[2-1-LTR] element, shown as triangles (as opposed to circles for SNPs). Variants are colored according to their LD (r^2) with the lead variant. The LD between the ERVK[2-1-LTR] element and the lead SNP was 1.00 (ERV = lead variant) for chromosome 19, 0.62 for chromosome 2, 0.84 for chromosome 4, and 0.94 for chromosome 21. All non-ERVK lead variants were imputed variants with imputation accuracy $\geq$ 0.95. Two-sided p-values were obtained with GEMMA as described in M. Experiment-wide 5% significance thresholds (dotted horizontal line in **a**)

account for the realization of 500,000 independent tests (see M). Source data are provided as a Source Data file.

Genome-wide landscape of polymorphic ERVK[2-1-LTR] elements in Belgian Blue cattle

This question prompted us to extensively characterize the polymorphic ERVK[2-1-LTR] repertoire in BB cattle. As mentioned before, PCIP coamplified a total of 309 endogenized ERVK[2-1-LTR] elements from the semen samples of the 430 BB bulls. We successfully PCR-amplified and completely sequenced 221 of these (Supplementary Data 8). This revealed two subclades, differing by poorly alignable central fragments of 6.2 (*C*-type) or 2.7 Kb (*D*-type), and accounting respectively for 15% and 85% of PCIP-detected ERVK[2-1-LTR] elements (Fig. 5a and Supplementary Fig. 8a). Intact ORFs for *GAG*, *PRO*, *POL* and *ENV* were present for ~50% of *C*-type members, hence assumed to be autonomous or *C*ompetent, while being absent for all *D*-type members, hence assumed to be non-autonomous or *D*efective (Supplementary Fig. 9).

We detected 216 substitutions (i.e. differences between ERVK[2-1-LTR] elements) within the *C*-specific fragment (of which 60 singletons), 187 substitutions within the *D*-specific fragment (45 singletons), and 391 substitutions in the 5' and 3' flanking regions (123 singletons). There were ~6 times more **C**pG to **T**pG and Cp**G** to Cp**A** substitutions than expected, reflecting the increased mutability of methylated cytosines. There were more G-to-A substitutions than C-to-T substitutions at non-CpG sites ($p = 0.0007$), supporting strand-specific APOBEC3-dependent G-to-A editing [21] (Fig. 5b,c). The latter was substantiated by the observation of clustered G-to-A singletons for several endogenized ERVK[2-1-LTR] elements (Supplementary Fig. 10a).

We designed probes targeting the insertion sites of 291 polymorphic ERVK[2-1-LTR] elements, added them to Illumina 50K arrays, and successfully re-genotyped the 430 bulls for 193 loci. Striking discrepancies between the array and PCIP genotypes were observed for at least five loci. Examination of the whole genome sequences (WGS) of the 40 bulls, revealed that this was due to the segregation of a third solo-LTR allele (in addition to "+" and full-length ERV, where "+" corresponds to the ancestral or wild-type allele) for these loci. Further examination of the WGS data with *LocaTER* output indicated that there exist at least 100 additional polymorphic ERVK[2-1-LTR] elements for which the only two segregating alleles are solo-LTR and "+", hence never captured by PCIP. The presumed solo-LTR status was confirmed by PCR for 84 tested elements, indicating that no important ERVK[2-1-LTR] subclade (other than *C* and *D*) was missed by PCIP (Supplementary Data 9). The distribution of allele frequencies in the 40 sequenced bulls was slightly shifted towards higher values for solo-LTRs when compared to *C*- and *D*-type ($p = 1.4 \times 10^{-4}$ and 2.0×10^{-8}), supporting their older age as expected (Supplementary Fig. 11).

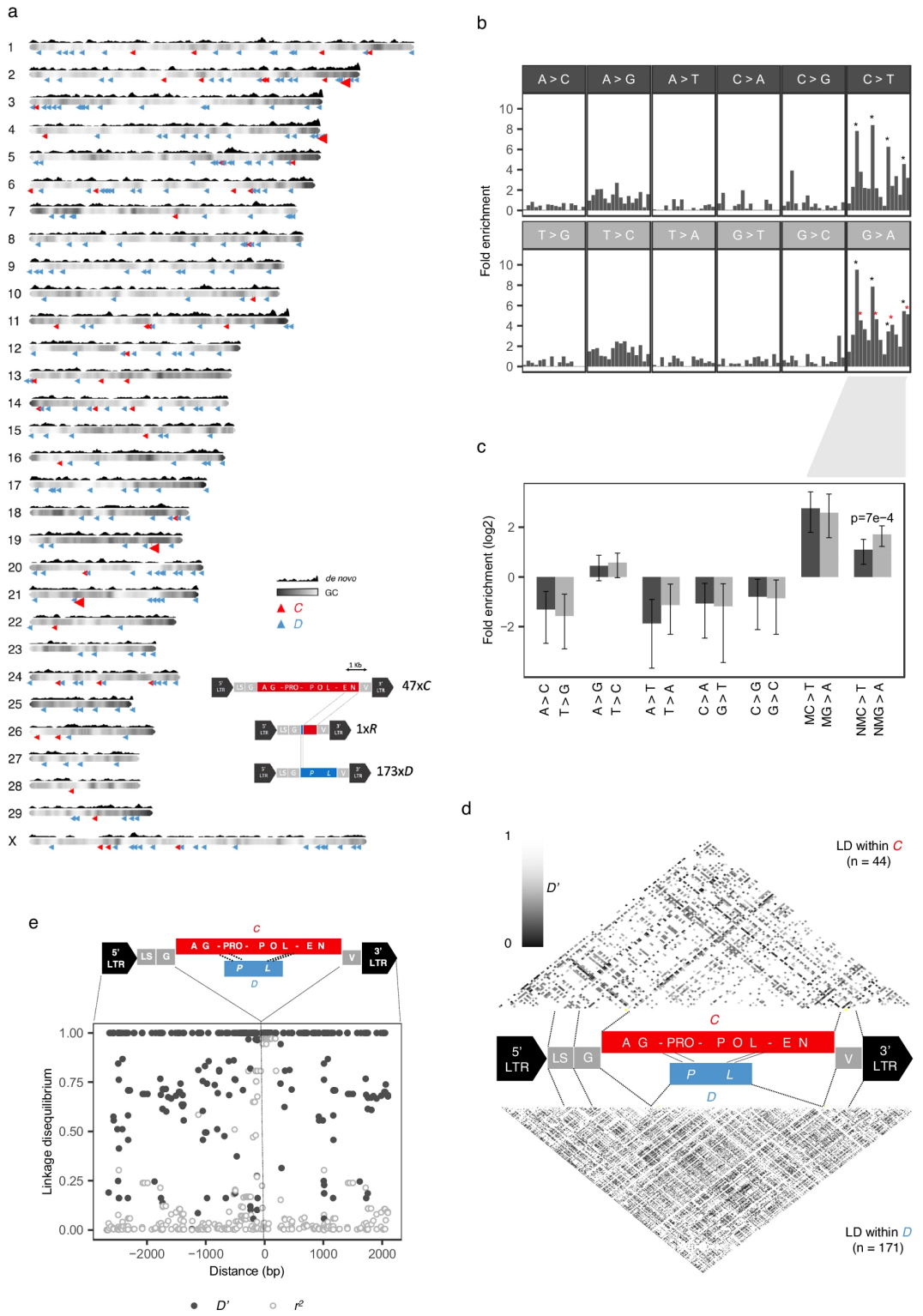


Figure 5: Catalogue of polymorphic ERVK[2-1-LTR] elements segregating in the Belgian Blue Cattle population.

(a) Resequencing 221 of 309 endogenized ERVK[2-1-LTR] elements, detected by applying PCIP to 430 bulls, revealed a subclade of ~10Kb *C*ompetent (*C*) (15%) and a subclade of ~6.8Kb *D*efective (*D*) elements (85%), as well as one *R*ecombinant (*R*) element (inset). Their genome-wide localization is represented by blue *(D)* or red *(C)* small triangles and the four *C* significant GWAS loci by larger red triangles. GC content and density of *de novo* ERVK[2-1-

LTR] insertions along chromosome lengths are shown. **(b)** 797 single base pair substitutions detected by aligning the sequences of the 221 ERVK[2-1-LTR] elements, sorted by type of substitution and trinucleotide context (Upper panel: ASA, ASC, ASG, AST, CSA, CSC, CSG, CST, …; lower panel: TST, GST, CST, AST, TSG, GSG, CSG, ASG, … where S = Substitution). The Y axis corresponds to the observed fold enrichment accounting for the abundance of the corresponding trinucleotide in the ERVK[2-1-LTR] consensus sequence. The strongest enrichments are observed for NCpG to NTpG and CpGN to CpAN (black asterisks), as expected given the increased mutability of methylated cytosines. The second strongest signal is AGN to AAN (red asterisks) which is likely an APOBEC3 G-to-A editing signature. **(c)** Comparison of the enrichment of "mirror" substitutions to test the strand-specificity of the underlying process. Only G-to-A vs C-to-T substitutions at non-CpG sites showed a significant difference in enrichment ($p = 0.0007$, obtained by bootstrapping, two-sided) in support of a role for the strand-specific APOBEC3 dependent G-to-A editing. The error bars mark the 95% confidence interval of the estimates obtained by bootstrapping. **(d)** Evidence (four gametes rule) of past recombination events between *C*-type (upper triangle) and *D*-type (lower triangle) ERVK[2-1-LTR] elements, respectively. *D'* values ≤ 1 for a pair of variants testify of past recombination events in the corresponding interval. Black or grey points in the upper and lower triangles mark variant pairs with evidence for recombination. **(e)** Evidence of past recombination events between a *C*- and a *D*-type elements in segments flanking the *C*- and *D*-type specific regions. Solid black dots - D': all values ≤ 1 are a testimony of past recombinations. Empty grey dots - r^2: a number of variants in the immediate vicinity of the *C*-*D* boundaries are in perfect LD ($r^2 \sim 1$) suggesting attrition of the boundary. Source data are provided as a Source Data file.

It is well established that recombination between ERV genomes can occur during reverse transcription of the pseudodiploid genomic RNAs (gRNAs) [22]. We examined the haplotype patterns of the 221 ERVK[2-1-LTR] elements to search for evidence of recombination. When a new variant arises by mutation (f.i. + becomes *M*), it is found in "complete" association ($D'=1$) with the variants (f.i. V_1) constituting the ERV element upon which it occurred: only three of the four possible haplotypes exist across all elements (V_1-*M*, V_1-+ and +-+). The appearance of the fourth haplotype (+-*M*) requires a recombination between a gRNA with V_1-*M* haplotype and one with +-+ haplotype. We first conducted this "four haplotypes" test for all pairs of variants separately for the *C* and *D* subclades. There was ample evidence for past recombinations having occurred throughout the ERVK[2-1-LTR] genome both within the *C* and within the *D* subclades, indicating that gRNA originating from distinct ERVK[2-1-LTR] elements (albeit from the same subclade) can be co-packaged in the same virus-like particle (VLP) and recombine (Fig. 5d). We then looked for evidence of recombination between *C* and *D*-type ERVK[2-1-LTR] elements. We did this by examining the linkage disequilibrium (LD) between *C* vs *D* genotype and all variants in the 5' and 3' flanking segments. Here also, there was ample evidence for past recombinations between *C*-and *D*-type gRNA indicating that these can also be co-packaged in VLP (Fig. 5e and Supplementary Fig. 10b). LD with C/D genotype was nearly perfect ($r^2 \sim 1$) for a set of variants immediately flanking the subclade-specific segments, suggesting that recombination between *C* and *D* gRNAs might be hampered in close proximity to the point of *C*-*D* divergence, leading to the progressive displacement of the *C*-*D* boundaries in a manner reminiscent of the "attrition" observed at mammalian pseudo-autosomal boundaries [23] (Fig. 5e and Supplementary Fig. 8a).

The consensus GAG, POL (reverse transcriptase domain) and ENV (transmembrane domain) sequences of *C*-type elements each cluster with betaretroviridae (Supplementary Fig. 12a). The ENV protein encompasses a surface unit (SU), proteolytic processing site, and transmembrane unit (TM) with fusion peptide, CX_7C cysteine motif, transmembrane region (TR) and cytoplasmic tail (CT) domains, typical of beta- and lenti-retroviruses [24]. The amino-terminal end of the GAG protein was predicted by Myristoylator [25] to correspond to a myristoylation site (high confidence score of 0.98) (Supplementary Fig. 12b). Taken together this suggests that ERVK[2-1-LTR] mobilize by inter-cellular reinfection [26]. The consensus primer binding site (PBS), shared by *C*-type and *D*-type elements (Supplementary Data 8), presents striking complementarities with the 3' extremities of the two bovine lysine tRNAs (CTT and TTT codons, respectively), supporting the ERV**K** denomination (Supplementary Fig. 12c). Observed mismatches might reflect the required balance between adequate tRNA priming for effective reverse transcription yet reduced complementarity to tRNA fragments mediating silencing [27].

The number of inherited competent ERVK[2-1-LTR] elements explains >25% of interindividual variation in male mobilization rate

Analysis of the ERVK[2-1-LTR] catalogue showed that the four ERV elements in the GWAS peaks were all of *C*-type ($p = 0.0008$). This suggests that the number of competent ERVK[2-1-LTR] elements in the genome is a major determinant of their mobilization rate. To more accurately evaluate the effect of the number of inherited *C*-type members on ERVK[2-1-LTR] mobilization rate, we compiled this number for the 430 BB bulls considering not only the four loci identified by GWAS, but all other (rarer and monomorphic) competent elements as well. It ranged from 8 to 19. We tested its effect on ERVK[2-1-LTR] mobilization rate using a linear model. It was highly significant ($p \leq 2.2 \times 10^{-16}$) and explained 24% of the variation for this trait (Fig. 6a). In striking contrast, the number of non-competent *D*-type elements, which ranged from 7 to 31, had no effect at all on ERVK[2-1-LTR] mobilization rate ($p = 0.87$; Fig. 6c).

Twenty-five *C*ompetent ERVK[2-1-LTR] elements harbor coding variants in at least one of the *GAG*, *PRO*, *POL* and/or *ENV* ORFs. This includes a stop codon at amino acid position 13 of the *POL* gene that is carried by the ERVK[2-1-LTR] element coinciding with the GWAS peak on chromosome 4 (Supplementary Data 10). We repeated the analysis of the effect of the number of inherited *C* elements on the ERVK[2-1-LTR] mobilization rate, considering only "intact" *C* elements (all are polymorphic). The number of intact *C* elements in the genome ranged from 0 to 6. The effect on mobilization rate became even stronger, now explaining 34% of the trait variance (Fig. 6b). To more accurately evaluate the effect of the coding variants on the *de novo* mobilization rate, we tested the effect of the number of such elements (with ORF-disrupting mutation in *GAG*, *PRO*, *POL* or *ENV*) in the genome (0, 1 or 2), conditional on having 0, 1, 2 or 3 copies of intact *C*ompetent elements (without ORF-disrupting mutation in *GAG*, *PRO*, *POL* and *ENV*). In the absence of any intact *C*ompetent element, increasing the number

of copies of ERVK[2-1-LTR] elements carrying a coding variant did not increase the mobilization rate. However, if the genome additionally harbored one or more intact *C*ompetent ERVK[2-1-LTR] elements, the copies with coding variants affected the *de novo* transposition rate ($p = 1.5 \times 10^{-4}$), pointing towards a possible ($p = 0.11$) epistatic interaction between mutated and intact *C*-type elements (Supplementary Fig. 13).

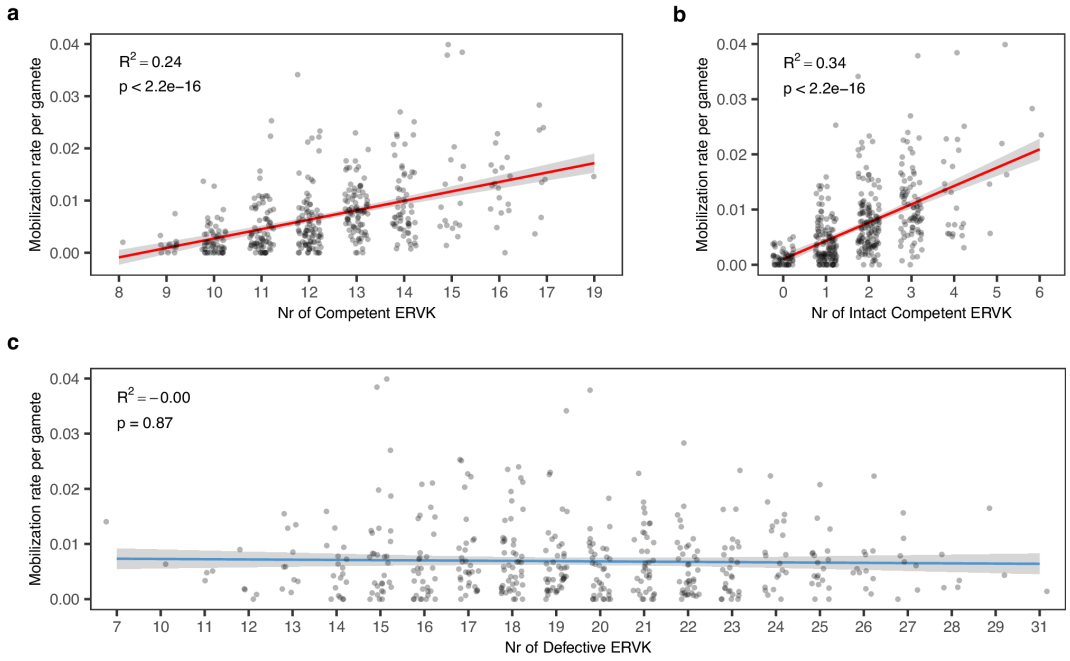


Figure 6: Effect of the number of inherited ERVK[2-1-LTR] elements in the genome on the *de novo* mobilization rate in the male germline.
(a) Effect of the total number of ~10 Kb *C*ompetent elements ($p < 2.2 \times 10^{-16}$). **(b)** Effect of the number of intact (i.e. without any coding variant in *GAG*, *PRO*, *POL* or *ENV*) ~10 Kb *C*ompetent elements ($p < 2.2 \times 10^{-16}$). **(c)** Effect of the total number of ~6.8Kb *D*efective elements (p=0.87). p: statistical significance (two-sided) of Spearman's correlation between the number of elements and the mobilization rate. R^2 = variance explained by the linear regression of number of elements on mobilization rate. The 95% confidence region for the regression fit was added using the stat_smooth() ggplot function. Source data are provided as a Source Data file.

Preferential de novo mobilization of defective ERVK[2-1-LTR] elements

The role of competent ERVK[2-1-LTR] elements in driving mobilization rate contrasts with the observation that the ERVK[2-1-LTR] element disrupting the *APOB* gene, and the five *de novo* mobilized ERVK[2-1-LTR] elements detected in the *Damona* pedigree are all of non-competent *D*-type. To verify whether this was mere coincidence, we selected three BB bulls for which the constitutive *C* and *D*-type ERVK[2-1-LTR] elements could be best discriminated based on the sequence tags obtained by PCIP (Fig. 3a and Supplementary Data 11-12). The three bulls harbored 26 (bull 1), 34 (bull 2) and 33 (bull 3) inherited (PCIP-detectable, i.e. ignoring solo-LTRs) ERVK[2-1-LTR] elements in their genome, of

which 12, 11 and 13 of *C*-type. PCIP tag information discriminated 13, 17 and 17 haplotypes, of which 2, 2 and 2 included *C* and *D* elements, while all others (11, 15 and 15) were either pure *C* or pure *D* (Fig. 7a-d). We performed 9, 6 and 9 PCIP experiments per bull, revealing respectively 178, 73 and 147 *de novo* insertions. We estimated the proportional contribution of the different inherited ERVK[2-1-LTR] elements to the *de novo* insertions by expectation-maximization (EM)(Methods). In all three bulls, *D* elements accounted for a significantly ($p \leq 0.01$) higher proportion of *de novo* insertions than of inherited *D*-type ERVK[2-1-LTR] elements (Fig. 7e-g). This suggests that *D* elements are able to outcompete *C* elements despite the fact that the latter are driving *de novo* mobilization supposedly by *trans*-complementation [28].

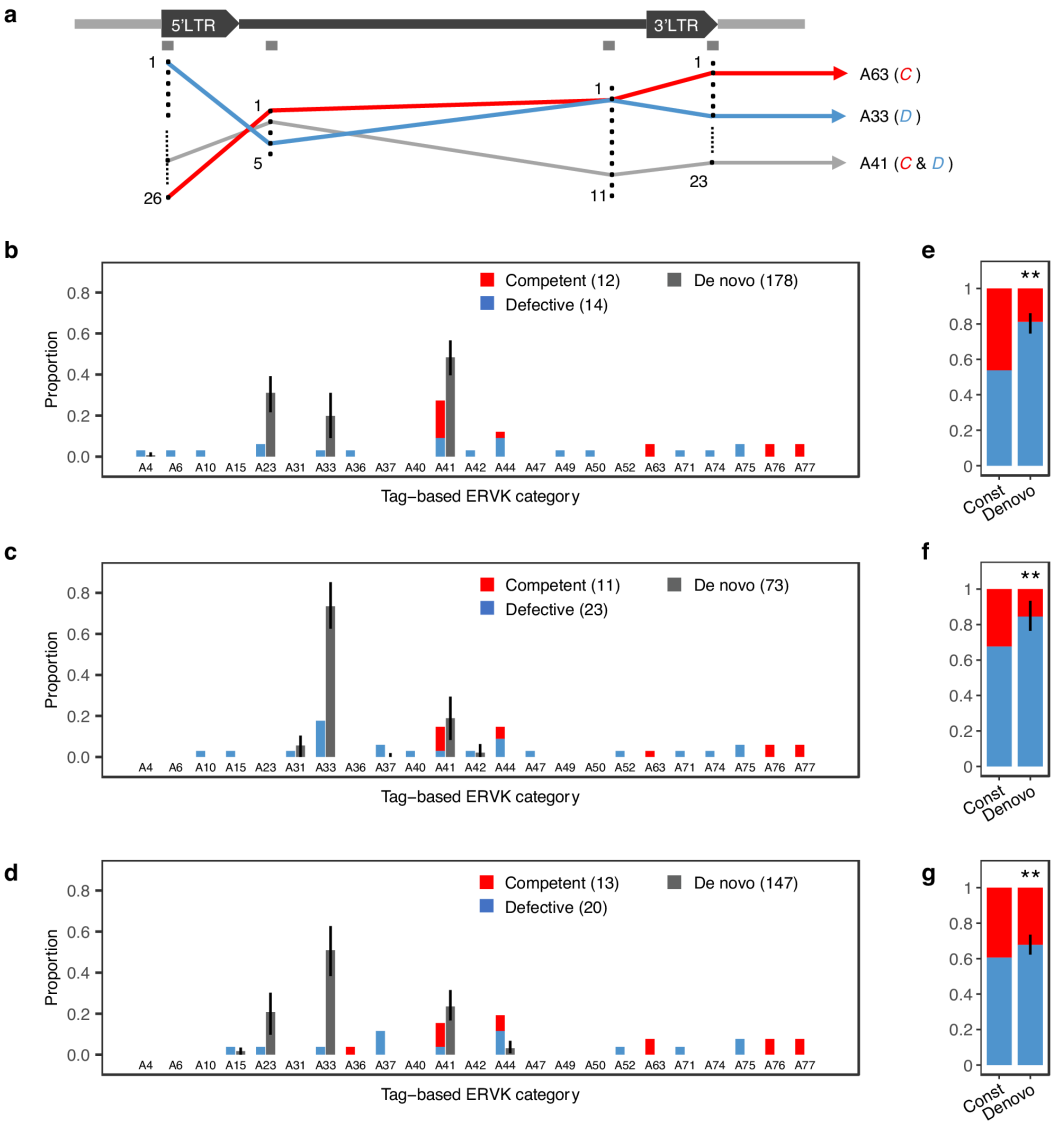


Figure 7: Preferential mobilization of *D*-type ERVK[2-1-LTR] elements.
(a) Position of - PCIP-tags. Each tag is characterized by multiple (from 5 (tag 2) to 26 (tag 1)) variants across the 309 detected ERVK[2-1-LTR] elements. The 309 elements are each characterized by a combination of variants, i.e. a haplotype. We observed 77 haplotypes (A1 to A77), indicating that different ERVK[2-1-LTR] elements share the same haplotype. For a given animal, some haplotypes correspond to only one ERVK[2-1-LTR] element in its genome (which

may be either *C* or *D*), while others may correspond to multiple ERVK[2-1-LTR] elements (which can be all *C*, all *D* or mixed *C* and *D* - such as A41 in the example). Different haplotypes may share the same variant for some tags (A63 and A33 share the same variant for tag 3 in the example). We often have partial PCIP-tag information for *de novo* insertions (f.i. only the 5' or 3' LTR) blurring the assignment of a *de novo* insertion to a specific haplotype. We used an expectation-maximization (EM) algorithm to probabilistically estimate the proportional contribution of the different haplotypes to the *de novo* insertions. **(b-d)** ERVK[2-1-LTR] elements in the genome of three bulls were assigned to 13 (A), 17 (B) and 17 (C) haplotypes, for a total of 23 distinct haplotypes. The proportion of inherited ERVK[2-1-LTR] elements in each haplotype, as well as the proportion of *C*- (red) and *D*-type (blue) elements in each haplotype, was deduced by combining PCIP and targeted sequencing data. The proportional contribution of each ERVK[2-1-LTR] haplotype to the *de novo* insertions was estimated by expectation-maximization (EM). As *C*- versus *D*-status cannot be deduced from PCIP-tag information alone, *de novo* insertions are in dark grey. The 95% confidence interval of the estimates (black lines) was determined by bootstrapping. (**e-g**) Proportion of *C* and *D* elements amongst endogenous ("Const(itutive)", red) and *de novo* inserted ("Denovo", blue) ERVK[2-1-LTR] elements. *De novo* insertions mapped to mixed (*C* and *D*) haplotypes were distributed amongst *C*- and *D*-types according to the corresponding ratio of inherited elements. The black vertical bars in "Denovo" correspond to the 95% confidence interval determined by bootstrapping. The difference (two-sided) between the C/D ratio for inherited and *de novo* insertions was significant (** meaning $p < 0.01$). Source data are provided as a Source Data file.

Strikingly, distinct haplotypes are not contributing equally to *de novo* events (Fig. 7b-d). For example, the *D*-only A23 and A33 haplotypes are jointly contributing 64.3% of *de novo* events while they are accounting for only 10.7% of inherited ERVK[2-1-LTR] elements. Also, the mixed A41 haplotype contributes 31.9% of *de novo* insertions, although it accounts for only 13.9% of inherited ERVK[2-1-LTR] elements. Of note, the A41 haplotype contributes 95.8% of putative *C*-type *de novo* insertions, while only accounting for 50% of constitutive *C*-type ERVK[2-1-LTR] alleles. Taken together, and even if we conservatively consider that all A41 *de novo* insertions are of *C*-type, A23 and A33 *D*-type *de novo* insertions are outcompeting A41 ones 2.6 to 1. It is noteworthy that we also observed five *de novo* insertions whose PCIP-tags didn't match any of the inherited ERVK[2-1-LTR] elements of the cognate bull, yet could have arisen by recombination.

Amongst the 398 *de novo* insertions detected in the three bulls, 304 were sampled once, 63 twice, 17 three times, 9 four times, and 5 more than four times (Fig. 8a and Supplementary Data 13). The frequency distribution of resampling rate informs about the developmental window during which *de novo* ERVK[2-1-LTR] mobilization is most likely to occur (Methods). We compared the real distribution, with the distribution obtained by simulations conducted under various timing scenarios, yet matching the real data for *de novo* insertion frequency (average number of *de novo* insertions per sperm cell) and number of explored haploid genomes. The results are compatible with *de novo* mobilization occurring during a window of ~5-9 consecutive cell divisions during the second half (in terms of number of cell division) of spermatogenesis, which may coincide with a period of reduced genome methylation [29] (Fig. 8a-b, Supplementary Fig. 14 and Supplementary Data 13). This is in striking contrast with *de*

novo point mutations of which at least 30% were shown to occur during early cleavage embryonic cell divisions in bulls [30].

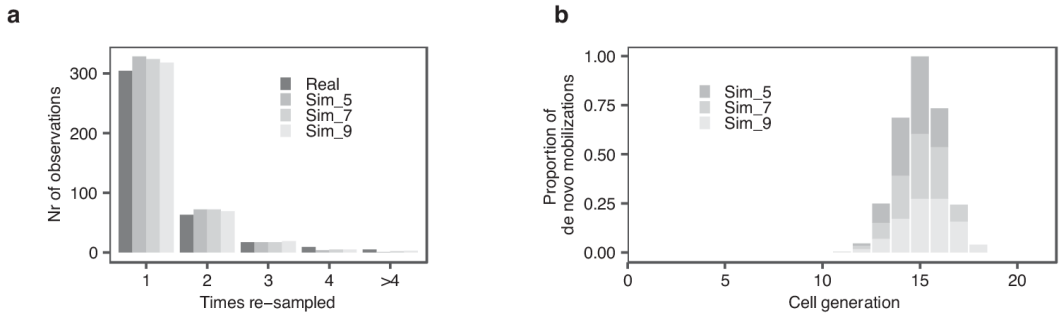


Figure 8: Determining the developmental window during which *de novo* ERVK mobilization occurs from the frequency distribution of resampling rate.
(**a**) Number of *de novo* insertions captured 1x, 2x, 3x, 4x and >4x for real data (darkest grey) and data simulated using the developmental windows shown in (b). (**b**) Stacked distributions of *de novo* mobilization rate at various cell generations of spermatogenesis (1 to 21, i.e. 20 cell divisions generating 2^{20}=1,048,576 "spermatogonial stem cells" from one ancestral primordial germ cell) that maximize the correspondence between simulated and real number and resampling rate of captured *de novo* insertions. Dark grey: 5-generation window; Grey: 7-generation window; Light grey: 9-generation window. The three window-sizes, when centered on cell generation 14, matched the real data (nearly) equally well (Supplementary Data 13). Source data are provided as a Source Data file.

Self-regulation of the number of ERVK[2-1-LTR] elements in the genome?

It is generally assumed that, after initial expansion, families of ERV elements progressively regress as host defense mechanisms emerge. We expected that GWAS might reveal components of these host defense systems, such as clusters of polymorphic piRNA targeting young ERVs. Yet, we could not identify obvious candidate defensive genes in the vicinity of the four associated loci without ERVK[2-1-LTR] elements (Supplementary Fig. 7). Moreover, the minor alleles (and hence more likely derived allele) in these four loci were systematically increasing rather than decreasing mobilization rate. These findings leave open the question of the nature of ERVK[2-1-LTR] neutralizing mechanisms.

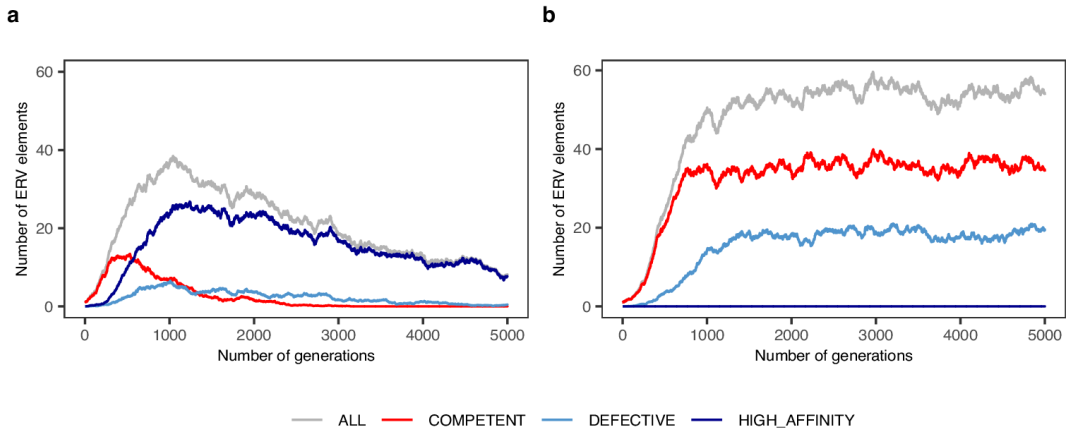


Figure 9: Possible contribution of defective yet high affinity elements to the demise of ERV families.
(a) Evolution of the average number of ERV elements in the genome (grey: total; red: *C*ompetent; light blue: *D*efective low affinity; dark blue: defective high affinity) of a panmictic population with 1,000 individuals, over 5,000 generations, with mutation and the possibility to generate *D*efective elements yet with high affinity for the mobilization machinery (provided in *trans* by *C*ompetent elements), and purifying selection. *C*ompetent elements (red) are rapidly overtaken by high-affinity *D*efective ones (dark blue) and die out. The remaining *D*efective elements (high- and low-affinity, dark and light blue) are progressively eliminated from the genome by purifying selection. **(b)** Same as in (a), yet without the possibility to generate high-affinity *D*efective elements. *C*ompetent elements are maintained in the population at average numbers that are determined by the stringency of the purifying selection. The ratio between *C*ompetent (red) and *D*efective ones (light blue) is determined by the mutation rate. Source data are provided as a Source Data file.

The observation that *D*-type ERVK[2-1-LTR] elements outcompete *C*-type ones, suggests an alternative mechanism that may contribute to the demise of ERV families. Indeed, as the number of defective ERV elements takes over the family at the expense of competent copies, which are themselves undergoing the assault of *de novo* mutations, the family may eventually lose all its competent elements and hence die out. We performed simulations to explore this hypothesis [31]. We assumed (i) that the mobilization rate in the germline of an individual is increasing with the number of competent elements (as observed), (ii) that elements undergo mutations that can only render them defective, but can either increase or decrease their affinity for the mobilization machinery, and (iii) that fitness is decreasing with or independent of the number of ERV elements in the genome (Methods). Under various combinations of parameter values, we systematically observed that defective elements with higher affinity for the mobilization machinery would emerge by mutation, and rapidly displace the competent elements, precluding further expansion of the family. From thereon, the impact on fitness determined the speed at which the family was eventually eliminated from the population (Fig. 9 and Supplementary Fig. 15).

Discussion

The polymorphic ERVK[2-1-LTR] clade occupies ~400 loci in the bovine genome. In this work, we have performed an in-depth characterization of a specific clade of bovine ERVs referred to by Repbase [15] as ERVK[2-1-LTR]. By combining whole genome and targeted sequencing, we show that members of this clade occupy ~ 400 loci in BB cattle, and that a large fraction of those are polymorphic in this population. We show that ERVK[2-1-LTR] elements come in three "morphs": solo-LTR, *C*(ompetent)-type elements and *D*(efective)-type elements. The majority of loci are biallelic, i.e. characterized by the wild-type "+" allele and one "ERV" allele which can be of *C*- (~9% of loci), *D*- (~47%), or solo-LTR-type (~41%). A minority of loci are triallelic (~2%), characterized by the cosegregation of "+" allele, a full-length (*C*- or *D*-type) ERV allele, and the derived solo-LTR allele (Supplementary Fig. 11). Of note, solo-LTRs are expected to be older than their full-length counterparts. Accordingly, their frequency spectrum is shifted upwards. The estimated proportions of *C*-, *D*-, and solo-LTR-type loci

therefore change with sample size: as sample size increases, more new *C*- and *D*-type than solo-LTR-type loci are uncovered, hence their proportion is increasing at the expense of the solo-LTR class.

The ERVK[2-1-LTR] clade derives from ancient endogenization of an unknown exogenous retrovirus. That a high proportion of ERVK[2-1-LTR] loci are polymorphic suggests that their colonization of the bovine genome is a relatively recent event. However, contrary to recent episodes of ERV endogenization reported in poultry, mice, sheep/goat, koala and cat [5], the identity of the exogenous retrovirus at the origin of the bovine ERVK[2-1-LTR] family remains unknown and may be extinct. Sequences with similarities ≥ 95% over ≥ 95% of the full length of the chromosome 19 *C*-type element exist in the reference genomes of other domesticated taurine (*Bos taurus*) and indicine breeds (*Bos indicus*), gaur (*Bos gaurus*), gayal (*Bos frontalis*), domestic yak (*Bos grunniens*), wild yak (*Bos mutus*), bison (*Bison bison*), water buffalo (*Bubalus bubalis*), and African buffalo (*Syncerus caffer*). A marked drop in similarity is observed when querying the genomes of caprinae, including sheep (*Ovis aries*) and goat (*Capra hircus*) (Supplementary Data 14). This suggests that ERVK[2-1-LTR] endogenization may have occurred in an ancestor of Bovinae, i.e. ~ 15 million years ago. To gain additional insights in the history of the ERVK[2-1-LTR] clade we compared the shared *GAG* and *ENV* sequences between *C*- and *D*-type elements. We restricted the analysis to regions immediately flanking the boundaries of the *C*- and *D*-specific segments and presenting little or no evidence of recombination (Supplementary Fig. 8). The average pairwise difference between *C*- and *D*-type elements was 38.4 in 250 base-pairs. Assuming a *de novo* mutation rate of $1x10^{-8}$ base pairs per generation [30], this would correspond to ~15 million generations or ~45 million years. However, this figure doesn't account for the extra mutations introduced during the undetermined number of reputedly low-fidelity reverse transcription steps that separate the different *C*- and *D*-type elements considered in this analysis. It is therefore bound to be overestimated, possibly grossly. We also analyzed the sequence divergence between the 5' and 3' LTRs of all *C* and *D*-type elements. Upon creation of a new insertion, 5' and 3'LTR have identical sequence. Observed differences therefore reflect the accumulation of *de novo* mutations and therefore presumably provide information about the time of insertion. The full-length ERVK[2-1-LTR] element with the most divergent 5' and 3'LTRs (98.3% similarity over 1,287 base pairs) is a *C*-type element on chromosome 6, present in the bovine reference genome and fixed in the Belgian Blue cattle population (Supplementary Data 8). Assuming a mutation rate of $1x10^{-8}$ base pairs per generation, this level of divergence is expected to accrue over ~622,000 generations or ~1.9 million years, hence supporting a more recent time of primordial endogenization of the ERVK[2-1-LTR] clade than the two previous estimates. It should be noted, however, that the eldest ERVK[2-1-LTR] elements, whose 5' and 3'LTR comparison would best inform about the time of origin of the clade, are more likely to have been reduced to solo-LTRs. This third figure is therefore liable to be an underestimate.

ERVK[2-1-LTR] elements are still actively mobilizing by reinfection of the bovine germline. Approximately 50% of *C*-type elements have intact ORF for *GAG*, *PRO*, *POL* and *ENV*, suggesting that ERVK[2-1-LTR] might still be active in the bovine germline. This prediction was further supported by the discovery that the insertion of an ERVK[2-1-LTR] element in the *APOB* gene underpins cholesterol deficiency in HF cattle. In this work, we unambiguously demonstrate that ERVK[2-1-LTR] are indeed still active in the bovine male and female germline, first by identifying five *de novo* mobilization events in whole genome sequenced three-generation pedigrees, and secondly by capturing thousands of *de novo* insertions in sperm cells by means of PCIP [18]. By analyzing the degree of mosaicism of captured *de novo* events, we show that ERVK[2-1-LTR] mobilization occurs in late spermatogenesis, yet can affect stem cells that persist throughout the entire life of the animal. The timing of mobilization could coincide with a phase of spermatogenesis during which genome methylation is at its lowest point [17,29]. The observation of an intact *ENV* ORF in a large proportion of *C*-type elements, combined with *in silico* prediction that the matrix domain (MA) of the GAG protein of *C*-type elements is a target for N-terminal myristoylation (Supplementary Fig. 12), strongly suggests that ERVK[2-1-LTR] elements still multiply by within-host, intercellular reinfection rather than by the supposedly more effective intracellular retro-transposition route. Which cells produce the viral-like particles (somatic or germline cells), which membrane receptor viral-like particles recognize in the recipient cells, and which reinfection path is used (free particles, virological synapses or microtubule-mediated transport) remains unknown.

The rate of ERVK[2-1-LTR] mobilization is determined by the number of inherited C-type elements. On average, ERVK[2-1-LTR] elements mobilize in the male germline at a rate of ~one in 150 sperm cells. Yet, this rate varies at least ten-fold between individuals, while remaining remarkably constant over time for a given animal (76% repeatability). We show in this work that the individual ERVK[2-1-LTR] mobilization rate of bulls is determined, to a large extent ($r^2 \approx 0.25$), by the number of inherited *C*-type elements. Intact *C*-type elements (i.e. with uninterrupted ORF for all four *GAG*, *PRO*, *POL* and *ENV* genes) have the most consistent effect on mobilization rate (Fig. 6b). However, even intact *C* elements differ in their effect. For example, an extra copy of the chr19:50466809 ERVK[2-1-LTR] element increases the mobilization rate by 5 events per 1,000 sperm cells, while the effects of the chr2:129626969 and chr21:19469667 elements are approximately half or 2.5 events per 1,000 sperm cells (Supplementary Fig. 7b). Non-intact *C*-type elements appear to have genuine, albeit more modest effects on mobilization rate, provided that there is at least one intact *C*-type element in the genome (Supplementary Fig. 13). Thus, the effect of the number of non-intact *C*-type elements on mobilization rate depends on the genotype of the animal for intact *C*-type elements, i.e. there is an epistatic interaction between intact and non-intact *C* elements with regards to their effect on mobilization rate. This probably indicates that the concentrations of at least some of the four gene products remain a limiting factor in the mobilization process.

De novo mobilizations are dominated by trans-complemented D-type insertions. We further show that – despite the driving role of *C*-type ERVK[2-1-LTR] elements – *de novo* mobilization events are dominated by the insertions of specific *D* elements (Fig. 7). We assume that this reflects (i) the occurrence of *trans*-complementation between *C*- and *D*-type elements, and (ii) a higher affinity of specific *D*- over *C*-type elements for the mobilization machinery. We don't know at this point what the molecular bases of the *trans*-complementation and differential affinity may be. However, the evidence for pervasive recombination between ERVK[2-1-LTR] elements, including between *C*- and *D*-type elements, suggests that *C*- and *D*-type gRNAs can pair in the pseudodiploid virus-like-particles and generate recombinant extrachromosomal DNA (ecDNA) molecules [22]. That *D*-type insertions are able to outnumber *C*-type insertions could either be due to the fact that *D*-type gRNAs are more abundant in cells than their *C*-type counterparts (f.i. because some of them are transcribed at a higher rate or are more stable), or because they are more effective in utilizing the *C*-type provided machinery to generate ecDNA[32]. Another intriguing possibility would be that "heterozygous" virus-like particles harboring a *C*- and a *D*-type gRNA would preferentially produce *D*-type double stranded extracellular DNA. *Trans*-complementation between mobile elements has been extensively documented before [28]. Examples of complementation in *trans* between morphs of the same ERV clade include ETn by MusD [33,34] and IAP IΔ1 by IAP [35,36] in mice, RecKoRV by KoRVA in koala [37,38], and possibly Type 1 by Type 2 HERV-K in human [39]. Strikingly, the bovine ERVK[2-1-LTR] *C/D* pair, murine MusD/ETn pair and koala KoRVA/RecKoRV pair share a pattern of swapping of a central segment of the competent ERV element with an old piece of retroelement, yet conservation of flanking sequences encompassing portions of the *GAG* and *ENV* (ERVK[2-1-LTR] and KoRVA) or *GAG* and *POL* (MusD) genes (in addition to the LTRs), which may be suggestive of *cis*-effects on the effectiveness of *trans*-complementation as reported for IAP IΔ1 [35].

D-type elements may act as parasite-of-parasite gene drives. We expected that GWAS for ERVK[2-1-LTR] mobilization rate would reveal emerging components of the host silencing machinery. Yet, there was no obvious evidence for genes participating in such mechanisms in any of the eight detected association peaks. In fact, the minor allele for all top variants (and hence more likely derived allele) was always increasing (rather than decreasing) mobilization rate. This probably indicates the ERVK[2-1-LTR] endogenization is too recent, and that specific silencing mechanisms are yet to evolve in the bovine population. However, our results reveal another mechanism that may precipitate the demise of ERV families. Indeed, *D*-type elements may act as parasite-of-parasite gene drives that may cause the spontaneous implosion of the ERVK[2-1-LTR] clade.

Possible phenotypic effects of ERVK[2-1-LTR] activity. We show that ERVK[2-1-LTR] mobilization generates deleterious mutations underpinning genetic defects in the cattle population. As shown by the outcome of knocking down host defense mechanisms against transposable elements [17], excessive ERV

mobilization rates may compromise fertility, which is an important concern in livestock breeding. To what extent differences in ERVK[2-1-LTR] mobilization rate correlate with fertility in cattle can now be addressed. Also, whether ERVK[2-1-LTR] mobilization affects the transcriptome and contributes to beneficial variation that can be exploited in breeding programs is another interesting question to pursue.

Material and methods

Ethics approval. We used biological materials provided by a breeding program, and the biological materials were collected by veterinarians as a part of routine animal breeding activities, not for experiments. All sperm straws were provided by commercial breeding companies. Hence, our work does not involve animal experiments.

The Damona pedigree for the detection of de novo ERVK mobilizations. The *Damona* pedigree comprises 743 Dutch Holstein-Friesian cattle assigned to 127 three generation pedigrees including a sire, a dam, an offspring, an average of 8 sibs of the offspring (range:0-17), and an average of 5 grand-offspring (range:1-11). Moreover, 1.4 grandparents per pedigree are available on average as well (5 pedigrees with 4 grandparents, 16 with 3, 38 with 2, 34 with 1, 34 with 0). The 127 pedigrees overlap: for instance, an animal can be offspring in one pedigree and parent in another. All animals from the *Damona* pedigree were whole genome sequenced (females: blood DNA; males: 24% blood, 76% sperm DNA). The average sequence depth is 26x for sire-dam-offspring trios, 17x for sibs, 10x for grand-offspring and 27.3x for grand-parents.

LocaTER. The *LocaTER* pipeline scans individual, whole genome sequences (short reads) for insertions of queried interspersed repeats (including ERV elements) that are present in the newly sequenced genome but not in the reference genome to which the reads are mapped (Supplementary Fig. 16). It searches for two features in the sequence data that are characteristic of such insertions. The first is an above normal concentration of discordant paired reads (i.e. the paired ends map to different genome locations) that assort in two adjacent sets. Set A comprises paired reads mapping respectively to the sense strand of the candidate location and to one end of the queried type of repeat (f.i. ERV elements), followed (in the 5' to 3' direction on the reference genome) by set B, comprising paired reads mapping respectively to the anti-sense strand of the candidate location and to the other end of the queried repeat. The second characteristic is an above normal concentration of soft- and hard-clipped reads, with (nearly) identical clip position located between sets A and B, and mapping the actual insertion site of the repeat. *LocaTER* also attempts to determine the genotype of the individual for the corresponding insertion, i.e. heterozygous or homozygous. Predictions of *LocaTER* were manually checked (using IGV) for a third characteristic of ERV insertions, namely a short, local target site duplication. A more detailed description of the functioning of *LocaTER* is provided in Supplementary Methods. The *LocaTER* pipeline is available from https://github.com/Lijingtangbo/TE-rate_manuscript.

Adapting PCIP to quantitatively estimate the rate of TE mobilization. Molecular biology: the modified PCIP reaction comprises six steps (Fig. 3a): (i) using 500 ng of genomic DNA as starting material, fragments containing ERVK[2-1-LTR] sequences were cleaved using a pair of single guide RNAs (Integrated DNA Technologies) targeting sequences at 429 and 374 bp from the 5'LTR and 3' LTR, respectively (Supplementary Data 15), and *S.pyogenes* Cas9 (New England Biolabs, M0386S). Of note, these sequences are the same for the *C*- and *D*-type elements. (ii) The digested DNA was further

mechanically sheared to ~3 Kb using a Megaruptor-1 (Diagenode), end-repaired using NEBNext EndRepair Module (New England Biolabs, E6050L), and purified with Agencourt AMPure XP beads (Beckman Coulter, A63881). (iii) The resulting DNA fragments were circularized using T4 DNA Ligase (New England Biolabs, M0202L), residual linear fragments eliminated with Plasmid-Safe-ATP-Dependent DNase (Epicentre, E3110K), purified, and reaction products split in two aliquots. (iv) Circular molecules encompassing ERVK[2-1-LTR] sequences were reopened with *S. pyogenes* Cas9 using distinct single guide RNAs (Integrated DNA Technologies) for the two aliquots (Supplementary Data 15), and purified. (v) ERVK[2-1-LTR] encompassing linear fragments were inverse PCR amplified using aliquot-specific primer pairs (Supplementary Data 15) with LongAmp Taq DNA Polymerase (New England Biolabs, M0533S), and purified. (vi) The purified amplicons were mechanically sheared to ~350 bp using a Bioruptor-pico (Diagenode), sequencing libraries generated using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, E7645L), and indexed libraries pooled and sequenced on a Novaseq 6000 sequencer (Illumina) targeting ~8 million 150 bp paired-end reads per library. Data processing: The ensuing sequence reads were demultiplexed, quality assessed using fastQC (https://github.com/s-andrews/FastQC), adapter sequences trimmed using Cutadapt (https://github.com/marcelm/cutadapt), and trimmed reads mapped to the bovine reference genome using BWA-MEM [40] and converted to BAM format using SAMtools [41]. Using a custom-made python script, we first identified clipped reads using CIGAR information. We selected clipped reads with mapping quality ≥ 40 and a minimum of 10 clipped bases. We then mapped the clipped reads to the segments of the ERVK[2-1-LTR] genome corresponding to the ERV-Tags in Fig. 3 (1 and 2 for the 5'LTR libraries and 3 and 4 for the 3'LTR libraries). We demanded an alignment score ≥ 0.6 to declare a hit, and labelled the read as either an insertion site (IS) or a shearing site (SS) read. When possible, we extended the alignment with ERVK[2-1-LTR] into non-clipped bases to refine the positions of the SS and IS. We then merged SS and IS site with same "breakpoint", thereby identifying candidate SS and IS supported by multiple concordant reads. We then paired IS with their cognate SS. The pairing was based on orientation (f.i. a 5' SS should be located upstream of the 5' IS for an ERV element in "sense" orientation, and downstream of the 5' IS for an ERV in "antisense" orientation; Fig. 3) and distance (the maximum distance between IS and SS was set at 5 Kb). One IS can be paired with one SS (typical for non-mosaic *de novo* insertions) or with multiple SS (typical for mosaic *de novo* insertions or inherited ERVK[2-1-LTR] elements). Additional filters were then applied to select candidate *de novo* insertions for manual inspection, including a distance of more than 3 Kb from constitutive ERVK[2-1-LTR] elements, and only observed in one animal. Data normalization: The above-mentioned procedure yields, for each individual s of T, the number of shearing sites for each i of 123 (5'LTR) or 189 (3'LTR) selected "constitutive" ERVK[2-1-LTR] elements (SSC_{is}), as well as (at least one) shearing site for N_s *de novo* ERVK[2-1-LTR] insertions. The SSC_{is}'s were modeled (with the R lm() function) as $SSC_{is} = ERV_i + ID_s + \varepsilon_{is}$, where ERV_i is the effect of locus i (not all ERVK[2-1-LTR] elements undergo PCIP as effectively), ID_s is the effect of individual s (we don't engage exactly the same

amount of DNA and of same quality in the PCIP reaction for all samples), and ε_{is} is the error term. We determine the values of ERV_i and ID_s that minimize the error sum of squares. Twice the ensuing ID_s (corresponding to the number of effectively captured haploid genomes for individual s) is used as normalization factor to estimate the individual-specific ERVK[2-1-LTR] mobilization rate as $TR_S = N_S/2ID_s$. Software: Analysis of the sequence reads was conducted using a mixed Python/R pipeline that uses raw sequence data as input and detects both constitutive and *de novo* ERV integration sites automatically. The pipeline consists of four modules: the mapping module, junction reads annotation module, clustering module and normalization module. It is available from https://github.com/Lijingtangbo/TE-rate_manuscript.

GWAS for ERVK[2-1-LTR] mobilization rate. SNP/Indel genotyping and imputation: Genomic DNA was extracted from sperm using standard procedures. The genome of 40 of 430 Belgian Blue bulls was whole genome-sequenced using Illumina S4 chemistry and a Novaseq 6000 instrument at average sequence depth of 40, and SNPs/Indels detected and genotyped using GATK[20]. The remaining 390 animals were genotyped using Illumina's medium density (~55K variants) SNP genotyping arrays. For the latter, genotype information was augmented by imputation in two steps (first to high density using 890 Belgian Blue animals genotyped with Illumina's high density (~770,000 variants) as reference, and then to whole genome using sequenced Belgian Blue animals as reference) using Shapit4 [42] for phasing, followed by Minimac4 [43] for actual imputation. We kept 10,875,490 variants with minor allele frequency over 2% and imputation accuracy over 90% for GWAS. Four hundred fifteen of the 430 utilized Belgian Blue animals were homozygous for the nt821(del11) mutation in the myostatin (*MSTN*) (gene causing double-muscling in homozygotes [44]), eight heterozygous, and seven homozygous wild-type. *MSTN* genotype had no effect on ERVK[2-1-LTR] mobilization rate (data not shown). Genotyping at 309 constitutive ERVK[2-1-LTR] loci: We applied the modified PCIP procedure to sperm DNA of all 430 Belgian Blue bulls. Three hundred and nine constitutive ERVK[2-1-LTR] loci were identified as loci marked by high numbers of clustered shearing sites (SS in Fig. 3a) relative to the number of explored haploid genomes (see description of PCIP method above) in at least some animals. Animals were genotyped for the corresponding ERVK[2-1-LTR] loci based on the within-locus (to account for the difference in PCIP efficiency for different ERVK[2-1-LTR]) distribution of the ratio between the number of observed SS and number of explored haploid genomes. GWAS: We conducted GWAS with GEMMA [45] using TR_S as phenotype. The model included a fixed regression on variant dosage (additive effect), as well as a random polygenic effect. Estimating the polygenic effects requires the additive genetic relationship matrix, which was computed from marker data using option "-gk1" in GEMMA. In a second round of GWAS, dosage of the chromosome 19 ERVK[2-1-LTR] locus (i.e. number of alleles with the ERVK[2-1-LTR] insertion) was added as an extra covariate in the model. Linkage disequilibrium in the Belgian Blue population results in the fact that the ~11 million variants behave as ~ 500,000 independent tests [46], yielding a genome-wide significance threshold of $0.05/500{,}000 = 10^{-7}$. Testing the effect of inbreeding on mobilization rate: The proportion of autosomal

SNPs (MAF ≥ 0.1) with homozygous genotype was used as proxy for a bull's coefficient of inbreeding. Its effect on mobilization rate was estimated by linear regression using the lm() R function.

Probability of coincidence of polymorphic ERVK[2-1-LTR] elements and association peaks. For four of the eight association peaks identified by GWAS, an ERVK[2-1-LTR] element was either the lead variant (1x), or in high LD ($r^2 \geq 0.62$) with the lead variant (3x). The eight corresponding lead variants were characterized by a minor allele frequency (MAF) in the 430 bulls which we denote f_i. To estimate the probability to observe this level of coincidence or more, fortuitously, we determined, for each lead variant i, the proportion of variants (across the entire genome) with $f_i - 0.025 < MAF < f_i + 0.025$ (0.05 bin) that would be in high LD ($r^2 \geq 0.62$) with anyone of the 87 polymorphic ERVK[2-1-LTR] elements with $MAF \geq 0.02$: p_i. We then sampled one object in each of eight "urns" (with respective probability of success = p_i), repeated this process 10^8 times, and counted how often we obtained 4 or more "successes". The probability to obtain 1, 2, 3 and > 3 successes was 0.009, 3.9×10^{-5}, 7.0×10^{-8}, and zero.

Establishing a catalogue of ERVK[2-1-LTR] elements in BB cattle. Reference sequences for *C*- and *D*-type ERVK[2-1-LTR]: We used ERVK[2-1-LTR] present in the bovine reference genome (ARS-UCD1.2 genome assembly) as reference (*C*-type: chr18-59909149-59919759; *D*-type: chr14-70257004-70263970). We identified the open reading frames (ORFs) corresponding to the *GAG*, *PRO*, *POL* and *ENV* genes, bounded on the amino-terminal end by the first ATG for *GAG*, *POL*, *ENV* genes, and by the first codon following the *GAG* stop codon for *PRO* (Supplementary Fig. 9). The corresponding protein sequences were blasted against 'viruses' (taxid: 10239) non-redundant protein sequences to annotate protein domains. Determining the full-length sequence of 221 constitutive ERVK[2-1-LTR] elements: We aimed at amplifying the full-length of all constitutive ERVK[2-1-LTR] detected by PCIP (see above). This was done by amplifying each ERVK[2-1-LTR] element as two overlapping fragments, jointly spanning its full length. Each primer pair (for the 5' half and 3' half of the element, respectively) comprised one primer targeting flanking sequences, and one primer in the body of the ERVK[2-1-LTR] element. As we did not know in advance whether a targeted ERVK[2-1-LTR] element was of *C*- or *D*-type, we tested each flanking primer with *C*-type and *D*-type specific "ERVK body" primers. We performed long-range PCR using the LongAmp Hot Start Taq 2× Master Mix (New England Biolabs, M0533S) and primers listed in Supplementary Data 15. We successfully amplified both left and right halves for 221 ERVK[2-1-LTR] elements, and one half (either left or right) for an additional 74 for a total of 295. This indicates that there are no other important ERVK[2-1-LTR] elements other than the ~6.8 Kb *D*-type and the ~10Kb *C*-type elements described in this work. The left and right amplicons of the 221 ERVK[2-1-LTR] elements were mechanically sheared to ~500 bp using a Bioruptor-pico (Diagenode), sequencing libraries generated using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, E7645L) with five cycles of PCR to amplify the adapter-ligated fragments. The quality of the libraries was checked using QIAxcel Advanced System (QIAGEN) and a qPCR. Indexed libraries were pooled and sequenced on a Miseq sequencer (Illumina) targeting 200×

coverage. Resulting reads were aligned to cognate (i.e. *C*- or *D*- type, 5' half or 3' half reference) ERVK[2-1-LTR] references using BWA-MEM [40]. We further checked, for each library, whether the corresponding flanking sequences were mapping to the expected genomic coordinates. Variant sites with respect to the reference sequence were annotated with BCFtools [47] and manually curated using IGV [48]. Identifying ERVK[2-1-LTR] loci with solo-LTR alleles: We designed probes specific for the + and ERV allele, respectively, for 291 polymorphic ERVK[2-1-LTR] elements, added them to a medium density (MD) SNP genotyping arrays (Illumina), and re-genotyped the 430 bulls. This yielded usable genotypes for 193 ERVK[2-1-LTR] elements. For at least five loci, bulls were genotyped as ERV/+ with the array, while appearing +/+ by PCIP. This suggested the co-occurrence of a solo-LTR allele for these loci, and this hypothesis was confirmed by PCR amplification of the predicted solo-LTR in at least one bull for each one of the five loci. Confronting the results of PCIP-based genotyping of the 40 whole genome sequenced bulls with the results of the analysis of their genome with *LocaTER* revealed 100 ERVK[2-1-LTR] insertions detected by *LocaTER* but never by PCIP. We hypothesized that only solo-LTR alleles would still be segregating in the Belgian Blue population at these loci (i.e. the full-length ERV allele would have been lost). We confirmed the solo-LTR status by PCR amplification and sequencing in at least one bull for 84 of these loci.

Estimating the proportional contribution of PCIP-tag-defined haplotypes of constitutive ERVK[2-1-LTR] elements to de novo insertions. PCIP yields two sequence tags of the captured ERVK[2-1-LTR] element on the 5'-side, and two on the 3'-side (Fig. 3a). Polymorphisms in these tags (amongst the 298 endogenized ERVK[2-1-LTR] elements) allow to distinguish 28 5'-side combinations and 39 3'-side combinations, which assort into 77 5'-3' combinations or haplotypes, referred to as A1, A2, … A77 in Fig. 7. The number of endogenized ERVK[2-1-LTR] elements sharing the same haplotype ranges from 1 to 76. PCIP typically generates 5'-side or 3'-side tag information for *de novo* insertions, exceptionally (∼4%) both (5'-3'). Also, PCIP may only provide information about one (of the two) 5'- or 3'-side tags. Thus, the assignment of a *de novo* ERVK[2-1-LTR] insertion to a specific existing haplotype is often ambiguous: multiple haplotypes remain possible given the available tag information. To nevertheless be able to accurately estimate the contribution of each of the haplotypes present in the genome of a bull (i.e. the A41, A44, A37's, etc. in Fig. 7) to the captured *de novo* insertions, we used an estimation-maximization (EM) approach. We started assuming uniform contributions of each endogenized haplotype. As an example, the 15 constitutive ERVK[2-1-LTR] elements inherited by bull A (Fig. 7b) fall into 13 haplotypes (A44, A41, A37, A63, A76, A77, A75, A36, A15, A23, A33, A52, A71). The 13 starting contributions were therefore set at 1/13 (= 0.077). If a *de novo* insertion "Z" is compatible with, say haplotype A41 and A44, it is assigned for half to A41 (coefficient $0.5 = 0.077/(0.077 + 0.077)$ and for half to A44 (coefficient 0.5). This process is repeated for all *de novo* insertions, and the contribution of each haplotype computed from the sum of coefficients across all *de novo* insertions (divided by the number of insertions). After this first iteration, the contributions will have shifted from uniformity. For instance, to p_{A41} for A41 and p_{A44} for A44. The coefficients, for

insertion "Z", now become $p_{A41}/(p_{A41}+p_{A44})$ for A44 and $p_{A44}/(p_{A41}+p_{A44})$ for A41. The same process is repeated over all insertions. New contributions are then computed from the sum of updated coefficients across *de novo* insertions. The process is repeated until convergence (~ 20 iterations).

Simulating frequency distributions of resampling rate (1,2, 3, .. times) of de novo insertions matching experimental data. When, during the development of the germline, the ERVK[2-1-LTR] elements are mobilized is an important question. The degree of mosaicism of *de novo* ERVK[2-1-LTR] insertions provides information about the timing of the mobilization event. If a *de novo* insertion has a "dosage" of 50% (i.e. detected once every two studied haploid genomes), it supposedly has occurred at the embryonic one cell stage, if 25% it supposedly has occurred at the two cell stage, etc. If the dosage is ~1/1000 and assuming that there are ~1 million spermatogonial stem cells [49], it suggests that it has occurred ten cell divisions before the spermatogonial stem cell stage, yielding a "clone" of ~1,000 (2^{10} = 1,024) spermatogonial stem cells sharing the same insertion. If the dosage is ~1/1,000,000, it suggests that the mobilization event occurred during the very last cell division(s). A problem is that we only explored between 5,000 and 10,000 haploid genomes for the three bulls for which multiple (6 to 9) PCIP experiments were conducted. Thus, we cannot accurately measure dosages below ~1/5,000. Any *de novo* insertion that is captured once will have an estimated dosage of $1/g$, where g is the number of effectively studied haploid genomes (see above). The actual dosage may be much lower; it could just be that this insertion was "by chance" one of those captured out of a very large pool of very rare *de novo* insertions. Of note, contrary to dosage for individual *de novo* insertions, which is poorly estimated if low, the number of *de novo* insertions per explored haploid genome is estimated in an unbiased fashion. Let us call the number of distinct *de novo* insertions captured (at least once) u, and the total number of captured *de novo* insertions (i.e. accounting for repeated capture) t. The number of *de novo* insertions per explored haploid genome is t/g. It was 0.030, 0.019 and 0.024 for bulls 1, 2 and 3, respectively. This is the phenotype that was used for the GWAS (see above).

There is, however, information to determine at which time during development *de novo* insertions occur, in the frequency distribution of insertions that are re-captured 1, 2, 3, … times. If all observed *de novo* insertions are only captured once ($t = u$), it means that there is a large pool of very rare insertions that must all have occurred at a very late stage of spermatogenesis. If most *de novo* insertions are recaptured multiple times ($t \gg u$), it means that *de novo* insertions occur at earlier stages of spermatogenesis. Thus, there is information about the developmental timing of *de novo* mobilization in the frequency distribution of the number of *de novo* insertions captured 1x, 2x, 3x, etc ($f_1, f_2, f_3, \ldots$). The corresponding distributions are shown in Supplementary Fig. 14 for the three bulls.

To gain insights in the developmental windows during which ERVK[2-1-LTR] mobilization takes place, we performed simulations under various scenarios to find those yielding recapturing frequency distributions that matched the real data best. We assumed a very simple model of gametogenesis in which 1,048,576 spermatogonial stem cells (=2^{20}) would derive from 20 successive binary cell divisions

starting from a single precursor cell. A *de novo* insertion occurring at cell generation 1, would be characterized by a dosage per haploid genome (in sperm) of 0.5. One occurring at cell generation 10, would have a dosage of 0.00097, and at generation 21, a dosage of 4.8×10^{-7}. We defined windows of cell generations during which *de novo* mobilization may occur. Windows were centered around one specific generation (focal generation), that was most mobilization prone. Windows could be narrow (f.i. occurring in one generation only) or broad (up to 9 generations centered on the focal one). For windows encompassing multiple generations, the mutability of cells in a given generation was modelled using a binomial distribution, i.e. the mutability was highest for the focal generation and decreased with increasing distance from the focal generation. Given a choice of multi-generational window, the probability for a mobilization event to have occurred at one of the spanned generations was not only a function of the mutability of that generation, but also of the number of cells in that generation (which doubles at every additional generation). Having chosen a given window, we simulated *de novo* mobilization events, spread across the qualifying generations according to their relative probabilities, until the sum of the corresponding dosages corresponded to the number of *de novo* insertions per explored haploid genome (i.e. t/g) for the studied bull (1,2 or 3). Every *in silico* generated insertion corresponded to a ball in an urn with a sampling probability corresponding to its dosage. A ball corresponding to an insertion free genome was added to the urn with a sampling probability of 1-t/g. We then sampled g balls from this urn with replacement, using the corresponding vector of sampling probabilities, where g is the actual number of studied haploid genomes for the three bulls (7690, 5386 and 9178). We then counted the number of non-insertion free balls (i.e. *in silico* insertions) that were sampled 1x, 2x, 3x, 4x and >4x. We performed 50 simulations per window and compiled the average number of observations for each category (1x, 2x, 3x, 4x and >4x). We compared this distribution with the corresponding real distribution (actual observations for bulls 1, 2 and 3). The quality of the match was quantified as the sum of the differences (absolute value) between simulated and real number of observations. Such simulations were conducted for window sizes ranging from 1 to 9, and sequentially considering every one of the 21 generations as focal one. Simulations were conducted separately by bull. As the results were very consistent for the three bulls, we summed the sum of the differences across bulls and selected the window that minimized the overall differences between simulated and real numbers across the three bulls (Supplementary Data 13). The corresponding window, as well as comparison between real and simulated frequency distributions, are shown in Fig. 8a-b.

Simulating the evolution of ERV families in a panmictic population. We simulated the evolution of a panmictic population of constant size N over G generations. The next generation (g_{i+1}) was obtained from the previous one (g_i) by sampling N times (with replacement) two parents (i.e. sex was not considered). The probability to become a parent was affected by the fitness of the individual (f_i, range: 0 – 1). The fitness of the individual was determined by the total number of ERV elements in its genome (t_i), and was modelled as: $f_i = 1 - {t_i}^p / {t_{max}}^p$. t_{max} is the maximum number of tolerated ERV elements in the genome, and was set at 250 (if $t_i \geq 250$, f_i equals 0). The stringency of purifying

selection was adjusted using p: $p = 1$ ("linear") > $p = 2$ ("quadratic") > $p = 3$ ("cubic"). In addition, we considered a scenario where fitness was not affected by the number of ERV elements ("none"; $f = 1$). At generation g_1, the population was "seeded" with one ERV element in the genome of a proportion s of individuals. These initiating ERV elements were considered to be competent (i.e. drive mobilization by producing the needed machinery), and have an affinity (a_{ij}) of 1 for this machinery. Once the parents of an offspring were selected, we simulated Mendelian transmission of the parental ERV elements to the offspring assuming that each parental ERV has a 50% chance to be inherited by the offspring. In addition, we allowed for *de novo* ERV mobilization in the gametes. Every parental ERV had a certain probability (q_{ij}) to generate a new copy of itself in the gamete inherited by the offspring: $q_{ij} = r_i \times a_{ij} / \sum_{j=1}^{t_i} a_{ij}$. In this, r_i is the *de novo* mobilization rate of parent i. It was determined by the number of competent ERV elements (c_i) in the parent's genome as follows: $r_i = (1 - e^{-0.1 c_i}) \times r_{max}$, where r_{max} is the highest possible mobilization rate set at 0.05. Once the offspring and its ERV elements were generated, the ERV elements were allowed to undergo mutation at a rate μ. We tested 0.0001 and 0.001 as values for μ. When an ERV element underwent mutation, (i) it would either become defective (if it was competent before) or stay defective (if it was already defective before), and – if the model allowed - (ii) its affinity (a_{ij}) might change (unless its affinity was already 0 before mutation). This was accomplished by sampling a z value from $N(0,1)$ and adding it to the affinity of the ERV element prior to mutation. If the sum was negative, a_{ij} was set at 0. The corresponding script was written in Perl and is made available in https://github.com/Lijingtangbo/TE-rate_manuscript.

Acknowledgments

This project was conducted with funding from the ERC (ERC AdG-GA323030 to M.G., *Damona* project), the Walloon Region (DOG6 to M.G., *Causel* project), the University of Liège (FSR to C.C., *RetroBlue* project) and the Fund for Scientific Research in Belgium (F.R.S.-FNRS, PDR to C.C., *TE-rate* project); L.T. was supported by a Scholarship of the Chinese Scholarship Council (CSC); C.H.'s PhD fellowship was funded by Livestock Improvement Corporation (LIC, Hamilton, New Zealand); G.C.M.M. is post-doctoral research assistant of the H2020 EU project *BovReg* (GA815668); C.C. is senior research associate from the Fund for Scientific Research in Belgium (F.R.S.-FNRS). We are grateful to CRV (Arnhem, the Netherlands) for providing us with biological material for the *Damona* pedigree, to Inovéo (Ciney, Belgium) and FABROCA (Porcheresse, Belgium) artificial insemination centers for contributing Belgian blue sperm straws. We thank anonymous reviewers for their helpful and constructive comments which have helped to considerably improve the paper.

References

1. Osmanski, A. B. *et al.* Insights into mammalian TE diversity via the curation of 248 mammalian genome assemblies. *Science.* **380**, eabn1430 (2023).
2. Dewannieux, M. & Heidmann, T. Endogenous retroviruses: Acquisition, amplification and taming of genome invaders. *Curr. Opin. Virol.* **3**, 646–656 (2013).
3. Mager, D. L. & Stoye, J. P. Mammalian endogenous retroviruses. in *Mobile DNA III* 1079–1100 (2015).
4. Johnson, W. E. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat. Rev. Microbiol.* **17**, 355–370 (2019).
5. Chiu, E. S. & Vandewoude, S. Endogenous Retroviruses Drive Resistance and Promotion of Exogenous Retroviral Homologs. *Annu. Rev. Anim. Biosci.* **9**, 225–248 (2021).
6. Blanco-Melo, D., Gifford, R. J. & Bieniasz, P. D. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *Elife* **6**, 1–19 (2017).
7. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
8. Feusier, J. *et al.* Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res.* **29**, 1567–1577 (2019).
9. Zhang, Y., Maksakova, I. A., Gagnier, L., Van De Lagemaat, L. N. & Mager, D. L. Genome-wide assessments reveal extremely high levels of polymorphism of two active families of mouse endogenous retroviral elements. *PLoS Genet.* **4**, (2008).
10. Elmer, J. L. & Ferguson-Smith, A. C. Strain-specific epigenetic regulation of endogenous retroviruses: The role of Trans-acting modifiers. *Viruses* **12**, 1–21 (2020).
11. Kipp, S. *et al.* A new Holstein Haplotype affecting calf survival. *Interbull Bull.* 49–53 (2015).
12. Becker, D. *et al.* Allele-biased expression of the bovine APOB gene associated with the cholesterol deficiency defect suggests cis-regulatory enhancer effects of the LTR retrotransposon insertion. *Sci. Rep.* **12**, 1–14 (2022).
13. Menzi, F. *et al.* A transposable element insertion in APOB causes cholesterol deficiency in Holstein cattle. *Anim. Genet.* **47**, 253–257 (2016).
14. Schütz, E. *et al.* Correction: 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**, 1–15 (2016).
15. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
16. Rosenkranz, D. piRNA cluster database: A web resource for piRNA producing loci. *Nucleic Acids Res.* **44**, D223–D230 (2016).
17. Zamudio, N. & Bourc'His, D. Transposable elements in the mammalian germline: A comfortable

niche or a deadly trap. *Heredity (Edinb).* **105**, 92–104 (2010).

18. Artesi, M. *et al.* PCIP-seq: simultaneous sequencing of integrated viral genomes and their insertion sites with long reads. *Genome Biol.* **22**, 1–24 (2021).
19. Kirk, P. D. W., Huvet, M., Melamed, A., Maertens, G. N. & Bangham, C. R. M. Retroviruses integrate into a shared, non-palindromic DNA motif. *Nat. Microbiol.* **2**, (2016).
20. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-11.10.33 (2013).
21. Ito, J., Gifford, R. J. & Sato, K. Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 610–618 (2020).
22. Drost, H. G., Sanchez, D. H. & Eyre-Walker, A. Becoming a Selfish Clan: Recombination Associated to Reverse-Transcription in LTR Retrotransposons. *Genome Biol. Evol.* **11**, 3382–3392 (2019).
23. Van Laere, A. S., Coppieters, W. & Georges, M. Characterization of the bovine pseudoautosomal boundary: Documenting the evolutionary history of mammalian sex chromosomes. *Genome Res.* **18**, 1884–1895 (2008).
24. Greenwood, A. D., Ishida, Y., O'Brien, S. P., Roca, A. L. & Eiden, M. V. Transmission, Evolution, and Endogenization: Lessons Learned from Recent Retroviral Invasions. *Microbiol. Mol. Biol. Rev.* **82**, e00044-17 (2017).
25. Bologna, G., Yvon, C., Duvaud, S. & Veuthey, A. L. N-terminal myristoylation predictions by ensembles of neural networks. *Proteomics* **4**, 1626–1632 (2004).
26. Ribet, D. *et al.* An infectious progenitor for the murine IAP retrotransposon: Emergence of an intracellular genetic parasite from an ancient retrovirus. *Genome Res.* **18**, 597–609 (2008).
27. Cullen, H. & Schorn, A. J. Endogenous Retroviruses Walk a Fine Line between Priming and Silencing. *Viruses* **12**, (2020).
28. Bannert, N. & Kurth, R. The evolutionary dynamics of human endogenous retroviral families. *Annu. Rev. Genomics Hum. Genet.* **7**, 149–173 (2006).
29. Saitou, M. & Miyauchi, H. Gametogenesis from Pluripotent Stem Cells. *Cell Stem Cell* **18**, 721–735 (2016).
30. Harland, C. *et al.* Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. *bioRxiv* 79863 (2017). doi:10.1101/079863
31. Katzourakis, A., Rambaut, A. & Pybus, O. G. The evolutionary dynamics of endogenous retroviruses. *Trends Microbiol.* **13**, 463–468 (2005).
32. Onafuwa-Nuga, A. & Telesnitsky, A. The Remarkable Frequency of Human Immunodeficiency Virus Type 1 Genetic Recombination. *Microbiol. Mol. Biol. Rev.* **73**, 451–480 (2009).
33. Mager, D. L. & Freeman, J. D. Novel Mouse Type D Endogenous Proviruses and ETn Elements Share Long Terminal Repeat and Internal Sequences. *J. Virol.* **74**, 7221–7229 (2000).
34. Ribet, D., Dewannieux, M. & Heidmann, T. An active murine transposon family pair:

Retrotransposition of 'master' MusD copies and ETn trans-mobilization. *Genome Res.* **14**, 2261–2267 (2004).

35. Saito, E. S., Keng, V. W., Takeda, J. & Horie, K. Translation from nonautonomous type IAP retrotransposon is a critical determinant of transposition activity: Implication for retrotransposon-mediated genome evolution. *Genome Res.* **18**, 859–868 (2008).
36. Dewannieux, M., Dupressoir, A., Harper, F., Pierron, G. & Heidmann, T. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nat. Genet.* **36**, 534–539 (2004).
37. Löber, U. *et al.* Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. *Proc. Natl. Acad. Sci.* **115**, 8609–8614 (2018).
38. Tarlinton, R. E. *et al.* Differential and defective transcription of koala retrovirus indicates the complexity of host and virus evolution. *J. Gen. Virol.* **103**, (2022).
39. Costas, J. Evolutionary Dynamics of the Human Endogenous Retrovirus Family HERV-K Inferred from Full-Length Proviral Genomes. *J. Mol. Evol.* **53**, 237–243 (2001).
40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
42. Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 24–29 (2019).
43. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
44. Grobet, L. *et al.* A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nat. Genet.* **17**, 71–74 (1997).
45. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
46. Gualdrón Duarte, J. L. *et al.* Sequenced-based GWAS for linear classification traits in Belgian Blue beef cattle reveals new coding variants in genes regulating body size in mammals. *Genet. Sel. Evol.* **55**, 83 (2023).
47. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).
48. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
49. Mamsen, L. S., Lutterodt, M. C., Andersen, E. W., Byskov, A. G. & Andersen, C. Y. 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 (2011).
50. Tang, L. GWAS reveals determinants of mobilization rate and dynamics of an active endogenous retrovirus of cattle. (2024). doi:10.5281/zenodo.10630335

Supplemental material

Supplementary material content

Supplementary Figure 1:
Identification and functional validation of the causative mutation for CD.

Supplementary Figure 2:
Genomic features of polymorphic ERV elements.

Supplementary Figure 3:
Pedigree-based identification of five *de novo* insertions of ERVK[2-1-LTR] elements.

Supplementary Figure 4:
Assessing the repeatability and robustness of PCIP and lack of evidence of an effect of a bull's inbreeding coefficient on the rate of ERVK[2-1-LTR] mobilization in its germline.

Supplementary Figure 5:
Genomic landscape of ERVK[2-1-LTR] *de novo* insertions.

Supplementary Figure 6:
Genomic features of ERVK[2-1-LTR] *de novo* insertions.

Supplementary Figure 7:
Additional GWAS loci and genotypic effect for the eight significant loci.

Supplementary Figure 8:
Dot plot between sequences representing the C clade (X-axis) and D clade (Y-axis) and neighbor-joining tree obtained with the concatenated *GAG*-shared (55 base pairs) and *ENV*-shared (195 base pairs) segments.

Supplementary Figure 9:
Protein sequences and domain annotation of a representative element of the C-clade (chr19: 50,466,809 bp).

Supplementary Figure 10:
Examples of APOBEC3 mutational signature, recombination events and attrition at the boundaries.

Supplementary Figure 11:
Distribution of allelic frequencies of ERVK[2-1-LTR] segregating in Belgian Blue cattle.

Supplementary Figure 12:
Phylogenetic relationship with exogenous retroviruses and functional sequence features of ERVK[2-1-LTR].

Supplementary Figure 13:
Epistatic interaction between ERVK[2-1-LTR] elements with coding variants in the *GAG*, *PRO*, *POL* or *ENV* gene, and ERVK[2-1-LTR] elements without.

Supplementary Figure 14:
Comparison between the real and simulated frequency distribution of the resampling rate of ERVK[2-1-LTR] *de novo* mobilization events in BB bulls BE157971524, BE187351114, BE63811423.

Supplementary Figure 15:
Unselected examples of the evolution of the number of ERV elements (per genome) in a panmictic population of 1,000 animals over a course of 5,000 generations.

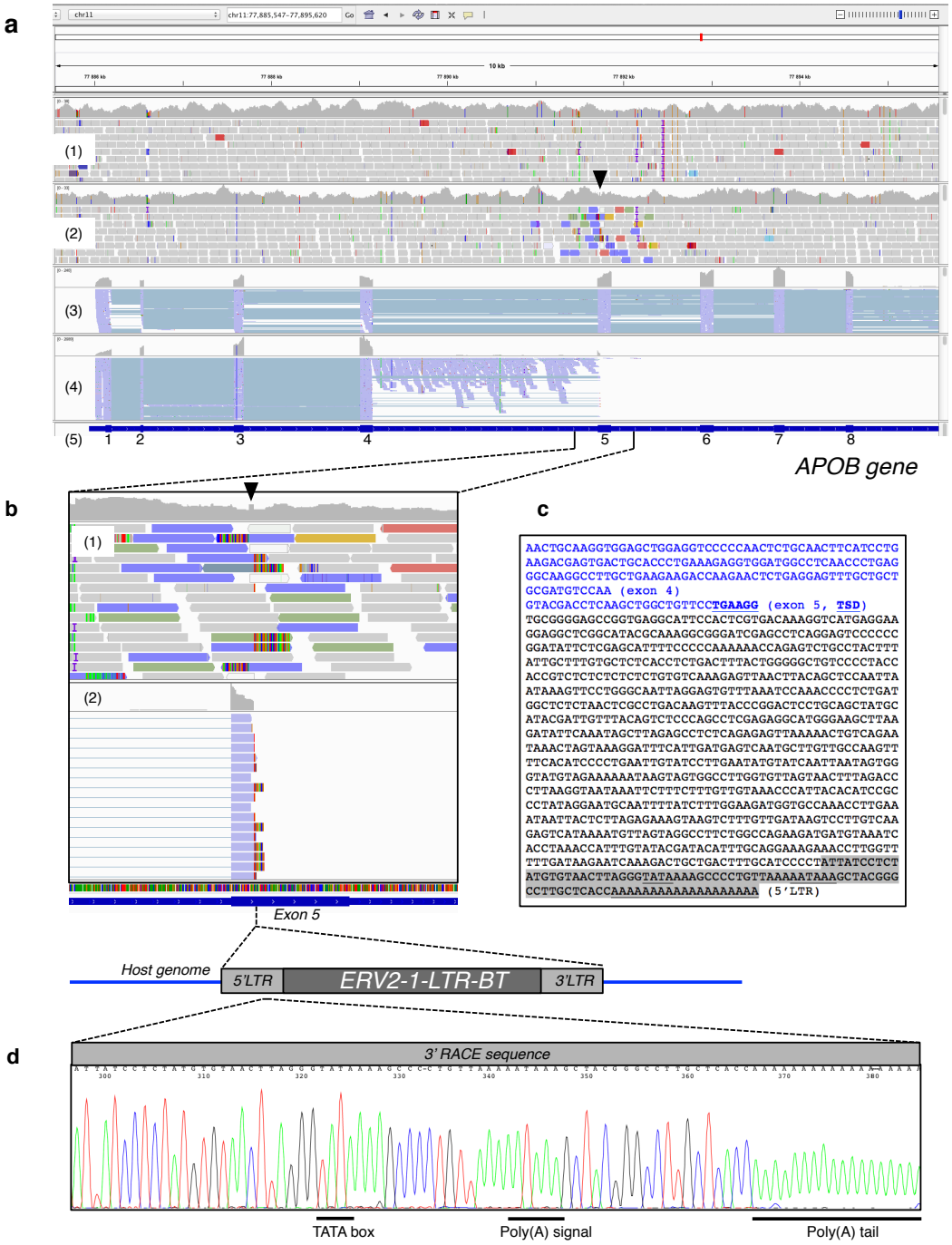
Supplementary Figure 16:
Schematic representation of the features exploited by *LocaTER* to identify polymorphic ERV element absent from the bovine reference sequence.

Supplementary Table 1:
Detailed domain annotation of a representative element of the C-clade (chr19: 50,466,809 bp).

Supplemental Method 1:
LocaTER for the detection of polymorphic and *de novo* repeat insertions.

Supplemental Data 1-15:
https://www.nature.com/articles/s41467-024-46434-1#Sec43

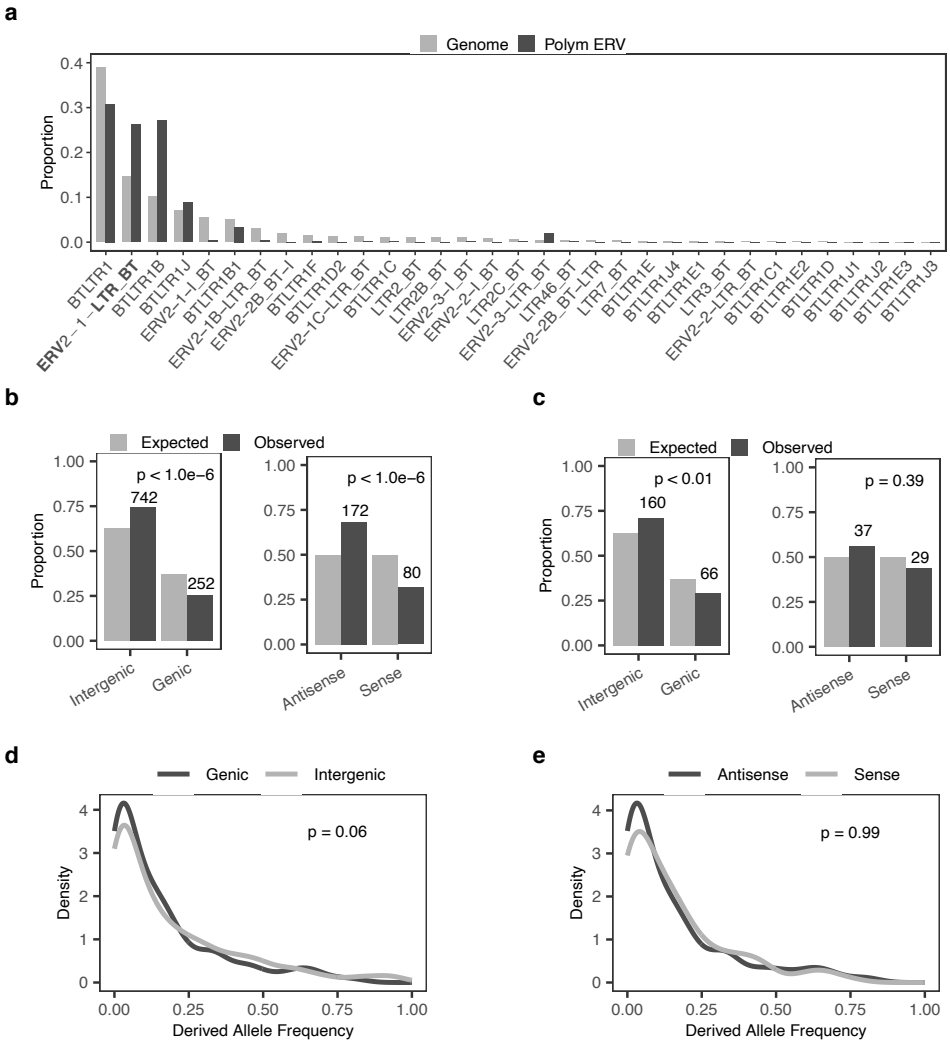
Source Data:
https://www.nature.com/articles/s41467-024-46434-1#Sec43



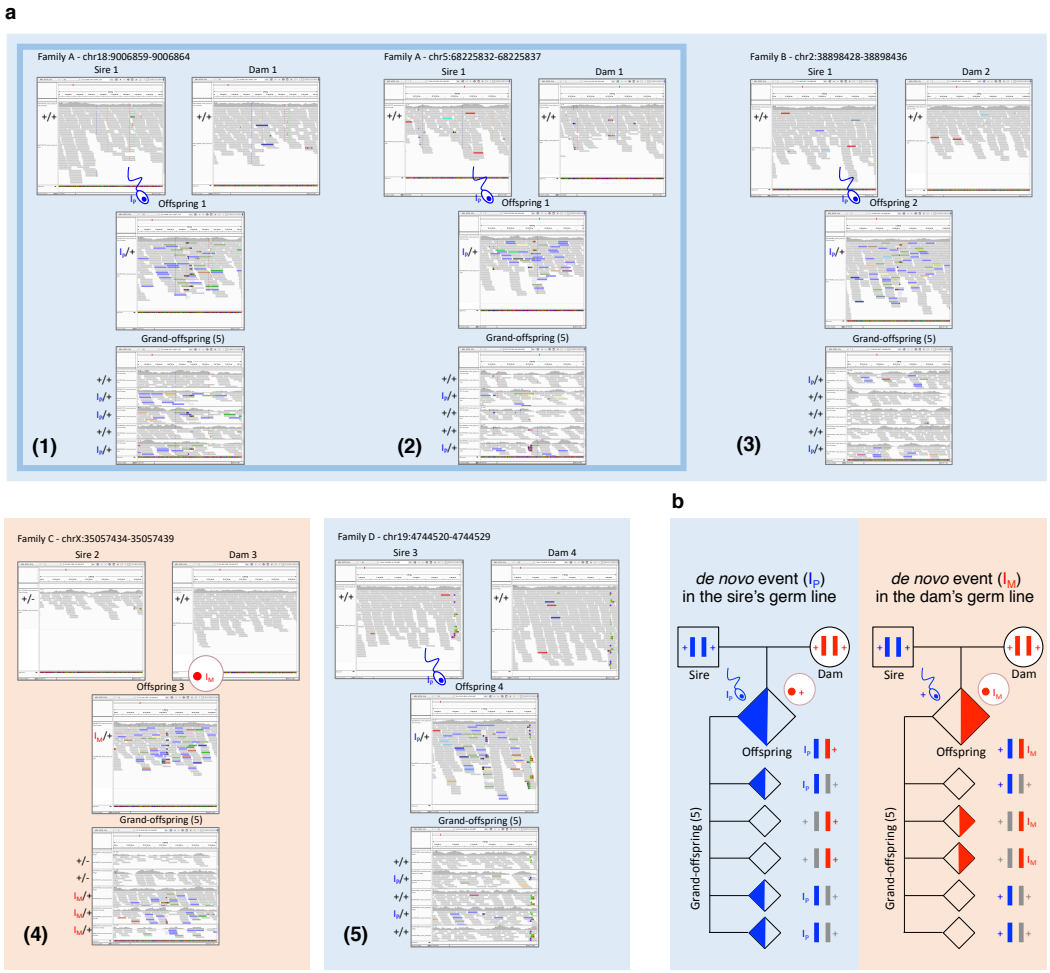
Supplementary Figure 1: Identification and functional validation of the causative mutation for CD.

(a) Integrative Genome Viewer (IGV) screen capture of a 10 kb genomic region (chr11:77,885,547-77,885,620bp) encompassing the ERVK insertion (black triangle) in the coding exon 5 of the *APOB* gene; with from top to bottom: the whole genome sequence (WGS) BAM files of a homozygous wild-type (1) and a heterozygous carrier (2) with reads from discordant read pairs surrounding the ERVK insertion showing a multicolor display; liver cDNA sequence reads (light blue) for a homozygous wild-type (3) and a homozygous mutant (4) highlighting the transcriptional shut-off in *APOB* exon 5; genomic organization of the *APOB* gene with exons and introns displayed as thick and thin dark blue rectangles respectively (5). **(b)** IGV screen capture of a zoomed 500 bp region surrounding the ERVK insertion site; with - from top to bottom: the WGS of a heterozygous carrier (1) with target site duplication (TSD) pinpointed by a black triangle; reads from the allele without the insertion are shown in grey;

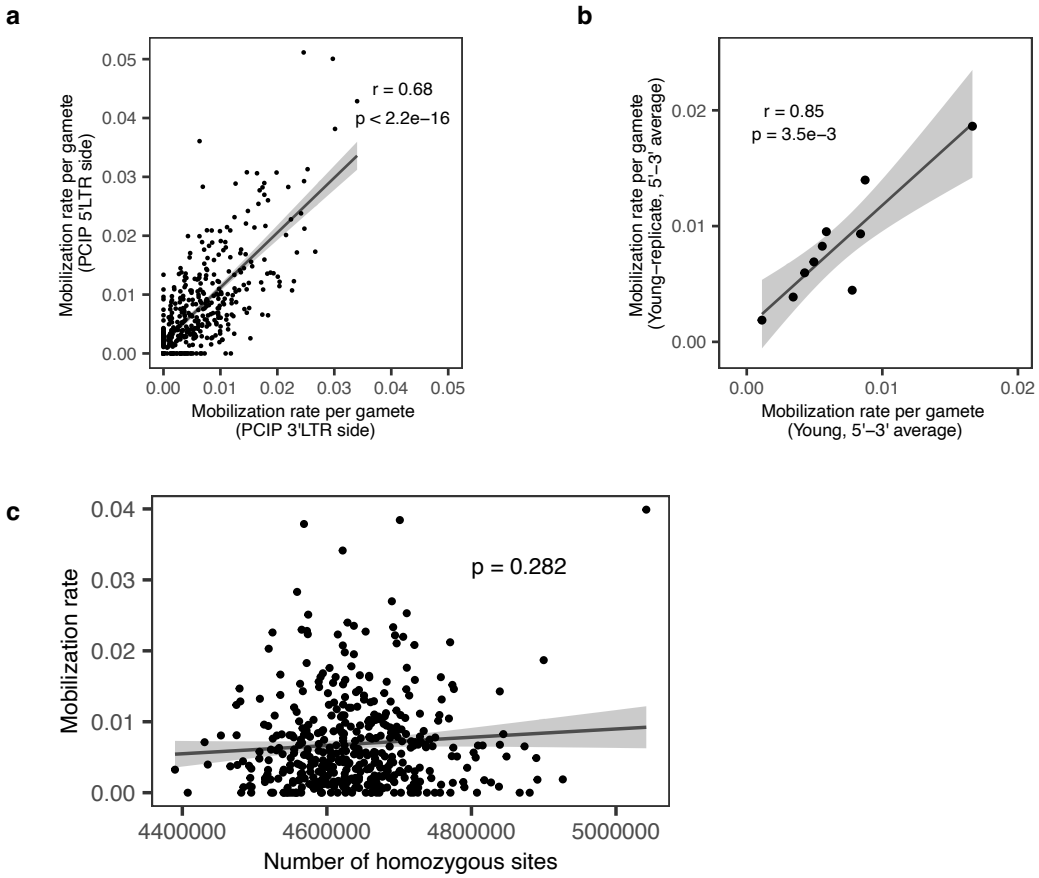
reads from discordant read pairs surrounding the breakpoints are showing a multicolor display and reads overlapping the breakpoints a characteristic soft-clipped feature (lack of homology with the reference sequence); cDNA sequence reads (light blue) for a homozygous mutant (2) highlighting the transcriptional shut-off in *APOB* exon 5 and reads overlapping the breakpoint showing a characteristic soft-clipped feature. **(c)** Liver cDNA sequence from a 3'RACE experiment performed on a homozygous mutant; with coding exon 4 and 5 in dark blue; the TSD is shown in blue, bold and underlined; from the ERVK insertion site, the transcribed part of its 5' LTR sequence is presented in black, ending with a poly(A) tail preceded by canonical TATA box and 'AATAAA' poly(A) signal (underlined); the sequence boxed in dark grey corresponds to the Sanger cDNA sequence trace displayed in panel d**.** **(d)** Sanger cDNA sequence trace from the above 3'RACE experiment, highlighting TATA box, poly(A) signal and poly(A) tail with black horizontal bars; a schematic representation of the full-length ERVK insertion in *APOB* exon 5 is depicted above the sequence trace.



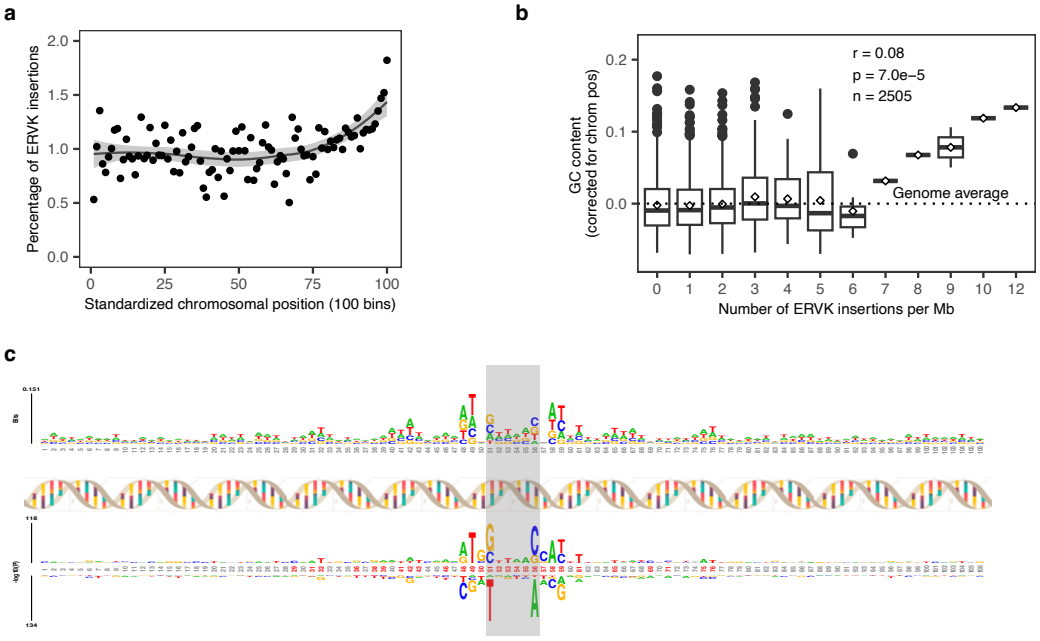
Supplementary Figure 2: Genomic features of polymorphic ERV elements.
(a) Proportional representation of ERVK sub-groups in genome space (light grey bars) and amongst polymorphic elements detected by *LocaTER* (dark grey bars). Repbase reports 33 subgroups of ERVK, of which the most abundant in the reference genome are BTLTR1 (38.9%), ERVK[2-1-LTR] (20.2%), BLTR1B (10.2%) and BLTR1J (7.1%). While BTLTR1 is underrepresented amongst polymorphic ERVK elements (30.1%), ERVK[2-1-LTR] (26.6%), BLTR1B (27.1%) and BLTR1J (8.8%) are respectively overrepresented 1.3, 2.6 and 1.2 times. Of note, the very rare ERVK[2-3-LTR] subgroup (0.5%) is 4 times overrepresented amongst polymorphic ERVK elements. This suggests that the latter four subgroups, especially, might still be active. **(b)** Genomic distribution of polymorphic ERVK elements (dark grey) compared to the corresponding genome space (light grey). Left: intergenic versus genic space. Right: antisense versus sense orientation for genic elements.
(c) Genomic distribution of polymorphic ERV elements other than the ERVK group (dark grey) compared to the corresponding genome space (light grey). Left: intergenic versus genic space. Right: antisense versus sense orientation for genic elements. The respective proportions did not differ significantly between ERVK and non-ERVK elements. See also main text. **(d)** Derived Allele Frequency (DAF) spectrum for genic (dark grey) and intergenic (light grey) ERV elements. DAF of genic insertions are slightly ($p = 0.06$) shifted towards lower values as expected under purifying selection. **(e)** Derived Allele Frequency (DAF) spectrum for genic antisense (dark grey) and genic sense (light grey) ERV elements. There is no significant evidence (Wilcoxon sum of rank test $p = 0.99$) for a shift of DAF of sense insertions towards lower values when compared to antisense, as expected if they would be subject to stronger purifying selection. Source data are provided as a Source Data file.



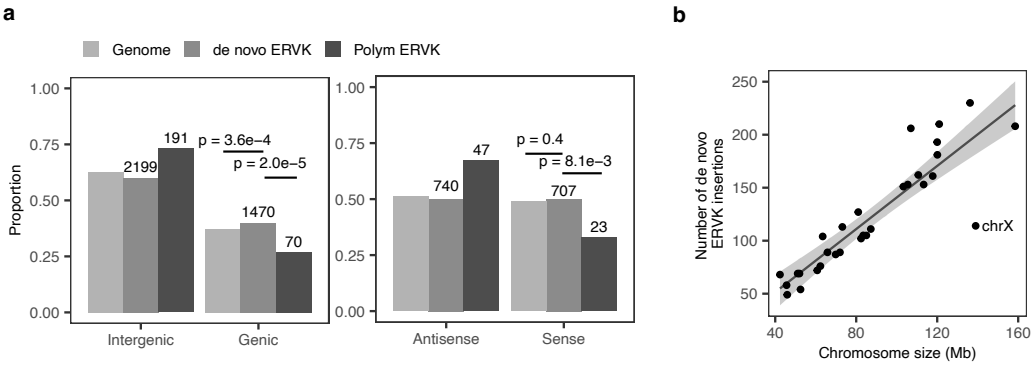
Supplementary Figure 3: Pedigree-based identification of five *de novo* insertions of ERVK[2-1-LTR] elements.
(a) IGV screen captures of the WGS corresponding to the respective genomic regions surrounding each of the five de novo insertions (1 to 5); with - from top to bottom - absence of the ERVK[2-1-LTR] element in the parents, presence in the offspring and transmission to grand-offspring in perfect linkage disequilibrium; a dark blue border highlights the family with two insertions transmitted by the same sperm cell; blue and pink backgrounds feature *de novo* events occurring in the paternal or maternal germline respectively; '+' for wild-type allele and 'I' for allele with insertion. Sires and Dams are numbered as in main Fig. 2d. **(b)** Schematic drawing of *de novo* events occurring either in the paternal (left) or maternal (right) germline.



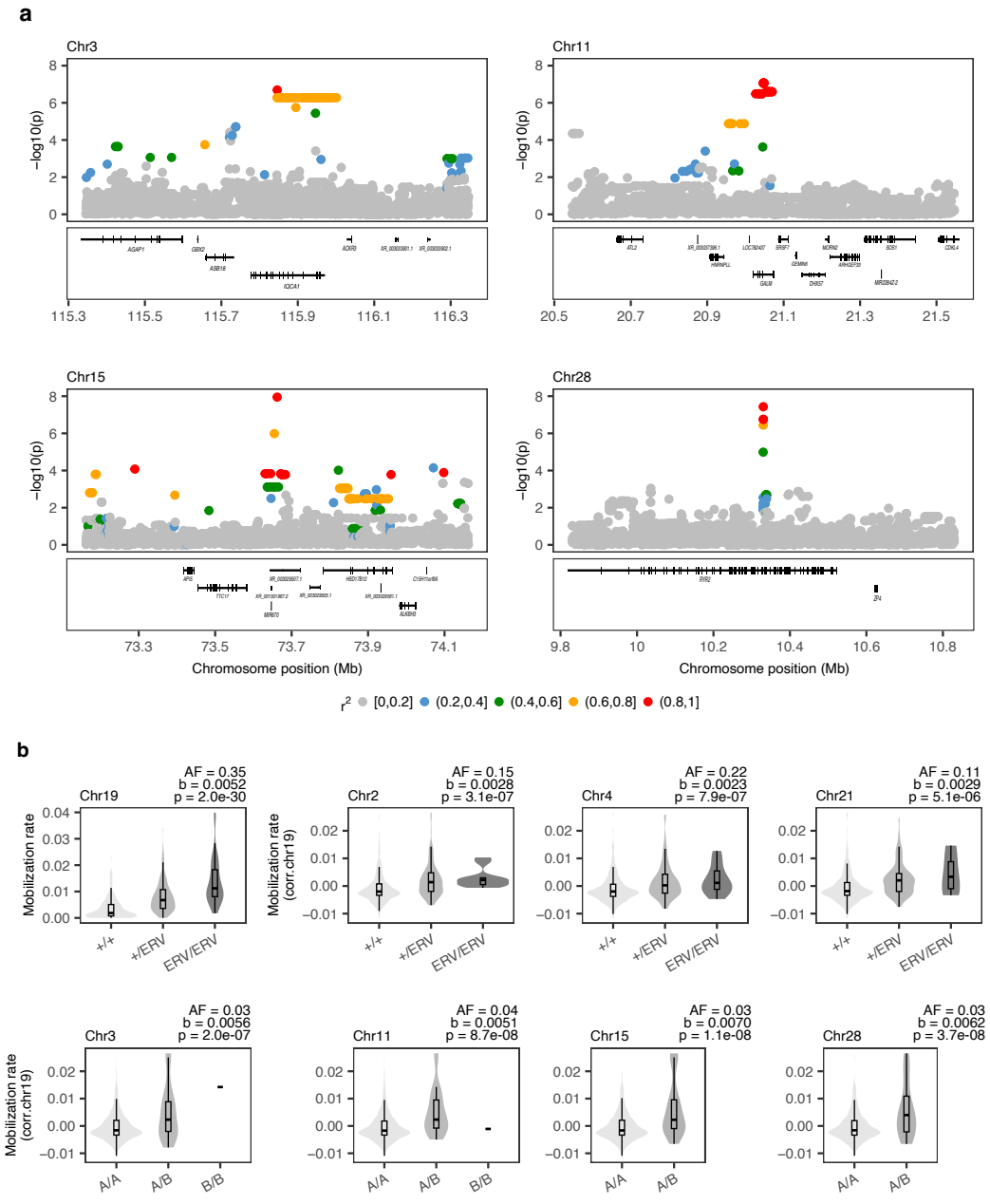
Supplementary Figure 4: Assessing the repeatability and robustness of PCIP and lack of evidence of an effect of a bull's inbreeding coefficient on the rate of ERVK[2-1-LTR] mobilization in its germline.
(a) Correlation between the 5' and 3' PCIP estimates of the mobilization rate of ERVK[2-1-LTR] in sperm DNA of 430 Belgian Blue sires. The correlation (r) of 0.68 indicates that PCIP is able to robustly measure differences in mobilization rate between samples. **(b)** Correlation between estimates of the mobilization rate of ERVK[2-1-LTR] elements in sperm samples for two biological replicates of ten young Belgian Blue bulls. Estimates correspond to the average of the 5'LTR and 3'LTR measures. Spearman's correlation (r) was 0.85. **(c)** The ERVK[2-1-LTR] mobilization rate measured in sperm of 430 Belgian Blue bulls (Y-axis) as a function of the number of homozygous sites out of 7,428,183 SNPs (X-axis) (MAF $\geq$ 0.1), used as proxy for their inbreeding coefficient. Source data are provided as a Source Data file.



Supplementary Figure 5: Genomic landscape of ERVK[2-1-LTR] *de novo* insertions. **(a)** Distribution of ERVK[2-1-LTR] insertions across standardized chromosome length. Chromosomes were split in 100 equally sized bins; the number ERVK[2-1-LTR] insertions counted for each bin, divided by the total number insertions on that chromosome, and multiplied by 100. Bin-specific percentages were then averaged across the 29 autosomes. A regression curve was fitted to the data using LOESS. ERVK[2-1-LTR] insertions appear to preferentially occur towards the telomere. Of note, all bovine autosomes are acrocentric. Thus, bin 1 is close to the centromere, and bin 100 close to the telomere. **(b)** Effect of local GC content on the rate of ERVK[2-1-LTR] insertion. The genome was subdivided in 2505 non-overlapping 1Mb bins. GC content was computed for each bin, and corrected (using a linear model) for distance from chromosome center. We then looked at the distribution of GC content for bins with 0, 1, 2, … 12 ERVK[2-1-LTR] insertions. There was a striking positive correlation between GC content and the number of insertions (r= 0.08, p=2.7x10^{-5}). A rectangle is drawn to represent the second and third quartiles with a horizontal line inside to indicate the median value. **(c)** 6-bp duplication (grey box) and 8-bp pseudo-palindromic motif at ERVK[2-1-LTR] insertion sites. Upper logo: nucleotide composition flanking 3669 ERVK[2-1-LTR] insertion sites. The height of the column corresponds to the departure from uniform nucleotide composition (measure of entropy). It shows strong signals for positions 1 (G) and 6 (C) of the duplication, as well as positions -3 (A), -2 (T), +2 (T) and +3 (A), as well as weaker up- and down-stream signals that appear to show a 10-bp (one helical turn) periodicity. Lower logo: log(1/p) value of the enrichment/depletion of the four nucleotides with respect to the average across a 100 bp window centered on the insertion. P-values were determined using z-scores computed with the mean and standard deviation of the abundance of the corresponding nucleotide in the window. The signal was strongest for positions 1 (G) and 6 (C) of the duplication, as well as positions -3 (A), -2 (T), -1 (G), +1 (C), +2 (T), +3 (A), and +5 (T). Although palindromic in appearance (5'ATGG….CCAT-3') when considering average base pair composition, there was no evidence for palindromicity when considering individual sequences as reported by Kirk et al. (2016). Logos were plotted using http://kplogo.wi.mit.edu/. Source data are provided as a Source Data file.



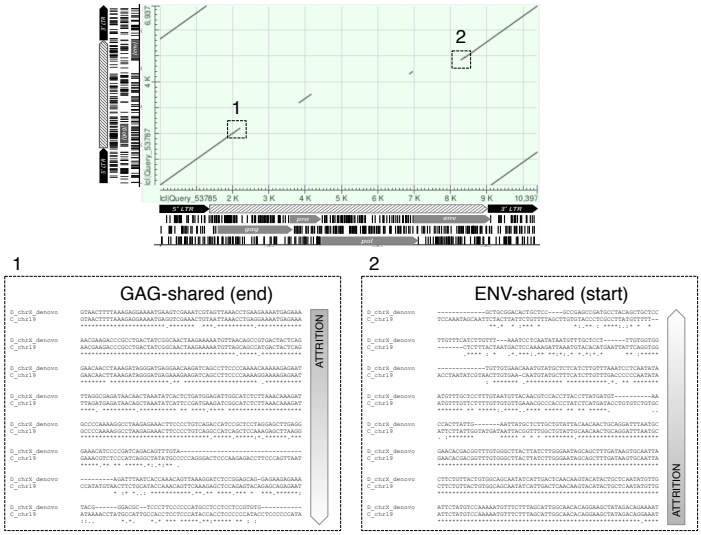
Supplementary Figure 6: Genomic features of ERVK[2-1-LTR] *de novo* insertions.
(a) Distribution of *de novo* ERVK[2-1-LTR] mobilization events with respect to genic vs intergenic and genic sense vs genic anti-sense space. We observed a modest yet significant (p = 0.0004) over-representation of genic *vs* intergenic insertions but not of anti-sense *vs* sense insertions. A number of factors may explain this observation including (i) more accessible genic than intergenic space, or (ii) overestimation of the genic space. The corresponding proportions are shown for polymorphic ERVK[2-1-LTR] elements (i.e. segregating in the Belgian Blue cattle population) for comparison. **(b)** Relationship between chromosome size and number of *de novo* ERVK[2-1-LTR] insertions. The chromosome that stands out the most is the X chromosome with nearly halve the number of insertions when compared to expectations. The most likely explanation of this observation is the hemizygosity of the X in male samples. Source data are provided as a Source Data file.



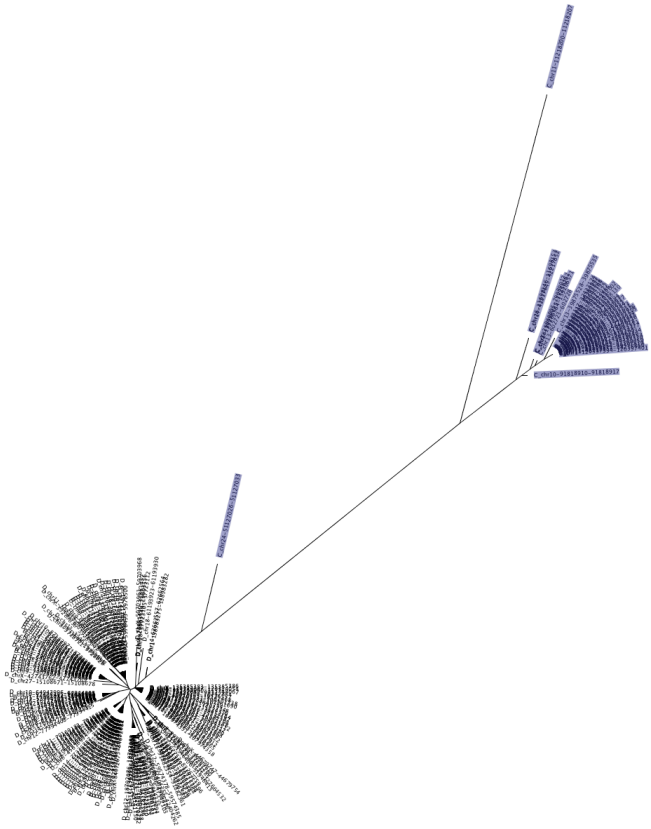
Supplementary Figure 7: Additional GWAS loci and genotypic effect for the eight significant loci.

(a) Zooms into the local association patterns of the four loci affecting ERVK[2-1-LTR] mobilization rate that do not encompass an ERV element. Variants are colored according to their LD (r^2) with the lead variant. Gene content of the corresponding window are shown below each panel. **(b)** Violin / box-plots showing the distribution of the ERVK[2-1-LTR] germline mobilization rates by marker genotype. For the four ERVK encompassing loci, bulls were sorted by ERV genotype (+/+, +/ERV, ERV/ERV). For the four other loci, bulls were sorted based on genotype at the lead SNP. AF: allelic frequency; b(eta): effect of the allele substitution on mobilization rate (slope of the linear regression on allelic dosage); p(value): statistical significance of the association. All non-ERVK lead SNPs were imputed variants with imputation accuracy $\geq$ 0.95. A rectangle is drawn to represent the second and third quartiles with a horizontal line inside to indicate the median value. Source data are provided as a Source Data file.

a

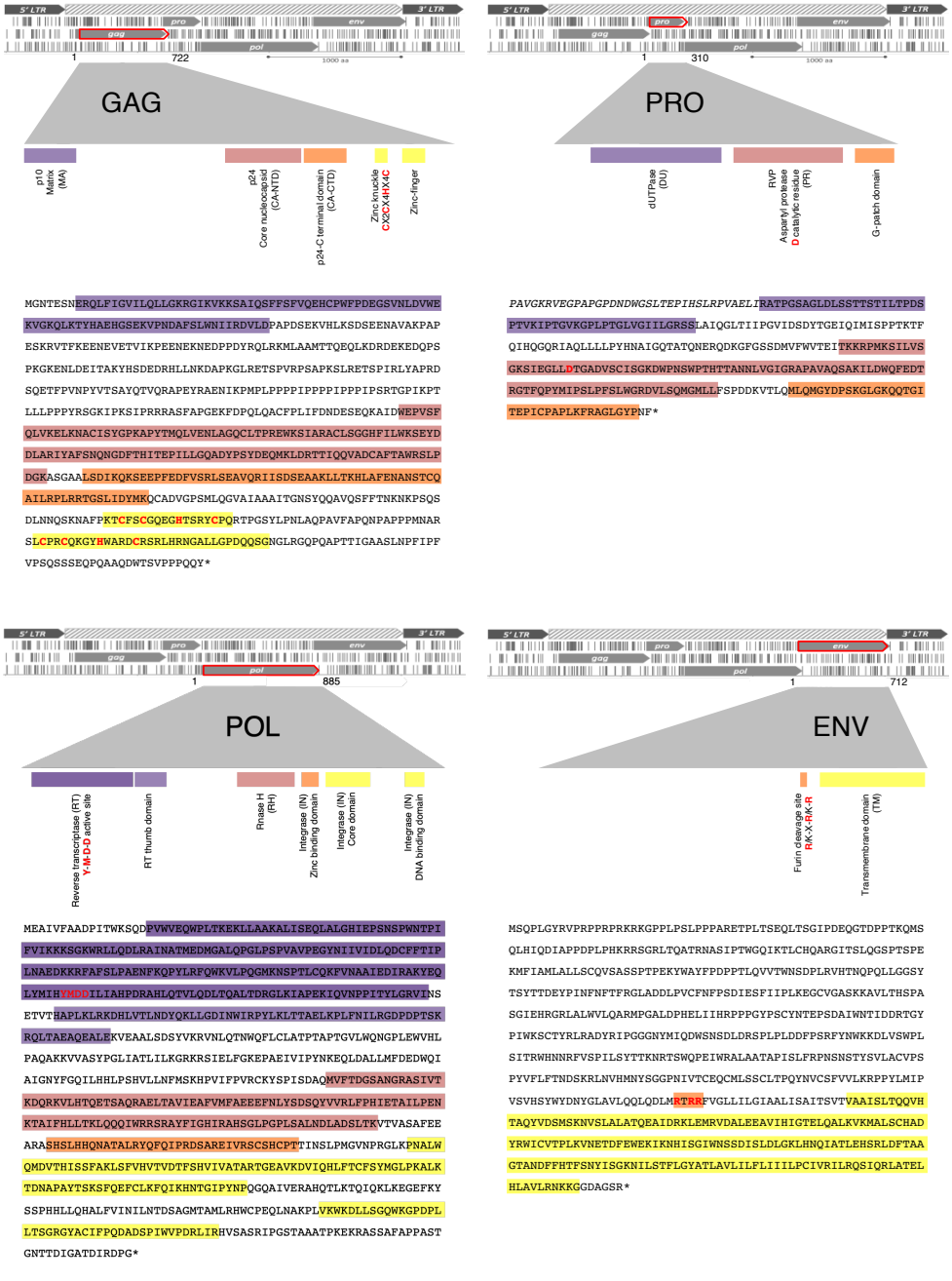


b

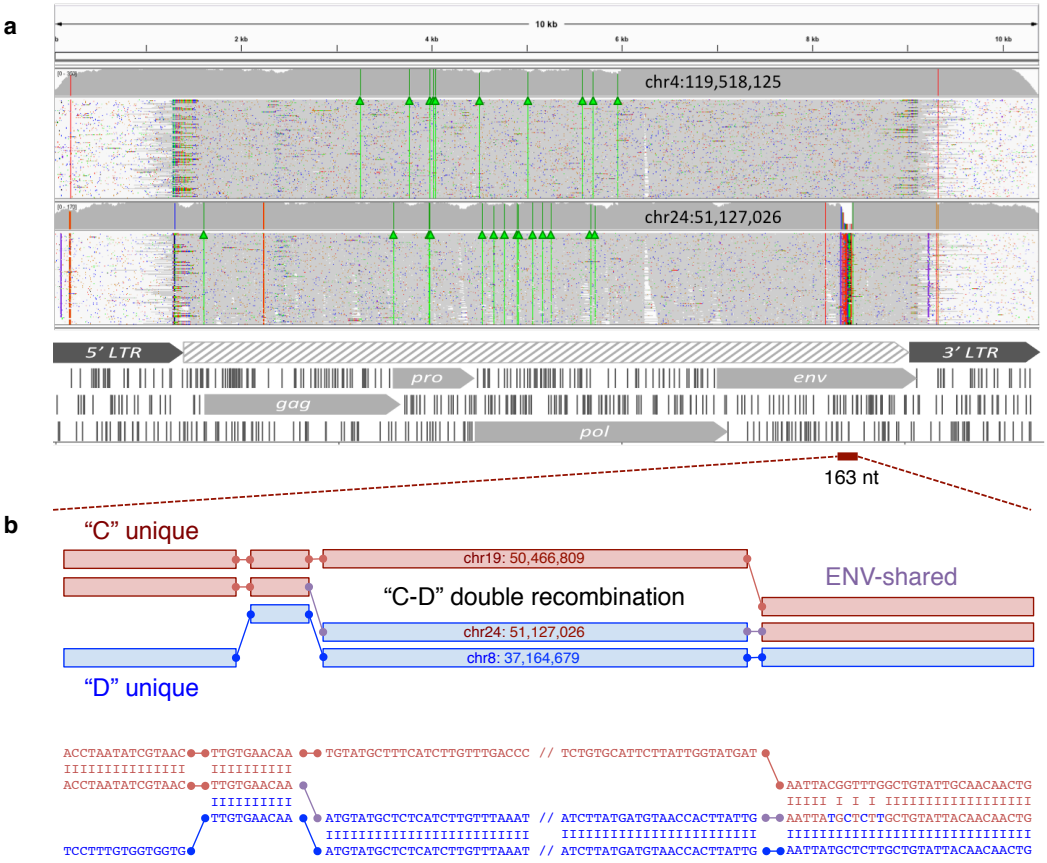


Supplementary Figure 8: Dot plot between sequences representing the C clade (X-axis) and D clade (Y-axis) and neighbor-joining tree obtained with the concatenated *GAG*-shared (55 base pairs) and *ENV*-shared (195 base pairs) segments.
(a) The diagonal lines correspond to nearly identical 5' and 3'LTR, shared *GAG* and *ENV* segments (extremities), as well fragments of *PRO* and *POL* (P and L in main Fig. 5). The limits of the *GAG*- (1) and *ENV*-shared (2) segments are boxed with dashed lines in the dot plot. The corresponding sequence alignments (1: *GAG*-shared; 2: *ENV*-shared) between a C and a D ERVK[2-1-LTR] element are displayed below the dot plot. Decreasing similarity (attrition) is

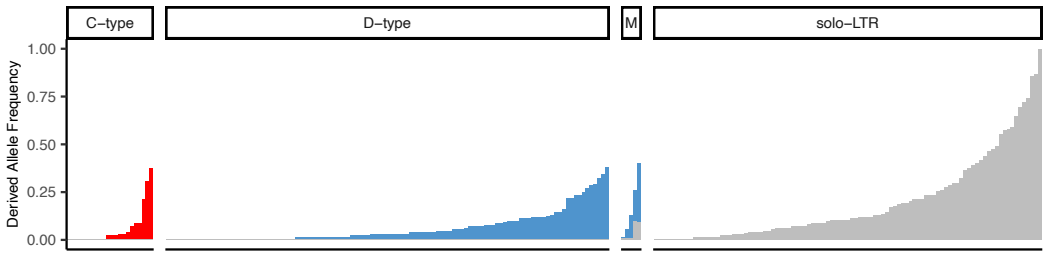
marked by arrows with gradients (from high to low similarity). The arrowheads mark the boundaries of the *GAG* and *ENV* shared sequences used to generate the tree shown in b. **(b)** Indeed, these segments do not show obvious "within segment" traces of recombination between C- and D-type elements so should provide the most correct relationship between C- and D-type (see hereafter for one "between segment" recombination). Insertion-deletions of more than one nucleotide were collapsed to single events. C-type elements are highlighted in blue. C- and D-type elements appear as clear distinct clades. One C-element (chr. 24) is closer to the D-clade. It corresponds to the only element that has a complete C-type *GAG*-shared sequence associated with a D-type *ENV*-shared sequence. The average pairwise difference between C- and D-type elements was 38.4 in 250 base-pairs. Assuming a *de novo* mutation rate of $1x10^{-8}$ base pairs per generation [30], this would correspond to ~15 million generations or ~50 million years, hence suggesting that the endogenization of the exogenous retrovirus precursor of the ERVK[2-1-LTR] element is a very ancient event. It is, however, possible that some of the differences between C- and D-type shared sequences were introduced upon creation of D-type elements, and/or that the mutation rate for ERV elements, influenced by the retro-transposition process, is higher than $1x10^{-8}$ base pairs per generation. Estimates of the divergence are also influenced by the somewhat arbitrary definition of the boundaries of the *GAG*-shared and *ENV*-shared segments.



Supplementary Figure 9: Protein sequences and domain annotation of a representative element of the C-clade (chr19: 50,466,809 bp).
Schematic representation of the four ORFs' domain annotation according to specific hits and super-families obtained by blasting each ORF against 'viruses' (taxid: 10239) non-redundant protein sequences. Number of amino-acids is given above each ORF (GAG, 722; PRO, 310; POL, 885; ENV, 712). The corresponding amino acid sequence is highlighted with the same color code below each domain scheme. Amino acids in italics at the N-terminal part of the PRO protein correspond to putative translation start sites. See also Supplementary Table 1 for a more detailed description of each protein domain.

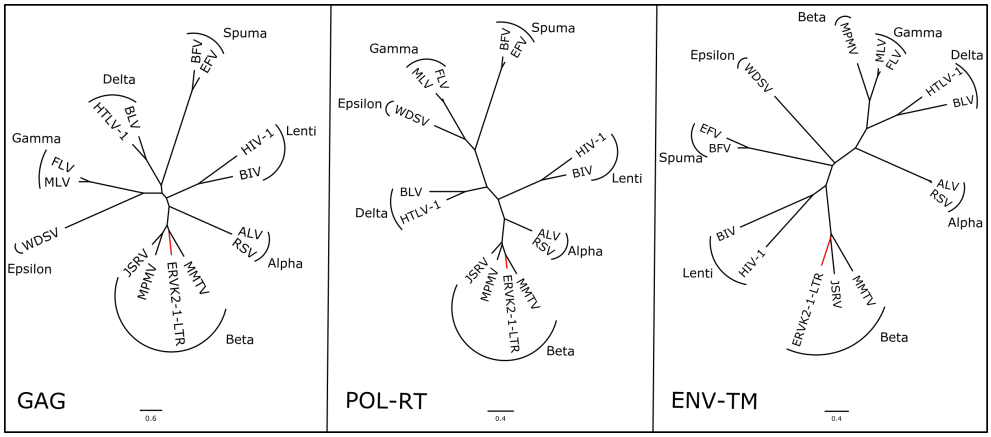


Supplementary Figure 10: Examples of APOBEC3 mutational signature, recombination events and attrition at the boundaries.
(a) Illustration of APOBEC3 induced G-to-A transitions on the plus-strand of the ERVK[2-1-LTR] genetic code: IGV screen capture of genomic sequence reads obtained from PCR amplicons of two clade ‘C’ ERVK[2-1-LTR] elements (chr4:119,518,125: top panel and chr24:51,127,026: bottom panel) mapped on the ~10Kb (5‘ to 3’) reference genomic sequence of a representative clade ‘C’ element (chr19: 50,466,809; top GWAS signal). Green vertical bars correspond to G-to-A transition and each one is highlighted by a green triangle (n = 11, chr4; n = 16, chr24). Other polymorphisms are shown with vertical color bars. Size and location of a double recombination are positioned by a red horizontal bar (163nt). **(b)** Schematic representation of a 163nt double recombination event between a clade ‘C’ (chr19: 50,466,809; top connected red rectangles) and a clade ‘D’ (chr8: 37,164, 679; bottom connected blue rectangles) ERVK[2-1-LTR] elements resulting in a recombinant ‘C’ element (chr24:51,127,026; middle connected red-blue-red rectangles). Their corresponding genomic sequence is given below with the same order and color code.



Supplementary Figure 11: Distribution of allelic frequencies of ERVK[2-1-LTR] segregating in Belgian Blue cattle.
ERVK[2-1-LTR] elements are sorted by "morph": C-type, D-type, solo-LTR. Within morph, the ERVK[2-1-LTR] elements are ranked by their frequency in the Belgian Blue population. At five loci, solo-LTR, D-type and wild-type (+, grey) alleles coexist at the shown frequencies (M(ultiple) morph). Source data are provided as a Source Data file.

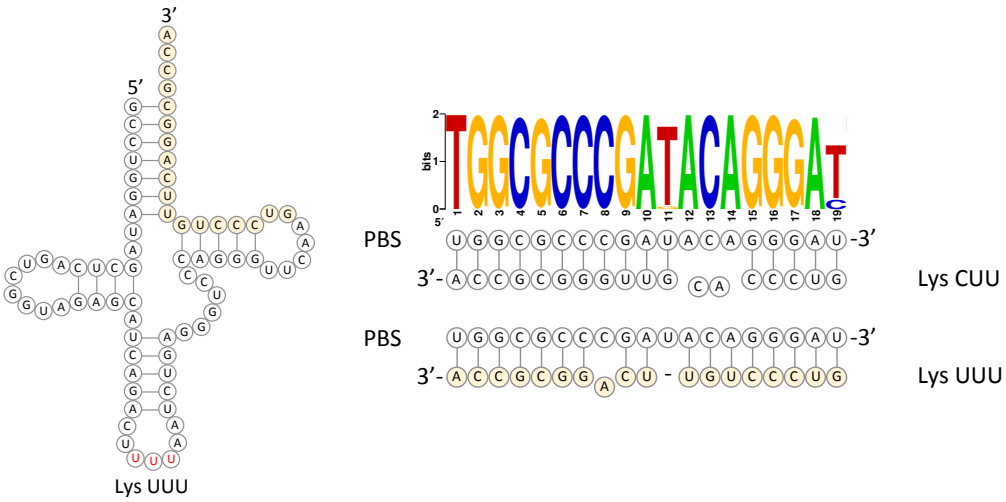
a



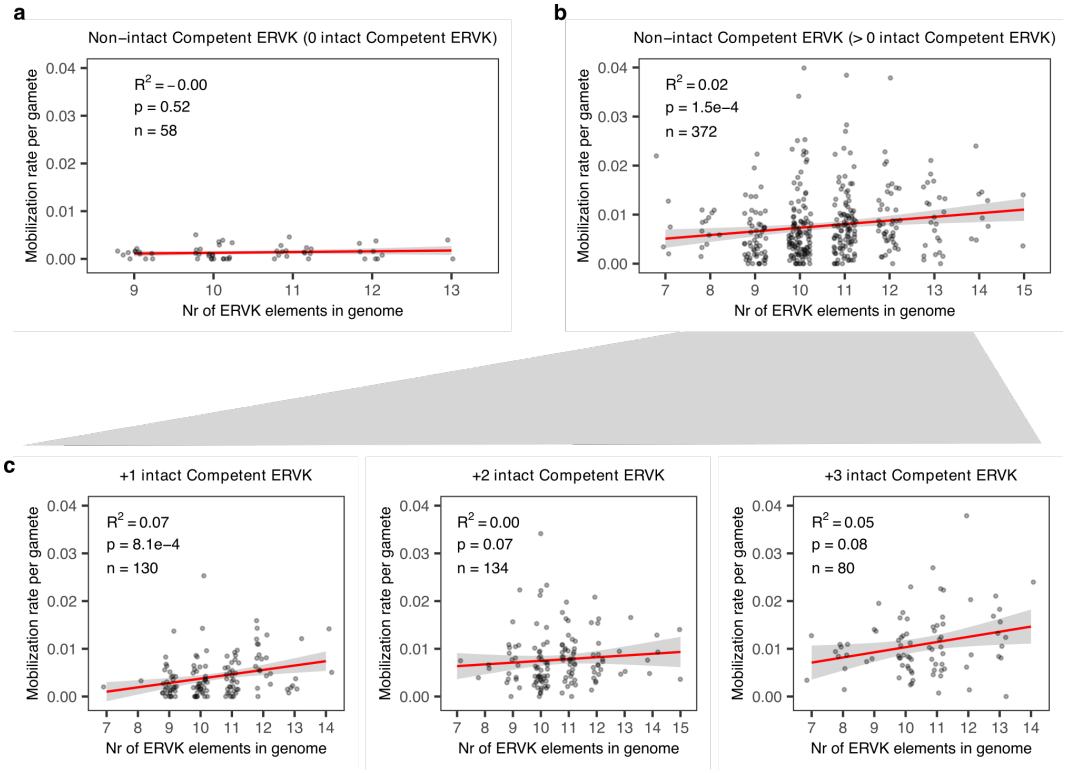
b

MMTV	M	G	V	S	G	S	K	G	Q	K	L	F	V	S	V	L	Q	R	L	L	-	-	0.9753
MPMV	M	G	Q	E	L	S	Q	-	H	E	R	Y	V	E	Q	L	K	Q	A	L	K	-	0.9839
JSRV	M	G	Q	T	H	S	-	-	R	Q	L	F	V	H	M	L	S	V	M	L	K	H	0.9781
ERVK2-1-LTR	M	G	N	T	E	S	N	E	R	Q	L	F	I	G	V	I	L	Q	L	L	-	-	0.9838

c

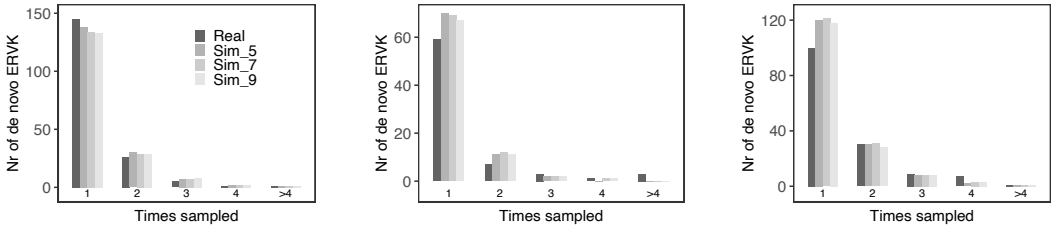


Supplementary Figure 12: Phylogenetic relationship with exogenous retroviruses and functional sequence features of ERVK[2-1-LTR].
(a) Unrooted phylogenetics trees of GAG, POL (reverse transcriptase domain) and ENV (transmembrane domain) sequences of ERVK[2-1-LTR] and representative Alpha, Beta, Gamma, Delta, Epsilon, Lenti and Spuma exogenous retroviruses. ALV: Avian leukosis virus; RSV: Rous sarcoma virus; MMTV: Mouse mammary tumor virus; MPMV: Mason-Pfizer monkey virus; JSRV: Jaagsiekte sheep retrovirus; HTLV-1: Human T-lymphotropic virus 1; BLV: Bovine leukemia virus; WDSV: Walleye dermal sarcoma virus; MLV: Moloney murine leukemia virus; FLV: Feline leukemia virus; HIV-1: Human immunodeficiency virus 1;BIV: Bovine immunodeficiency virus; BFV: Bovine foamy virus; EFV: Equine foamy virus. **(b)** Sequence of the N-terminal domain of the representative beta retroviruses and ERVK[2-1-LTR] GAG protein disclosing a consensus sequence required for myristoylation ([M]GXXXS/T) with the first M corresponding to the GAG initiation codon and a domain rich in positively charged basic residues (basic domain in red) interacting with the membrane phospholipids. The score for myristoylation was predicted by Myristoylator (https://web.expasy.org/cgi-bin/myristoylator). **(c)** Consensus primer binding site (PBS) shared by C and D-type elements aligned to the 3' end of the bovine CTT and TTT lysine tRNA genes.

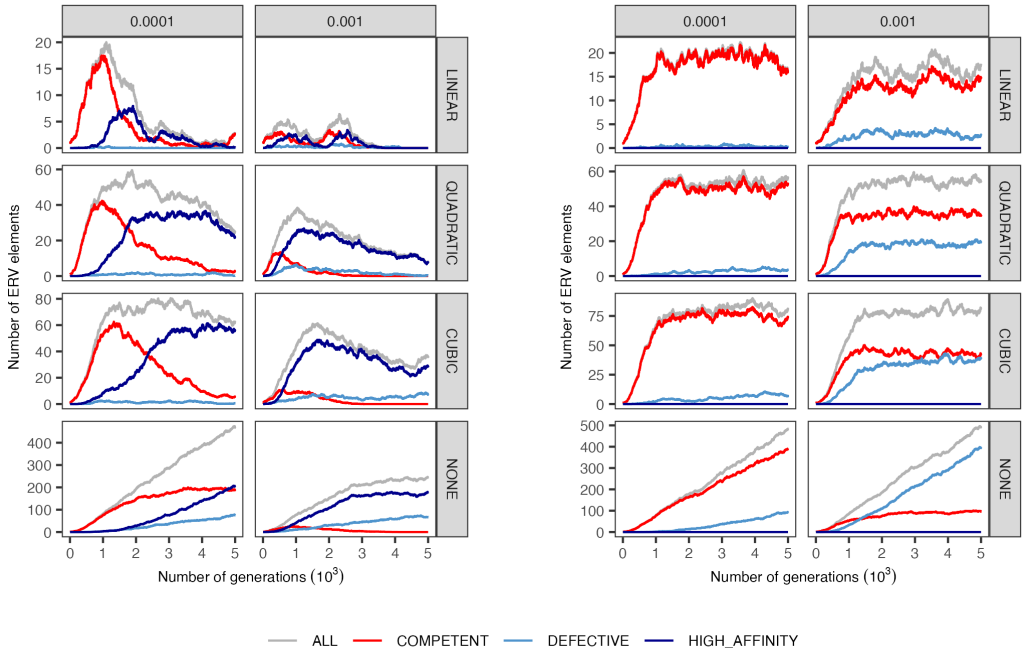


Supplementary Figure 13: Epistatic interaction between ERVK[2-1-LTR] elements with coding variants in the *GAG*, *PRO*, *POL* or *ENV* gene, and ERVK[2-1-LTR] elements without.

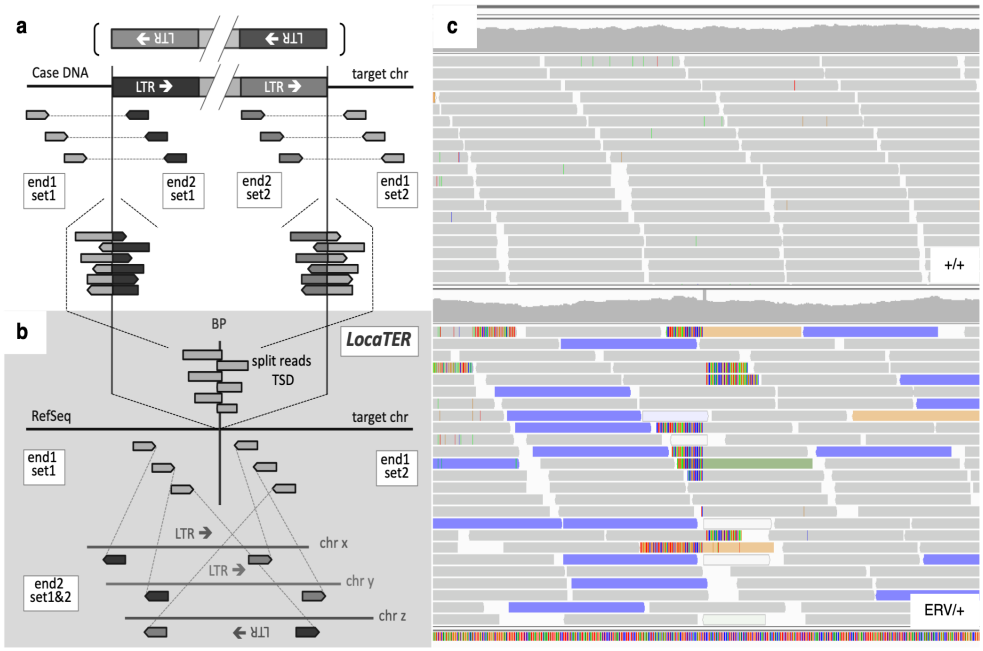
(a) In the absence of any intact Competent ERVK[2-1-LTR] element, the dosage (0, 1 or 2) of ERVK[2-1-LTR] elements with coding variants has no effect (p = 0.693) on the *de novo* mobilization rate. **(b)** With one or more intact Competent ERVK[2-1-LTR] elements in the genome, the dosage (0, 1 or 2) of ERVK[2-1-LTR] elements with coding variants has a significant effect (p = 7.0e-04) on the *de novo* mobilization rate. The probability to have a regression slope as low as the one observed without intact Competent ERVK[2-1-LTR] elements, assuming that the effect of dosage is in fact the same as in the presence of intact Competent ERVK[2-1-LTR] elements, i.e. the probability that it is a false negative was shown to be 0.11. **(c)** Effects of the number of non-intact ERVK[2-1-LTR] elements in the presence of 1, 2 or 3 intact ERVK[2-1-LTR] elements are shown separately. Source data are provided as a Source Data file.



Supplementary Figure 14: Comparison between the real (dark grey) and simulated (lighter greys) frequency distribution of the resampling rate of ERVK[2-1-LTR] *de novo* mobilization events in BB bulls BE157971524, BE187351114, BE63811423.
The simulations assumed that mobilization occurred in a developmental window centered on cell generation 14 of spermatogenesis of 21 (yielding 1,048,576 spermatogonial stem cells). Windows spanning 5, 7 or 9 cell generations, with mobilization rates distributed as shown in Fig. 8b, matched the real data (3 bulls combined) equally well. The simulations further matched the real data with regards to bull-specific *de novo* insertion frequency (average number of *de novo* insertions per sperm cell) and number of explored haploid genomes (see Methods). Source data are provided as a Source Data file.



Supplementary Figure 15: Unselected examples of the evolution of the number of ERV elements (per genome) in a panmictic population of 1,000 animals over a course of 5,000 generations.
In the left panel, mutation may cause competent elements to become defective and either increase or decrease the affinity for the mobilization machinery (provided in *trans* by competent elements). In the right panel, mutation can only cause competent elements to become defective and lose their affinity for the mobilization machinery. The mutation rate was set at 0.0001 and 0.001 per element per generation. Purifying selection was set from severe ("linear") to none ("none") with two intermediate intensities ("quadratic" and "cubic"). Source data are provided as a Source Data file.



Supplementary Figure 16: Schematic representation of the features exploited by *LocaTER* to identify polymorphic ERV element absent from the bovine reference sequence.
(a) Representation of the features of an ERV insertion when the WGS paired reads are (hypothetically) mapped on the case DNA sequence. **(b)** Representation of the features of an ERV when the WGS paired reads are mapped on the reference sequence. **(c)** IGV screen capture of homozygous wide type (upper) and one heterozygous ERV insertion (bottom) detected by *LocaTER*. *LocaTER* exploits three distinctive features of ERV element insertion: (i) paired sets of discordant paired-ends (with respect to ARS-UCD1.2 genome assembly) mapping, respectively, to the sense strand upstream the insertion site (end 1, set 1) and one end of the ERV LTR (end 2, set 1), and to the antisense strand downstream of the insertion of the insertion site (end 2, set 2) and the other end of the LTR (end 1, set 2), (ii) the presence of split sense and antisense reads consistently bridging the insertion site and the ERV LTR, (iii) the presence of the signature target site duplication (TSD). Breakpoint (BP) at the insertion site is represented by a vertical bar.

Supplementary Table 1: Detailed domain annotation of a representative element of the C-clade (chr19: 50,466,809 bp).
Specific hits and super-families were obtained by blasting each ORF (GAG, PRO, POL, ENV) against 'viruses' (taxid: 10239) non-redundant protein sequences.

GAG_722	Name	Accession	Description	Interval	E-value
	Gag_p24	pfam00607	gag gene protein p24 (core nucleocapsid protein, CA) ;p24 forms inner protein layer of the nucleocapsid	348-539	5.06E-50
	Gag_p10	pfam02337	Retroviral GAG p10 protein; This family consists of various retroviral GAG (core) polyproteins and encompasses the p10 region producing the p10 protein upon proteolytic cleavage of GAG by retroviral protease. The p10 or matrix protein (MA) is associated with the virus envelope glycoproteins in most mammalian retroviruses and may be involved in virus particle assembly, transport and budding.	10-92	3.87E-37
	zf-CCHC_5	pfam14787	GAG-polyprotein viral zinc-finger.	640-672	5.00E-09
	PTZ00368	PTZ00368	universal minicircle sequence binding protein (UMSBP).	592-654	4.63E-05
	ZnF_C2HC	smart00343	zinc finger;	593-609	4.12E-03

PRO_310	Name	Accession	Description	Interval	E-value
	RVP	pfam00077	Retroviral aspartyl protease; Single domain aspartyl proteases (PR) from retroviruses, retrotransposons, and badnaviruses (plant dsDNA viruses). These proteases are generally part of a larger polyprotein; usually pol, more rarely gag.	163-261	3.91E-27
	HIV_retropepsin_like	cd05482	Retropepsins, pepsin-like aspartate proteases (PR); This is a subfamily of retropepsins. The family includes pepsin-like aspartate proteases from retroviruses, retrotransposons and retroelements. While fungal and mammalian pepsins are bilobal proteins with structurally related N- and C-termini, retropepsins are half as long as their fungal and mammalian counterparts. The monomers are structurally related to one lobe of the pepsin molecule and retropepsins function as homodimers. The active site aspartate (D) occurs within a motif (Asp-Thr/Ser-Gly), as it does in pepsin. Retroviral aspartyl protease is synthesized as part of the POL polyprotein that contains an aspartyl protease, a reverse transcriptase, RNase H, and an integrase. The POL polyprotein undergoes specific enzymatic cleavage to yield the mature proteins. In aspartate peptidases, Asp residues are ligands of an activated water molecule in all examples where catalytic residues have been identified.	170-255	3.15E-26
	G-patch	pfam01585	G-patch domain; This domain is found in a number of RNA binding proteins, and is also found in proteins that contain RNA binding domains. This suggests that this domain may have an RNA binding function. This domain has seven highly conserved glycines.	272-308	1.34E-06
	dUTPase	pfam00692	dUTPase (DU); dUTPase hydrolyzes dUTP to dUMP and pyrophosphate.	36-153	2.28E-35

POL_886	Name	Accession	Description	Interval	E-value
	RT_Rtv	cd01645	RT_Rtv: Reverse transcriptases (RTs) from retroviruses (Rtvs). RTs catalyze the conversion of single-stranded RNA into double-stranded viral DNA for integration into host chromosomes. Proteins in this subfamily contain long terminal repeats (LTRs) and are multifunctional enzymes with RNA-directed DNA polymerase, DNA directed DNA polymerase, and ribonuclease hybrid (RNase H) activities. The viral RNA genome enters the cytoplasm as part of a nucleoprotein complex, and the process of reverse transcription generates in the cytoplasm forming a linear DNA duplex via an intricate series of steps.	18-230	8.70E-119
	RVT_thumb	pfam06817	Reverse transcriptase thumb domain; This domain is known as the thumb domain. It is composed of a four helix bundle.	237-302	5.43E-36
	RNase_HI_RT_Bel	cd09273	Bel/Pao family of RNase HI in long-term repeat retroelements; Ribonuclease H (RNase H) enzymes are divided into two major families, Type 1 and Type 2, based on amino acid sequence similarities and biochemical properties. RNase H is an endonuclease that cleaves the RNA strand of an RNA/DNA hybrid in a sequence non-specific manner in the presence of divalent cations. RNase H is widely present in various organisms, including bacteria, archaea and eukaryote. RNase HI has also been observed as adjunct domains to the reverse transcriptase gene in retroviruses, in long-term repeat (LTR)-bearing retrotransposons and non-LTR retrotransposons. RNase HI in LTR retrotransposons perform degradation of the original RNA template, generation of a polypurine tract (the primer for plus-strand DNA synthesis), and final removal of RNA primers from newly synthesized minus and plus strands. The catalytic residues for RNase H enzymatic activity, three aspartatic acids and one glutamic acid residue (DEDD), are unvaried across all RNase H domains. Phylogenetic patterns of RNase HI of LTR retroelements is classified into five major families, Ty3/Gypsy, Ty1/Copia, Bel/Pao, DIRS1 and the vertebrate retroviruses. Bel/Pao family has been described only in metazoan genomes. RNase H inhibitors have been explored as an anti-HIV drug target because RNase H inactivation inhibits reverse transcription.	449-571	1.63E-26
	rve	pfam00665	Integrase core domain (IN);Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of three domains. The amino-terminal domain is a zinc binding domain pfam02022. This domain is the central catalytic domain. The carboxyl terminal domain that is a non-specific DNA binding domain pfam00552. The catalytic domain acts as an endonuclease when two nucleotides are removed from the 3' ends of the blunt-ended viral DNA made by reverse transcription. This domain also catalyzes the DNA strand transfer reaction of the 3' ends of the viral DNA to the 5' ends of the integration site	633-727	7.17E-24
	IN_DBD_C	pfam00552	Integrase DNA binding domain (IN); Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of three domains. The amino-terminal domain is a zinc binding domain. The central domain is the catalytic domain pfam00665. This domain is the carboxyl terminal domain that is a non-specific DNA binding domain.	796-839	5.02E-18

ENV_713	Name	Accession	Description	Interval	E-value
	GP41	pfam00517	Retroviral envelope protein; This family includes envelope protein from a variety of retroviruses. It includes the GP41 subunit of the envelope protein complex from human and simian immunodeficiency viruses (HIV and SIV) which mediate membrane fusion during viral entry. The family also includes bovine immunodeficiency virus, feline immunodeficiency virus and Equine infectious anaemia (EIAV). The family also includes the Gp36 protein from mouse mammary tumor virus (MMTV) and human endogenous retroviruses (HERVs).	512-706	1.34E-35

Supplemental Method 1: *LocaTER* for the detection of polymorphic and *de novo* repeat insertions.

The *LocaTER* pipeline is designed to identify candidate polymorphic ERV insertions from next generation WGS data (Supplementary Fig. 16). It requires a database of locations for annotated ERV 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 analyzing each read to identify the signatures of an ERV insertion. For every read in the genome it checks to determine if it is cleanly aligned (mapping quality 20-60), if the read is properly paired (SAM flag is Properly Paired), is not aligned to a known ERV and that the mate is aligned to a known ERV of a specific class. When a read is detected that matches these criteria a 1.5kb window (3x the library 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 ERV they are aligned to, and the ERV family. 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 analyzed 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 ERV (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 target site 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 reanalyzed for the new window collecting the data as described above. In addition, the number of reads that completely bridge the insertion site are determined and used to estimate if the site is heterozygous (1 or more read completely bridging the 5' and 3' insertion sites) or homozygous (no reads bridging 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 ERV class with the most mates aligned to it is selected as the likely class of the new ERV insertion. The pedigree for the population is then analyzed 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 their likely genotype.

Experimental Section

Study 2

An active *Helitron* transposon family in wheat and its role in genome evolution

Experimental Section

Study 2

Contribution to

An active *Helitron* transposon family in wheat and its role in genome evolution

In revision

Haoran Peng[1,6], Lijing Tang[2], Nataliya Hrunyk[1], Roman Kellenberger[1], Dario Fossati[3], Hélène Rimbert[4], Pierre Sourdille[4], Alison B. Hickman[5], Fred Dyda[5], Carole Charlier[2], Frédéric Choulet[4] and Etienne Bucher[1*]

[1]Crop Genome Dynamics Group, Agroscope, 1260 Nyon, Switzerland; [2]Unit of Animal Genomics, GIGA & Faculty of Veterinary Medicine, University of Liège, Liège, Belgium; [3]Field Crop Breeding and Genetic Resources, Agroscope, 1260 Nyon, Switzerland; [4]INRAE, GDEC, Université Clermont Auvergne, 63000 Clermont-Ferrand, France; [5]Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA; 6Present address: Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam, Germany

*Correspondence: etienne.bucher@agroscope.admin.ch

Abstract

Transposons play a pivotal role in genome evolution and phenotypic variation in numerous eukaryotic species, including plants and animals[1]. *Helitrons*, a recently identified category of transposons, remain poorly understood in terms of epigenetic regulation and real-time mobilization[2,3]. Here, we introduce a *Helitron* family named *Xuan–Feng* (Chinese 'whirlwind'), which can be mobilized in wheat. Under epigenetic impairment, heat stress induces the formation of extrachromosomal circular DNA (eccDNA) and *de novo* insertions of the non-autonomous *Xuan* element. Combining whole-genome sequencing with genetics allowed the identification of the functional autonomous *Feng* element required for *Xuan* mobilization in the wheat genome. Notably, this *Feng* element is stress-responsive and encodes an intact *Helitron* transposase which catalyzes eccDNA formation and integration. Our evolutionary analysis across 17 *Triticeae* (wheat and wheat relatives) genomes revealed a constant copy number of autonomous *Feng* members, supporting the notion of an active maintenance of a steady state of transposable elements insertions[4]. These findings represent a step forward in studying active *Helitrons* in virtually any organism which will contribute to a better understanding of their role in genome evolution.

Main

The genomes of most plant species predominantly consist of transposable elements (TEs)[5]. Indeed, hexaploid bread wheat (*Triticum aestivum*), a prominent crop worldwide, exhibits a large genome size (16 Gb), with more than 85% originating from TEs and other repetitive sequences[6]. The significance of TEs in the functional and evolutionary dynamics of genomes is now widely acknowledged. They contribute to the regulation of gene expression, stress responses, and adaptation and play key roles during crop domestication[7-10]. *Helitrons*, a distinct category of TEs, eluded discovery until the early 2000s when bioinformatics methodologies enabled their identification over 50 years after the initial discovery of TEs[11]. *Helitrons* have been suggested to employ a rolling-circle transposition mode reminiscent of plasmid replication and single-stranded DNA viruses[12]. This transposition mode is distinct, as it does not generate target site duplications, a characteristic footprint of many TEs[13].

A canonical *Helitron* has conserved termini: a 5′ TC, a 3′ CTRR sequence (where R stands for A or G), and a 16–20 bp hairpin structure located 10–15 bp from the 3′ end[2,14]. Due to the high diversity of their internal sequences, specific *Helitron* family-defining sequences of 30 bases are present in both termini and are referred to as the left terminal sequence (LTS) and right terminal sequence (RTS) [15,16]. Whereas their relics are abundant across the eukaryotic kingdom and indicative of historical activity[17], direct evidence of real-time *Helitron* mobility is scarce. Indeed, the observation of controllable active TE mobility is rarely observed in plants, largely due to the concerted action of multiple epigenetic mechanisms that silence them[18,19]. Notably, DNA methylation and small RNA pathways robustly suppress their mobility[20,21]. Given that *Helitrons* employ extrachromosomal circular DNA (eccDNA) intermediates during their transposition process[22,23], these eccDNAs may serve as valuable diagnostic markers for assessing *Helitrons* activity. However, only directly capturing *de novo* insertions can provide the ultimate proof of their transpositional activity.

To promote TE mobilization using a similar approach as previously reported[24], we treated wheat seedlings with a combination of two chemicals that interfere with TE silencing: the RNA polymerase I, II, and III inhibitor juglone (J)[25] and the DNA methylation inhibitor zebularine (Z)[26]. Since numerous TEs are stress-responsive[27], we combined the drug treatments (JZ) with heat stress (H). In the wheat cultivar Arina, the JZH-treated seedlings exhibited significantly reduced vigor, characterized by shorter roots and delayed leaf development (Fig. 1a). To identify activated *Helitrons*, eccDNAs from these plants were extracted, amplified by rolling circle amplification, and subjected to long-read sequencing[28,29]. We observed that JZH-treated samples generated a unique eccDNA, precisely overlapping with a *Helitron* TE annotation (2,513 bp, chr6D: 152,366,979-152,369,491 of the Arina*LrFor* reference genome[30]), whereas this signal was absent in the control samples (Fig. 1b). Dot-plot graphs generated from the long reads revealed numerous reads containing multiple tandem repeat

copies of the *Helitron* sequence. These repeats were the result of the *in vitro* RCA reaction that was used to amplify the eccDNA signal (Extended Data Fig. 1a).

Further, to independently confirm the presence of *Helitron* eccDNA, we conducted inverse PCR utilizing primers designed to span the eccDNA junction (Fig. 1c). We ruled out any nonspecific signals potentially originating from genomic tandem repeat sequences in the Arina genome, as we did not detect any PCR products when genomic DNA was used as a PCR template (Extended Data Fig. 1b). To determine the requirements for efficient *Helitron* eccDNA production, we subjected plants to individual treatments with drugs alone (JZ), heat stress alone (H), and their combination (JZH). *Helitron*-derived eccDNA was exclusively detected via inverse PCR on RCA-amplified DNA from the JZH group (Extended Data Fig. 1c). These results demonstrate that Helitron-derived eccDNA accumulation is heat stress responsive but repressed by silencing.

We then wanted to test whether eccDNA accumulation varied between distantly related wheat cultivars. We tested four European cultivars (Arina, Renan, CH Nara and Rosatch) and two Asian cultivars (Chinese Spring and Norin61). JZH-treated Arina, Renan, CH Nara, and Rosatch produced detectable *Helitron*-derived eccDNAs but not the untreated control group (Fig. 1c). Conversely, the cultivars Chinese Spring and Norin61 did not produce detectable levels of Helitron-derived eccDNAs following JZH treatment (Fig. 1c). To delineate the exact junction between the RTS and LTS of this *Helitron*, the PCR fragments were subjected to Sanger sequencing. We found that the junctions presented the exact expected RTS and LTS junctions in all tested varieties that produced *Helitron*-derived eccDNA (Extended Data Fig. 1d).

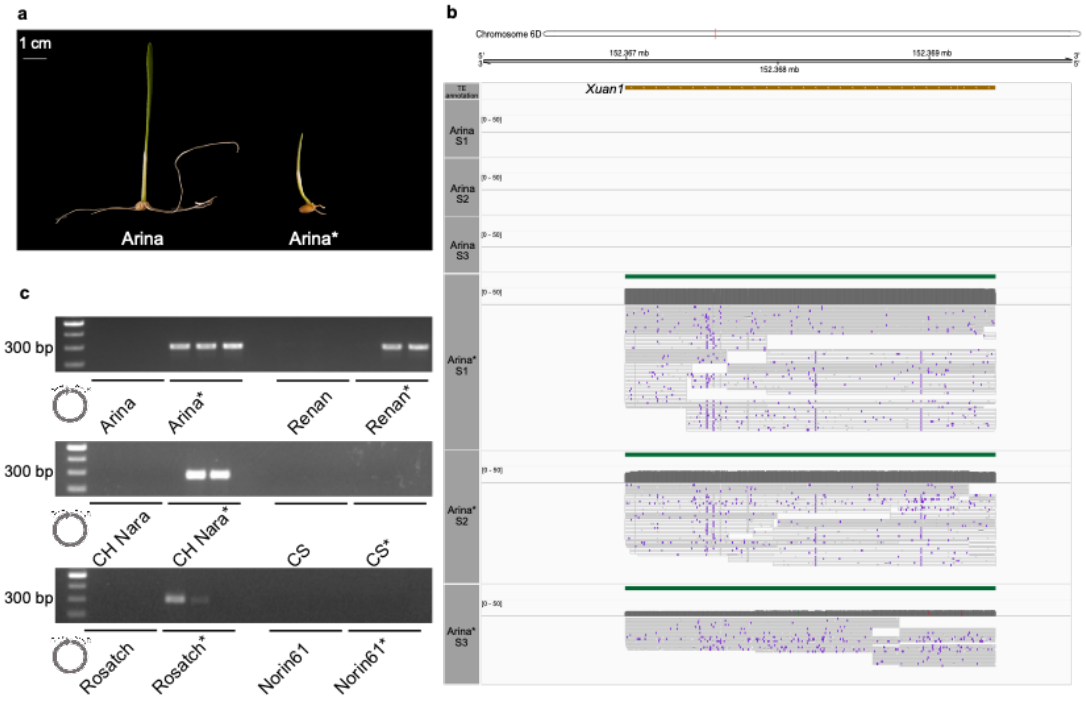


Figure 1. Detection of *Helitron*-derived eccDNA in wheat.
a, Morphological comparison of a 7-day Arina control (left) and a JZH-treated seedling (* right). **b,** Nanopore eccDNA sequencing analysis of untreated and treated Arina seedlings. The chromosomal position is shown on the top, the first row presents our *Helitron* annotation of the Arina*LrFor* genome[30]. Each row below the TE annotation represents an individual sample with three biological replicates from the control and JZH treated wheat seedlings. The green lines depict the eccDNA loci that were detected by ecc_finder[28], the dark gray oblong shape indicates normalized read coverage and below are the individual Nanopore reads mapped to this region (gray reads are single-mapped and white reads mapped to multiple loci in the genome). **c,** Detection of *Helitron* eccDNA using inverse PCR on RCA-treated DNA as a template. Primer localizations are depicted on the left (F, R). The first lane shows the GeneRuler 100 bp DNA ladder, each cultivar was tested with three biological replicates issued from the control and JZH-treated groups. The asterisk (*) indicates JZH-treated samples.

The eccDNA-producing *Helitron* exhibited canonical features, initiating with TC dinucleotides and terminating with CTAG tetranucleotides. We also identified a 17-nt hairpin structure, including a 5-nt loop and a 6-nt stem, located 11-nt upstream of the RTS (Fig. 2a and Extended Data Fig. 2a). Based on the absence of transposase coding capability (Extended Data Fig. 2a), we concluded that this element was a non-autonomous *Helitron* and designated it *Xuan1*. Non-autonomous TEs do not encode a functional transposase necessary for self-transposition. Instead, they rely on trans-mobilization facilitated by the enzymatic machinery of their autonomous counterparts from the same family[13]. To delineate this *Helitron* family, we employed a classification approach based on the 30 bp RTS region[15,16]. We designated this family *Xuan–Feng* (meaning 'whirlwind' in Chinese, *Xuan* signifies non-autonomous *Helitrons* and *Feng* denotes potentially autonomous *Helitrons*). The automated annotation of this family identified *Xuan1-79* and *Feng1-6* (Extended Data Table 1).

Furthermore, through an overall sequence homology search, we identified *Feng7*, which was initially overlooked due to its truncated LTS. By employing whole genome sequencing with Nanopore reads, we discovered *Feng8*, which was not completely assembled in the Arina*LrFor* reference genome[30] (Fig. 2a and Extended Data Fig. 2b, Extended Data Table 1). The *Xuan1* element and putatively autonomous *Feng* members exhibited conserved LTS and RTS, except *Feng6* and *Feng7*, which had accumulated mutations in their LTS (Fig. 2a). Using a multiple alignment approach, *Feng* members were clustered into two groups: one group that included *Feng1*, *Feng2*, and *Feng7* and the second group that contained all other *Feng* copies. Notably, *Feng4*, *Feng6* and *Feng8* showed over 99% nucleotide sequence identity (Extended Data Fig. 2c). We also detected a predicted heat response element (HRE) with the consensus sequence nTTCnnGAAn in the promoter regions of *Feng3-6* and *Feng8* (Extended Data Fig. 2d). Notably, the Arabidopsis retrotransposon *ONSEN* (*ATCOPIA78*) was found to include the same sequence motif in its heat-stress responsive promotor[31]. We also observed a preference for insertions of this *Helitron* family at AT dinucleotides (Extended Data Fig. 2e), which is a characteristic observed in numerous *Helitrons* across different species[12].

To identify the autonomous *Feng* copy that facilitated the biosynthesis of *Xuan1* eccDNA, we examined the transcriptional activity of all *Feng* members using RNA-seq to determine whether they responded to the JZH treatment. We observed only minimal, statistically non-significant, transcript levels for *Xuan1* in response to the JZH treatment (Fig. 2b). By contrast, the *Feng* elements displayed varying transcriptional responses following JZH treatment. *Feng2*, and *Feng5* did not exhibit a significant transcriptional activation, while *Feng1*, *Feng3*, *Feng4*, and *Feng6-Feng8* showed significantly increased transcript levels upon JZH treatment (Fig. 2b and Extended Data Fig. 3). *Feng1* produced only a short 1284 bp transcript following JZH treatment. Although *Feng3* and *Feng7* exhibited a partial constitutive transcriptional signal under control conditions, complete transcripts and significantly increased transcript levels were only observed in the JZH-treated group (Fig. 2b and Extended Data Fig. 3). A

close inspection of the transcripts produced by *Feng3, Feng4*, *Feng6*, *Feng7*, and *Feng8* revealed that all *Feng* copies, except *Feng8*, carried nonsense mutations or a frameshift mutation in the coding sequence (Extended Data Fig. 2f). Thus, we concluded that only *Feng8* had the capacity to code for a functional transposase.

Based on the RNA sequencing data, we found that *Feng8* contained a coding sequence (CDS) spanning 5,028 bp and encoded a transposase of 1,676 amino acids (Fig. 2c, Extended Data Fig. 4). This CDS consisted of seven exons, and the first five introns start with a GT donor site and end with the AG acceptor site, except for the last intron, which starts with GC. To verify the functional protein domains, we aligned FENG8 with the Helraiser protein (an *in silico* reconstituted *Helitron* transposase from *Myotis lucifugus*)[32]. We noticed that the FENG8 transposase had a larger N-terminal domain (M1–L638) and harbored a potential zinc-finger-like motif (R311–V337). The RepHel core of FENG8 comprised a HUH endonuclease domain spanning 289 amino acids (P668–M956), featuring the conserved HUH motif (H767, H769), as well as two tyrosine residues (Y902, Y906) known to form the active site in Helraiser. Additionally, it encompassed a helicase domain spanning 483 (R1194–V1676) amino acids, containing the eight conserved motifs that are characteristic of the SF1 superfamily of DNA helicases (Extended Data Fig. 4).

The structure predicted by AlphaFold2[33] revealed the presence of an N-terminal domain (ND) and a helicase domain connected by two intermediate domains, INT1 and INT2, as observed in the experimentally determined structure of Helraiser, and similarly forming a large tightly knit molecule. The first 264 amino acids of the ND are predicted to be disordered except for two α-helices. This is consistent with the cryo-electron microscopy structure of Helraiser, where the first 110 amino acids were found to be disordered[32]. Given the amino acid identity (33.92%) between FENG8 and Helraiser, the predicted and experimental protein structures both exhibited high overall similarity, suggesting a similar function (Extended Data Fig. 5).

However, there were topological differences in the N-terminal half of the predicted FENG8 structure when compared with the experimental structure of Helraiser. Interestingly, the prediction in this region was very similar to the structure of Helraiser predicted when AlphaFold2[33] was still unaware of the experimental structure. We note that this region of Helraiser contained a β-sheet structure, with neighboring strands bring far-away residues in the primary sequence into each other's proximity. Whether the resulting topological differences between the AlphaFold2[33] prediction of FENG8 and the Helraiser experimental structure are real or a prediction artifact, we cannot tell in the absence of a FENG8 experimental structure.

To investigate the prevalence of the FENG8 transposase in other species, we extracted a total of 83 *Helitron*-like element coding sequences from the NCBI database plus Helraiser (Extended Data Table

2). The phylogenetic tree indicated that FENG8 is more closely related to *Helitrons* present in monocotyledonous plants, whereas Helraiser clustered with animal and insect sequences, suggesting a long-standing evolutionary divergence (Extended Data Fig. 6).

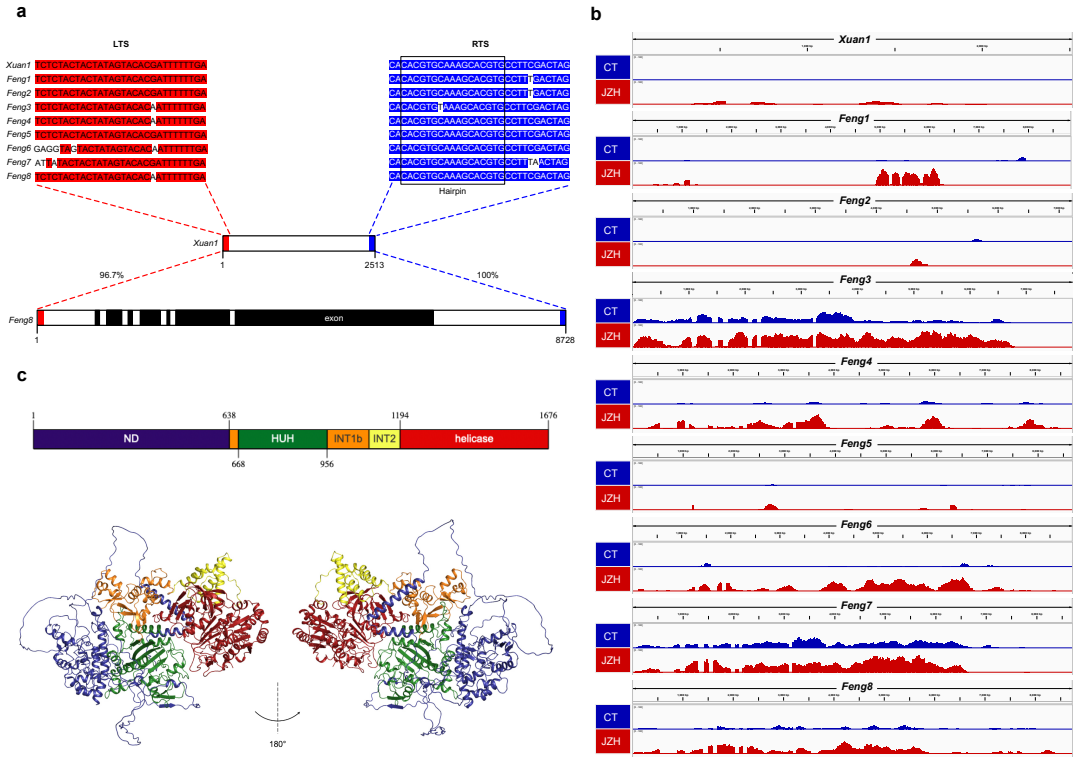


Figure 2. Characterization of the *Xuan–Feng Helitron* family.
a, Alignments of the LTS (with red shades) and RTS (with blue shades) sequences of *Xuan1* and the eight *Feng* members. The consensus bases are highlighted. **b,** Genome browser views showing the normalized transcript levels assessed by RNA-seq for *Xuan1* and the eight *Feng* sequences under control conditions (blue) and under JZH treatment (red). One out of three biological replicates is shown here (additional replicates can be found in Extended Data Fig. 3) **c,** Schematic diagram depicting the domains of the FENG8 transposase and two views of the predicted 3D structure below (rotated by 180°).

To test whether *Feng8* is necessary for *Xuan1* to spawn eccDNAs and facilitate *de novo* insertion, we took advantage of wheat recombinant inbred lines (RILs) that were obtained from an initial cross with two wheat cultivars—Chinese Spring and Renan—followed by eight generation of self-pollination (Fig. 3a). *Feng8* was absent in the genome of Chinese Spring[34] and this variety did not spawn *Xuan1* eccDNA following JZH treatment. The Renan cultivar, however, carried an intact *Feng8* copy in its genome[35] and spawned *Xuan1* eccDNAs upon JZH treatment (Fig. 1c, Extended Data Fig. 7). We cloned and sequenced *Feng8* from the Arina and Renan genomic DNA and found that they showed 100% identity in the nucleotide sequence. Similarly, our comparison of the *Xuan1* sequence of Arina, Chinese Spring, and Renan showed 100% identity among these lines. Based on single nucleotide polymorphism (SNP) genotyping data, 21 RILs were selected: 10 lines with the *Feng8* (Renan allele in the region) and 11 lines without (Chinese Spring allele in the region). The presence or absence of *Feng8* was confirmed by PCR for all RILs (Fig. 3c). To test the capacity of these RILs to produce *Xuan1* eccDNA, they were all subjected to JZH treatment. We found that only the RILs with *Feng8* produced *Xuan1* eccDNA, as confirmed by inverse PCR (Fig. 3b, 3c). As eccDNA levels were very low in these lines, we consistently performed these assays with three biological replicates, and we never detected false positives.

To capture the *de novo* insertions of *Xuan1* in the RILs, we took advantage of the high sensitivity of pooled CRISPR inverse PCR sequencing (PCIP-Seq, Methods and Extended Data Fig. 8a)[36]. We blindly applied this method to the 21 RILs, and captured a total of 49 somatic *de novo Xuan1* insertions in 8 out of 10 RILs with *Feng8* and none for the 11 RILs without *Feng8* (Fig. 3d, Extended Data Fig. 8 and Table 3). We extracted 20 bp of flanking host DNA sequence from these 49 *de novo* insertion sites and confirmed a preference for insertions at AT dinucleotides (Fig. 3e and Extended Data Table 3).

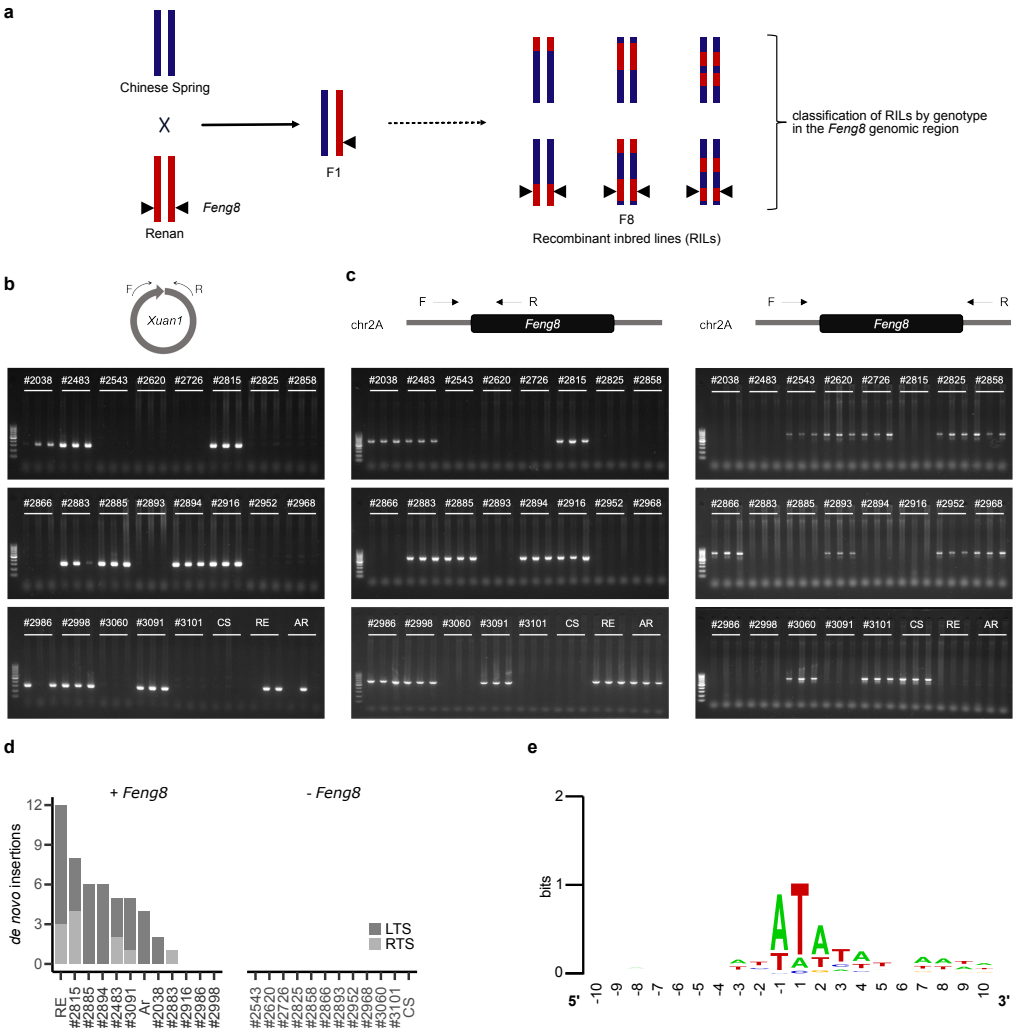


Figure 3. *Feng8* is necessary for *Xuan1* eccDNA biosynthesis and *de novo* insertion.
a, Schematic diagram showing the crossing scheme used to obtain Chinese Spring X Renan recombinant inbred lines (RILs). Blue segments highlight chromosome segments originating from Chinese Spring and red ones those originating from Renan. The black triangles indicate the presence of *Feng8*. Each RIL was first analyzed by SNP genotyping, and then the presence or absence of *Feng8* was confirmed by PCR (see **c**). **b,** Inverse PCR detection of *Xuan1* eccDNA for each JZH-treated RIL (number indicated on top) with three biological replicates (CS, Chinese Spring; RE, Renan; and AR, Arina). **c,** *Feng8* genotyping by PCR for the presence (left) and absence (right) of the *Helitron*. **d,** Quantification of somatic *Xuan1 de novo* insertion events in the genomes of RILs that have *Feng8* (+) and lines that do not (-). **e,** Sequence logo of 49 *de novo Xuan1* integration sites (insertion site between -1 and 1 base position). The logo was created using WebLogo (http://weblogo.berkeley.edu).

Since we demonstrated that *Feng8* can mobilize non-autonomous related elements, we wondered about the extent of presence/absence variations (PAVs) originating from lineage-specific insertions of *Feng* among wheat and wild-related species. We investigated *Feng* copy-number variations by scanning the available genome sequences of 17 *Triticeae* genotypes, including 13 *T. aestivum* (cultivated bread wheat, AABBDD) accessions[30,34,35], *T. durum* (cultivated durum wheat, AABB)[37], *T. dicoccoides* (wild emmer wheat, AABB)[38], *T. urartu* (wild diploid AA)[39], and *Ae. tauschii* (wild diploid DD)[40]. We used *Feng8* as a template for a similarity search against these high-quality genome assemblies while considering only complete copies (minimum size of 6 kb). Although repeated, each copy was inserted at a specific locus and we used junctions between LTS/RTS and their flanking site to identify orthologous copies between these 17 genomes (Fig. 4a). *Feng* copies maintained a low-copy number in *T. aestivum*, ranging from 4 to 8 copies.

We did not find any case of a recent massive increase in copy number in the analyzed *Triticeae* genomes. By contrast, *Feng* copies were generally ancestrally conserved (i.e., present at an orthologous location) across accessions and even with wild tetraploid (*T. dicoccoides*) and diploid species (*T. urartu* and *Ae. tauschii*). *Feng7* on chr7A was conserved at an orthologous position across hexaploids, tetraploids, and in the donor of the A subgenome *T. urartu* (which diverged from *T. aestivum*^{AA} ~1.3 Mya). *Feng7* orthologs shared 99.6–100% identity (over 8697 bp) between *T. aestivum* genotypes and tetraploid genomes. The identity was slightly lower (98.7%) with *Feng7* of *T. urartu* which was close to the expected divergence considering a rate of 1.3×10^{-8} mutation[41]. This revealed that *Feng7* inserted before the divergence of A subgenomes (ca. 1.3 Mya).

Similarly, *Feng5* on chr7D was present in diploid *Ae. tauschii* revealing that it inserted before hexaploidization and was lost afterwards in Chinese Spring. *Feng1*,*3*,*8* (A subgenome) and *Feng2*,*6* (B subgenome) were present in tetraploids, showing they did not insert recently but were present ancestrally before hexaploidization. *Feng4* (on chr6B) was the only example of a copy that was specific to *T. aestivum*, suggesting that it was the most recently inserted copy. It shared the highest identity (99.7%) with *Feng6*, suggesting that *Feng6* was the original copy that transposed. However, it is likely that the *Feng6* and *Feng4* transposases are nonfunctional because they carry the same early stop codons in their CDS (Extended Data Fig. 2). A plausible model would be that *Feng8* mobilized *Feng6* to generate the new *Feng4* insertion. We also found one complete copy called *Feng9*, which was specific to *T. dicoccoides* (8,730 bp, chr1A: 547,390,398-547,399,127) and absent in all other genomes (empty insertion site). Thus, *Feng9* originated from a recent insertion in *T. dicoccoides*. It shared the highest identity (99.6%) with *Feng8*. Notably, *Feng6* was absent in *T. dicoccoides* suggesting a recent deletion event.

Altogether, the PAV landscape revealed that *Feng* copies were present in all A–B–D lineages. It was present before A–B–D diverged ~5 Mya and maintained a low copy number (<10 per diploid genome).

A similarity search against the rye genome[42], which diverged even earlier (i.e., 7 Mya), also revealed the presence of 6 *Feng* fragments (sharing >90% identity over >1 kb), confirming the copy-number of this element in this lineage as well. Taken together, we conclude that this *Helitron* family rarely transposes but largely maintains the capability to do so. These results are in line with previous conclusions[4] reached by comparing the full TE repertoire through whole-genome alignments: In wheat, TE dynamics follow an equilibrium model of evolution, with numerous families being active yet maintaining a constant copy number rather than experiencing rapid amplification by bursts.

In summary, our study demonstrates that an epigenetically silenced and heat-stress responsive *Helitron* can be activated in wheat, as evidenced by transcriptional activity, eccDNA formation, and novel insertions. We genetically confirmed that the mobilization of the non-autonomous *Helitron* (*Xuan1*) relied on a single intact autonomous *Helitron* copy (*Feng8*), which may not be present in all *Triticeae* genomes (see model in Fig. 4b). This highlights a delicate balance between the gain and loss of *Helitron* copies, potentially regulated by the TE itself. Together, our findings pave the way for exploring *Helitrons* in virtually any species and an active TE in wheat, offering promising applications for mutation breeding.

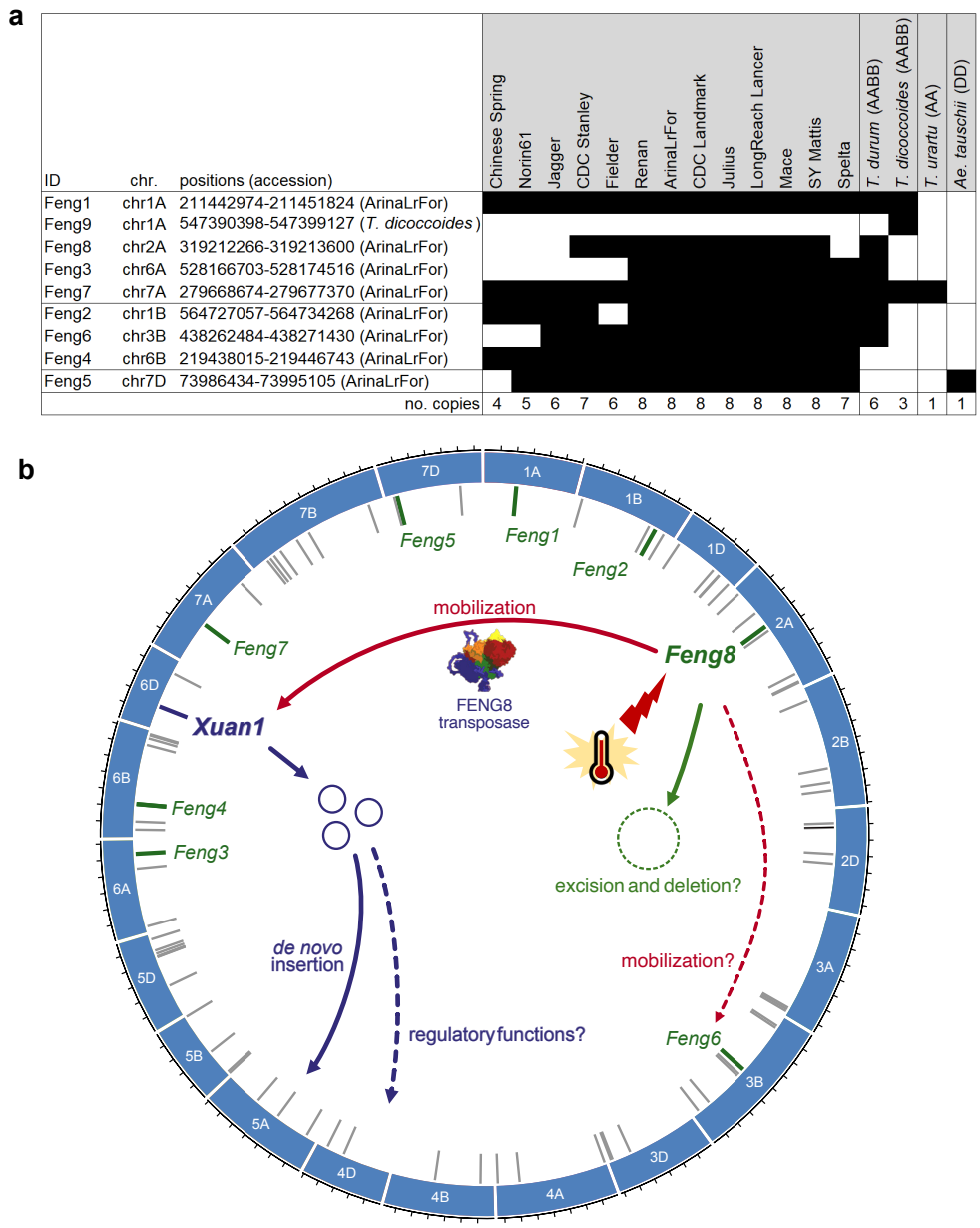


Figure 4. *Feng* copies in *Triticeae* genomes and their copy number regulation.
a, Presence/absence variations of *Feng* copies across 17 *Triticeae* genomes. Presence: black; absence: white. **b,** Mode of the actions of the *Xuan1–Feng8 Helitron* two-component system. Loss of DNA methylation and heat stress activate the expression of the FENG8 transposase. The transposase recognizes *Xuan1* and catalyzes the production of extrachromosomal circles. The high conservation of *Xuan1–Feng8* in *Triticeae* suggests that the *Xuan1*-derived circles could serve a biological function. The genome comparisons suggest that *Feng8* was involved in mobilizing *Feng6* and that it may have catalyzed its own deletion in certain wheat varieties. The blue outer ring represents the Arina*LrFor* wheat chromosomes (outside ticks 100 Mb), and the inner ring depicts *Xuan-Feng* Helitron annotations (unnamed gray bars are *Xuans*).

Material and methods

Plant materials and growth conditions

All cultivars of wheat seeds (Arina, CH Nara, Rosatch, Chinese Spring and Norin61) were obtained from Agroscope GenBank. Recombinant inbred lines were obtained from INRAE GenBank. Prior to planting, the wheat seeds were soaked overnight in sterilized water and subsequently sterilized using a 10 min heat shock at 50°C[43]. Following surface sterilization, three seeds were planted on 10 mL of agar media (0.55%) in a PHCbi MLR-352 growth chamber. The growth chamber was set to long-day conditions, with 16 h of light at 20°C during the day and 18°C at night, and a light level of 4 was maintained throughout the experiment.

Chemical treatment and stress

The treatment group seeds were cultured on agar media containing two drugs (as suggested by epibreed AG, Switzerland), namely 50 µM Juglone (Sigma, H47003) and 40 µM zebularine (Sigma, Z4775) dissolved in DMSO. An equal volume of DMSO (Fisher Scientific, BP231-100) was administered to the control group seeds. Heat stress was induced by exposing 6-day-old seedlings to 40°C for 24 h.

DNA extraction and eccDNA sequencing

The uppermost 3 cm section of the stem (coleoptile) was harvested from three individual seedlings, and coleoptile sheath was carefully removed. The three harvested shoot sections were combined into one tube immediately at the end of the stress period and snap-frozen in liquid nitrogen, followed by storage at -80°C until DNA extraction. The total DNA was then extracted from the frozen samples using the modified CTAB method[44].

The eccDNA-seq method was employed for all samples, as previously described[45]. Briefly, 2 µg of total DNA was digested with 10 U of PlasmidSafe (LubioScience, E3101K) for 16 h at 37°C to eliminate linear DNA, followed by enzyme denaturation at 70°C for 30 min. Next, 100 ng of the digested DNA was precipitated with ethanol supplemented with 1 µL of GlycoBlue (Fisher Scientific, 10391565). Circular DNA was then amplified through rolling circle amplification using the Illustra TempliPhi kit (GE Healthcare, 25-6400-10) according to the manufacturer's instructions, with the reaction left for 16 h at 30°C. The RCA reaction was inactivated by heating at 65°C for 10 min, followed by cooling at 4°C. The amplified DNA was once again precipitated with ethanol and debranched using 10 U of T7 Endonuclease I (New England Biolabs, M0302S) by incubating at 37°C for 30 min. After ethanol precipitation, 400 ng of debranched DNA was used for library preparation with the Nanopore Rapid Barcoding Sequencing Kit (SQK-RBK004). The eccDNA-seq was carried out using PromethION 2 Solo with flow cell FLO-PRO114M and basecalling was performed by Guppy v7.1.4.

Wheat whole-genome sequencing

The high-molecular weight genomic DNA from the leaf samples was fragmented using a gTube to select fragments of approximately 20 kb. The library preparation was carried out using a ligation sequencing kit (SQK-LSK109), and sequencing was performed on the FLO-PRO002 flow cell of PromethION by Biomarker (BMKGENE) Europe. The clean data obtained from nanopore sequencing were aligned to the wheat reference genome (Triticum_aestivum_ArinaLrFor_v3.0, IPK) using minimap2[46] v2.26 with default parameters. Subsequently, we employed nanomonsv [47] v0.7 to detect structural variations.

eccDNA loci identification

Basecalled, demultiplexed, and trimmed eccDNA-seq reads from the same sample were concatenated. To identify eccDNA formation regions in JZH-treated Arina seedlings, reads were aligned to the wheat reference genome (Triticum_aestivum_ArinaLrFor_v3.0, IPK) using ecc_finder v.1.0.0 in *map-ont* mode with the parameter *MAX_SIZE* in the Genrich.h dependency script altered to 1000000 to enable processing of long-read alignments.

Inverse PCR to test for eccDNA circle junctions

A total of 100 ng of genomic DNA and 50 ng of RCA-amplified DNA were utilized as separate templates to generate inverse PCR. Specific primers were designed to span the junction of the *Xuan1* circle. The PCR reaction was performed in a 20 µL volume according to the manufacturer's protocol of GoTaq® G2 DNA Polymerase (Promega, M7841). The PCR amplification conditions consisted of an initial denaturation step at 95°C for 2 min, followed by 30 cycles of denaturation at 94°C for 30 s, annealing at 58°C for 30 s, extension at 72°C for 2 min, and a final extension step at 72°C for 10 min. The purification of PCR products was carried out using the Wizard® PCR clean-up system (Promega, A9281), following the Sanger sequencing service provided by Microsynth AG.

TE annotation and Helitron family identity

Genome-wide *de novo* TE annotation was produced by EDTA[48] v1.7 with default parameters. For *Helitron* annotation, we initially filtered out all candidate sequences annotated by HelitronScanner[14] v1.0 with parameter '*minlength=200, max_length=20000, head_threshold=5, tail_threshold=5, flankh_len=50, flankt_len=50*' that contained "N" bases. Next, using the 30 bp RTS of the *Xuan1* as a reference, we retained candidates with an identity above 80% to establish them as members of the same family by SeqKit[49] v2.0.0, with the parameter '*grep -j 6 -s -i -m 5 -P -p*'. For sequence alignment, we employed the MAFFT[50] tool.

Transcript analysis

Total RNA extractions were carried out using the NucleoSpin® RNA kit (Macherey-Nagel, 740955.50) for three biological replicate samples under each condition. The extracted RNA samples were subjected to Illumina 150 bp paired-end sequencing using a stranded poly-A directional library at Novogene. Raw mRNA-seq reads were aligned to the complete gene sequences of *Feng1-8* and *Xuan1* using STAR[51] v.2.7.10a with parameter --*outFilterMismatchNmax* 0 to disallow mismatches. Alignment BAM files were filtered with samtools[52] v.1.18 to keep only primary mappings, and normalized and converted to bigwig format with the *scale* function of BAMscale[53] v.1.0 using the parameters --*operation rna* --*scale custom* and --*factor* with custom normalization factors for each library. Calculation of normalization factors was conducted by assembling a *de-novo* reference transcriptome with Trinity[54] v.2.15.1 using all reads and default parameters, quantifying transcripts with RSEM[55] v.1.3.3 and bowtie2[56] as aligners and applying the function calcNormFactors from the R library edgeR[57] v.4.0.3 on the posterior mean counts in R v.4.3.2.

Cloning and sequencing of PCR fragments

Feng8 cloning from Arina and Renan gDNA was produced by PCR using TOPO™ XL-2 complete PCR cloning kit (Invitrogen, K805010 and K805020). The purified PCR products were ligated with the pCR-XL-2-TOPO™ vector at a molar ratio of 1:1 (transcript: vector). The ligation reaction was carried out at 25°C for 30 min. Subsequently, 2 µL of the ligation product was transformed into One Shot™ OmniMAX™ 2 T1^R chemically competent *Escherichia coli* using a heat shock method. The transformed cells were plated on selective agar plates, and individual colonies were selected for further analysis. The presence of the desired inserts in the colonies was confirmed by PCR, and the plasmids were subsequently purified using the PureLink™ quick plasmid miniprep kit (Invitrogen, K210010). To ensure the integrity of the cloned inserts, the entire plasmid sequence was determined using Microsynth AG.

Transposase analyses and structure prediction

The physicochemical properties of proteins were analyzed using the ProtParam tool[58]. The 3D structure prediction of the protein was obtained using AlphaFold[33] v2.3.1. Five parallel model predictions were conducted for each sequence, and the structures with the highest predicted local distance difference test scores (pLDDTs) were selected for analysis and visualization using PyMOL software v2.1.

Phylogenetic tree construction

To generate the maximum-likelihood phylogenetic tree for transposase, the amino acid sequence of FENG8 was employed as the query for a blastp search against the NCBI database. Top-scoring hits from 83 different species were chosen to construct the phylogenetic tree. Multiple sequence alignment was carried out using MUSCLE[59] v5.1, and the alignment was subsequently converted to phylip format using

trimAl[60] v1.4. The phylogenetic tree was constructed using IQ-TREE[61] v2.12 with a bootstrap value of 1000 replicates.

Selection of 21 Chinese Spring × Renan recombination inbred lines based on SNP genotyping data

We analyzed SNP genotyping data (Axiom array TaBW420k) available for 282 individuals originating from a cross between Chinese Spring (CS) and Renan[62], to select 21 individuals that carry either CS or Renan alleles at both *Feng8* (chr2A) and *Xuan1* (chr6D) loci. The positions of the markers were identified by mapping SNP context sequences on the CS IWGSC RefSeq v2.1 using pipeline nf_remapMarkers (https://forgemia.inra.fr/umr-gdec/nf-remapmarkers). Although *Feng8* is absent from CS, its insertion site is at position 323,357,927 on chr2A. The position of *Xuan1* is 169,657,404 on chr6D. We retrieved SNP alleles in a window of 100 Mb centered on each of the two loci on chr2A and chr6D. It represented 496 and 163 SNPs on these two chromosomes, respectively. We kept only 132 individuals, all alleles within 100 Mb, encompassing each locus derived from the same parent. We classified the 132 individuals into four haplotypes: A/A, A/B, B/B, B/A (allele A=CS, allele B=Renan; *Feng8*/*Xuan1*). B/B and B/A individuals were expected to carry a functional *Feng8* originating from the Renan parent, while *Feng8* was predicted to be absent for A/B and A/A individuals. We eventually randomly selected 21 individuals for their ability to produce eccDNAs so that 5, 5, 5, and 6 were of the following haplotypes: B/B (#2038, #2483, #2883, #2986, #2998), B/A (#2815, #2885, #2894, #2916, #3091), AB (#3101, #2620, #2866, #2893, #3060), and AA (#2968, #2726, #2825, #2858, #2952, #2543), respectively.

Modified PCIP-seq to capture de novo Xuan1 insertions

The method has been described in detail in Tang *et al*[36]. Molecular biology: (i) Using 500 ng of genomic DNA as starting material, fragments containing *Xuan1* sequences were cleaved using single-guide RNAs (Integrated DNA Technologies) targeting sequences at 1357 bp from the LTS (see primer list) and *S. pyogenes* Cas9 (New England Biolabs, M0386S). (ii) The digested DNA was further mechanically sheared to ~350 bp using a Bioruptor-pico (Diagenode), sequencing libraries were generated using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, E7645L), and indexed libraries pooled and sequenced on a Novaseq 6000 sequencer (Illumina) targeting ~8 million 150 bp paired-end reads per library. Data processing: The ensuing sequence reads were demultiplexed, quality assessed using fastQC[63], adapter sequences trimmed using Cutadapt[64], and trimmed reads mapped to the Chinese Spring IWGSC RefSeq genome[6] using BWA-MEM[65] and converted to BAM format using SAMtools[66]. Using a custom-made Python script, we first identified clipped reads using CIGAR information. We selected clipped reads with mapping quality ≥ 40 and a minimum of 10 clipped bases. We then mapped the clipped reads to the segments of the corresponding *Xuan1* genome. We specified an alignment score ≥ 0.6 to declare a hit and labeled the read as either an insertion site (IS) or a shearing site (SS) read. When possible, we extended the alignment with *Xuan1*

into non-clipped bases to refine the positions of the SS and IS. We then merged SS and IS sites with the same "breakpoint," thereby identifying candidate SS and IS supported by multiple concordant reads. We then paired the IS with their cognate SS. The pairing was based on orientation (*e.g.*, a 5' SS should be located upstream of the 5' IS for a *Xuan1* element in "sense" orientation, and downstream of the 5' IS for a *Xuan1* in "antisense" orientation; and distance (the maximum distance between IS and SS was set at 5 Kb). The pipeline is available here: https://github.com/Lijingtangbo.

Data availability

The *Xuan1* and *Feng8* sequences have been submitted to NCBI GenBank with accession numbers PP376094 and PP376095. The sequencing data from this study have been submitted to the European Nucleotide Archive (ENA, www.ebi.ac.uk/ena/, accessed on ERP157467) under the project PRJEB72688. The whole genome sequencing raw reads are under fastq accessions ERR12723706-ERR12723708, raw reads of eccDNA-Seq for Arina are under the accessions ERR12724408-ERR12724413. The raw reads of the RNA sequencing are under the accessions ERR12724677-ERR12724682. Positions of TaBW420K SNP markers on CS IWGSC RefSeq v2.1 are available under DOI: https://doi.org/10.57745/YXNWZK. The raw fastq files for the PCIP-seq experiment applied to the 24 RILs have been submitted to the European Nucleotide Archive (ENA, www.ebi.ac.uk/ena) under the project PRJEB75199.

Funding

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program [725701, BUNGEE, to E.B.], by a grant from the Agence Nationale de la Recherche and Fonds National Suisse [ANR-21-PRCI-CE02, FNS 310030E_205554 "CropCircle" to E.B.] and by the China Scholarship Council Grant [201806990012 to H.P.]. We are grateful to the Mésocentre Clermont-Auvergne and the platform AuBi of the University Clermont Auvergne for providing computing and storage resources. F.D. and A.B.H. were supported by the Intramural Program of the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health. We thank the GIGA Genomics platform for producing the NGS data used for *de novo* insertion detection.

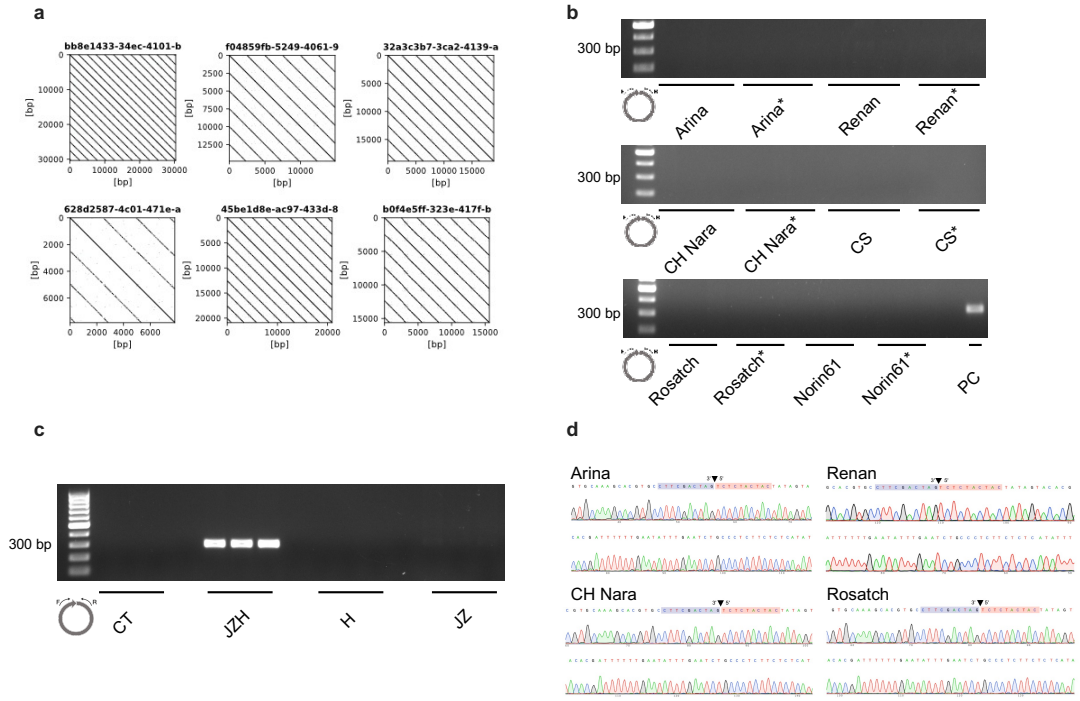
Contributions

H.P., L.T., and N.H. conducted the experiments; H.P., R.K., H.R., A.H., F.D., C.C., F.C., and E.B. analyzed the data; D.F., P.S. and F.C. provided materials; H.P., F.C. and E.B. designed the experiments and wrote the paper.

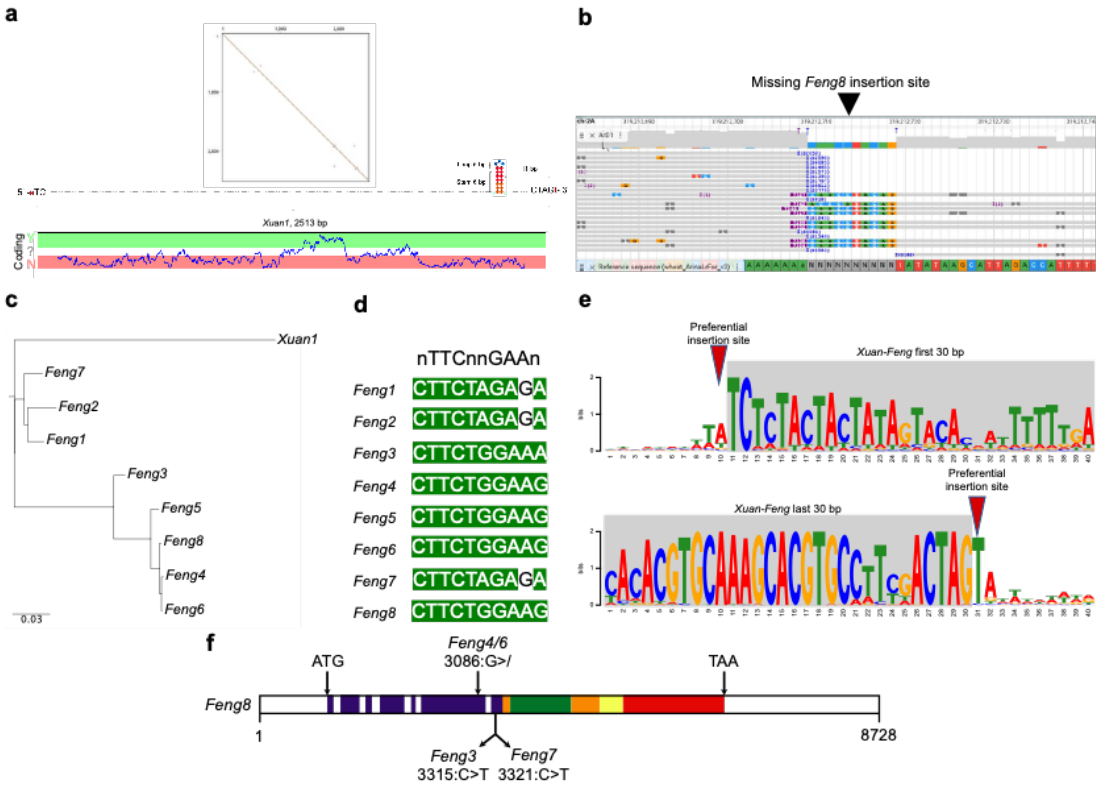
Competing interests

E.B. is a member of the board of epibreed AG, a Spin-Off company involved in plant breeding.

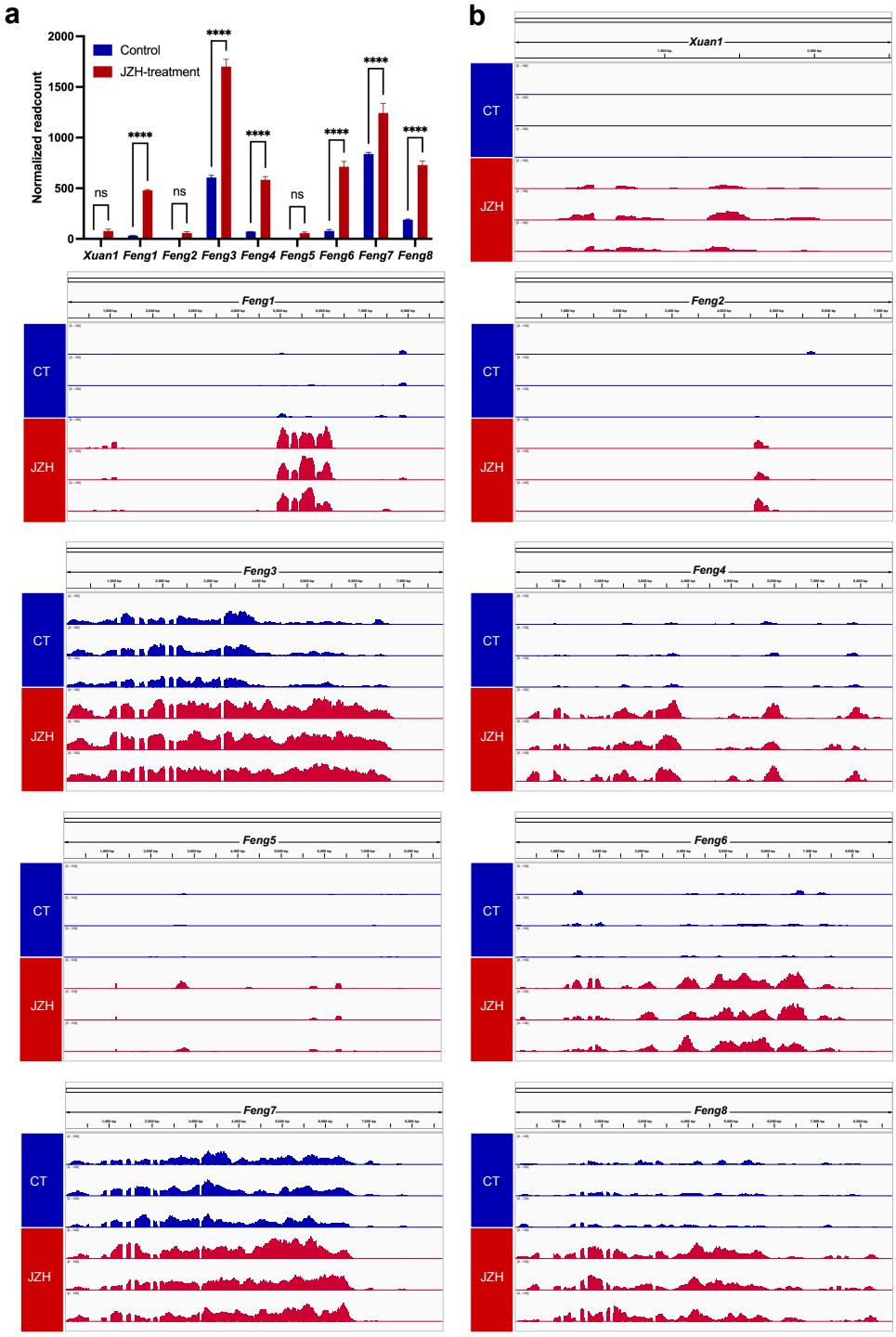
Supplemental material



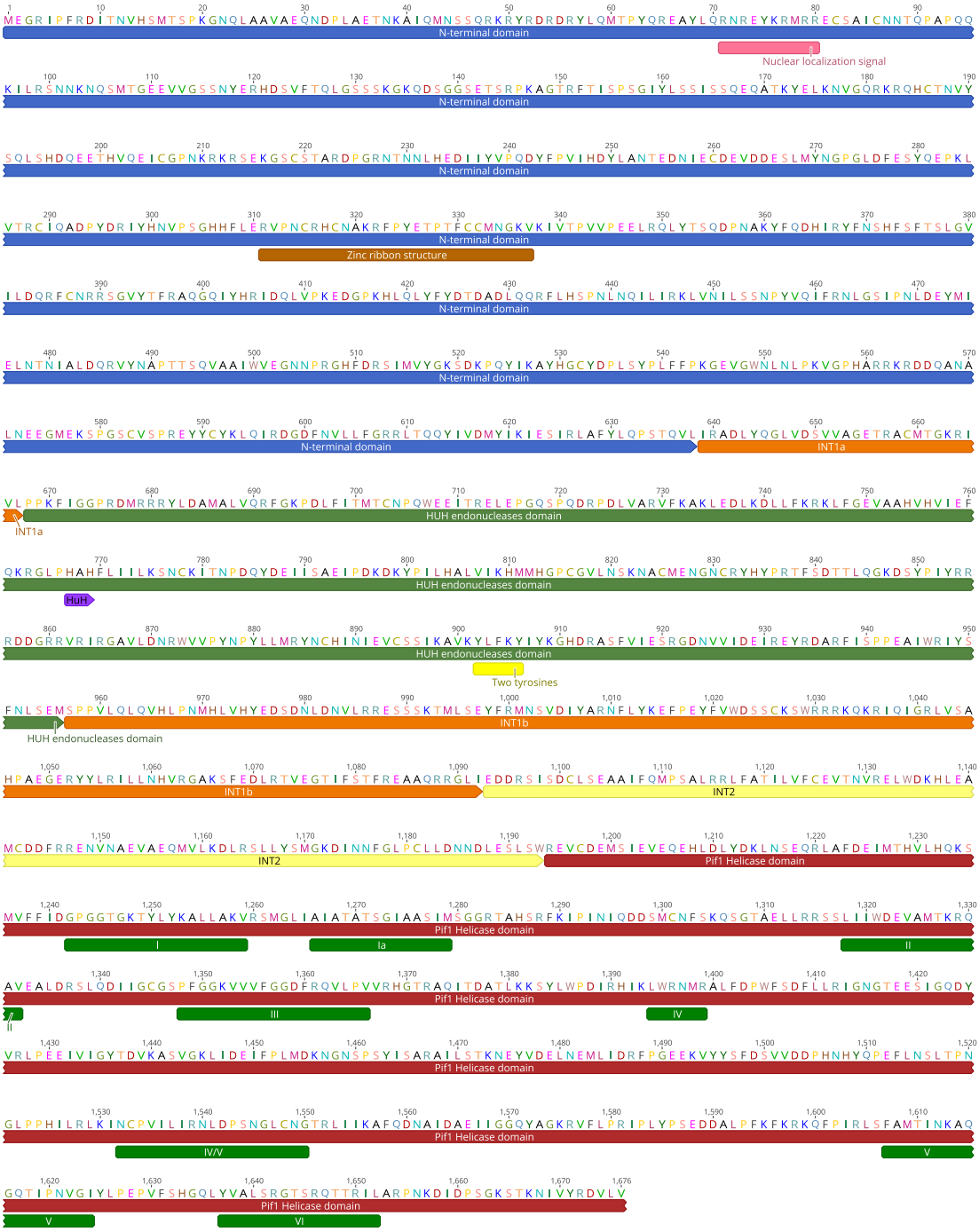
Extended Data Figure 1. Identification and verification of *Helitron* eccDNA.
a, Dot-plot analyses visualizing the self-similarity of selected single long reads containing tandem repeats of the *Helitron* in JZH-treated Arina. **b,** Confirmation of absence of genomic *Xuan1* tandem repeats using inverse PCR with genomic DNA as template (no RCA), the primer localization depicted on the left (F, R). The first lane shows the GeneRuler 100 bp DNA ladder, each cultivar was tested using three biological replicates issued from the control and JZH-treated groups. RCA eccDNA from treated Arina plants was used as a template for the positive controls (PC). *Xuan1* eccDNA can only be detected following RCA. The asterisk (*) indicates JZH-treated samples. **c,** Detection of *Helitron* eccDNA in the Arina variety using inverse PCR using eccDNA-amplified DNA as a template; the primer localization depicted on the left (F, R). The first lane shows the GeneRuler 100 bp DNA ladder; each treatment was tested with three biological replicates issued from the control (CT), JZH treatment, heat treatment (H), and JZ-treatment groups. **d,** Chromatograms from Sanger sequencing, the blue background highlights the 10 bp of *Helitron* RTS, and the red background highlights the 10 bp of *Helitron* LTS, and the inverted black triangle indicates the circle junction.



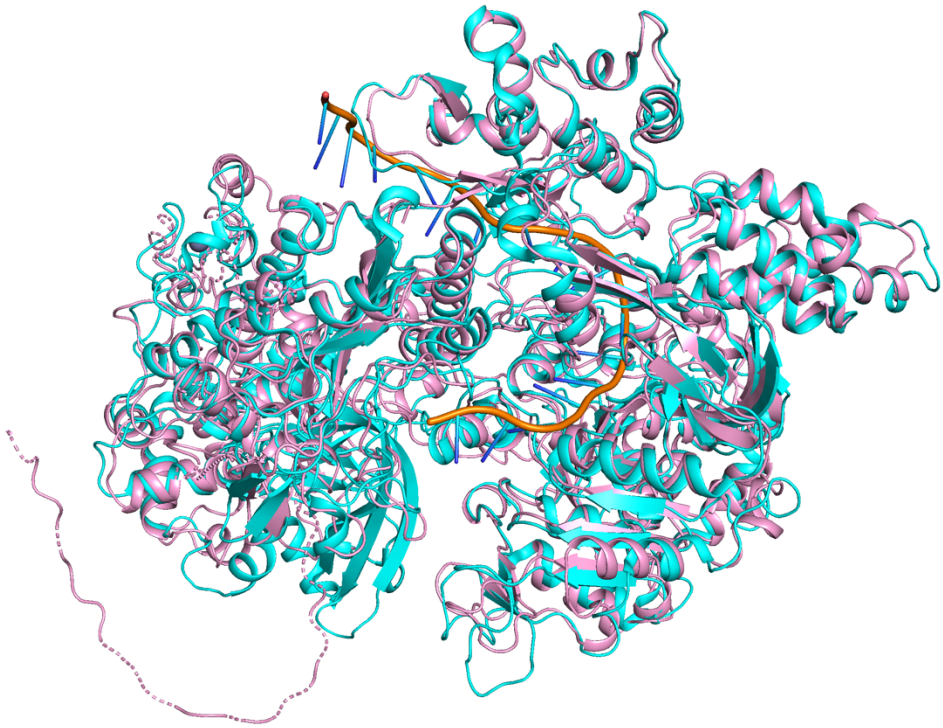
Extended Data Figure 2. Sequence features of the *Xuan–Feng Helitron* family.
a, Self-dot plot of the *Xuan1* sequence and schematic representation of the molecular features. The graph below indicates the protein coding capacity of the sequence, with red indicating low and green indicating high. The hairpin structure is depicted as colored rings. **b,** Genome browser view of the *Feng8* locus on chr2A:319,212,266 of the Arina*LrFor* reference genome where the *de novo* assembly failed to assemble *Feng8* (as indicated by the multiple N's in the reference sequence shown on the bottom). The gray horizontal bars represent individual nanopore sequencing reads and insertions (numbers with magenta backgrounds) and mismatches (colored bases) compared to the reference sequence. **c,** Phylogenetic tree depicting the relationship between *Feng* members, *Xuan1* was included as an outlier. **d,** Comparison of the predicted heat response element (HRE) motif in the eight *Feng* members. The consensus bases are highlighted. **e,** Sequence motif logo illustrating the 40 bp boundaries of the *Xuan–Feng* family members, including 10 bp upstream and downstream sequences for each 5' and 3' ends. The gray background highlights the *Helitron* sequence. **f,** Schematic diagram depicting mutations that impact the protein coding sequences in *Feng3* (nonsense mutation), *Feng4* (frameshift mutation), *Feng6* (frameshift mutation), and *Feng7* (frameshift mutation) represented on the intact *Feng8* sequence. The colors on the exons correspond to the different transposase domains shown in Fig. 2.



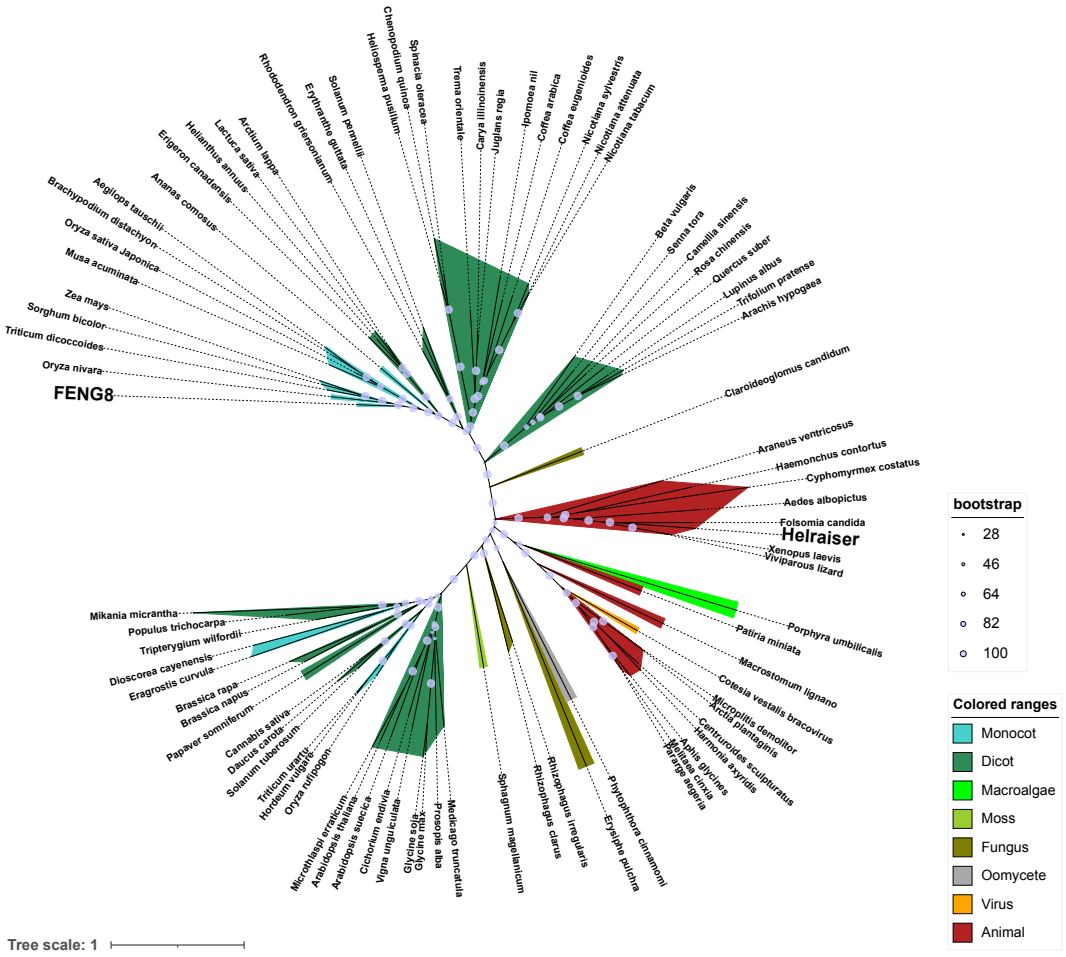
Extended Data Figure 3. Transcript levels of *Xuan1* and *Feng1-8* assessed by RNA-seq. **a,** Quantification of the RNA-seq data obtained from control (blue) and JZH- treated (red) Arina seedlings. The statistics use two-way ANOVA of normalized read count mean with SEM; the stars indicate the adjusted P value < 0.0001. **b,** Genome browser views showing three biological replicates of normalized transcript levels assessed by RNA-seq for *Xuan1* and the eight *Feng* sequences under control conditions (blue) and resulting from JZH treatment (red).



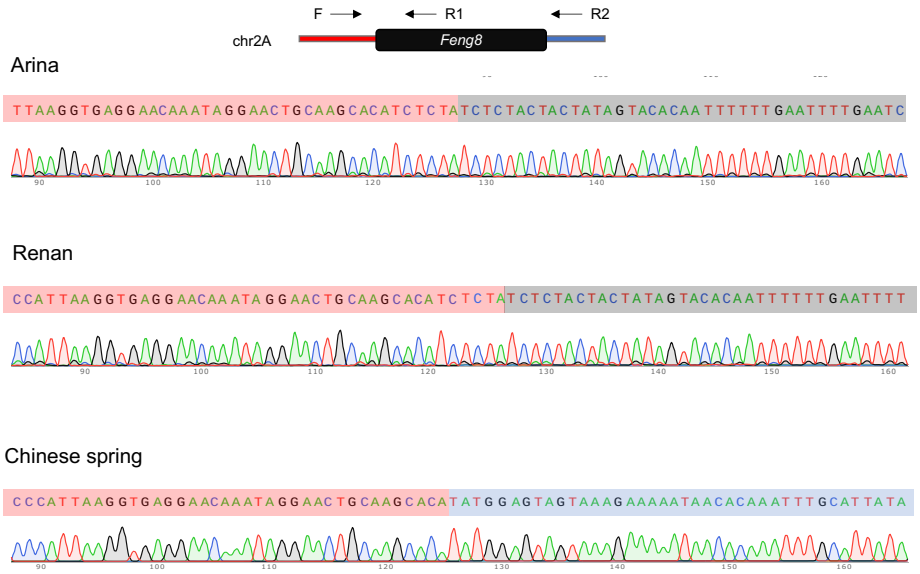
Extended Data Figure 4. Graphical representation of the FENG8 protein sequence highlighting the conserved domains of the transposase.
Residues involved in the zinc ribbon are highlighted in brown, the HUH motif is highlighted in purple, the Y2 motif in yellow, and the eight Helicase motifs are indicated in green.



Extended Data Figure 5. Graphical representation of the overlaid structures of FENG8 and Helraiser.
Comparison of the predicted FENG8 transposase structure (in pink) and the experimentally determined Helraiser structure (PDB: 7LCC, in cyan). Single-stranded transposon left-end DNA bound to Helraiser is shown in orange.

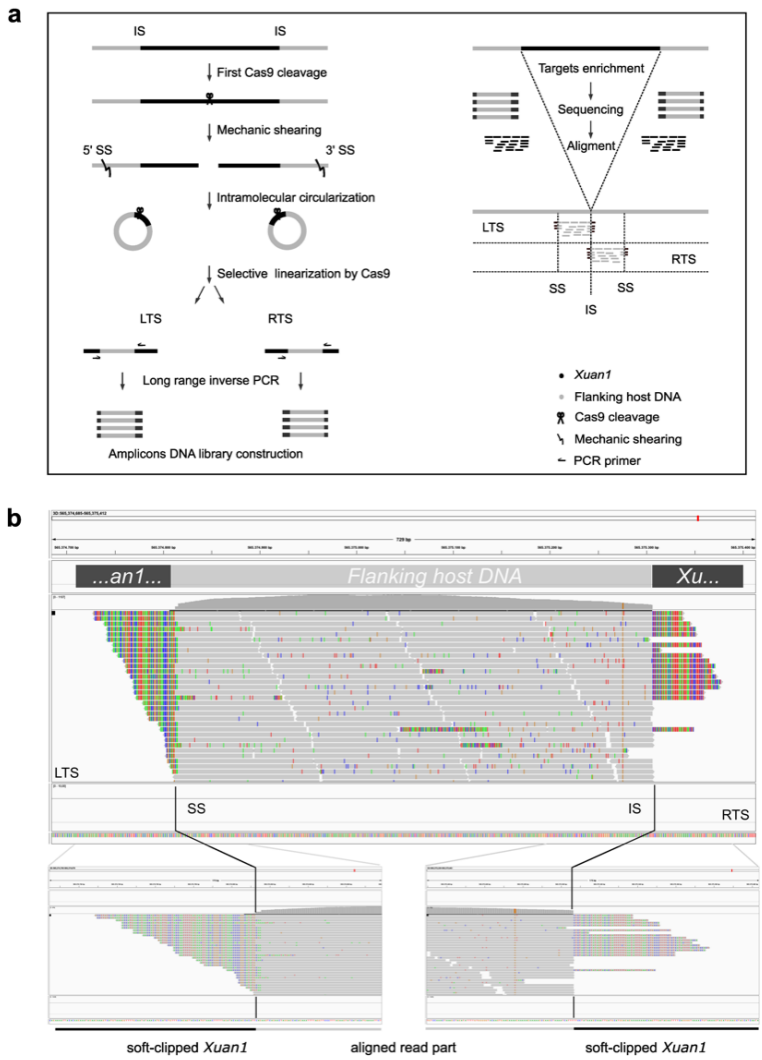


Extended Data Figure 6. Phylogeny of FENG8 and other Helicases.
Maximum-likelihood phylogeny of the FENG8 transposase with Helitron-like transposases from 84 species. Detailed information for each protein can be found in the Extended Data Table 2.



Extended Data Figure 7. Sequence analysis of the *Feng8* insertion site in three wheat varieties.

Chromatograms from Sanger sequencing for genotyping *Feng8* in cultivars Arina, Renan and Chinese Spring. The red background highlights the upstream sequence of *Feng8*, the gray background highlights the *Feng8* sequence, and the blue background highlights the downstream sequence of *Feng8*. Note that Chinese Spring does not carry a *Feng8* insertion at this locus.



Extended Data Figure 8. Identification of *de novo Xuan1* insertions using a modified PCIP-seq. **a,** Graphical representation of PCIP-seq procedures. Genomic DNA is attacked with a CRISPR/Cas9 targeting the *Xuan1* genomic sequence (1357 bp). The DNA is then mechanically sheared (generating shearing sites 5'SS and 3'SS), end-repaired, circularized, and attacked separately by a second pair of CRISPR/Cas9 targeting *Xuan1* sequences adjacent to *Xuan1* extremities to specifically reopen circles encompassing insertion sites (IS) in two separated pools (left terminal site [LTS] and right terminal site [RTS]). Fragments encompassing the IS and SS are then amplified by long-range inverse PCR using divergent *Xuan1* targeting primer pairs. The resulting PCR products are subject to NGS and the obtained sequence reads are mapped on the Chinese Spring IWGSC RefSeq[6], revealing *Xuan1* insertion sites (IS) and shearing sites (SS) used as molecular identifiers. **b,** Integrative Genome Viewer (IGV) screen capture of a sense *de novo Xuan1* insertion captured in sample (RE) using PCIP-seq[36]. The upper panel shows the *de novo* insertion captured from LTS but not from RTS. The bottom panels display zoomed regions surrounding the SS (left) and IS (right), highlighting characteristic soft-clipped read features.

Primer list
F: forward primer, R: reverse primer

For *Xuan1* circle junction check

Xuan1-JC-F	ACTGCAAATCAGCAAGAGGTTG	
Xuan1-JC-R	TTTTCAGGTTTGGCGGGAGT	301 bp

For *Feng8* cloning

Feng8C-1F	GAAGAAACAAACACATCTCCATACC	1F-2R: 4597 bp
Feng8C-2F	ACAACATTACAAGGAAAGGACTCA	2F-1R: 4826 bp
Feng8C-1R	CGCTGAAACCACGATCTAATACAT	1F-1R: 9008 bp
Feng8C-2R	TATCGGAGTCTTCATAATGCACAA	

For *Feng8* genotyping

Feng8G-F	GTACTATGGGAAGTATAGATTATGAAG	F-2R: 599 bp
Feng8G-1R	TTTCTTCACCACGAGCACGAGCA	F-1R: 737 bp
Feng8G-2R	GAAGCTGCAATCTGTACGCTGAGG	

For modified PCIP-seq

1st sgRNA	ATAAACCCCGGCATATGTTGGGG
2nd sgRNA (LTS)	ACCATAGTGCCGTTGACCACAGG
2nd sgRNA (RTS)	CCCAAAACATCTGCCGTACGTGG
Inverse PCR (LTS)-F	CAACAATGGAGGGACATGCC
Inverse PCR (LTS)-R	TTGGCAGGGAATCGATTTGC
Inverse PCR (RTS)-F	GCATTACATCCACCTACGTCC
Inverse PCR (RTS)-R	GAGCCTGTACAACTCGTGC

Extended Data Table 1. The *Xuan–Feng Helitron* family members.

Locus in ArinaLrFor	Size	Forward /Reverse	Name ID
chr1A:211442974-211451824	8851	R	*Feng1*
chr1B:564727057-564734268	7212	F	*Feng2*
chr6A:528166703-528174516	7814	F	*Feng3*
chr6B:219438015-219446743	8729	R	*Feng4*
chr7D:73986434-73995105	8672	R	*Feng5*
chr3B:438262484-438271430	8947	F	*Feng6*
chr7A:279668674-279677370	8697	R	*Feng7*
Insert chr2A:319,212,266	8728	R	*Feng8*
chr6D:152366979-152369491	2513	R	*Xuan1*
chr1D:456280548-456282897	2350	F	*Xuan10*
chr2A:208439032-208441843	2812	R	*Xuan11*
chr2A:353767939-353776535	8597	R	*Xuan12*
chr2A:38197790-38198994	1205	F	*Xuan13*
chr2A:692869009-692869881	873	F	*Xuan14*
chr2A:747304360-747305851	1492	R	*Xuan15*
chr2A:754818057-754819542	1486	R	*Xuan16*
chr2B:554476594-554479324	2731	F	*Xuan17*
chr2B:649006919-649008141	1223	R	*Xuan18*
chr2B:703622127-703623847	1721	R	*Xuan19*
chr2D:297603786-297606305	2520	F	*Xuan2*
chr2B:82701048-82706557	5510	F	*Xuan20*
chr2D:117893488-117894727	1240	F	*Xuan21*
chr2D:119750147-119751538	1392	R	*Xuan22*
chr2D:362840816-362841731	916	R	*Xuan23*
chr2D:91352056-91355610	3555	R	*Xuan24*
chr3A:700614554-700619462	4909	F	*Xuan25*
chr3A:713985640-713998115	12476	R	*Xuan26*
chr3A:727875872-727878600	2729	F	*Xuan27*
chr3B:479425126-479426481	1356	F	*Xuan28*
chr3B:495367829-495382659	14831	F	*Xuan29*
chr1B:34980517-34983105	2589	R	*Xuan3*
chr3B:69526011-69530491	4481	F	*Xuan30*
chr3B:754698060-754717253	19194	F	*Xuan31*
chr3B:844521550-844525941	4392	R	*Xuan32*
chr3B:85936148-85937567	1420	R	*Xuan33*
chr3D:464414808-464418403	3596	F	*Xuan34*
chr3D:592105962-592107184	1223	R	*Xuan35*
chr3D:604278422-604279660	1239	R	*Xuan36*
chr4A:588759809-588767616	7808	R	*Xuan37*
chr4A:6382660-6384077	1418	R	*Xuan38*
chr4A:734689786-734690987	1202	R	*Xuan39*
chr1B:517734864-517736406	1543	R	*Xuan4*
chr4B:390762811-390779619	16809	F	*Xuan40*
chr4B:85619121-85620881	1761	R	*Xuan41*
chr4D:326316056-326316589	534	R	*Xuan42*
chr4D:445637985-445638896	912	F	*Xuan43*
chr5A:303372529-303373375	847	R	*Xuan44*
chr5A:472545693-472562978	17286	F	*Xuan45*
chr5A:72889358-72890092	735	F	*Xuan46*
chr5B:18871653-18872588	936	R	*Xuan47*

chr5B:193098391-193104115	5725	R	*Xuan48*
chr5B:4544975-4548925	3951	F	*Xuan49*
chr1B:628948431-628949311	881	F	*Xuan5*
chr5B:478530669-478532160	1492	F	*Xuan50*
chr5D:213328934-213330835	1902	R	*Xuan51*
chr5D:418966751-418971078	4328	R	*Xuan52*
chr5D:421464392-421465613	1222	F	*Xuan53*
chr5D:424488613-424496819	8207	R	*Xuan54*
chr5D:443918268-443920263	1996	F	*Xuan55*
chr5D:466003370-466019944	16575	F	*Xuan56*
chr5D:545135720-545137231	1512	R	*Xuan57*
chr6A:37130689-37146999	16311	F	*Xuan58*
chr6A:428508580-428511300	2721	F	*Xuan59*
chr1D:31808203-31809559	1357	F	*Xuan6*
chr6B:102207187-102208097	911	R	*Xuan60*
chr6B:220339019-220341829	2811	R	*Xuan61*
chr6B:48589565-48590973	1409	F	*Xuan62*
chr6B:620475696-620476164	469	R	*Xuan63*
chr6B:665528436-665529672	1237	F	*Xuan64*
chr6B:687459974-687462785	2812	F	*Xuan65*
chr6B:695493524-695494858	1335	F	*Xuan66*
chr6D:391293775-391295220	1446	F	*Xuan67*
chr7A:279668674-279684268	15595	R	*Xuan68*
chr7A:645893950-645913273	19324	F	*Xuan69*
chr1D:318661651-318674519	12869	R	*Xuan7*
chr7B:124289500-124292228	2729	R	*Xuan70*
chr7B:154715831-154734905	19075	R	*Xuan71*
chr7B:208308241-208313609	5369	R	*Xuan72*
chr7B:340012740-340013888	1149	R	*Xuan73*
chr7B:434513281-434526327	13047	R	*Xuan74*
chr7B:873536841-873538256	1416	R	*Xuan75*
chr7B:99708278-99709724	1447	F	*Xuan76*
chr7D:500468419-500472732	4314	F	*Xuan77*
chr7D:52694625-52696909	2285	R	*Xuan78*
chrUn:53397574-53400421	2848	R	*Xuan79*
chr1D:369134526-369140034	5509	R	*Xuan8*
chr1D:372467999-372469215	1217	R	*Xuan9*

Extended Data Table 2. The 83 *Helitron*-like elements identified by homology search.

Identification	Accession	Species
Monocots	XP_037473821	*Triticum dicoccoides*
Monocots	EMS67201	*Triticum urartu*
Monocots	XP_040246309	*Aegilops tauschii*
Monocots	KAE8788045	*Hordeum vulgare*
Monocots	XP_010229584	*Brachypodium distachyon*
Monocots	BBF89892	*Oryza nivara*
Monocots	XP_015613561	*Oryza sativa Japonica*
Monocots	EEC77085	*Oryza sativa Indica*
Monocots	BBF89964	*Oryza rufipogon*
Monocots	XP_021302777	*Sorghum bicolor*
Monocots	ONL95700	*Zea mays*
Monocots	TVU44908	*Eragrostis curvula*
Monocots	XP_020100028	*Ananas comosus*
Monocots	ABF70031	*Musa acuminata*
Monocots	XP_039138709	*Dioscorea cayenensis*
Eudicots	XP_022004216	*Helianthus annuus*
Eudicots	KAI3684494	*Arctium lappa*
Eudicots	XP_023760549	*Lactuca sativa*
Eudicots	KAG5563862	*Rhododendron griersonianum*
Eudicots	XP_012841028	*Erythranthe guttata*
Eudicots	XP_027767532	*Solanum pennellii*
Eudicots	XP_042968932	*Carya illinoinensis*
Eudicots	XP_035539682	*Juglans regia*
Eudicots	XP_043615266	*Erigeron canadensis*
Eudicots	KAI3507904	*Cichorium endivia*
Eudicots	PON70500	*Trema orientale*
Eudicots	XP_030497675	*Cannabis sativa*
Eudicots	XP_027090168	*Coffea arabica*
Eudicots	XP_027167959	*Coffea eugenioides*
Eudicots	XP_024175667	*Rosa chinensis*
Eudicots	KAH0670706	*Solanum tuberosum*
Eudicots	XP_019181597	*Ipomoea nil*
Eudicots	XP_021864545	*Spinacia oleracea*
Eudicots	XP_045789330	*Trifolium pratense*
Eudicots	XP_009757892	*Nicotiana sylvestris*
Eudicots	XP_029148282	*Arachis hypogaea*
Eudicots	XP_019106803	*Beta vulgaris*
Eudicots	KAF1862196	*Lupinus albus*
Eudicots	KAH9608115	*Heliosperma pusillum*
Eudicots	XP_027911672	*Vigna unguiculata*
Eudicots	XP_028111973	*Camellia sinensis*
Eudicots	KAF7814059	*Senna tora*
Eudicots	XP_023916657	*Quercus suber*
Eudicots	XP_021747554	*Chenopodium quinoa*
Eudicots	KAD5961964	*Mikania micrantha*
Eudicots	XP_016510487	*Nicotiana tabacum*
Eudicots	XP_019261841	*Nicotiana attenuata*
Eudicots	XP_028752888	*Prosopis alba*
Eudicots	XP_017245330	*Daucus carota*
Eudicots	XP_024449084	*Populus trichocarpa*
Eudicots	XP_024630833	*Medicago truncatula*

Eudicots	XP_038693332	*Tripterygium wilfordii*
Eudicots	RZB57599	*Glycine soja*
Eudicots	XP_033137139	*Brassica rapa*
Eudicots	BAB02793	*Arabidopsis thaliana*
Eudicots	KAG7557012	*Arabidopsis suecica*
Eudicots	CAA7047626	*Microthlaspi erraticum*
Eudicots	KAH1254932	*Glycine max*
Eudicots	XP_013650967	*Brassica napus*
Eudicots	XP_026446488	*Papaver somniferum*
Mosses	KAH9559289	*Sphagnum magellanicum*
Mosses	KAG0630621	*Ceratodon purpureus*
Fungus	CAG8650759	*Claroideoglomus candidum*
Fungus	CAB5196742	*Rhizophagus irregularis*
Fungus	GBB95559	*Rhizophagus clarus*
Animal	XP_038075107	*Patiria miniata*
Animal	GBL92227	*Araneus ventricosus*
Animal	XP_034965595	*Viviparous lizard*
Animal	XP_021958962	*Folsomia candida*
Animal	XP_041441500	*Xenopus laevis*
Animal	XP_039754310	*Pararge aegeria*
Animal	KAE9523080	*Aphis glycines*
Oomycete	KAG6612311	*Phytophthora cinnamomi*
Nematodes	CDJ84245	*Haemonchus contortus*
Animal	XP_023214081	*Centruroides sculpturatus*
Animal	XP_045450619	*Melitaea cinxia*
Animal	XP_029728066	*Aedes albopictus*
Animal	XP_008555449	*Microplitis demolitor*
Animal	CAB3235853	*Arctia plantaginis*
Animal	XP_045477963	*Harmonia axyridis*
Animal	PAA68602	*Macrostomum lignano*
Animal	KYM96549	*Cyphomyrmex costatus*
Fungus	POS82697	*Erysiphe pulchra*
Red algae	OSX80228	*Porphyra umbilicalis*
Virus	AEE09607	*Cotesia vestalis bracovirus*

Extended Data Table 3. Detailed information related to the 49 *de novo Xuan1* insertions captured by PCIP-Seq.

From left to right: RIL sample name; chromosomal position of the *de novo Xuan1* insertion site in the Chinese Spring IWGSC RefSeq[6]; orientation of the *de novo* insertion; library of origin (LTS or RTS); chromosome, start, and end position of the 20 bp flanking host genomic sequence.

Sample	Chr:position	Ori	Lib	Chr	Start	End	Flanking host DNA seq (5'-3')
RE	1B:404773519	plus	LTS	1B	404773509	404773529	ATTCACACCATTCTTTATGG
RE	3A:460109534	plus	LTS	3A	460109524	460109544	TGCGTAGTTATGTATGCGCA
RE	3D:565375306	plus	LTS	3D	565375296	565375316	TAAAATCATTTTTCCTATAA
RE	5A:502158364	plus	LTS	5A	502158354	502158374	TAAAGATTTATATATAGGAA
RE	7B:14701923	plus	LTS	7B	14701913	14701933	TATCTATAAATATCTAGTCT
#3091	5D:470437802	plus	LTS	5D	470437792	470437812	ATTAGCCTATTATTGATAAT
#2483	2D:285512	plus	LTS	2D	285502	285522	CATACATACATACATAAATA
#2483	4A:84514810	plus	LTS	4A	84514800	84514820	ATTTAGAATATATAGTATGT
#2894	5D:246012192	plus	LTS	5D	246012182	246012202	AATTGGTTATAGCAAAGGTA
#2894	6B:534523048	plus	LTS	6B	534523038	534523058	ACCGAAGTTTAAGATGTAGT
#2894	7D:71544346	plus	LTS	7D	71544336	71544356	TGTTCATATATAGCTCATTA
#2885	2D:38764438	plus	LTS	2D	38764428	38764448	CTCCTGCTAATAATAGTAGT
#2885	2D:581406954	plus	LTS	2D	581406944	581406964	ATATCTATAATACTCTATAG
#2885	3D:1756269	plus	LTS	3D	1756259	1756279	GTTGGAAATATGGTCTAAAG
#2038	2D:155415132	plus	LTS	2D	155415120	155415140	AGTAGCAAAATNNNNNNNN
#2815	5A:407974758	plus	LTS	5A	407974748	407974768	GCAACCTTTATTATGTAGTC
RE	5A:454943288	plus	RTS	5A	454943279	454943299	CTGACACGTATACTATATAG
RE	Un:43310594	plus	RTS	Un	43310585	43310605	CTAGGCTTTATTATTCTAGT
#3091	3A:535003756	plus	RTS	3A	535003747	535003767	ACACAACATATTATTTTAGT
#2483	7A:14843130	plus	RTS	7A	14843121	14843141	GTCCTGCTACTATAGGTAGG
#2883	5A:665794321	plus	RTS	5A	665794312	665794332	AGGTAAAGAATTATATATAG
#2815	2A:544305974	plus	RTS	2A	544305965	544305985	TTTATGAACATACTATAAGT
#2815	5B:132290922	plus	RTS	5B	132290913	132290933	GCTGGGGATATATAGCCTTA
Ar	1B:171552833	minus	LTS	1B	171552825	171552845	GAGGTGAAGATTTATATATA
Ar	2D:445208049	minus	LTS	2D	445208041	445208061	CAAATACCATAAACTCATAA
Ar	5D:192266178	minus	LTS	5D	192266170	192266190	CGACGTCTCTTATAGAGGAT
Ar	7A:259256189	minus	LTS	7A	259256181	259256201	ACCCGAGATTAACCAGATAG
RE	2B:528574043	minus	LTS	2B	528574035	528574055	ATCAATAACACATATATCTA
RE	5A:76472095	minus	LTS	5A	76472087	76472107	CGATATAGTATTTATGAGTT
RE	7A:218026437	minus	LTS	7A	218026429	218026449	AAACTAGACTTACAAATTAG
RE	7D:548604417	minus	LTS	7D	548604409	548604429	ATACCTCACATATCAGATAT
#3091	3B:64309540	minus	LTS	3B	64309532	64309552	CTTGGTAACACATAATAATT
#3091	4A:148261036	minus	LTS	4A	148261028	148261048	CCTCAAGCTTACCATCGAGG
#3091	7B:588861232	minus	LTS	7B	588861224	588861244	TTTCTGTTGACATACGAATA
#2483	3A:131008225	minus	LTS	3A	131008217	131008237	GAGTCCTATATATATATATA
#2894	2D:544790320	minus	LTS	2D	544790312	544790332	CTACTGGTCATAATTTACTA
#2894	6D:456677245	minus	LTS	6D	456677237	456677257	CGGATTTACCTGTAATACCC

#2894	7A:63504052	minus	LTS	7A	63504044	63504064	TGGGCCTTTATTTCTTTATT
#2885	5D:399224574	minus	LTS	5D	399224566	399224586	ACAAGCGCTATTTAAGTAAG
#2885	7B:507952028	minus	LTS	7B	507952020	507952040	TGTCTCCTGATTAGTAGTAG
#2885	7D:168420688	minus	LTS	7D	168420680	168420700	CAAGAAGTTATATACCTTTA
#2038	5D:245375729	minus	LTS	5D	245375721	245375741	CTAGTAGTTTATGTTACAGT
#2815	2A:601710764	minus	LTS	2A	601710756	601710776	CCATAGGAATTATTCTAATG
#2815	4D:323092433	minus	LTS	4D	323092425	323092445	TCCTCGAGCTAATGTCAAAC
#2815	5D:50490230	minus	LTS	5D	50490222	50490242	TGGGGGAGTATATATAGTAT
RE	4B:236340475	minus	RTS	4B	236340466	236340486	TAATGATGCATATGTAGATA
#2483	3A:625437519	minus	RTS	3A	625437510	625437530	TAAAGAATTATATATAGGAA
#2815	3D:16015605	minus	RTS	3D	16015596	16015616	TTTATATACATATAGATATA
#2815	6B:663706606	minus	RTS	6B	663706597	663706617	GCACAGACGATGTAGTACTA

References

1. Lisch, D. How important are transposons for plant evolution? Nature Reviews Genetics 14, 49–61 (2013).
2. Kapitonov, V. V. & Jurka, J. Rolling-circle transposons in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America 98, 8714–8719 (2001).
3. Barro-Trastoy, D. & Köhler, C. Helitrons: genomic parasites that generate developmental novelties. Trends in Genetics (2024) doi:10.1016/j.tig.2024.02.002.
4. Papon, N. et al. All families of transposable elements were active in the recent wheat genome evolution and polyploidy had no impact on their activity. The Plant Genome 16, e20347 (2023).
5. Vitte, C., Fustier, M. A., Alix, K. & Tenaillon, M. I. The bright side of transposons in crop evolution. Briefings in Functional Genomics and Proteomics 13, 276–295 (2014).
6. The International Wheat Genome Sequencing Consortium (IWGSC) et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361, eaar7191 (2018).
7. Naito, K. et al. Dramatic amplification of a rice transposable element during recent domestication. Proceedings of the National Academy of Sciences 103, 17620–17625 (2006).
8. Wicker, T. et al. Transposable Element Populations Shed Light on the Evolutionary History of Wheat and the Complex Co-Evolution of Autonomous and Non-Autonomous Retrotransposons. Advanced Genetics 3, 2100022 (2021).
9. Klein, S. P. & Anderson, S. N. The evolution and function of transposons in epigenetic regulation in response to the environment. Current Opinion in Plant Biology 69, 102277 (2022).
10. Gebrie, A. Transposable elements as essential elements in the control of gene expression. Mobile DNA 14, 9 (2023).
11. McClintock, B. Controlling Elements and the Gene. Cold Spring Harbor Symposia on Quantitative Biology 21, 197–216 (1956).
12. Thomas Jainy, Pritham Ellen J., Chandler Mick, & Craig Nancy. Helitrons, the Eukaryotic Rolling-circle Transposable Elements. Microbiology Spectrum 3, 3.4.03 (2015).
13. Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. Annual Review of Genetics 54, 539–561 (2020).
14. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proceedings of the National Academy of Sciences 111, 10263–10268 (2014).
15. Yang, L. & Bennetzen, J. L. Structure-based discovery and description of plant and animal Helitrons. Proceedings of the National Academy of Sciences 106, 12832–12837 (2009).
16. Yang, L. & Bennetzen, J. L. Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proc. Natl. Acad. Sci. U.S.A. 106, 19922–19927 (2009).

17. Li, S.-F. et al. Landscape and evolutionary dynamics of Helitron transposons in plant genomes as well as construction of online database HelDB. Journal of Systematics and Evolution 61, 919–931 (2023).
18. Mirouze, M. et al. Selective epigenetic control of retrotransposition in Arabidopsis. Nature 461, 427–430 (2009).
19. Ito, H. et al. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature 472, 115–120 (2011).
20. Liu, P., Cuerda-Gil, D., Shahid, S. & Slotkin, R. K. The Epigenetic Control of the Transposable Element Life Cycle in Plant Genomes and Beyond. Annu. Rev. Genet. 56, 63–87 (2022).
21. Lisch, D. Epigenetic Regulation of Transposable Elements in Plants. Annu. Rev. Plant Biol. 60, 43–66 (2009).
22. Grabundzija, I. et al. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nature Communications 7, 10716 (2016).
23. Grabundzija, I., Hickman, A. B. & Dyda, F. Helraiser intermediates provide insight into the mechanism of eukaryotic replicative transposition. Nature Communications 9, 1278 (2018).
24. Thieme, M. et al. Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding. Genome Biology 18, 1–10 (2017).
25. Chao, S.-H., Greenleaf, A. L. & Price, D. H. Juglone, an inhibitor of the peptidyl-prolyl isomerase Pin1, also directly blocks transcription. Nucleic Acids Research 29, 767–773 (2001).
26. Zhou, L. et al. Zebularine: A Novel DNA Methylation Inhibitor that Forms a Covalent Complex with DNA Methyltransferases. Journal of Molecular Biology 321, 591–599 (2002).
27. Horváth, V., Merenciano, M. & González, J. Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. Trends in Genetics 33, 832–841 (2017).
28. Zhang, P., Peng, H., Llauro, C., Bucher, E. & Mirouze, M. ecc_finder: A Robust and Accurate Tool for Detecting Extrachromosomal Circular DNA From Sequencing Data. Frontiers in Plant Science 12, (2021).
29. Peng, H., Mirouze, M. & Bucher, E. Extrachromosomal circular DNA: A neglected nucleic acid molecule in plants. Current Opinion in Plant Biology 69, 102263 (2022).
30. Walkowiak, S. et al. Multiple wheat genomes reveal global variation in modern breeding. Nature 588, 277–283 (2020).
31. Cavrak, V. V. et al. How a Retrotransposon Exploits the Plant's Heat Stress Response for Its Activation. PLoS Genetics 10, (2014).
32. Kosek, D. et al. The large bat Helitron DNA transposase forms a compact monomeric assembly that buries and protects its covalently bound 5′-transposon end. Molecular Cell 81, 4271-4286.e4 (2021).
33. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
34. Zhu, T. et al. Optical maps refine the bread wheat Triticum aestivum cv. Chinese Spring genome assembly. The Plant Journal 107, 303–314 (2021).

35. Aury, J.-M. et al. Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. GigaScience 11, giac034 (2022).
36. Tang, L. et al. GWAS reveals determinants of mobilization rate and dynamics of an active endogenous retrovirus of cattle. Nature Communications 15, 2154 (2024).
37. Maccaferri, M. et al. Durum wheat genome highlights past domestication signatures and future improvement targets. Nature Genetics 51, 885–895 (2019).
38. Zhu, T. et al. Improved Genome Sequence of Wild Emmer Wheat Zavitan with the Aid of Optical Maps. G3 Genes|Genomes|Genetics 9, 619–624 (2019).
39. Ling, H.-Q. et al. Genome sequence of the progenitor of wheat A subgenome Triticum urartu. Nature 557, 424–428 (2018).
40. Wang, L. et al. Aegilops tauschii genome assembly Aet v5.0 features greater sequence contiguity and improved annotation. G3 Genes|Genomes|Genetics 11, jkab325 (2021).
41. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. Proceedings of the National Academy of Sciences 101, 12404–12410 (2004).
42. Rabanus-Wallace, M. T. et al. Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential. Nature Genetics 53, 564–573 (2021).
43. Watts, J. E., de Villiers, O. T. & Watts, L. Sterilization of wheat seeds for tissue culture purposes. South African Journal of Botany 59, 641–642 (1993).
44. Jannatul Ferdous. A quick DNA extraction protocol: Without liquid nitrogen in ambient temperature. African Journal of Biotechnology 11, 6956–6964 (2012).
45. Roquis, D. et al. Genomic impact of stress-induced transposable element mobility in Arabidopsis. Nucleic acids research 49, 10431–10447 (2021).
46. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).
47. Shiraishi, Y. et al. Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. Nucleic Acids Research gkad526 (2023) doi:10.1093/nar/gkad526.
48. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology 20, 1–18 (2019).
49. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE 11, e0163962 (2016).
50. Katoh, K., Asimenos, G. & Toh, H. Multiple Alignment of DNA Sequences with MAFFT. in Bioinformatics for DNA Sequence Analysis (ed. Posada, D.) 39–64 (Humana Press, Totowa, NJ, 2009). doi:10.1007/978-1-59745-251-9_3.
51. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
52. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
53. Pongor, L. S. et al. BAMscale: quantification of next-generation sequencing peaks and generation of scaled coverage tracks. Epigenetics & Chromatin 13, 21 (2020).

54. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29, 644–652 (2011).
55. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).
56. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357–359 (2012).
57. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).
58. Gasteiger, E. et al. Protein Identification and Analysis Tools on the ExPASy Server. in The Proteomics Protocols Handbook (ed. Walker, J. M.) 571–607 (Humana Press, Totowa, NJ, 2005). doi:10.1385/1-59259-890-0:571.
59. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792–1797 (2004).
60. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973 (2009).
61. Minh, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37, 1530–1534 (2020).
62. Darrier, B. et al. High-Resolution Mapping of Crossover Events in the Hexaploid Wheat Genome Suggests a Universal Recombination Mechanism. Genetics 206, 1373–1388 (2017).
63. Andrews, S. A Quality Control Tool for High Throughput Sequence Data [Online]. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
64. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12 (2011).
65. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
66. Danecek, P. et al. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008 (2021).

Discussion - Perspectives

Starting from in-house generated WGS of cattle (DAMONA) (Lee et al., 2021), we established a catalogue of polymorphic ERV in the Holstein Friesian (HF) population which comprises more than 1,200 sites. We developed array-based assays to interrogate a subset of these polymorphic insertion sites in routine genotyping conducted for genomic selection and genetic testing for defects. These assays were released as public probes in the last version of the bovine EuroGenomics custom arrays (Illumina). By detecting germline *de novo* ERV insertions in the same large extended bovine pedigree (DAMONA), we have shown evidence that ERVK[2-1-LTR], one subfamily of ERVs, is still mobile in both the male and female germline of cattle. We developed a novel, capture-based sequencing method which allows us to directly and reproducibly measure the rate of ERVK[2-1-LTR] mobilization in a sample of DNA. We demonstrated that the method is robust and quantitative. Applying this method to sperm DNA extracted from 430 bulls, we showed that the mobilization rate averages 1 per ~150 sperm cells, but varies more than ten-fold between bulls. We further showed that this rate behaves like a molecular trait that differs dramatically between individuals yet remains stable within an individual over many years. We generated a comprehensive ERVK[2-1-LTR] catalogue for the bovine encompassing around 300 full-length elements that are segregating in the 430 bulls, of which 15% are ~10 Kb *C*ompetent type (*C*-type) autonomous elements encoding *GAG*, *PRO*, *POL* and *ENV* genes, and 85% are ~6.8 Kb *D*efective type (*D*-type) non-autonomous elements devoid of functional genes. In addition, we mapped more than 3,600 germline *de novo* ERVK[2-1-LTR] insertions that occurred *in vivo* in the male germline. These insertions are not randomly distributed along the genome, but are enriched in subtelomeric and GC-rich regions. We uncovered the genetic factors underlying the observed inter-individual heterogeneity in the rate of ERVK[2-1-LTR] mobilization by performing a genome wide association study. We showed that more than one quarter of the variance of ERVK[2-1-LTR] mobilization rate is determined by the number of *C*-type elements in the genome of the individual, which positively correlated with rate. We also showed evidence for an epistatic effect between non-intact and intact *C*-type elements on ERVK[2-1-LTR] mobilization rate. Of note, there was no evidence from the GWAS for emerging host silencing mechanisms (i.e. derived alleles) decreasing mobilization rate. We showed that *de novo* ERVK[2-1-LTR] insertions are dominated by *D*-type elements, suggesting that these elements are hijacking the machinery of *C*-type elements. These findings suggest that the ERVK[2-1-LTR] elements may self-inactivate in the future by a parasite (*D*-type elements)-of-parasite (*C*-type elements) gene drive. We successfully, and for the first time, captured active DNA transposon *Helitron* in the large wheat genome *in vivo*, using the method developed for ERVK[2-1-LTR], which shows that the method can be used to study other active transposons in various species.

4.1 Germline mutation rate as a heritable and quantitative trait

What defines us as individuals is how we differ from each other. We often ask how tall a person is. Height is one of many quantifiable trait. Most of the inter-individual differences for a trait are not

stochastic. Traits are largely determined by heredity and environment. The genomes we inherit from our parents provide a set of instructions for development, and the environment in which we develop profoundly shapes our identity. Understanding how the genome inherited from our parents contributes to our phenotypic differences is one of the core missions of genetics. Recent advances in genomic techniques have accelerated the progress in understanding the molecular mechanisms underlying disease and traits of importance for animal production. This may contribute to the development of means of disease diagnosis and treatment, and means of selection for livestock. To identify the variants in the genome associated with traits, we generally divide the problem into two parts: (i) examining difference in the individuals' genomes, and (ii) measuring the phenotype of interest. All of us inherit one genome copy from our father and one genome copy from our mother. These are very similar, but not identical. Aligning these two genomes would reveal a different base pair approximately every 1,000 sites. This type of differences are referred to as single nucleotide variants (SNVs), of which there are around 3.9 million per diploid human genome (Ebert et al., 2021). Compiling all SNVs at the population level reveals hundreds of millions of SNVs.

In animal breeding, we are interested in understanding the genetics of the traits related to animal productivity and fertility, such as carcass weight of meat-producing animals, milk yield of dairy cows and egg production of hens (Georges et al., 2019). Recently, cellular traits including gene expression level, protein expression level, etc. have been examined using similar forward genetic approaches (Aguet et al., 2017). Such studies provide insights in gene regulation and molecular mechanisms underlying traits of interests. All types of genetic variants mentioned above originate from germline *de novo* mutations (DNMs). This ultimate source of genetic variants provides the substrate for both Darwinian and artificial selection. One aspect of the process of germline *de novo* mutation is its rate, that is how frequent a mutation occurs in the germline for instance per gamete. It is very appealing to treat the mutation rate as a cellular trait, to examine whether and how it varies between individuals.

The first line of evidence of mutation rate variation has come from population genetic studies that have found difference in the mutation spectrum between human populations. For example the TCC > TTC mutation rate is increased in Europeans compared to other populations (Harris and Pritchard, 2017). These differences might be explained by population-specific genetic variation in genes involved in this type of mutation. Consistent with this notion, Sasani et. al. identified genetic variants in the *Mutyh* gene associated with a 50% higher C>A mutation rate in a panel of recombinant inbred mouse lines, providing evidence that common genetic variants may modulate germline mutagenesis in mammalian species (Sasani et al., 2022). Moreover, significant differences in the magnitude of the paternal age effect on DNMs between families have been reported in two independent studies (Rahbari et al., 2016; Sasani et al., 2019). The number of mutations gained per year ranged from 0.19 to 3.24 among 40 human families (Sasani et al., 2019). Despite the small number of data points underpinning these observations, it

suggests that the DNM rate could be subject to family-specific factors, including the parental genetic make-up. Of note, most studies conducted in humans used two-generation trios, which allows to trace the parent-of-origin for only a fraction (~ 20%) of DNMs. Moreover, these studies include some DNMs that occurred in the early development of the offspring (rather than transmitted by gametes). These factors complicate the estimation of individual mutation rates. In addition, in typical studies in human, only one offspring per couple is sequenced, thus only one oocyte of the mother and one sperm of the father are evaluated. It is therefore not possible to estimate the "repeatability" of the number of transmitted DNMs (i.e. is there any evidence for an individual effect, whether due to genetics and/or permanent environmental effects). All these issues must be addressed when studying single nucleotide mutation rate as a heritable trait in order to have a chance to identify mutator and/or anti-mutator alleles in human populations.

To study the process of DNM in the bovine germline, Harland et. al. (2017) sequenced the whole genome of ~750 individuals constituting 131 sire-dam-offspring trios with an average of five grand-offspring each. This unique design enabled them to (i) assign parent-of-origin for nearly all DNMs, (ii) classify DNMs according to the developmental stage at which it occurred (showing distinct features and implying variable molecular mechanisms), and (iii) obtain single nucleotide mutation rates for 131 sperms and 131 oocytes. They estimated the "repeatability" of mutation rate estimates by exploiting the fact that multiple gametes were examined for some sires and dams. They found outlier individuals with abnormal number of DNM of specific classes, and suggested genetic explanations for some. This study strongly argues that to fully understand the process of mutation it is important to understand and study the interactions between DNMs and the development of the germline.

Although most studies mentioned above primarily focus on SNVs and small INDELs, there are a few studies that have expanded their analyses into other types of DNMs (Kloosterman et al., 2015; Mitra et al., 2021; Turner et al., 2017). Short tandem repeats (STRs), consisting of repeated sequence motifs of 1-6 base pairs, exhibit high mutation rates. STR mutations result in the expansion or contraction of the number of repeats, some of which are well known to cause human disease, such as Huntington's disease. Using more than 6,000 Icelandic parent-offspring trios, Snaedis et, al. (2023) estimated the *de novo* STR mutation rate per generation at 63.7 on average (Kristmundsdottir et al., 2023). Notably, they found two independent coding variants, respectively in *MSH2* (a mismatch repair gene) and *NEIL2* (a DNA damage repair gene), that increase the number of parentally transmitted *de novo* STRs (Kristmundsdottir et al., 2023). This discovery suggests that the STR mutation rates is partially under genetic control in humans. Similarly, mutations in the murine *MSH3* gene are associated with STRs expansion (Maksimov et al., 2023). Inheritance of such defective mismatch repair genes results in cancer predisposition syndromes, most notably Lynch syndrome, which causes increased susceptibility to a number of cancers including early onset colorectal cancer and endometrial cancer (Peltomäki, 2003). These studies

demonstrate that the STRs mutation rate is a heritable trait, and that it is feasible to discover associated genetic factors, providing striking examples of the influence of inherited variants on germline mutation properties. Structural variants (SVs) are defined as variants that affect > 50 bp. They include deletions, duplications, inversions, insertions, translocations and complex variants that combine more than one of these. *De novo* SVs are shown to occur around hundred-fold less frequently than point mutations (Belyeu et al., 2021; Lee et al., 2023). The inherent challenge of accurately identifying SVs further complicates their study. In our study, we also observed potential inter-individual difference in ERV activity as three out of five *de novo* events occurred in the germline of one bull. However, despite the fact that this ERV family is among the most active mammalian TEs and the relatively large number of sequenced trios, it was not possible to rigorously quantify the inter-individual variation in mobilization rate from this pedigree alone. Thus, the method we have developed to quantitatively estimate the individual rate of TE mobilization, without relying on observations in offspring could mitigate the challenges. We argue that this method could be widely applied to any active TEs in any species both in the germline and soma, opening up possibilities to study TE transposition rates as a quantitative trait amenable to genetic analysis.

4.2 Timing of mobilization and mosaicism

The integration events of ERV captured in sperm DNA may have occurred any time before the formation of spermatozoa. The fraction of cells affected by the insertion (hence degree of mosaicism) depends on the mutational timing. Insertions occurring very early in embryogenesis will affect a large fraction of cells, while insertions arising after tissue differentiation will only affect a subset of cells of that specific tissue. The degree of mosaicism measured for each integration site was used to infer the timing of mobilization. There are several observations in our study that inform about the time of mobilization. Firstly, the data suggest that the five germline *de novo* insertions occurred late during gametogenesis. Specifically, (i) the allelic dosage in the probands is very close to 50% suggesting that they are all "constitutive" heterozygotes; (ii) the five insertions are transmitted to the next generations in complete linkage with a parental haplotype; and (iii) none was detectable in parental DNA. The first two arguments argue against the possibility of an early embryonic origin of those five insertions in the probands, while the third rules out an early event in the parents. Those *de novo* ERV insertions present distinct features compared to some *de novo* SVs that occurred during early embryonic development and detected in the same dataset (Lee et al., 2023). Secondly, the *de novo* ERV mobilization rate in sperm did not increase with age. This is in contrast with mutation rates for SNP and indels which increase with paternal age. The paternal age effect on mutation rate is in part attributed to the increased number of replications in the germline and hence increased number of replicative errors. No significant increase in *de novo* ERV rate with age is compatible with the fact that ERV mobilization primarily occurs before puberty. This is in line with the observation that a fraction of *de novo* ERV insertions are found both in

young and old sperm samples of the same bull. This suggests that ERVs insertions are already present in spermatogonia stem cells (SSCs) which will give rise to sperm throughout reproductive life. Lastly, the frequency distribution of resampling rate of *de novo* insertions detected by deep (6-9 times) PCIP experiments on three bulls informs us about the developmental window during which *de novo* ERV mobilization is most likely to have occurred. We compared the real distribution, with the distribution obtained by simulations conducted under various timing scenarios, yet matching the real data for *de novo* insertion frequency (average number of *de novo* insertions per sperm cell) and number of explored haploid genomes. The results are compatible with *de novo* mobilization occurring during a window of ~5-9 consecutive cell divisions during the second half (in terms of number of cell divisions) of spermatogenesis. Intriguingly, this may coincide with the second wave of epigenome reprogramming during which methylation, histone modification and chromatin states are largely erased and rebuilt. In summary, the available evidence points towards a peak of mobilization of ERVs in the male cattle germline during a narrow developmental window between primordial germ cell (PGC) proliferation and the formation of spermatogonial stem cells (SSCs).

However, caution is needed when interpreting these results. The initial number of PGCs and what fraction of PGCs have actually contributed to cells that give rise to SSCs is variable and dynamic. The process from PGCs to SSCs is not a simple, continuous series of mitoses but is rather dynamic and affected by several intrinsic and extrinsic factors. Actually, the precursor cells undergo proliferation and elimination during embryonic development. So, selection at this step could be highly influential in shaping the genetic pool of final gametes. Apoptosis is a consistent feature of PGC development. Indeed, blocking apoptotic pathways *in vivo* can cause infertility (Aitken et al., 2011; Knudson et al., 1995). Mouse male germ cells, for instance, exhibit two waves of apoptosis during development. The first occurs around 13 days of gestation, coinciding with the time of migration of PGCs into the gonads. A second peak is observed around 10 days after birth when the first wave of spermatogenesis has started and active spermatogonial proliferation is ongoing (Ueno et al., 2009; Wang et al., 1998). Lineage tracing confirmed that mice male germ cells die as clones, suggesting shared intrinsic determinants. Intriguingly, transcriptional activity of retrotransposons (LINE1 in this case) plays a role in this process (Nguyen et al., 2020). It is tempting to examine to what extent the final pool of germ cells is affected by the process of ERV mobilization.

Clonal selection occurs in spermatogonia, where it has been termed “selfish spermatogonial selection”. Although most new mutations are neutral or deleterious to sperm, certain mutations may confer selective advantages. It has been shown that mutations in a limited number of human genes, among which *FGFR2*, *FGFR3* and *HRAS*, can give the sperm cell a competitive advantage (Goriely and Wilkie, 2012). This leads to an unusually high number of sperm cells carrying these mutations and an increase in their proportions with paternal age. Of note, the degree of mosaicism for the very same *de novo* ERV

insertions that were captured in this study in young and old semen samples did not change over time. This supports the notion that bull spermatogonial stem cells carrying those insertions do not dramatically expand.

4.3 How are the ERVK[2-1-LTR] multiplying?

This study also raises an important question: how do the ERVK[2-1-LTR] mobilize *in vivo*. Studies of several ERV families in mice have led to the identification of coding-competent copies that are able to mobilize *in vivo*. Whereas some ERVs behave as *bona fide* retroviruses, undergoing a replicative cycle that includes the generation of extracellular viral like particles (VLPs) and cell entry through reinfection, others elements exhibit a strictly intracellular amplification cycle, with VLPs accumulating inside the cell, leading to an increase of copy numbers through retrotransposition. Two groups have independently reconstructed a putative ancestral "progenitor" element for one of the most recently amplified ERV families (HERVK-HML2) in human by deriving a consensus sequence from extant elements (Dewannieux et al., 2006; Young and Bieniasz, 2007). The reconstructed element behaved like a retrovirus producing VLPs and amplified via an extracellular pathway involving reinfection. These efforts were in agreement with evolutionary analyses that suggested that the proliferation of this family was due to germline reinfection (Belshaw et al., 2004). Comparing the sequence of an infectious progenitor murine IAP (IAPE) with that of the more abundant IAP elements (lacking ENV) suggested that it was the alteration of a signal peptide targeting the GAG protein to the plasma membrane, followed by the decay of the ENV coding sequence, that caused the transition from extracellular reinfection to intracellular retrotransposition with intracellular sequestration of VLPs (Ribet et al., 2008, 2007). Accordingly, replacement of the N-terminal GAG domain of IAP and MusD by that of infectious retroviruses restored the targeting of VLPs to the plasma membrane and their release in the cell supernatant. These particles could be made infectious by pseudotyping with a functional Env protein, thus reconstructing a *bona fide* retrovirus. Conversely, the IAPE replicating via reinfection could be converted into a functional 'intracellularized' element by modifying its N-terminal GAG domain (Ribet et al., 2007, 2008). We have shown that around half of *C*-type ERVK[2-1-LTR] elements have intact ORF for GAG, PRO, POL and ENV. Combined with the *in silico* prediction that the matrix domain of the GAG protein is a target for N-terminal myristoylation, these strongly suggest that ERVK[2-1-LTR] elements still mobilize by within-host intercellular reinfection rather than by the supposedly more effective intracellular retrotransposition route (Magiorkinis et al., 2012). To the best of our knowledge, how intercellular mobilization of VLPs results in germline *de novo* insertions has not been elucidated *in vivo* in any mammalian species. Nevertheless, germ cell to germ cell and somatic cell to germ cell transmission of VLPs of LTR retrotransposons (Errantiviruses; sharing structural and functional characteristics with vertebrates ERV) resulting in germline insertions has been demonstrated in *Drosophila* ovaries (Wang et al., 2018; Yoth et al., 2023). ENV independent microtubule-mediated

transport was shown to underpin germ cell to germ cell transmission (Chalvet et al., 1999; Wang et al., 2018). It would be very interesting to determine which cells produce the VLPs of ERVK[2-1-LTR] (soma, germline or both), which membrane receptor ERVK[2-1-LTR] VLPs recognize in the recipient germ cells, whether they enter cells using an ENV dependent pathway, or which reinfection path they use (free particles, virological synapses or microtubule-mediated transport).

4.4 Where are the brakes?

ERV clades arise by retroviral infection of the germline followed by expansion as endogenized proviruses generate copies of themselves by reinfection and/or retrotransposition. The ERV activity in the germline eventually fades, leaving tens to hundreds of degraded copies of each ERV clade in the genome. It is assumed that this entails the emergence of host defense systems arising *de novo* and/or by re-cycling available cellular machinery. Indeed, as ERVs settle, expand and then regress in a species' genome, the rate of ERV mobilization is bound to be an evolving phenotype as a result of this tug-of-war. We expected that our GWAS studies on this mobilization phenotype would inform us about the brakes against ERV mobilization assuming two possible situations (not mutually exclusive). First, silencing machineries such as KRAB-ZFPs and piRNA pathways against mobilization of this family of ERV are still evolving. Modifiers of epigenetic modification and expression of specific ERV elements have been mapped in mice and polymorphisms in KRAB-ZFP clusters were nominated as plausible candidates (Bertozzi et al., 2020). The second possible situation is that the silencing machinery has been already established and that derived genetic mutations in component of silencing machinery may disrupt this machinery, which results in derepression of ERV activity. Against expectations, we didn't map any brakes related to either of the two situations. Instead, we have shown that the copy number of competent ERVK[2-1-LTR] elements is positively associated with mobilization rate. With hindsight, this is a rather trivial result. The discrepancy between expectation and observation could be explained by many possibilities. This is possibly because sample size of the cohort is too small to detect other genetic effects with minor contributions to the phenotype. As shown by the outcome of knocking down host defense mechanisms against transposable elements, excessive ERV mobilization rates may compromise fertility (Zamudio and Bourc'His, 2010). Our GWAS study failed to detect these effects because the bulls we worked with are healthy in terms of fertility. By applying PCIP-seq on animals with extreme phenotypes (i.e. subfertility) we might be able to map this type of genetic effects using GWAS.

Our findings point towards a possible self-regulating mechanism that may contribute to the demise of ERVs. In this, *D*-type elements would outcompete *C*-type elements in generating *de novo* insertions and act as parasite-of-parasite gene-drive that may cause the spontaneous implosion of the ERVK[2-1-LTR] clade. Defective elements complemented in *trans* by competent elements are a common phenomenon for ERVs and non-LTR retrotransposons. Alu and SVA elements in the human genome, for instance, copy themselves using the proteins encoded by retrotransposition-competent LINE1 elements

(Dewannieux et al., 2003). But it seems that, in contrast to the ERV elements studied in this thesis, LINE1 elements preferably use their encoded RNA as a template for reverse transcription and make new copies in *cis*, hence rendering a gene-drive mode of action of Alu/SVA less likely (Kulpa and Moran, 2006). The essential requirement is therefore that *D*-type elements use the mobilization machinery encoded by *C*-type elements in *trans* more effectively that the *C*-type elements themselves. The mechanisms that would confer this advantage are currently unknown. A number of hypotheses are interesting to consider. The *D*-type genomic RNAs (gRNAs) could be more abundant in cells than that of *C*-type because some of them are transcribed at a higher rate and/or more stable. Also, *C*-type gRNAs serve two functions, as a translation template and as a viral genome. A subset of *C*-type gRNAs will be used for splicing and translation and this could advantage *D*-type elements. Another possible explanation is that *D*-type gRNAs can utilize the machinery provided by *C*-type elements more efficiently (i.e. be preferentially packaged into viral like particles) due to the fact that they are smaller (~3kb difference) or have a different RNA secondary structure that favors their recognition and packaging (Nikolaitchik et al., 2013). Other steps where *C*- and *D*-elements may be differentiated include RNA dimerization, RNA modification, and binding to Gag or cellular splicing factors or ribosomes.

Examples of complementation in *trans* between morphs of the same ERV clade include ETn by MusD (Mager and Freeman, 2000; Ribet et al., 2004) and IAP IΔ1 by IAP (Saito et al., 2008) in mice, RecKoRV by KoRV in koala, and possible type 1 by type 2 HERV-K in human. There are likely other pairs of ERVs akin to *C/D*-type bovine ERVK[2-1-LTR] in other species. Strikingly, the bovine ERVK[2-1-LTR] *C/D* pair, murine MusD/ETn pair and koala KoRV/RecKoRV pair share a pattern of swapping of a central segment of the competent ERV element with an old piece of retroelement, yet conservation of flanking sequences encompassing portions of the *GAG* and *ENV* (ERVK[2-1-LTR] and KoRV) or *GAG* and *POL* (MusD) genes (in addition to the LTRs), which may be suggestive of *cis*-effects on the effectiveness of *trans*-complementation as reported for IAP IΔ1 and ETn.

The observation of the *C/D*-type and similar pairs of ERV elements in other species raises a number of interesting questions. First, how did the *D*-type elements form? RecKoRV derives from the recombination between KoRV and another retroelement that has no intact protein coding regions. This is probably also the case for bovine *D*-type ERVK[2-1-LTR] and mice *D*-type ETn, which were formed by recombination between corresponding *C*-type elements with non-related retroelements. This is supported by the fact that internal pieces of *D*-type ERVK[2-1-LTR] elements contain sequences showing little homology with retroelements. A second interesting question is why *D*-type elements were retained in the genome? The fact that *D*-type elements can mobilize by *trans*-complementation suggests that the retained sequences with homology to *C*-type elements are required. The packaging signal of retroviruses, for example, maps just downstream to the 5'LTR, and is required for selective

interaction and packaging of the GAG protein, securing the formation of viral like particles for further mobilization. However, studies of KoRV, that is presently undergoing endogenization in Koala, found at least 16 forms of recombinant KoRVs that appear to have arisen independently, suggesting that recombination with other retroelements is common. We speculate that there has been more than one form of *D*-type elements ERVK[2-1-LTR] that arose by various recombination events, but that the form described in our work out-competed the others by hijacking *C*-type elements during evolution. Taking together, *D*-type elements arose from recombination with existing retroelements and may act as brakes that may constraint post-endogenization copy numbers and even may lead to the clade's demise.

4.5 Functional impact of polymorphic ERVs

Ongoing germline mobilization of ERVs generates insertional polymorphisms in current cattle populations. We have established a catalogue of polymorphic endogenous retroviruses (pERVs) by mining a dataset of whole genome sequences (WGS) of Holstein Frisian cattle. We showed that those pERVs are significantly depleted from genic regions. Moreover, when genic, a significant bias against sense insertion was observed. The depletion from genic regions was not observed for *de novo* ERV insertions, suggesting that genomic distribution was shaped by post-integration purifying selection against deleterious genic insertions. Indeed, we showed that one ERVK[2-1-LTR] insertion in exon of *APOB* gene generates a deleterious mutation underpinning a genetic defect in this population. We have identified other examples of exonic insertions that disrupt canonical gene expression in the catalogue of PCIP detected polymorphic ERVK[2-1-LTR] in the Belgian Blue cattle population. For instance, a rare antisense pERV insertion in the coding exon 6 of the *Centrosomal Protein 126* gene (*CEP126*) leads to allele specific transcriptional arrest. It has been reported that intronic pERVs insertions in mouse can disrupt transcript processing, notably splicing and poly-adenylation (Gagnier et al., 2019). The LTR contains promoter and *cis*-regulatory elements which can lead to increased or altered expression of nearby genes (Zhou et al., 2021).

To fully assess *cis*-regulatory effect of pERVs on transcriptomics, we generated RNA-seq data comprising multiple tissues including skeletal muscle, liver, blood and testis of two cattle breeds. Genotypes at pERV loci were obtained for all samples using a combination of direct array-genotyping and/or pERV imputation using our reference WGS dataset. At each pERV locus, systematic *cis*-perturbation analyses will be performed, including differential expression resulting in allelic imbalance, alternative promoter usage, alternative splicing with exon-skipping or exon-capture, or creation of new poly-adenylation signal. A comprehensive analysis of the species-specific re-wiring potential of these young ERV insertions is expected.

The profound phenotypic effects of pERVs in other livestock species have been documented in the literature. For instance, the henny feathering trait in chicken is associated with the insertion of an intact

avian leukosis virus in the 5'UTR of *CYP19A1* (Li et al., 2019). Insertion of an endogenous Jaagsiekte sheep retrovirus element into the *BCO2* gene abolishes its function and leads to yellow discoloration of adipose tissue in sheep (Kent et al., 2021). With the ability to directly array-genotype pERVs and the possibility to incorporate this information into future genetic analyses, we could examine their associations with any trait of interest and pave the way for their potential use in animal breeding.

4.6 Somatic ERV activation and immune response

In this study, we focused on the germline activity of ERV elements in cattle. In somatic cells, ERVs are typically kept in check by epigenetic silencing mechanisms, such as DNA methylation and histone modifications. However, under certain conditions such as aging, diseases, or environmental stresses, these regulatory mechanisms may fail, leading to the reactivation of ERVs. For instance, HERV-K elements can be unlocked to transcribe viral genes and even produce retrovirus-like particles in human senescent cells (Xiaoqian Liu et al., 2023). ERV activation has also been observed in the organs of aged primates, mice, and human (Zhang et al., 2023).

Exogenous retroviral infection has been found to induce HERV transactivation; for example, HIV-1 infection activates HERV-K (Contreras-Galindo et al., 2013). Nevertheless, both innate and adaptive immune responses against ERV activation have been observed (Kassiotis, 2023). Similar to exogenous retroviruses, individual steps in the replication cycle of ERVs produce nucleic acid intermediates capable of engaging innate immune sensors. Single-stranded RNA (ssRNA) produced by ERV transcription can be detected by ssRNA sensors like Toll-like receptor 7 (Yu et al., 2012). Reverse transcription of ERV genomes may also trigger innate sensors that detect cytosolic DNA (Lima-Junior et al., 2021). Additionally, bidirectional expression of ERVs can generate complementary RNA, forming double-stranded RNA that triggers immune responses (Kassiotis, 2023).

As part of the host's genetic material, proteins encoded by ERVs are generally tolerated by the adaptive immune system. However, specific features of ERV proteins can significantly alter their immunogenic or tolerogenic properties (Kassiotis, 2023). For example, the formation of particles by specific HERV-K proviruses can enhance the immunogenicity of HERV-K proteins. The most frequently targeted ERV-encoded proteins in humans belong to the HML-2 family, which includes the most recent and intact proviruses. As a result, it is challenging to determine which specific copy induces a T cell or B cell response. Nevertheless, antibody responses to HML-2 envelope glycoprotein and Gag precursor protein have been observed in several cancers (Ng et al., 2023; Wang-Johanning et al., 2008).

Looking ahead, it will be interesting to investigate whether the ERVs are transcribed and/or remain transpositionally active in particular somatic tissues using the method developed for germline samples.

It will also be tempting to examine how the host immune system responds to their activation and how this interaction impacts host health, if at all.

ERVs comprise a significant portion of vertebrate genomes, yet we have only recently begun to understand how these elements impact genome evolution and biological function. Enabled by the method developed in this thesis, we studied one of the most intrinsic aspects of ERVs – germline transposition – in cattle in detail, and investigated how and why transposition rates can vary between individuals. This study provides us with a snapshot of the dynamics of ERV transposition post endogenization. The framework and method are expected to be widely applied to study other active families of transposable elements.

References

Adelson, D.L., Raison, J.M., Edgar, R.C., 2009. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. Proc. Natl. Acad. Sci. U. S. A. 106, 12855–12860. https://doi.org/10.1073/pnas.0901282106

Adrion, J.R., Song, M.J., Schrider, D.R., Hahn, M.W., Schaack, S., 2017. Genome-wide estimates of transposable element insertion and deletion rates in drosophila melanogaster. Genome Biol. Evol. 9, 1329–1340. https://doi.org/10.1093/gbe/evx050

Aguet, F., Ardlie, K.G., Cummings, B.B., Gelfand, E.T., Getz, G., Hadley, K., Handsaker, R.E., Huang, K.H., Kashin, S., Karczewski, K.J., Lek, M., Li, Xiao, MacArthur, D.G., Nedzel, J.L., Nguyen, D.T., Noble, M.S., Segrè, A. V., Trowbridge, C.A., Tukiainen, T., Abell, N.S., Balliu, B., Barshir, R., Basha, O., Battle, A., Bogu, G.K., Brown, A., Brown, C.D., Castel, S.E., Chen, L.S., Chiang, C., Conrad, D.F., Cox, N.J., Damani, F.N., Davis, J.R., Delaneau, O., Dermitzakis, E.T., Engelhardt, B.E., Eskin, E., Ferreira, P.G., Frésard, L., Gamazon, E.R., Garrido-Martín, D., Gewirtz, A.D.H., Gliner, G., Gloudemans, M.J., Guigo, R., Hall, I.M., Han, B., He, Y., Hormozdiari, F., Howald, C., Kyung Im, H., Jo, B., Yong Kang, E., Kim, Y., Kim-Hellmuth, S., Lappalainen, T., Li, G., Li, Xin, Liu, B., Mangul, S., McCarthy, M.I., McDowell, I.C., Mohammadi, P., Monlong, J., Montgomery, S.B., Muñoz-Aguirre, M., Ndungu, A.W., Nicolae, D.L., Nobel, A.B., Oliva, M., Ongen, H., Palowitch, J.J., Panousis, N., Papasaikas, P., Park, YoSon, Parsana, P., Payne, A.J., Peterson, C.B., Quan, J., Reverter, F., Sabatti, C., Saha, A., Sammeth, M., Scott, A.J., Shabalin, A.A., Sodaei, R., Stephens, M., Stranger, B.E., Strober, B.J., Sul, J.H., Tsang, E.K., Urbut, S., van de Bunt, M., Wang, G., Wen, X., Wright, F.A., Xi, H.S., Yeger-Lotem, E., Zappala, Z., Zaugg, J.B., Zhou, Y.-H., Akey, J.M., Bates, D., Chan, J., Chen, L.S., Claussnitzer, M., Demanelis, K., Diegel, M., Doherty, J.A., Feinberg, A.P., Fernando, M.S., Halow, J., Hansen, K.D., Haugen, E., Hickey, P.F., Hou, L., Jasmine, F., Jian, R., Jiang, L., Johnson, A., Kaul, R., Kellis, M., Kibriya, M.G., Lee, K., Billy Li, J., Li, Q., Li, Xiao, Lin, J., Lin, S., Linder, S., Linke, C., Liu, Y., Maurano, M.T., Molinie, B., Montgomery, S.B., Nelson, J., Neri, F.J., Oliva, M., Park, Yongjin, Pierce, B.L., Rinaldi, N.J., Rizzardi, L.F., Sandstrom, R., Skol, A., Smith, K.S., Snyder, M.P., Stamatoyannopoulos, J., Stranger, B.E., Tang, H., Tsang, E.K., Wang, L., Wang, M., Van Wittenberghe, N., Wu, F., Zhang, R., Nierras, C.R., Branton, P.A., Carithers, L.J., Guan, P., Moore, H.M., Rao, A., Vaught, J.B., Gould, S.E., Lockart, N.C., Martin, C., Struewing, J.P., Volpi, S., Addington, A.M., Koester, S.E., Little, A.R., Brigham, L.E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W.F., Lonsdale, J.T., McDonald, A., Mestichelli, B., Myer, K., Roe, Brian, Salvatore, M., Shad, S., Thomas, J.A., Walters, G., Washington, M., Wheeler, J., Bridge, J., Foster, B.A., Gillard, B.M., Karasik, E., Kumar, R., Miklos, M., Moser, M.T., Jewell, S.D., Montroy, R.G., Rohrer, D.C., Valley, D.R., Davis, D.A., Mash, D.C., Undale, A.H., Smith, A.M., Tabor, D.E., Roche, N. V., McLean, J.A., Vatanian, N., Robinson, K.L., Sobin, L., Barcus, M.E., Valentino, K.M., Qi, L., Hunter, S., Hariharan, P., Singh, S., Um, K.S., Matose, T., Tomaszewski, M.M., Barker, L.K., Mosavel, M., Siminoff, L.A., Traino, H.M., Flicek, P., Juettemann, T., Ruffier, M., Sheppard, D., Taylor, K., Trevanion, S.J., Zerbino, D.R., Craft, B., Goldman, M., Haeussler, M., Kent, W.J., Lee, C.M., Paten, B., Rosenbloom, K.R., Vivian, J., Zhu, J., Aguet, F., Brown, A.A., Castel, S.E., Davis, J.R., He, Y., Jo, B., Mohammadi, P., Park, YoSon, Parsana, P., Segrè, A. V., Strober, B.J., Zappala, Z., Cummings, B.B., Gelfand, E.T., Hadley, K., Huang, K.H., Lek, M., Li, Xiao, Nedzel, J.L., Nguyen, D.Y., Noble, M.S., Sullivan, T.J., Tukiainen, T., MacArthur, D.G., Getz, G., Addington, A., Guan, P., Koester, S., Little, A.R., Lockhart, N.C., Moore, H.M., Rao, A., Struewing, J.P., Volpi, S., Brigham, L.E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W.F., Lonsdale, J.T., McDonald, A., Mestichelli, B., Myer, K., Roe, Bryan, Salvatore, M., Shad, S., Thomas, J.A., Walters, G., Washington, M., Wheeler, J., Bridge, J., Foster, B.A., Gillard, B.M., Karasik, E., Kumar, R., Miklos, M., Moser, M.T., Jewell, S.D., Montroy, R.G., Rohrer, D.C., Valley, D., Mash, D.C., Davis, D.A., Sobin, L., Barcus, M.E., Branton, P.A., Abell, N.S., Balliu, B., Delaneau, O., Frésard, L., Gamazon, E.R., Garrido-Martín, D., Gewirtz, A.D.H., Gliner, G., Gloudemans, M.J., Han, B., He, A.Z., Hormozdiari, F., Li, Xin, Liu, B., Kang, E.Y., McDowell, I.C., Ongen, H.,

Palowitch, J.J., Peterson, C.B., Quon, G., Ripke, S., Saha, A., Shabalin, A.A., Shimko, T.C., Sul, J.H., Teran, N.A., Tsang, E.K., Zhang, H., Zhou, Y.-H., Bustamante, C.D., Cox, N.J., Guigó, R., Kellis, M., McCarthy, M.I., Conrad, D.F., Eskin, E., Li, G., Nobel, A.B., Sabatti, C., Stranger, B.E., Wen, X., Wright, F.A., Ardlie, K.G., Dermitzakis, E.T., Lappalainen, T., Battle, A., Brown, C.D., Engelhardt, B.E., Montgomery, S.B., 2017. Genetic effects on gene expression across human tissues. Nature 550, 204–213. https://doi.org/10.1038/nature24277

Aitken, R.J., Findlay, J.K., Hutt, K.J., Kerr, J.B., 2011. Apoptosis in the germ line. Reproduction 141, 139–150. https://doi.org/10.1530/REP-10-0232

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. ming, Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Sollewijn Gelpke, M.D., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J.K., Doggett, N., Zharkikh, A., Tavtigian, S. V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., Brenner, S., 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science (80-.). 297, 1301–1310. https://doi.org/10.1126/science.1072104

Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T., Hannon, G.J., 2008. A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. Mol. Cell 31, 785–799. https://doi.org/10.1016/j.molcel.2008.09.003

Armezzani, A., Varela, M., Spencer, T.E., Palmarini, M., Arnaud, F., 2014. "Ménage à Trois": The evolutionary interplay between JSRV, enJSRVs and domestic sheep. Viruses 6, 4926–4945. https://doi.org/10.3390/v6124926

Ashley, J., Cordy, B., Lucia, D., Fradkin, L.G., Budnik, V., Thomson, T., 2018. Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. Cell 172, 262-274.e11. https://doi.org/10.1016/j.cell.2017.12.022

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P., Rizzu, P., Smith, S., Fell, M., Talbot, R.T., Gustincich, S., Freeman, T.C., Mattick, J.S., Hume, D.A., Heutink, P., Carninci, P., Jeddeloh, J.A., Faulkner, G.J., 2011. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479, 534–537. https://doi.org/10.1038/nature10531

Baltimore, D., 1970. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses. Nature 226, 1209–1211. https://doi.org/10.1038/2261209a0

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, 11. https://doi.org/10.1186/s13100-015-0041-9

Barro-Trastoy, D., Köhler, C., 2024. Helitrons: genomic parasites that generate developmental novelties. Trends Genet. xx, 1–12. https://doi.org/10.1016/j.tig.2024.02.002

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Pačes, J., Burt, A., Tristem, M., 2004. Long-term reinfection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. U. S. A. 101, 4894–4899. https://doi.org/10.1073/pnas.0307800101

Belyeu, J.R., Brand, H., Wang, H., Zhao, X., Pedersen, B.S., Feusier, J., Gupta, M., Nicholas, T.J., Brown, J., Baird, L., Devlin, B., Sanders, S.J., Jorde, L.B., Talkowski, M.E., Quinlan, A.R., 2021. De novo structural mutation rates and gamete-of-origin biases revealed through genome

sequencing of 2,396 families. Am. J. Hum. Genet. 108, 597–607. https://doi.org/10.1016/j.ajhg.2021.02.012

Bergeron, L.A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M.F., Quintard, B., Hoffman, J.I., Li, Z., St. Leger, J., Shao, C., Stiller, J., Gilbert, M.T.P., Schierup, M.H., Zhang, G., 2023. Evolution of the germline mutation rate across vertebrates. Nature 615, 285–291. https://doi.org/10.1038/s41586-023-05752-y

Bertozzi, T.M., Elmer, J.L., Macfarlan, T.S., Ferguson-Smith, A.C., 2020. KRAB zinc finger protein diversification drives mammalian interindividual methylation variability. Proc. Natl. Acad. Sci. U. S. A. 117, 31290–31300. https://doi.org/10.1073/pnas.2017053117

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., Feschotte, C., 2018. Ten things you should know about transposable elements 1–12. https://doi.org/10.1186/s13059-018-1577-z

Brind'Amour, J., Kobayashi, H., Richard Albert, J., Shirane, K., Sakashita, A., Kamio, A., Bogutz, A., Koike, T., Karimi, M.M., Lefebvre, L., Kono, T., Lorincz, M.C., 2018. LTR retrotransposons transcribed in oocytes drive species-specific and heritable changes in DNA methylation. Nat. Commun. https://doi.org/10.1038/s41467-018-05841-x

Carmell, M.A., Girard, A., van de Kant, H.J.G., Bourc'his, D., Bestor, T.H., de Rooij, D.G., Hannon, G.J., 2007. MIWI2 Is Essential for Spermatogenesis and Repression of Transposons in the Mouse Male Germline. Dev. Cell 12, 503–514. https://doi.org/10.1016/j.devcel.2007.03.001

Chalvet, F., Teysset, L., Terzian, C., Prud'homme, N., Santamaria, P., Bucheton, A., Pélisson, A., 1999. Proviral amplification of the Gypsy endogenous retrovirus of Drosophila melanogaster involves env-independent invasion of the female germline. EMBO J. 18, 2659–2669. https://doi.org/10.1093/emboj/18.9.2659

Chameettachal, A., Mustafa, F., Rizvi, T.A., 2023. Understanding Retroviral Life Cycle and its Genomic RNA Packaging. J. Mol. Biol. 435, 167924. https://doi.org/https://doi.org/10.1016/j.jmb.2022.167924

Chameettachal, A., Vivet-Boudou, V., Pitchai, F.N.N., Pillai, V.N., Ali, L.M., Krishnan, A., Bernacchi, S., Mustafa, F., Marquet, R., Rizvi, T.A., 2021. A purine loop and the primer binding site are critical for the selective encapsidation of mouse mammary tumor virus genomic RNA by Pr77Gag. Nucleic Acids Res. 49, 4668–4688. https://doi.org/10.1093/nar/gkab223

Chelmicki, T., Roger, E., Teissandier, A., Dura, M., Bonneville, L., Rucli, S., Dossin, F., Fouassier, C., Lameiras, S., Bourc'his, D., 2021. m6A RNA methylation regulates the fate of endogenous retroviruses. Nature 591, 312–316. https://doi.org/10.1038/s41586-020-03135-1

Chen, J.Q., Zhang, M.P., Tong, X.K., Li, J.Q., Zhang, Z., Huang, F., Du, H.P., Zhou, M., Ai, H.S., Huang, L.S., 2022. Scan of the endogenous retrovirus sequences across the swine genome and survey of their copy number variation and sequence diversity among various Chinese and Western pig breeds. Zool. Res. 43, 423–441. https://doi.org/10.24272/J.ISSN.2095-8137.2021.379

Chen, Z., Chen, Z., Chen, Z., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., 2020. Role of Mammalian DNA Methyltransferases in Development. Annu. Rev. Biochem. https://doi.org/10.1146/annurev-biochem-103019-102815

Cherepanov, P., Maertens, G., Proost, P., Devreese, B., Van Beeumen, J., Engelborghs, Y., De Clercq, E., Debyser, Z., 2003. HIV-1 Integrase Forms Stable Tetramers and Associates with LEDGF/p75 Protein in Human Cells*. J. Biol. Chem. 278, 372–381. https://doi.org/https://doi.org/10.1074/jbc.M209278200

Chessa, B., Pereira, F., Arnaud, F., Amorim, A., Goyache, F., Mainland, I., Kao, R.R., Pemberton, J.M., Beraldi, D., Stear, M.J., Alberti, A., Pittau, M., Iannuzzi, L., Banabazi, M.H., Kazwala, R.R., Zhang, Y.P., Arranz, J.J., Ali, B.A., Wang, Z., Uzun, M., Dione, M.M., Olsaker, I., Holm, L.E., Saarma, U., Ahmad, S., Marzanov, N., Eythorsdottir, E., Holland, M.J., Paolo, A.M., Bruford, M.W., Kantanen, J., Spencer, T.E., Palmarini, M., 2009. Revealing the history of sheep domestication using retrovirus integrations. Science (80-.). 324, 532–536. https://doi.org/10.1126/science.1170587

Chiu, E.S., Vandewoude, S., 2021. Endogenous Retroviruses Drive Resistance and Promotion of Exogenous Retroviral Homologs. Annu. Rev. Anim. Biosci. 9, 225–248. https://doi.org/10.1146/annurev-animal-050620-101416

Chuong, E.B., Elde, N.C., Feschotte, C., 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science (80-.). 351, 1083–1087. https://doi.org/10.1126/science.aad5497

Chuong, E.B., Rumi, M.A.K., Soares, M.J., Baker, J.C., 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat. Genet. 45, 325–329. https://doi.org/10.1038/ng.2553

Coffin, J.M., Fan, H., 2016. The Discovery of Reverse Transcriptase. Annu. Rev. Virol. 3, 29–51. https://doi.org/10.1146/annurev-virology-110615-035556

Coffin, J.M., Hughes, S.H., Varmus, H.E., 1997. Retroviruses.

Cohen, J.C., Varmus, H.E., 1979. Endogenous mammary tumour virus DNA varies among wild mice and segregates during inbreeding. Nature 278, 418–423. https://doi.org/10.1038/278418a0

Cornelis, G., Heidmann, O., Bernard-Stoecklin, S., Reynaud, K., Véron, G., Mulot, B., Dupressoir, A., Heidmann, T., 2012. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. Proc. Natl. Acad. Sci. U. S. A. 109. https://doi.org/10.1073/pnas.1115346109

Cornelis, G., Heidmann, O., Degrelle, S.A., Vernochet, C., Lavialle, C., Letzelter, C., Bernard-Stoecklin, S., Hassanin, A., Mulot, B., Guillomot, M., Hue, I., Heidmann, T., Dupressoir, A., 2013. Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. Proc. Natl. Acad. Sci. U. S. A. 110. https://doi.org/10.1073/pnas.1215787110

DeRijck, J., deKogel, C., Demeulemeester, J., Vets, S., ElAshkar, S., Malani, N., Bushman, F.D., Landuyt, B., Husson, S.J., Busschots, K., Gijsbers, R., Debyser, Z., 2013. The BET Family of Proteins Targets Moloney Murine Leukemia Virus Integration near Transcription Start Sites. Cell Rep. 5, 886–894. https://doi.org/10.1016/j.celrep.2013.09.040

Dewannieux, M., Dupressoir, A., Harper, F., Pierron, G., Heidmann, T., 2004. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. Nat. Genet. 36, 534–539. https://doi.org/10.1038/ng1353

Dewannieux, M., Esnault, C., Heidmann, T., 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. 35, 41–48. https://doi.org/10.1038/ng1223

Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., Heidmann, T., 2006. Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res. 16, 1548–1556. https://doi.org/10.1101/gr.5565706

Dewannieux, M., Heidmann, T., 2013. Endogenous retroviruses: Acquisition, amplification and taming of genome invaders. Curr. Opin. Virol. 3, 646–656. https://doi.org/10.1016/j.coviro.2013.08.005

Douek, D.C., Brenchley, J.M., Betts, M.R., Ambrozak, D.R., Hill, B.J., Okamoto, Y., Casazza, J.P., Kuruppu, J., Kunstman, K., Wolinsky, S., Grossman, Z., Dybul, M., Oxenius, A., Price, D.A., Connors, M., Koup, R.A., 2002. HIV preferentially infects HIV-specific CD4+ T cells. Nature 417, 95–98. https://doi.org/10.1038/417095a

Drost, H.G., Sanchez, D.H., Eyre-Walker, A., 2019. Becoming a Selfish Clan: Recombination Associated to Reverse-Transcription in LTR Retrotransposons. Genome Biol. Evol. 11, 3382–3392. https://doi.org/10.1093/gbe/evz255

Dupressoir, A., Heidmann, T., 1996. Germ Line-Specific Expression of Intracisternal A-Particle Retrotransposons in Transgenic Mice. Mol. Cell. Biol. 16, 4495–4503. https://doi.org/10.1128/mcb.16.8.4495

Dupressoir, A., Vernochet, C., Harper, F., Guégan, J., Dessen, P., Pierron, G., Heidmann, T., 2011. A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. Proc. Natl. Acad. Sci. U. S. A. 108. https://doi.org/10.1073/pnas.1112304108

Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., Santamarina, M., Höps, W., Ashraf, H., Chuang, N.T., Yang, X., Munson, K.M., Lewis, A.P., Fairley, S., Tallon, L.J., Clarke, W.E., Basile, A.O., Byrska-Bishop, M., Corvelo, A., Evani, U.S., Lu, T.-Y., Chaisson, M.J.P., Chen, J., Li, C., Brand, H., Wenger, A.M., Ghareghani, M., Harvey, W.T., Raeder, B., Hasenfeld, P., Regier, A.A., Abel, H.J., Hall, I.M., Flicek, P., Stegle, O., Gerstein, M.B., Tubio, J.M.C., Mu, Z., Li, Y.I., Shi, X., Hastie, A.R., Ye, K., Chong, Z., Sanders, A.D., Zody, M.C., Talkowski, M.E., Mills, R.E., Devine, S.E., Lee, C., Korbel, J.O., Marschall, T., Eichler, E.E., 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science (80-.). 372, eabf7117. https://doi.org/10.1126/science.abf7117

Elleder, D., Kim, O., Padhi, A., Bankert, J.G., Simeonov, I., Schuster, S.C., Wittekindt, N.E., Motameny, S., Poss, M., 2012. Polymorphic Integrations of an Endogenous Gammaretrovirus in the Mule Deer Genome. J. Virol. 86, 2787–2796. https://doi.org/10.1128/jvi.06859-11

Engels, W.R., Preston, C.R., 1981. Identifying P factors in Drosophila by means of chromosome breakage hotspots. Cell 26, 421–428. https://doi.org/https://doi.org/10.1016/0092-8674(81)90211-7

Ernst, C., Odom, D.T., Kutter, C., 2017. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. Nat. Commun. 8, 1–9. https://doi.org/10.1038/s41467-017-01049-7

Evrony, G.D., Lee, E., Park, P.J., Walsh, C.A., 2016. Resolving rates of mutation in the brain using single-neuron genomics. Elife 5, 1–32. https://doi.org/10.7554/eLife.12966

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A.R.R., Suzuki, H., Hayashizaki, Y., Hume, D.A., Orlando, V., Grimmond, S.M., Carninci, P., 2009. The regulated retrotransposon transcriptome of mammalian cells. Nat. Genet. 41, 563–571. https://doi.org/10.1038/ng.368

Feschotte, C., Pritham, E.J., 2007. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41, 331–368. https://doi.org/10.1146/annurev.genet.40.110405.090448

Feusier, J., Watkins, W.S., Thomas, J., Farrell, A., Witherspoon, D.J., Baird, L., Ha, H., Xing, J., Jorde, L.B., 2019. Pedigree-based estimation of human mobile element retrotransposition rates. Genome Res. 29, 1567–1577. https://doi.org/10.1101/gr.247965.118

Finnegan, D.J., 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. 5, 103–107. https://doi.org/https://doi.org/10.1016/0168-9525(89)90039-5

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. U. S. A. 117, 9451–9457. https://doi.org/10.1073/pnas.1921046117

Frank, J.A., Singh, M., Cullen, H.B., Kirou, R.A., Benkaddour-Boumzaouad, M., Cortes, J.L., Pérez, J.G., Coyne, C.B., Feschotte, C., 2022. Evolution and antiviral activity of a human protein of retroviral origin. Science (80-.). 378, 422–428. https://doi.org/10.1126/science.abq7871

Frost, J.M., Amante, S.M., Okae, H., Jones, E.M., Ashley, B., Lewis, R.M., Cleal, J.K., Caley, M.P., Arima, T., Maffucci, T., Branco, M.R., 2023. Regulation of human trophoblast gene expression by endogenous retroviruses. Nat. Struct. Mol. Biol. 30, 527–538. https://doi.org/10.1038/s41594-023-00960-6

Gagnier, L., Belancio, V.P., Mager, D.L., 2019. Mouse germ line mutations due to retrotransposon insertions. Mob. DNA 10, 1–22. https://doi.org/10.1186/s13100-019-0157-4

Gao, Z., Moorjani, P., Sasani, T.A., Pedersen, B.S., Quinlan, A.R., Jorde, L.B., Amster, G., Przeworski, M., 2019. Overlooked roles of DNA damage and maternal age in generating human germline mutations. Proc. Natl. Acad. Sci. U. S. A. 116, 9491–9500. https://doi.org/10.1073/pnas.1901259116

Genet, M., Torres-Padilla, M.E., 2020. The molecular and cellular features of 2-cell-like cells: a reference guide. Dev. 147. https://doi.org/10.1242/dev.189688

Georges, M., Charlier, C., Hayes, B., 2019. Harnessing genomic information for livestock improvement. Nat. Rev. Genet. 20, 135–156. https://doi.org/10.1038/s41576-018-0082-2

Gifford, R., Tristem, M., 2003. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26, 291–315. https://doi.org/10.1023/A:1024455415443

Gkountela, S., Zhang, K.X., Shafiq, T.A., Liao, W.W., Hargan-Calvopiña, J., Chen, P.Y., Clark, A.T., 2015. DNA demethylation dynamics in the human prenatal germline. Cell 161, 1425–1436. https://doi.org/10.1016/j.cell.2015.05.012

Goerner-Potvin, P., Bourque, G., 2018. Computational tools to unmask transposable elements. Nat. Rev. Genet. 19, 688–704. https://doi.org/10.1038/s41576-018-0050-x

Göke, J., Lu, X., Chan, Y.S., Ng, H.H., Ly, L.H., Sachs, F., Szczerbinska, I., 2015. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. Cell Stem Cell 16, 135–141. https://doi.org/10.1016/j.stem.2015.01.005

Goldmann, J.M., Wong, W.S.W., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A.B., Glusman, G., Vissers, L.E.L.M., Hoischen, A., Roach, J.C., Vockley, J.G., Veltman, J.A., Solomon, B.D., Gilissen, C., Niederhuber, J.E., 2016. Parent-of-origin-specific signatures of de novo mutations. Nat. Genet. 48, 935–939. https://doi.org/10.1038/ng.3597

Goriely, A., Wilkie, A.O.M., 2012. Paternal age effect mutations and selfish spermatogonial selection: Causes and consequences for human disease. Am. J. Hum. Genet. 90, 175–200. https://doi.org/10.1016/j.ajhg.2011.12.017

Gozashti, L., Feschotte, C., Hoekstra, H.E., 2023. Transposable Element Interactions Shape the Ecology of the Deer Mouse Genome. Mol. Biol. Evol. 40, 1–17. https://doi.org/10.1093/molbev/msad069

Grabundzija, I., Hickman, A.B., Dyda, F., 2018. Helraiser intermediates provide insight into the mechanism of eukaryotic replicative transposition. Nat. Commun. 9. https://doi.org/10.1038/s41467-018-03688-w

Grabundzija, I., Messing, S.A., Thomas, J., Cosby, R.L., Bilic, I., Miskey, C., Gogol-Doring, A., Kapitonov, V., Diem, T., Dalda, A., Jurka, J., Pritham, E.J., Dyda, F., Izsvak, Z., Ivics, Z., 2016. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat. Commun. 7. https://doi.org/10.1038/ncomms10716

Grow, E.J., Flynn, R.A., Chavez, S.L., Bayless, N.L., Wossidlo, M., Wesche, D.J., Martin, L., Ware, C.B., Blish, C.A., Chang, H.Y., Pera, R.A.R., Wysocka, J., 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature 522, 221–246. https://doi.org/10.1038/nature14308

Gruhn, W.H., Tang, W.W.C., Dietmann, S., Alves-Lopes, J.P., Penfold, C.A., Wong, F.C.K., Ramakrishna, N.B., Azim Surani, M., 2023. Epigenetic resetting in the human germ line entails histone modification remodeling. Sci. Adv. 9. https://doi.org/10.1126/sciadv.ade1257

Guan, Y., Wang, P.J., 2021. Golden opportunity for piRNA in female fertility. Nat. Cell Biol. 23, 936–938. https://doi.org/10.1038/s41556-021-00749-z

Guo, F., Li, X., Liang, D., Li, T., Zhu, P., Guo, H., Wu, X., Wen, L., Gu, T.P., Hu, B., Walsh, C.P., Li, J., Tang, F., Xu, G.L., 2014. Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. Cell Stem Cell 15, 447–459. https://doi.org/10.1016/j.stem.2014.08.003

Guo, F., Yan, L., Guo, H., Li, L., Hu, B., Zhao, Y., Yong, J., Hu, Y., Wang, X., Wei, Y., Wang, W., Li, R., Yan, J., Zhi, X., Zhang, Y., Jin, H., Zhang, W., Hou, Y., Zhu, P., Li, J., Zhang, L., Liu, S., Ren, Y., Zhu, X., Wen, L., Gao, Y.Q., Tang, F., Qiao, J., 2015. The transcriptome and DNA methylome landscapes of human primordial germ cells. Cell 161, 1437–1452. https://doi.org/10.1016/j.cell.2015.05.015

Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Yan, J., Ren, X., Lin, S., Li, J., Jin, X., Shi, X., Liu, P., Wang, X., Wang, W., Wei, Y., Li, X., Guo, F., Wu, X., Fan, X., Yong, J., Wen, L., Xie, S.X., Tang, F., Qiao, J., 2014. The DNA methylation landscape of human early embryos. Nature 511, 606–610. https://doi.org/10.1038/nature13544

Guo, J., Nie, X., Giebler, M., Mlcochova, H., Wang, Y., Grow, E.J., Kim, R., Tharmalingam, M., Matilionyte, G., Lindskog, C., Carrell, D.T., Mitchell, R.T., Goriely, A., Hotaling, J.M., Cairns, B.R., 2020. The Dynamic Transcriptional Cell Atlas of Testis Development during Human Puberty. Cell Stem Cell 26, 262-276.e4. https://doi.org/10.1016/j.stem.2019.12.005

Guo, J., Sosa, E., Chitiashvili, T., Nie, X., Rojas, E.J., Oliver, E., Plath, K., Hotaling, J.M., Stukenborg, J.B., Clark, A.T., Cairns, B.R., 2021. Single-cell analysis of the developing human testis reveals somatic niche cell specification and fetal germline stem cell establishment. Cell Stem Cell 28, 764-778.e4. https://doi.org/10.1016/j.stem.2020.12.004

Harada, K., Yukuhiro, K., Mukai, T., 1990. Genetics Transposition rates of movable genetic elements in Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. 87, 3248–3252. https://doi.org/10.1073/pnas.87.8.3248

Harland, C., Charlier, C., Karim, L., Cambisano, N., Deckers, M., Mni, M., Mullaart, E., Coppieters, W., Georges, M., 2017. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. bioRxiv 79863. https://doi.org/10.1101/079863

Harris, K., Pritchard, J.K., 2017. Rapid evolution of the human mutation spectrum. Elife 6, 1–17. https://doi.org/10.7554/eLife.24284

Hayward, A., Cornwallis, C.K., Jern, P., 2015. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. Proc. Natl. Acad. Sci. 112, 464–469. https://doi.org/10.1073/pnas.1414980112

Hecht, S.J., Stedman, K.E., Carlson, J.O., DeMartini, J.C., 1996. Distribution of endogenous type B and type D sheep retrovirus sequences in ungulates and other mammals. Proc. Natl. Acad. Sci. U. S. A. 93, 3297–3302. https://doi.org/10.1073/pnas.93.8.3297

Heidmann, O., Vernochet, C., Dupressoir, A., Heidmann, T., 2009. Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: A new “syncytin” in a third order of mammals. Retrovirology 6, 1–11. https://doi.org/10.1186/1742-4690-6-107

Heldmann, O., Heidmann, T., 1991. Retrotransposition of a mouse IAP sequence tagged with an indicator gene. Cell 64, 159–170. https://doi.org/https://doi.org/10.1016/0092-8674(91)90217-M

Hill, P.W.S., Leitch, H.G., Requena, C.E., Sun, Z., Amouroux, R., Roman-Trufero, M., Borkowska, M., Terragni, J., Vaisvila, R., Linnett, S., Bagci, H., Dharmalingham, G., Haberle, V., Lenhard, B., Zheng, Y., Pradhan, S., Hajkova, P., 2018. Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. Nature 555, 392–396. https://doi.org/10.1038/nature25964

Horns, F., Martinez, J.A., Fan, C., Haque, M., Linton, J.M., Tobin, V., Santat, L., Maggiolo, A.O., Bjorkman, P.J., Lois, C., Elowitz, M.B., 2023. Engineering RNA export for measurement and manipulation of living cells. Cell 1–17. https://doi.org/10.1016/j.cell.2023.06.013

Huang, T.C., Wang, Y.F., Vazquez-Ferrer, E., Theofel, I., Requena, C.E., Hanna, C.W., Kelsey, G., Hajkova, P., 2021. Sex-specific chromatin remodelling safeguards transcription in germ cells.

Nature 600, 737–742. https://doi.org/10.1038/s41586-021-04208-5

Imbeault, M., Helleboid, P.Y., Trono, D., 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature 543, 550–554. https://doi.org/10.1038/nature21683

Ito, J., Gifford, R.J., Sato, K., 2020. Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. Proc. Natl. Acad. Sci. U. S. A. 117, 610–618. https://doi.org/10.1073/pnas.1914183116

Ivancevic, A.M., Kortschak, R.D., Bertozzi, T., Adelson, D.L., 2018. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. Genome Biol. 19, 1–13. https://doi.org/10.1186/s13059-018-1456-7

Jaenisch, R., 1977. Germ line integration of moloney leukemia virus: Effect of homozygosity at the M-MuLV locus. Cell 12, 691–696. https://doi.org/https://doi.org/10.1016/0092-8674(77)90269-0

Jaenisch, R., 1976. Germ line integration and Mendelian transmission of the exogenous Moloney leukemia virus. Proc. Natl. Acad. Sci. U. S. A. 73, 1260–1264. https://doi.org/10.1073/pnas.73.4.1260

Jern, P., Sperber, G.O., Blomberg, J., 2005. Use of Endogenous Retroviral Sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology 2, 1–12. https://doi.org/10.1186/1742-4690-2-50

Johnson, W.E., 2019. Origins and evolutionary consequences of ancient endogenous retroviruses. Nat. Rev. Microbiol. 17, 355–370. https://doi.org/10.1038/s41579-019-0189-2

Johnson, W.E., 2015. Endogenous Retroviruses in the Genomics Era. Annu. Rev. Virol. 2, 135–159. https://doi.org/10.1146/annurev-virology-100114-054945

Jones, P.A., Ohtani, H., Chakravarthy, A., De Carvalho, D.D., 2019. Epigenetic therapy in immune-oncology. Nat. Rev. Cancer 19, 151–161. https://doi.org/10.1038/s41568-019-0109-9

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T., Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A., Ward, L.D., Arnadottir, G.A., Helgason, E.A., Helgason, H., Gylfason, A., Jonasdottir, Adalbjorg, Jonasdottir, Aslaug, Rafnar, T., Frigge, M., Stacey, S.N., Th. Magnusson, O., Thorsteinsdottir, U., Masson, G., Kong, A., Halldorsson, B. V., Helgason, A., Gudbjartsson, D.F., Stefansson, K., 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature 549, 519–522. https://doi.org/10.1038/nature24018

Kapitonov, V. V., Jurka, J., 2001. Rolling-circle transposons in eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 98, 8714–8719. https://doi.org/10.1073/pnas.151269298

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L.A., Bourque, G., Yandell, M., Feschotte, C., 2013. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genet. 9. https://doi.org/10.1371/journal.pgen.1003470

Kapusta, A., Suh, A., Feschotte, C., 2017. Dynamics of genome size evolution in birds and mammals. Proc. Natl. Acad. Sci. U. S. A. 114, E1460–E1469. https://doi.org/10.1073/pnas.1616702114

Kazazian, H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., Antonarakis, S.E., 1988.

Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332, 164–166. https://doi.org/10.1038/332164a0

Kelly, C.J., Chitko-McKown, C.G., Chuong, E.B., 2022. Ruminant-specific retrotransposons shape regulatory evolution of bovine immunity. Genome Res. 32, 1474–1487. https://doi.org/10.1101/gr.276241.121

Kent, M., Moser, M., Boman, I.A., Lindtveit, K., Árnyasi, M., Sundsaasen, K.K., Våge, D.I., 2021. Insertion of an endogenous Jaagsiekte sheep retrovirus element into the BCO2 - gene abolishes its function and leads to yellow discoloration of adipose tissue in Norwegian Spælsau (Ovis aries). BMC Genomics 22, 1–8. https://doi.org/10.1186/s12864-021-07826-5

Kloosterman, W.P., Francioli, L.C., Hormozdiari, F., Marschall, T., Hehir-Kwa, J.Y., Abdellaoui, A., Lameijer, E.W., Moed, M.H., Koval, V., Renkens, I., Van Roosmalen, M.J., Arp, P., Karssen, L.C., Coe, B.P., Handsaker, R.E., Suchiman, E.D., Cuppen, E., Thung, D.T., McVey, M., Wendl, M.C., Uitterlinden, A., Van Duijn, C.M., Swertz, M.A., Wijmenga, C., Van Ommen, G.J.B., Slagboom, P.E., Boomsma, D.I., Schönhuth, A., Eichler, E.E., De Bakker, P.I.W., Ye, K., Guryev, V., Van Ommen, G.J.B., Bovenberg, J.A., De Craen, A.J.M., Beekman, M., Hofman, A., Willemsen, G., Wolffenbuttel, B., Platteel, M., Du, Y., Chen, R., Cao, H., Cao, R., Sun, Y., Cao, J.S., Van Dijk, F., Neerincx, P.B.T., Deelen, P., Dijkstra, M., Byelas, G., Kanterakis, A., Bot, J., Vermaat, M., Laros, J.F.J., Den Dunnen, J.T., De Knijff, P., Van Leeuwen, E.M., Amin, N., Rivadeneira, F., Estrada, K., De Ligt, J., Hottenga, J.J., Kattenberg, V.M., Van Enckevort, D., Mei, H., Santcroos, M., Van Schaik, B.D.C., McCarroll, S.A., Ko, A., Sudmant, P., Nijman, I.J., 2015. Characteristics of de novo structural changes in the human genome. Genome Res. 25, 792–801. https://doi.org/10.1101/gr.185041.114

Knudson, C.M., Tung, K.S.K., Tourtellotte, W.G., Brown, G.A.J., Korsmeyer, S.J., 1995. Bax-Deficient Mice with Lymphoid Hyperplasia and Male Germ Cell Death, Science. American Association for the Advancement of Science. https://doi.org/10.1126/science.270.5233.96

Kojima, S., Koyama, S., Ka, M., Saito, Y., Parrish, E.H., Endo, M., Takata, S., Mizukoshi, M., Hikino, K., Takeda, A., Gelinas, A.F., Heaton, S.M., Koide, R., Kamada, A.J., Noguchi, M., Hamada, M., Matsuda, K., Yamanashi, Y., Furukawa, Y., Morisaki, T., Murakami, Y., Muto, K., Nagai, A., Obara, W., Yamaji, K., Takahashi, K., Asai, S., Takahashi, Y., Suzuki, T., Sinozaki, N., Yamaguchi, H., Minami, S., Murayama, S., Yoshimori, K., Nagayama, S., Obata, D., Higashiyama, M., Masumoto, A., Koretsune, Y., Kamatani, Y., Murakawa, Y., Ishigaki, K., Nakamura, Y., Ito, K., Terao, C., Momozawa, Y., Parrish, N.F., 2023. Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk. Nat. Genet. 55, 939–951. https://doi.org/10.1038/s41588-023-01390-2

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, Aslaug, Jonasdottir, Adalbjorg, Wong, W.S.W., Sigurdsson, G., Walters, G.B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D.F., Helgason, A., Magnusson, O.T., Thorsteinsdottir, U., Stefansson, K., 2012. Rate of de novo mutations and the importance of father-s age to disease risk. Nature 488, 471–475. https://doi.org/10.1038/nature11396

Kristmundsdottir, S., Jonsson, H., Hardarson, M.T., Palsson, G., Beyter, D., Eggertsson, H.P., Gylfason, A., Sveinbjornsson, G., Holley, G., Stefansson, O.A., Halldorsson, G.H., Olafsson, S., Arnadottir, G.A., Olason, P.I., Eiriksson, O., Masson, G., Thorsteinsdottir, U., Rafnar, T., Sulem, P., Helgason, A., Gudbjartsson, D.F., Halldorsson, B. V., Stefansson, K., 2023. Sequence variants affecting the genome-wide rate of germline microsatellite mutations. Nat. Commun. 14, 1–12. https://doi.org/10.1038/s41467-023-39547-6

Kruse, K., Díaz, N., Enriquez-Gasca, R., Gaume, X., Torres-Padilla, M.-E., Vaquerizas, J.M., 2019. Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. bioRxiv 523712. https://doi.org/10.1101/523712

Kulpa, D.A., Moran, J. V, 2006. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat. Struct. Mol. Biol. 13, 655–660. https://doi.org/10.1038/nsmb1107

Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T.W., Hata, K., Li, E., Matsuda, Y., Kimura, T., Okabe, M., Sakaki, Y., Sasaki, H., Nakano, T., 2008. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev. 22, 908–917. https://doi.org/10.1101/gad.1640708

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, Christina, Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Hong, M.L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., De La Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M. V., Kaul, R., Raymond, Christopher, Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., 2001. Initial sequencing and analysis of the human genome. Nature 412, 565–566. https://doi.org/10.1038/35087627

Lee, Y., Bouwman, A.C., Harland, C., Bosse, M., Costa, G., Moreira, M., Veerkamp, R.F., Mullaart, E.,

Cambisano, N., Groenen, M.A.M., Karim, L., Coppieters, W., Georges, M., Charlier, C., 2023. The rate of de novo structural variation is increased in in vitro – produced offspring and preferentially affects the paternal genome 1–10. https://doi.org/10.1101/gr.277884.123.

Lee, Y.L., Takeda, H., Moreira, G.C.M., Karim, L., Mullaart, E., Coppieters, W., Appeltant, R., Veerkamp, R.F., Groenen, M.A.M., Georges, M., Bosse, M., Druet, T., Bouwman, A.C., Charlier, C., 2021. A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle. PLoS Genet. 17, 1–27. https://doi.org/10.1371/journal.pgen.1009331

Li, J., Davis, B.W., Jern, P., Dorshorst, B.J., Siegel, P.B., Andersson, L., 2019. Characterization of the endogenous retrovirus insertion in CYP19A1 associated with henny feathering in chicken. Mob. DNA 10, 1–8. https://doi.org/10.1186/s13100-019-0181-4

Li, S.F., Zhang, X.Y., Yang, L.L., Jia, K.L., Li, J.R., Lan, L.N., Zhang, Y.L., Li, N., Deng, C.L., Gao, W.J., 2023. Landscape and evolutionary dynamics of Helitron transposons in plant genomes as well as construction of online database HelDB. J. Syst. Evol. 61, 919–931. https://doi.org/10.1111/jse.12929

Liu, Xiaoqian, Liu, Z., Wu, Z., Ren, J., Fan, Y., Sun, L., Cao, G., Niu, Y., Zhang, B., Ji, Q., Jiang, X., Wang, C., Wang, Q., Ji, Z., Li, L., Esteban, C.R., Yan, K., Li, W., Cai, Yusheng, Wang, S., Zheng, A., Zhang, Y.E., Tan, S., Cai, Yingao, Song, M., Lu, F., Tang, F., Ji, W., Zhou, Q., Belmonte, J.C.I., Zhang, W., Qu, J., Liu, G.H., 2023. Resurrection of endogenous retroviruses during aging reinforces senescence. Cell 186, 287-304.e26. https://doi.org/10.1016/j.cell.2022.12.017

Liu, Xuexue, Zhang, Y., Pu, Y., Ma, Y., Jiang, L., 2023. Whole-genome identification of transposable elements reveals the equine repetitive element insertion polymorphism in Chinese horses. Anim. Genet. 54, 144–154. https://doi.org/10.1111/age.13277

Löber, U., Hobbs, M., Dayaram, A., Tsangaras, K., Jones, K., Alquezar-Planas, D.E., Ishida, Y., Meers, J., Mayer, J., Quedenau, C., Chen, W., Johnson, R.N., Timms, P., Young, P.R., Roca, A.L., Greenwood, A.D., 2018. Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. Proc. Natl. Acad. Sci. 115, 8609–8614. https://doi.org/10.1073/pnas.1807598115

Lyden, T.W., Johnson, P.M., Mwenda, J.M., Rote, N.S., 1994. Ultrastructural characterization of endogenous retroviral particles isolated from normal human placentas. Biol. Reprod. 51, 152–157. https://doi.org/10.1095/biolreprod51.1.152

Mager, D.L., Freeman, J.D., 2000. Novel Mouse Type D Endogenous Proviruses and ETn Elements Share Long Terminal Repeat and Internal Sequences. J. Virol. 74, 7221–7229. https://doi.org/10.1128/jvi.74.16.7221-7229.2000

Mager, D.L., Freeman, J.D., 1995. HERV-H Endogenous Retroviruses: Presence in the New World Branch but Amplification in the Old World Primate Lineage. Virology 213, 395–404. https://doi.org/10.1006/viro.1995.0012

Mager, D.L., Goodchild, N.L., 1989. Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings. Am. J. Hum. Genet. 45, 848–854.

Mager, D.L., Stoye, J.P., 2015. Mammalian Endogenous Retroviruses. Microbiol. Spectr. 3, 10.1128/microbiolspec.mdna3-0009–2014. https://doi.org/10.1128/microbiolspec.mdna3-0009-

2014

Magiorkinis, G., Gifford, R.J., Katzourakis, A., De Ranter, J., Belshaw, R., 2012. Env-less endogenous retroviruses are genomic superspreaders. Proc. Natl. Acad. Sci. U. S. A. 109, 7385–7390. https://doi.org/10.1073/pnas.1200913109

Maksimov, M.O., Wu, C., Ashbrook, D.G., Villani, F., Colonna, V., Mousavi, N., Ma, N., Lu, L., Pritchard, J.K., Goren, A., Williams, R.W., Palmer, A.A., Gymrek, M., 2023. A novel quantitative trait locus implicates Msh3 in the propensity for genome-wide short tandem repeat expansions in mice. Genome Res. 33, 689–702. https://doi.org/10.1101/gr.277576.122

Malki, S., vanderHeijden, G.W., O'Donnell, K.A., Martin, S.L., Bortvin, A., 2014. A Role for retrotransposon LINE-1 in fetal oocyte attrition in mice. Dev. Cell 29, 521–533. https://doi.org/10.1016/j.devcel.2014.04.027

Marchi, E., Kanapin, A., Magiorkinis, G., Belshaw, R., 2014. Unfixed Endogenous Retroviral Insertions in the Human Population. J. Virol. 88, 9529–9537. https://doi.org/10.1128/JVI.00919-14

Martin, M.A., Bryan, T., Rasheed, S., Khan, A.S., 1981. Identification and cloning of endogenous retroviral sequences present in human DNA. Proc. Natl. Acad. Sci. U. S. A. 78, 4892–4896. https://doi.org/10.1073/pnas.78.8.4892

Mcclintock, B., 1950. The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. U. S. A. 36, 344–355. https://doi.org/10.1073/pnas.36.6.344

Mehta, J., Starmer, C., Sugden, R., Schelling, T., Kahneman, D., Stanovich, K., West, R., Rubinstein, A., Jung, R., Haier, R., Goel, V., Dolan, R., Noveck, I., Smith, K., Kyllonen, P., Christal, R., Baddeley, A., Smith, E., Jonides, J., Knight, R., Wager, T., 2009. The Genome Sequence of Taurine: A Window to Ruminant Biology and Evolution. Science (80-.). 522–528.

Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.-Y., Edouard, P., Howes, S., Keith, J.C., McCoy, J.M., 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis . Nature 403, 785–789. https://doi.org/10.1038/35001608

Mitra, I., Huang, B., Mousavi, N., Ma, N., Lamkin, M., Yanicky, R., Shleizer-Burko, S., Lohmueller, K.E., Gymrek, M., 2021. Patterns of de novo tandem repeat mutations and their role in autism. Nature 589, 246–250. https://doi.org/10.1038/s41586-020-03078-7

Modzelewski, A.J., Gan Chong, J., Wang, T., He, L., 2022. Mammalian genome innovation through transposon domestication. Nat. Cell Biol. https://doi.org/10.1038/s41556-022-00970-4

Modzelewski, A.J., Shao, W., Chen, J., Lee, A., Qi, X., Noon, M., Tjokro, K., Sales, G., Biton, A., Anand, A., Speed, T.P., Xuan, Z., Wang, T., Risso, D., He, L., 2021. A mouse-specific retrotransposon drives a conserved Cdk2ap1 isoform essential for development. Cell 184, 5541-5558.e22. https://doi.org/10.1016/j.cell.2021.09.021

Muller, T.G., Zila, V., Muller, B., Krausslich, H.G., 2022. Nuclear Capsid Uncoating and Reverse Transcription of HIV-1. Annu. Rev. Virol. 9, 261–284. https://doi.org/10.1146/annurev-virology-020922-110929

Mustafa, F., Al Amri, D., Al Ali, F., Al Sari, N., Al Suwaidi, S., Jayanth, P., Philips, P.S., Rizvi, T.A.,

2012. Sequences within Both the 5′ UTR and Gag Are Required for Optimal In Vivo Packaging and Propagation of Mouse Mammary Tumor Virus (MMTV) Genomic RNA. PLoS One 7. https://doi.org/10.1371/journal.pone.0047088

Nam, C.H., Youk, J., Kim, J.Y., Lim, J., Park, Jung Woo, Oh, S.A., Lee, H.J., Park, Ji Won, Won, H., Lee, Y., Jeong, S.Y., Lee, D.S., Oh, J.W., Han, J., Lee, J., Kwon, H.W., Kim, M.J., Ju, Y.S., 2023. Widespread somatic L1 retrotransposition in normal colorectal epithelium. Nature 617, 540–547. https://doi.org/10.1038/s41586-023-06046-z

Nguyen, D.H., Soygur, B., Peng, S.P., Malki, S., Hu, G., Laird, D.J., 2020. Apoptosis in the fetal testis eliminates developmentally defective germ cell clones. Nat. Cell Biol. 22, 1423–1435. https://doi.org/10.1038/s41556-020-00603-8

Nikolaitchik, O.A., Dilley, K.A., Fu, W., Gorelick, R.J., Tai, S.H.S., Soheilian, F., Ptak, R.G., Nagashima, K., Pathak, V.K., Hu, W.S., 2013. Dimeric RNA Recognition Regulates HIV-1 Genome Packaging. PLoS Pathog. 9. https://doi.org/10.1371/journal.ppat.1003249

Osmanski, A.B., Paulat, N.S., Korstian, J., Grimshaw, J.R., Halsey, M., Sullivan, K.A., Moreno-Santillán, D.D., Crookshanks, C., Roberts, J., Garcia, C., Johnson, M.G., Densmore, L.D., Consortium, Z., Rosen, J., Storer, J.M., Hubley, R., Dávalos, L.M., Lindblad-Toh, K., Karlsson, E.K., Ray, D.A., 2023. Insights into mammalian TE diversity via the curation of 248 mammalian genome assemblies. Science (80-.). 380, eabn1430. https://doi.org/10.1126/science.abn1430

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., Jiang, N., Hirsch, C.N., Hufford, M.B., 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20, 1–18. https://doi.org/10.1186/s13059-019-1905-y

Ozata, D.M., Gainetdinov, I., Zoch, A., O'Carroll, D., Zamore, P.D., 2019. PIWI-interacting RNAs: small RNAs with big functions. Nat. Rev. Genet. 20, 89–108. https://doi.org/10.1038/s41576-018-0073-3

Palmarini, M., Sharp, J.M., de las Heras, M., Fan, H., 1999. Jaagsiekte Sheep Retrovirus Is Necessary and Sufficient To Induce a Contagious Lung Cancer in Sheep. J. Virol. 73, 6964–6972. https://doi.org/10.1128/jvi.73.8.6964-6972.1999

Pastuzyn, E.D., Day, C.E., Kearns, R.B., Kyrke-Smith, M., Taibi, A. V., McCormick, J., Yoder, N., Belnap, D.M., Erlendsson, S., Morado, D.R., Briggs, J.A.G., Feschotte, C., Shepherd, J.D., 2018. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. Cell 172, 275-288.e18. https://doi.org/10.1016/j.cell.2017.12.024

Peltomäki, P., 2003. Role of DNA Mismatch Repair Defects in the Pathogenesis of Human Cancer. J. Clin. Oncol. 21, 1174–1179. https://doi.org/10.1200/JCO.2003.04.060

Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Al Turki, S., Dominiczak, A., Morris, A., Porteous, D., Smith, B., Stratton, M.R., Hurles, M.E., 2016. Timing, rates and spectra of human germline mutation. Nat. Genet. 48, 126–133. https://doi.org/10.1038/ng.3469

Rebollo, R., Galvão-Ferrarini, M., Gagnier, L., Zhang, Y., Ferraj, A., Beck, C.R., Lorincz, M.C., Mager, D.L., 2020. Inter-Strain Epigenomic Profiling Reveals a Candidate IAP Master Copy in C3H Mice. Viruses. https://doi.org/10.3390/v12070783

Ribet, D., Dewannieux, M., Heidmann, T., 2004. An active murine transposon family pair: Retrotransposition of "master" MusD copies and ETn trans-mobilization. Genome Res. 14, 2261–2267. https://doi.org/10.1101/gr.2924904

Ribet, D., Harper, F., Dewannieux, M., Pierron, G., Heidmann, T., 2007. Murine MusD Retrotransposon: Structure and Molecular Evolution of an "Intracellularized" Retrovirus. J. Virol. 81, 1888–1898. https://doi.org/10.1128/jvi.02051-06

Ribet, D., Harper, F., Dupressoir, A., Dewannieux, M., Pierron, G., Heidmann, T., 2008. An infectious progenitor for the murine IAP retrotransposon: Emergence of an intracellular genetic parasite from an ancient retrovirus. Genome Res. 18, 597–609. https://doi.org/10.1101/gr.073486.107

Richardson, S.R., Gerdes, P., Gerhardt, D.J., Sanchez-Luque, F.J., Bodea, G.O., Muñoz-Lopez, M., Jesuadian, J.S., Kempen, M.J.H.C., Carreira, P.E., Jeddeloh, J.A., Garcia-Perez, J.L., Kazazian, H.H., Ewing, A.D., Faulkner, G.J., 2017a. Heritable L1 retrotransposition in the mouse primordial germline and early embryo. Genome Res. 27, 1395–1405. https://doi.org/10.1101/gr.219022.116

Richardson, S.R., Gerdes, P., Gerhardt, D.J., Sanchez-Luque, F.J., Bodea, G.O., Muñoz-Lopez, M., Jesuadian, J.S., Kempen, M.J.H.C., Carreira, P.E., Jeddeloh, J.A., Garcia-Perez, J.L., Kazazian, H.H., Ewing, A.D., Faulkner, G.J., 2017b. Heritable L1 retrotransposition in the mouse primordial germline and early embryo. Genome Res. 27, 1395–1405. https://doi.org/10.1101/gr.219022.116

Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J., Santamarina, M., Ju, Y.S., Temes, J., Garcia-Souto, D., Detering, H., Li, Y., Rodriguez-Castro, J., Dueso-Barroso, A., Bruzos, A.L., Dentro, S.C., Blanco, M.G., Contino, G., Ardeljan, D., Tojo, M., Roberts, N.D., Zumalave, S., Edwards, P.A.W., Weischenfeldt, J., Puiggròs, M., Chong, Z., Chen, K., Lee, E.A., Wala, J.A., Raine, K., Butler, A., Waszak, S.M., Navarro, F.C.P., Schumacher, S.E., Monlong, J., Maura, F., Bolli, N., Bourque, G., Gerstein, M., Park, P.J., Wedge, D.C., Beroukhim, R., Torrents, D., Korbel, J.O., Martincorena, I., Fitzgerald, R.C., Van Loo, P., Kazazian, H.H., Burns, K.H., Akdemir, K.C., Alvarez, E.G., Baez-Ortega, A., Beroukhim, R., Boutros, P.C., Bowtell, D.D.L., Brors, B., Burns, K.H., Campbell, P.J., Chan, K., Cortés-Ciriano, I., Dueso-Barroso, A., Dunford, A.J., Edwards, P.A., Estivill, X., Etemadmoghadam, D., Feuerbach, L., Fink, J.L., Frenkel-Morgenstern, M., Garsed, D.W., Gerstein, M., Gordenin, D.A., Haan, D., Haber, J.E., Hess, J.M., Hutter, B., Imielinski, M., Jones, D.T.W., Kazanov, M.D., Klimczak, L.J., Koh, Y., Korbel, J.O., Kumar, K., Lee, E.A., Lee, J.J.K., Li, Y., Lynch, A.G., Macintyre, G., Markowetz, F., Martinez-Fundichely, A., Meyerson, M., Miyano, S., Nakagawa, H., Navarro, F.C.P., Ossowski, S., Pearson, J. V., Pearson, J. V., Rippe, K., Roberts, N.D., Roberts, S.A., Rodriguez-Martin, B., Rodriguez-Martin, B., Schumacher, S.E., Shackleton, M., Sidiropoulos, N., Sieverling, L., Stewart, C., Tubio, J.M.C., Villasante, I., Waddell, N., Wala, J.A., Weischenfeldt, J., Yang, L., Yao, X., Yoon, S.S., Zamora, J., Zhang, C.Z., Campbell, P.J., Tubio, J.M.C., Schumacher, S.E., Scully, R., Rodriguez-Martin, B., 2020. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat. Genet. 52, 306–319. https://doi.org/10.1038/s41588-019-0562-0

Rous, P., 1911. A Sarcoma of the Fowl Transmissible by an Agent Separable from the Tumor Cells. J. Exp. Med. 142, 397. https://doi.org/10.1097/00000441-191108000-00079

Rubin, H., 2011. The early history of tumor virology: Rous, RIF, and RAV. Proc. Natl. Acad. Sci. U. S. A. 108, 14389–14396. https://doi.org/10.1073/pnas.1108655108

Saito, E.S., Keng, V.W., Takeda, J., Horie, K., 2008. Translation from nonautonomous type IAP retrotransposon is a critical determinant of transposition activity: Implication for retrotransposon-mediated genome evolution. Genome Res. 18, 859–868. https://doi.org/10.1101/gr.069310.107

Saitou, M., Miyauchi, H., 2016. Gametogenesis from Pluripotent Stem Cells. Cell Stem Cell 18, 721–735. https://doi.org/10.1016/j.stem.2016.05.001

Sakashita, A., Kitano, T., Ishizu, H., Guo, Y., Masuda, H., Ariura, M., Murano, K., Siomi, H., 2023. Transcription of MERVL retrotransposons is required for preimplantation embryo development. Nat. Genet. 55, 484–495. https://doi.org/10.1038/s41588-023-01324-y

Sakashita, A., Maezawa, S., Takahashi, K., Alavattam, K.G., Yukawa, M., Hu, Y.C., Kojima, S., Parrish, N.F., Barski, A., Pavlicev, M., Namekawa, S.H., 2020. Endogenous retroviruses drive species-specific germline transcriptomes in mammals. Nat. Struct. Mol. Biol. 27, 967–977. https://doi.org/10.1038/s41594-020-0487-4

Sang, Q., Ray, P.F., Wang, L., 2023. Understanding the genetics of human infertility. Science (80-.). 380, 158–163. https://doi.org/10.1126/science.adf7760

Sasani, T.A., Ashbrook, D.G., Beichman, A.C., Lu, L., Palmer, A.A., Williams, R.W., Pritchard, J.K., Harris, K., 2022. A natural mutator allele shapes mutation spectrum variation in mice. Nature 605, 497–502. https://doi.org/10.1038/s41586-022-04701-5

Sasani, T.A., Pedersen, B.S., Gao, Z., Baird, L., Przeworski, M., Jorde, L.B., Quinlan, A.R., 2019. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. Elife 8, 1–24. https://doi.org/10.7554/eLife.46922

Schorn, A.J., Gutbrod, M.J., LeBlanc, C., Martienssen, R., 2017. LTR-Retrotransposon Control by tRNA-Derived Small RNAs. Cell 170, 61-71.e11. https://doi.org/10.1016/j.cell.2017.06.013

Schorn, A.J., Martienssen, R., 2018. Tie-Break: Host and Retrotransposons Play tRNA. Trends Cell Biol. 28, 793–806. https://doi.org/10.1016/j.tcb.2018.05.006

Seczynska, M., Bloor, S., Cuesta, S.M., Lehner, P.J., 2022. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. Nature 601, 440–445. https://doi.org/10.1038/s41586-021-04228-1

Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F., Popp, C., Thienpont, B., Dean, W., Reik, W., 2012. The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. Mol. Cell 48, 849–862. https://doi.org/10.1016/j.molcel.2012.11.001

Sexton, C.E., Tillett, R.L., Han, M. V., 2021. The essential but enigmatic regulatory role of HERVH in pluripotency. Trends Genet. 1–10. https://doi.org/10.1016/j.tig.2021.07.007

Sharmistha, M., C., R.D., 2015. P Transposable Elements in Drosophila and other Eukaryotic Organisms. Microbiol. Spectr. 3, 10.1128/microbiolspec.mdna3-0004–2014. https://doi.org/10.1128/microbiolspec.mdna3-0004-2014

Shen, L., Inoue, A., He, J., Liu, Y., Lu, F., Zhang, Y., 2014. Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. Cell Stem Cell 15, 459–471. https://doi.org/10.1016/j.stem.2014.09.002

Shinn, P., Chen, H., Berry, C., Ecker, J.R., Bushman, F., Jolla, L., 2002. HIV-1 Integration in the Human Genome Favors Active Genes and Local Hotspots. Cell 110, 521–529.

Singh, P., Li, A.X., Tran, D.A., Oates, N., Kang, E.R., Wu, X., Szabó, P.E., 2013. De Novo DNA Methylation in the Male Germ Line Occurs by Default but Is Excluded at Sites of H3K4 Methylation. Cell Rep. 4, 205–219. https://doi.org/10.1016/j.celrep.2013.06.004

Smith, L.B., Walker, W.H., O'Donnell, L., 2015. 6 - Hormonal regulation of spermatogenesis through Sertoli cells by androgens and estrogens, in: Griswold, M.D.B.T.-S.C.B. (Second E. (Ed.), . Academic Press, Oxford, pp. 175–200. https://doi.org/https://doi.org/10.1016/B978-0-12-417047-6.00006-5

Smith, Z.D., Chan, M.M., Humm, K.C., Karnik, R., Mekhoubad, S., Regev, A., Eggan, K., Meissner, A., 2014. DNA methylation dynamics of the human preimplantation embryo. Nature 511, 611–615. https://doi.org/10.1038/nature13581

Staub, C., Johnson, L., 2018. Review: Spermatogenesis in the bull. Animal 12, s27–s35. https://doi.org/10.1017/S1751731118000435

Stoye, J.P., Coffin, J.M., 1988. Polymorphism of murine endogenous proviruses revealed by using virus class-specific oligonucleotide probes. J. Virol. 62, 168–175. https://doi.org/10.1128/jvi.62.1.168-175.1988

Suh, D.S., Choi, E.H., Yamazaki, T., Harada, K., 1995. Studies on the transposition rates of mobile genetic elements in a natural population of Drosophila melanogaster. Mol. Biol. Evol. 12, 748–758. https://doi.org/10.1093/oxfordjournals.molbev.a040253

Surani, M.A., Durcova-Hills, G., Hajkova, P., Hayashi, K., Tee, W.W., 2008. Germ line, stem cells, and epigenetic reprogramming. Cold Spring Harb. Symp. Quant. Biol. 73, 9–15. https://doi.org/10.1101/sqb.2008.73.015

Tarlinton, R.E., Meers, J., Young, P.R., 2006. Retroviral invasion of the koala genome. Nature 442, 79–81. https://doi.org/10.1038/nature04841

Taylor, A.L., 1963. Bacteriophage-Induced Mutation in Escherichia Coli. Proc. Natl. Acad. Sci. United States 50, 1043–1051. https://doi.org/10.1073/pnas.50.6.1043

Tchasovnikarova, I.A., Timms, R.T., Matheson, N.J., Wals, K., Antrobus, R., Göttgens, B., Dougan, G., Dawson, M.A., Lehner, P.J., 2015. Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. Science (80-.). 348, 1481–1485. https://doi.org/10.1126/science.aaa7227

Temin, H.M., Satoshi, M., 1970. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus. Nature 226, 1211–1213. https://doi.org/10.1038/2261211a0

Turner, T.N., Coe, B.P., Dickel, D.E., Hoekzema, K., Nelson, B.J., Zody, M.C., Kronenberg, Z.N., Hormozdiari, F., Raja, A., Pennacchio, L.A., Darnell, R.B., Eichler, E.E., 2017. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell 171, 710-722.e12. https://doi.org/10.1016/j.cell.2017.08.047

Ueno, H., Turnbull, B.B., Weissman, I.L., 2009. Two-step oligoclonal development of male germ cells. Proc. Natl. Acad. Sci. U. S. A. 106, 175–180. https://doi.org/10.1073/pnas.0810325105

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sánchez-Luque, F.J., Bodea, G.O.,

Ewing, A.D., Salvador-Palomeque, C., Van Der Knaap, M.S., Brennan, P.M., Vanderver, A., Faulkner, G.J., 2015. Ubiquitous L1 mosaicism in hippocampal neurons. Cell 161, 228–239. https://doi.org/10.1016/j.cell.2015.03.026

Vendrell-Mir, P., Barteri, F., Merenciano, M., González, J., Casacuberta, J.M., Castanera, R., 2019. A benchmark of transposon insertion detection tools using real data. Mob. DNA 10, 1–19. https://doi.org/10.1186/s13100-019-0197-9

Veselovska, L., Smallwood, S.A., Saadeh, H., Stewart, K.R., Krueger, F., Maupetit Méhouas, S., Arnaud, P., Tomizawa, S. ichi, Andrews, S., Kelsey, G., 2015. Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. Genome Biol. 16, 1–17. https://doi.org/10.1186/s13059-015-0769-z

Wang, J., Xie, G., Singh, M., Ghanbarian, A.T., Raskó, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N. V., Schumann, G.G., Chen, W., Lorincz, M.C., Ivics, Z., Hurst, L.D., Izsvák, Z., 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature 516, 405–409. https://doi.org/10.1038/nature13804

Wang, L., Dou, K., Moon, S., Tan, F.J., Zhang, Z.Z., 2018. Hijacking Oogenesis Enables Massive Propagation of LINE and Retroviral Transposons. Cell 174, 1082-1094.e12. https://doi.org/10.1016/j.cell.2018.06.040

Wang, R.A., Nakane, P.K., Koji, T., 1998. Autonomous cell death of mouse male germ cells during fetal and postnatal period. Biol. Reprod. 58, 1250–1256. https://doi.org/10.1095/biolreprod58.5.1250

Weiss, R.A., 2006. The discovery of endogenous retroviruses. Retrovirology 3, 1–11. https://doi.org/10.1186/1742-4690-3-67

Wells, J.N., Feschotte, C., 2020. A Field Guide to Eukaryotic Transposable Elements. Annu. Rev. Genet. 54, 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M., Coffin, J.M., 2016. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc. Natl. Acad. Sci. 113, E2326–E2334. https://doi.org/10.1073/pnas.1602336113

Wolf, D., Goff, S.P., 2009. Embryonic stem cells use ZFP809 to silence retroviral DNAs. Nature 458, 1201–1204. https://doi.org/10.1038/nature07844

Wolf, D., Goff, S.P., 2007. TRIM28 Mediates Primer Binding Site-Targeted Silencing of Murine Leukemia Virus in Embryonic Cells. Cell 131, 46–57. https://doi.org/10.1016/j.cell.2007.07.026

Wolf, D., Hug, K., Goff, S.P., 2008. TRIM28 mediates primer binding site-targeted silencing of Lys1,2 tRNA-utilizing retroviruses in embryonic cells. Proc. Natl. Acad. Sci. U. S. A. 105, 12521–12526. https://doi.org/10.1073/pnas.0805540105

Wolf, G., de Iaco, A., Sun, M.A., Bruno, M., Tinkham, M., Hoang, D., Mitra, A., Ralls, S., Trono, D., Macfarlan, T.S., 2020. Krab-zinc finger protein gene expansion in response to active retrotransposons in the murine lineage. Elife 9, 1–22. https://doi.org/10.7554/eLife.56337

Xu, Q., Xie, W., 2018. Epigenome in Early Mammalian Development: Inheritance, Reprogramming and Establishment. Trends Cell Biol. 28, 237–253. https://doi.org/10.1016/j.tcb.2017.10.008

Xu, R., Li, S., Wu, Q., Li, C., Jiang, M., Guo, L., Chen, M., Yang, L., Dong, X., Wang, H., Wang, C., Liu, X., Ou, X., Gao, S., 2022. Stage-specific H3K9me3 occupancy ensures retrotransposon silencing in human pre-implantation embryos. Cell Stem Cell 29, 1051-1066.e8. https://doi.org/10.1016/j.stem.2022.06.001

Yamashita, M., Emerman, M., 2006. Retroviral infection of non-dividing cells: Old and new perspectives. Virology 344, 88–93. https://doi.org/https://doi.org/10.1016/j.virol.2005.09.012

Yang, L., Malhotra, R., Chikhi, R., Elleder, D., Kaiser, T., Rong, J., Medvedev, P., Poss, M., 2021. Recombination Marks the Evolutionary Dynamics of a Recently Endogenized Retrovirus. Mol. Biol. Evol. 1–33. https://doi.org/10.1093/molbev/msab252

Yang, P., Wang, Y., Macfarlan, T.S., 2017. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. Trends Genet. 33, 871–881. https://doi.org/10.1016/j.tig.2017.08.006

Yang, Q.-E., Oatley, J.M., 2015. 3 - Early postnatal interactions between Sertoli and germ cells, in: Griswold, M.D.B.T.-S.C.B. (Second E. (Ed.), . Academic Press, Oxford, pp. 81–98. https://doi.org/https://doi.org/10.1016/B978-0-12-417047-6.00003-X

Yao, H.H.-C., Ungewitter, E., Franco, H., Capel, B., 2015. 2 - Establishment of fetal Sertoli cells and their role in testis morphogenesis, in: Griswold, M.D.B.T.-S.C.B. (Second E. (Ed.), . Academic Press, Oxford, pp. 57–79. https://doi.org/https://doi.org/10.1016/B978-0-12-417047-6.00002-8

Yeo, J.Y., Goh, G.-R., Su, C.T., Gan, S.K., 2020. The Determination of HIV-1 RT Mutation Rate, Its Possible Allosteric Effects, and Its Implications on Drug Resistance. Viruses. https://doi.org/10.3390/v12030297

Yoth, M., Maupetit-Méhouas, S., Akkouche, A., Gueguen, N., Bertin, B., Jensen, S., Brasset, E., 2023. Reactivation of a somatic errantivirus and germline invasion in Drosophila ovaries. Nat. Commun. 14, 1–15. https://doi.org/10.1038/s41467-023-41733-5

Young, N.L., Bieniasz, P.D., 2007. Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog. 3, 0119–0130. https://doi.org/10.1371/journal.ppat.0030010

Zábranský, A., Hadravová, R., Štokrová, J., Sakalian, M., Pichová, I., 2009. Premature processing of mouse mammary tumor virus Gag polyprotein impairs intracellular capsid assembly. Virology 384, 33–37. https://doi.org/https://doi.org/10.1016/j.virol.2008.10.038

Zamudio, N., Bourc'His, D., 2010. Transposable elements in the mammalian germline: A comfortable niche or a deadly trap. Heredity (Edinb). 105, 92–104. https://doi.org/10.1038/hdy.2010.53

Zhang, Y., Li, T., Preissl, S., Amaral, M.L., Grinstein, J.D., Farah, E.N., Destici, E., Qiu, Y., Hu, R., Lee, A.Y., Chee, S., Ma, K., Ye, Z., Zhu, Q., Huang, H., Fang, R., Yu, L., Izpisua Belmonte, J.C., Wu, J., Evans, S.M., Chi, N.C., Ren, B., 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nat. Genet. 51, 1380–1388. https://doi.org/10.1038/s41588-019-0479-7

Zhang, Y., Maksakova, I.A., Gagnier, L., Van De Lagemaat, L.N., Mager, D.L., 2008. Genome-wide assessments reveal extremely high levels of polymorphism of two active families of mouse endogenous retroviral elements. PLoS Genet. 4. https://doi.org/10.1371/journal.pgen.1000007

Zheng, J., Wei, Y., Han, G.-Z., 2022. The diversity and evolution of retroviruses: Perspectives from viral "fossils." Virol. Sin. 37, 11–18. https://doi.org/https://doi.org/10.1016/j.virs.2022.01.019

Zhou, X., Sam, T.W., Lee, A.Y., Leung, D., 2021. Mouse strain-specific polymorphic provirus functions as cis-regulatory element leading to epigenomic and transcriptomic variations. Nat. Commun. 12. https://doi.org/10.1038/s41467-021-26630-z

Zhu, P., Guo, H., Ren, Y., Hou, Y., Dong, J., Li, R., Lian, Y., Fan, X., Hu, B., Gao, Y., Wang, X., Wei, Y., Liu, P., Yan, J., Ren, X., Yuan, P., Yuan, Y., Yan, Z., Wen, L., Yan, L., Qiao, J., Tang, F., 2018. Single-cell DNA methylome sequencing of human preimplantation embryos. Nat. Genet. 50, 12–19. https://doi.org/10.1038/s41588-017-0007-6

Zirkin, B.R., Papadopoulos, V., 2018. Leydig cells: Formation, function, and regulation. Biol. Reprod. 99, 101–111. https://doi.org/10.1093/biolre/ioy059

Zoch, A., Auchynnikava, T., Berrens, R. V., Kabayama, Y., Schöpp, T., Heep, M., Vasiliauskaitė, L., Pérez-Rico, Y.A., Cook, A.G., Shkumatava, A., Rappsilber, J., Allshire, R.C., O'Carroll, D., 2020. SPOCD1 is an essential executor of piRNA-directed de novo DNA methylation. Nature 584, 635–639. https://doi.org/10.1038/s41586-020-2557-5