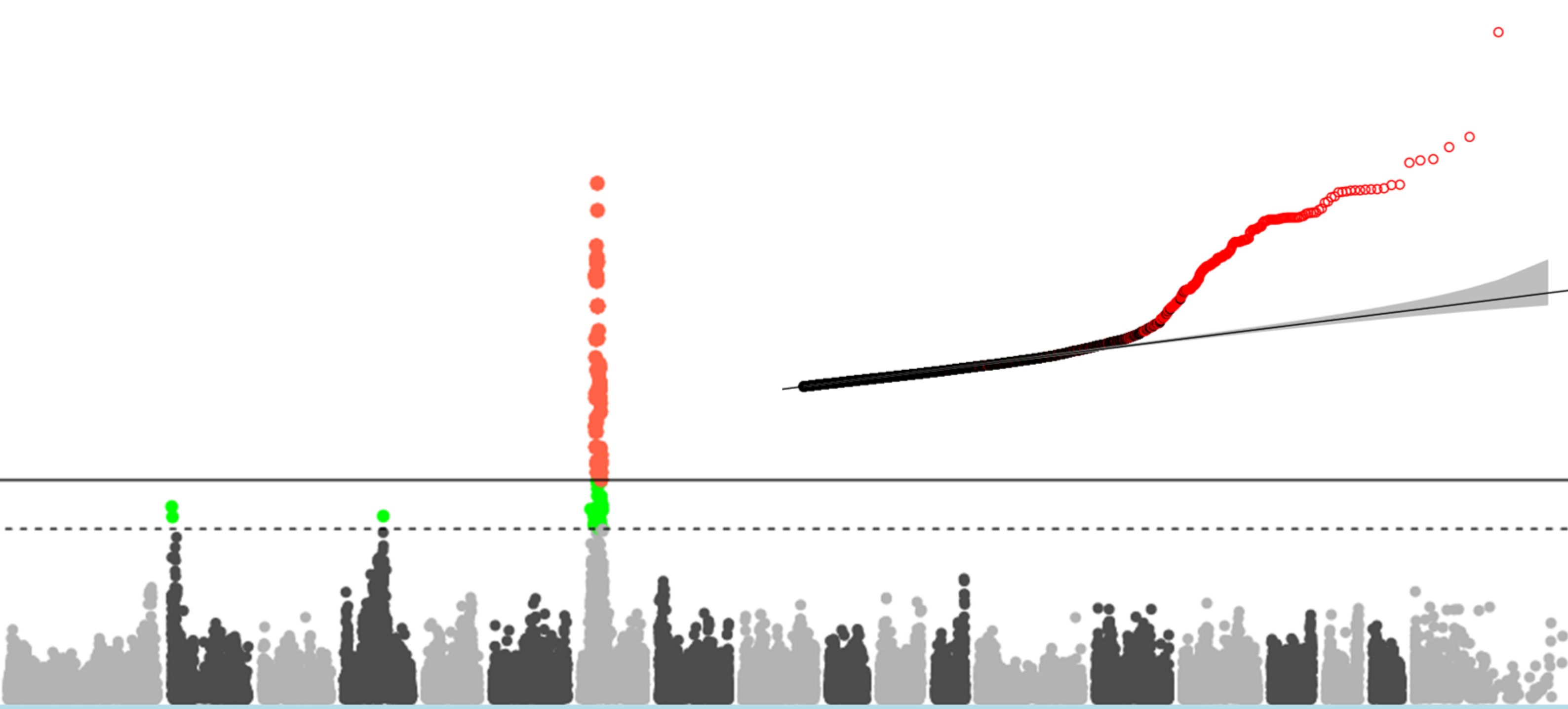
ACADEMIE UNIVERSITAIRE WALLONIE-
EUROPE UNIVERSITE DE LIEGE
FACULTE DE MEDECINE VETERINAIRE
DEPARTEMENT DES PRODUCTIONS
ANIMALES UNITE DE GENOMIQUE

CONTRIBUTION A LA CARTOGRAPHIE, PAR ETUDES DE LIAISON ET D'ASSOCIATION, DE LOCI D'INTERET CHEZ LES ANIMAUX DOMESTIQUES.

CONTRIBUTION TO LINKAGE AND ASSOCIATION MAPPING OF TRAIT LOCI IN LIVESTOCK.



Zhiyan ZHANG

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

ANNEE ACADEMIQUE 2013-2014

Acknowledgments

I would like to express my gratitude to my supervisor, Professor Michel GEORGES, whose expertise, understanding, and patience, added considerably to my graduate experience. I appreciate his vast knowledge and skill in many areas (e.g., statistics, biology, bioinformatics, genetics), and he not only impact me on the scientific parts but also on the truth of human being. He also helps to correct and even almost rewritten the entire papers. A sincerely thanks to my co-promoter Lusheng HUANG, you guide me into the road of scientific research and always correct my way when I feeling lost.

A very great and important thanks goes out to Dr. Tom DRUET, without your step by step teaching and details explanations, I would not be able finished all this works. Dr. Tom DRUET is very careful and preciseness for even small errors. With his teaching by personal example as well as verbal instruction, I've now adjust some careless and hot-headed problems. Lots of statistical knowledge, genetics knowledge and programming skills were learned from Dr. Tom DRUET.

I would like to thank the other members of my Lab, Dr Carole CHARLIER and Dr. Wouter COPPIETERS for the great help in statistics and genetics, especially show me the really amazing things in genetics. Also I would like to thank Dr. Haruko TAKEDA, who just like my old sister takes care of me for lot of things, when I just first time come to the Lab, preparing everything and with great encouragement when I feel discouraged. I would like to thank Olivier STERN, always encourage me to keep fighting during stay at liege, and in my birthday, give me a special beer party.

I must also acknowledge my friends PhD. Students Huijun CHENG, Xin ZHANG, Wanbo Li, and Dr. Li LIN, Lixin ZHANG, Xuewen XU, Ming FANG, for many helps not in living stuff, sharing nearly each noon happy with me, but also in deeply discussion of ideas and tough issue in biology and quantitative genetics. This help to exchanges of knowledge, skills, and venting of frustration during my graduate program and enrich the experience. Without your help, obviously life would more hard and no such happiness.

I would also like to thank my family my parents, do anything they can to support my studies and give every through my entire life. In particular, I must acknowledge my wife and best friend, Jie LIU without whose love, encouragement and editing assistance, I would not finish this thesis. A particular thanks to my son, Jianghan ZHANG, your smiles, the voice calling papa, is my best incentive.

The smooth progress and successful completion of my experiment and paper is never possible without the help of members in this lab. f.i. Nathalie FAUST help to prepare documents for my VISA application, Philippe GAMBRON help to set up my printer, Rodrigo GULARTE MERIDA helps in R parallelization programming….. Sincerely thanks for all kinds of help during these years from you.

In conclusion, I recognize that my living would not possible without the support of Chinese scholarship council (File No. 2008836040); I recognize that this research would not have been possible without the financial assistance of by grants of the Service Public de Wallonie and from the Communauté Française de Belgique (Biomod ARC). I also acknowledge University of Liège (SEGI and GIGA bioinformatics platform) for the use of NIC3 and GIGA-grid supercomputers.

Abbreviations

A.I.	Artificial insemination
BBC	Belgian Blue Cattle
BLAD	Bovine leukocyte deficiency
BLUP	Best linear unbiased prediction
BS	Brachyspina
BVs	Breeding values
CC	Collaborative cross
CI	Confidence interval
CMT	Charcot-Marie-Tooth disease
CNV	Copy number variation
CNVRs	CNV regions
CRC	Calcium release channel
Cs29	Chromosome29
Cs6	Chromosome6
CVM	Complex Vertebral Malformation
DAG	Directed acyclic graph
DH	Draft horses
DNA	Deoxyribonucleic acid
DUMPS	Deficiency in uridine monophosphate synthetase
eQTL	Expression QTL
FDR	False discover rate
GLMMs	Generalized linear mixed models
GRAN	Granulocyte count
GRAR	Granulocyte count percentage
GS	Genomic selection
GWAS	Genome-wide association study
HCT	Hematocrit
HGB	Hemoglobin
HGP	Human genome project
IBD	Identical-by-descent
LB	Lysogeny broth
LD	Linkage disequilibrium
LMMs	Linear mixed-models
LYM	Lymphocyte count
LYMA	Lymphocyte count percentage
MAAT	Marker-assisted association test
MAF	Minor allele frequency
MASA	Marker assisted segregation analysis
MCH	Mean corpuscular hemoglobin
MCHC	Mean corpuscular hemoglobina concentration
MCV	Mean corpuscular volume

MDS	Multidimensional scaling
MLE	Maximum likelihood estimation
MMBIR	Microhomology-mediated break-induced replication
MME	Mixed model equations
MON	Monocyte count
MONR	Monocyte count percentage
MPV	Mean platelet volume
MSTN	Myostatin
NGS	Next-generation sequencing
NMRD	Non-sense mediated RNA decay
OR	Odd ratios
ORF	Open reading frame
PCR	Polymerase chain reaction
PCT	Plateletcrit
PDW	Platelet distribution width
PLT	Platelet count
QCT	Quantitative complementation test
QQ Plots	Quantile-quantile plots
QTL	Quantitative trait loci
RBC	Red blood cell count
RDW	Red blood cell volume distribution width
REML	Restricted maximum likelihood
RLN	Recurrent laryngeal neuropathy
RNF11	RING finger protein 11
SCC	Somatic cell counts
SNP	Single-nucleotide polymorphism
TH	Thoroughbreds
TR	Trotters
UTR	Untranslated region
W	Warmbloods
WBC	White blood cell count

Table of contents

Résumé

Description du sujet de recherche abordé

Jusqu'il y a peu, les valeurs d'élevage des animaux étaient estimées sur base de données phénotypiques mesurées sur l'individu et/ou ses apparentés, et la notion que la covariance entre valeurs d'élevages est proportionnelle au coefficient de parenté entre individus. L'essor de la génomique permet maintenant l'analyse directe du génome et l'identification des loci qui déterminent les valeurs d'élevage des individus. En conséquence, la sélection « assistée par marqueurs » ou « génomique », plus performante, est en passe de remplacer la sélection phénotypique.

L'identification des régions génomiques et des variants génétiques qui contrôlent les phénotypes d'intérêts requiert des méthodes statistiques avancées en constante évolution. Dans le cadre de cette thèse, nous avons (i) contribué au développement de méthodes de cartographie génétique, (ii) appliqué ces méthodes pour cartographier des loci influençant des phénotypes d'intérêt, tant métriques que méristiques, et (iii) contribué au développement de méthodes pour l'utilisation d'information génomique en sélection et production animales.

Résultats

Les méthodes de cartographie que nous avons contribué à développer se distinguent principalement pas le fait que (i) elles exploitent la structure haplotypique du génome (à l'aide d'un modèle markovien caché) ce qui devrait augmenter le déséquilibre de liaison avec les variants causaux et ainsi la puissance de détection, (ii) elles exploitent simultanément l'information de liaison génétique dans les familles et d'association à l'échelle de la population, (iii) elles corrigent pour la stratification en modélisant un effet polygénique aléatoire, et (iv) elles s'appliquent aussi bien à des phénotypes quantitatifs que binaires.

Nous avons ensuite appliqué les méthodes développées (et d'autres) pour la cartographie de loci influençant (i) des paramètres hématologiques chez le porc, et (ii) des caractères binaires comprenant des maladies héréditaires simples ou complexes et des variations génomiques structurelles de type Copy Number Variants (CNV) chez le bovin et le cheval.

In fine, nous avons contribué au développement de méthodes pour l'utilisation d'information génomique en production animale. Nous avons contribué à l'extension de la méthode de cartographie basée sur des haplotypes à des fins d'imputation et avons évalué la précision de celle-ci dans des scénarios proches de la réalité. En outre, nous avons contribué au développement d'une méthode permettant d'identifier des vaches atteintes de mammites dans l'exploitation, par génotypage d'un échantillon de lait de la cuve (mélange de laits de toutes les vaches de l'exploitation).

Conclusions et Perspectives

En conclusion, nos travaux ont mené au développement d'un logiciel (« GLASCOW ») qui est utilisé de façon croissante par la communauté scientifique pour la localisation de gènes influençant des phénotypes à déterminisme complexe, en particulier binaire. Nous avons, en utilisant la méthode développée, contribué à la localisation de régions génomiques influençant plusieurs caractères d'intérêt chez le porc, le bovin et le cheval. Et – in fine - nous avons contribué au développement de méthodes permettant de réduire des coûts d'accès à la technologie génomique, d'une part en complétant du génotypage réel par du génotypage in silico par le procédé d'imputation, et d'autre part en développant une méthode de déconvolution de génotypes obtenus sur mélanges d'ADN.

Summary

Description of the research project

Until recently, breeding values were estimated based on phenotypes measured on the individual and its relatives, and the notion that the covariance between breeding values is proportionate to the kinship coefficient. Advances in genomics now allow for direct analysis of the genome and identification of the loci that determine the breeding values of individuals. As a consequence, marker assisted selection and genomic selection have become more effective and are replacing conventional selection.

The identification of loci influencing the traits of interest requires the use of advanced statistical methods that are constantly evolving. In the context of this thesis, we have (i) contributed to the development of gene mapping methods, (ii) applied these methods to map loci influencing both metric and meristic traits, and (iii) contributed to the development of methods for the integration of genomic information in livestock breeding and management.

Results

The mapping methods that we have helped developing distinguish themselves mainly by the fact that (i) they exploit haplotype information (by means of a hidden markov model) which should increase the linkage disequilibrium with causative variants and hence detection power, (ii) they can simultaneously extract linkage information within families, and linkage disequilibrium information across the population, and (iii) they correct for population stratification by means of a random polygenic effect, and (iv) they can be applied to binary as well as quantitative traits.

We have applied these and other methods to map loci influencing (i) quantitative hematological parameters in a porcine line-cross, and (ii) binary traits including diseases in bovine and non-syntenic Copy Number Variants in cattle, horse and human.

In fine, we have contributed to the development of methods for the utilization of marker information in animal selection and production. We have extended the haplotype-based mapping method to allow imputation and have evaluated the utility of this approach in scenarios mimicking reality. We have also contributed to the development of a method to quantify somatic cell counts

in the milk of individual cows by genotyping a sample of milk from the farm's tank (hence a mixture of milk from all cows on the farm).

Conclusions and Perspectives

Our work has resulted in the development of a software package ("GLASCOW") that is increasingly used by the community to map genes influencing complex traits, primarily binary. By using this tool, we have contributed to the localization of several trait loci in pig, cattle, horse and human. We have contributed to the development of approaches that reduce the costs of genomic analyses in livestock by, on the one hand, complementing real SNP genotypes with genotypes obtained *in silico* by means imputation, and, on the other hand, by developing a method to deconvolute genotypes obtained on DNA pools.

Introduction

Key concepts in association mapping and the use of marker information in livestock production.

Heritable traits

Vive la différence! What defines us as individuals is how we differ from each other. "*How tall am I*?" essentially asks whether I am amongst the tall, average or small ones of my class. Very early in life, we learn to recognize a myriad of the distinctive features of our contemporaries, which we use to recognize, seek contact or rather avoid them.

We also know intuitively that most of these differences do not occur just randomly, as if a deity had bestowed each one of us with a random assortment of features, but that they are largely determined by and hence define our origins. Origins shape distinctive features in three ways: (i) the ***environment*** in which we develop profoundly affects our identity, (ii) our "way of life", or ***cultural heritage***, determines much of whom we are, and (iii) the ***genome*** we inherited from our parents provides each of us with a unique blueprint, or set of instructions that guides our development.

How does the genome inherited by our parents contribute to our phenotypic differences? All of us inherit one genome copy from our father (sometimes referred to as "padumnal"), and genome copy from our mother (sometimes referred to as "madumnal"). They obviously are both "human genomes", being very similar to each other and certainly more similar to each other than to the genome of any other species. Yet they are not identical: aligning our padumnal and madumnal genomes – a still mostly virtual exercise which will however soon become practical – would reveal a different base pair approximately every 1,000 residues. Such differences are referred to as Single Nucleotide Polymorphisms or SNPs of which there are approximately 3 million in a typical human genome. Compiling all SNPs at the population level reveals tens of millions of SNPs. SNPs are characterized by at least two alleles of which one is typically less common than the other: the minor allele. The population frequency of the minor allele (or MAF) allows one to make the rather arbitrary distinction between common SNPs (f.i. MAF ≥ 5%), low frequency SNPs (f.i. 0.5% < MAF < 5%), and rare SNPs (MAF ≤ 0.5%). Approximately 38 million SNPs have now been reported of which ~7 million are common SNPs (f.i. Frazer *et al.* 2009; The 1000 Genomes Project Consortium 2012). SNPs, which include transitions, transversions and single base-pair insertion deletions (indels), are only one type of ***genetic variants***. Others include larger indels, simple sequence repeats (including micro- and minisatellites), copy number variants (CNVs), inversions, and translocations. Copy Number Variants are large genome segment whose copy number varies between individuals. They often coincide with segmental duplications. Known CNV affect an estimated ~4% of our genome and ~13% of our genes (Conrad *et al.* 2010). All genetic variants - whichever their type - originate from germ-line mutations. Every gamete carries of the order 50 to 100 *de novo* point mutations (generating a "derived" from an "ancestral" allele), which primarily results from errors of DNA replication in the germline. Sperm cells, particularly from older men, carry more *de novo* mutations than oocytes as the number of cell divisions to produce spermatozoa are larger than for oocytes (Hurles 2012). The fate of the *de novo* mutations inherited by a conceptus is determined by drift, and – for non-neutral mutations (see hereafter) – by selection. The balance between the gain of new variants by mutation, and the loss of variants by drift, results in a steady state equilibrium characterized – for neutral variants – by a predictable rate of polymorphism with predictable distribution of MAF (Kimura 1983). The expected homozygosity at equilibrium is:

$$H = \frac{1}{4Nem+1}$$

where *Ne* is the effective population size and m the mutation rate per gamete. The majority of SNPs are thought to be largely neutral with respect to phenotype. A minority is assumed to affect gene function, either by altering the gene's expression profile, or by changing the three-dimensional structure and hence function of the gene product. The latter are susceptible to affect the individual's phenotype – they are said to be "causative" SNPs. As many of them encompass genes, CNVs are thought to more often than SNPs affect gene function and hence make a significant impact on phenotypic variation. Causal SNPs may undergo effect of selection (f.i. Bamshad & Wooding 2003; Sabeti *et al.* 2006; Cutter & Payseur 2013). Negative selection against deleterious variants will reduce the level of polymorphism and shift the MAF distribution to lower values. Positive selection and balancing selection will leave their own signatures on the genome. Selection on causative variants may affect the fate of their neural neighbors.

Alleles may – in principle – differ "epigenetically" (i.e. by virtue of distinct heritable DNA or chromatin modifications) rather than in their sequence. Such ***metastable epialleles*** have been shown to segregate in plant populations and contribute to phenotypic differences (f.i. Hauser *et al.* 2011). It is generally believed however that the epigenetic status of genes is largely reset in the mammalian germline, precluding the widespread occurrence of epialleles in mammals, except for the marks that differentiate the padumnal and madumnal alleles of parentally imprinted genes (f.i. Morgan & Whitelaw 2008).

For a minority of traits, inter-individual differences are entirely determined by genetic variants at one gene. Such "Mendelian" traits are said to be ***monogenic***. The causative variants may be (completely or partially) recessive or dominant. The vast majority of monogenic traits are inherited diseases (including "inborn errors of metabolism") in humans, and inherited diseases and coat color variants in domestic animals. They commonly involve severe, recessive loss-of-function variants dominated by nonsense, frameshift, splice-site and damaging missense variants in protein coding genes.

The vast majority of phenotypes, including common diseases and agricultural important phenotypes, have a multifactorial or complex determinism. Inter-individual differences are determined by environmental, cultural and genetic factors. Genetic effects are generally assumed to be "***polygenic***", i.e. depend on multiple genetic variants affecting multiple genes. The number of genes involved remains largely unknown for most traits. The distribution of allele-substitution effects appears to be exponential, i.e. variants with large effects are less numerous than variants with small effects (f.i. Hayes & Goddard 2001). There also appears to be an inverse correlation between effect size and MAF, which is thought to primarily reflect purifying selection against variants with large effects (f.i. Manolio *et al.* 2009). Molecular evidence from model organisms suggests that epistatic interactions between polygenes might be commonplace, i.e. that the effect of a genotype at one locus is dependent on the genotype at another locus (f.i. Bloom *et al.* 2013). Yet, initial studies have not revealed a major contribution of epistatic effects to the variance of most studied traits in human and livestock (f.i. (Cordell 2009). Complex phenotypes include continuously distributed quantitative traits (f.i. most production traits in agriculture), as well as binary traits (f.i. common complex diseases). It is noteworthy that a detailed analysis of inherited defects that are generally labeled monogenic, often reveals instances of incomplete penetrance and variable expressivity. Monogenic traits therefore often appear only simple on the surface.

The proportion of the inter-individual variation – for a trait of interest – that is due to genomic differences is called the ***heritability*** (H^2) of the trait (Visscher *et al.* 2008). H^2 is a population-specific parameter, i.e. the same phenotype may have different H^2 in different populations or even in the same populations at different times. This is due to the fact that the panoply of segregating sequence variants (and their MAF) as well as of non-genetic factors influencing the phenotype are most likely to differ between populations. A common way to estimate H^2 in humans is to compare the resemblance between monozygotic and dizygotic twins. Monozygotic and dizygotic twin pairs are assumed to be equally exposed to environmental and cultural influences, but differ in their degree of genetic resemblance: monozygotic twins are genetically identical while dizygotic twins are genetically as related as non-twin sibs. A higher phenotypic resemblance between monozygotic than between dizygotic twins does support a quantifiable contribution of genomic polymorphisms to trait variation. In domestic animals, the heritability is typically estimated from the correlation between the phenotypic and genetic resemblance or kinship, increasingly using the mixed "individual animal" model. It is assumed in these studies that genetic resemblance is not correlated with environmental resemblance. The individual animal model rests on Fisher's mathematically convenient infinitesimal model, i.e. the trait is influenced by an infinitely large number of variants with individually minute effects that are evenly scattered throughout the entire genome. The "broad sense" H^2 heritability can be partitioned in an additive component (narrow sense h^2) and a non-additive residual. h^2 is of particular interest in agriculture as it constrains the success of selection programs. The heritability of a complex binary trait (such as a common complex disease) is generally estimated by assuming the existence of an underlying (non-observed), continuously distributed ***"liability" with threshold*** value separating affected from non-affected individuals.

Identifying causative variants

Motivation and basic principles. Identifying the causative variants influencing heritable traits of medical and agronomic importance is one of the most active areas of research in the life-sciences. This is due to the fact that recent advances in genomic technologies makes this one of the most effective experimental designs to improve our understanding of the molecular mechanisms underpinning disease and agricultural production, which may contribute to the development of improved methods of diagnosis (medicine) and selection (agriculture), as well as of treatment (medicine) and production (agriculture).

The basic principles of the "forward genetic" approach towards identifying causative variants influencing a trait of interest are extremely simple and based on the examination of the correlation between phenotype and genotype (for a given variant). Thus, assume a population of individuals that have been (i) evaluated for the phenotype of interest, and (ii) for which the entire genomic sequence has been determined. In principle one can measure the correlation between phenotype and genotype for all variants. Practically this is done either by sorting the individuals by phenotype (f.i. cases vs controls) and checking for different genotype frequencies between groups, or by sorting the individuals by genotype and checking for different phenotype means between groups. One expects such a correlation to exist for causative variants.

Avoiding spurious associations. The issue is that such phenotype-genotype correlation may also exist for "passenger" (i.e. non causative) variants. This will be the case if the genotype at the passenger variants is correlated with the genotype at causative variants, or with environmental or cultural effects that influence the phenotype (f.i. Platt *et al.* 2010).

Correlation between genotype at passenger and causative variants is expected for closely linked variants. The corresponding correlation is referred to as "***linkage disequilibrium***" (LD) or "gametic association". *De novo* mutations are initially completely associated with the haplotype characterizing the chromosome upon which they occurred (a haplotype is a combination of alleles for a set of neighboring variants). If the newly derived allele spreads in the population it will progressively re-assort with distinct haplotypes by meiotic recombination. With time the initial association should erode and equilibrium attained (i.e. independent genotypes at neighboring variants). Yet, drift continuously regenerates LD. At equilibrium, the expected squared correlation between the genotypes of adjacent variants is:

$$r^2 = \frac{1}{4Neq+1} + \frac{1}{n}$$

where q is the recombination rate between the variants, and *n* the sample size in which r^2 is measured. Unless LD between passenger and causative variants is perfect (r^2=1), the correlation between phenotype and genotype should be highest for the causative variant. Yet, this prediction may not apply if multiple closely linked causative variants co-segregate in the population. Such "allelic heterogeneity" appears to be the rule rather than the exception, at least in human populations. Passenger variants that are in LD with multiple causative variants may by chance be more strongly associated with the phenotype than the individual causative variants, a phenomenon referred to as "***synthetic association***" (f.i. Dickson *et al.* 2010; Platt *et al.* 2010). At present the best way to untangle such dependencies is to simultaneously fit multiple if not all variants in a "multivariate" analysis. Thus, one wants to estimate the effect of each variant on phenotype conditional on the genotype of all other variants in the vicinity. This approach is only applicable for variants that are not in perfect LD with each other (i.e. it is impossible – using this approach - to differentiate passenger and causative variants that are in perfect LD in the studied population). Multi-colinearity issues make it sometimes even difficult to differentiate variants that are in high, although not perfect LD.

Correlation may also exist between passenger variants and non-syntenic (markers located on a different chromosome) causative variants. This is a very common occurrence in domestic animals, particularly in cattle population relying extensively on artificial insemination. Assume a polygenic trait such as milk production. Highly significant "sire effects" are commonplace. Any rare variant carried by a sire with superior breeding value would "tag" its descendants and be associated with increased milk production because of its association with the polygenic background underlying the high breeding value. This is one example of spurious association due to "***population stratification***". In this example, the trait of interest is directly affected by the causal polygenes. In another form of stratification, the studied population comprises sub-groups exposed to distinct environments or cultural influences, which are influencing the phenotype of interest. Variants tagging sub-populations will show association with the phenotype by virtue of their correlation with the sub-group specific non-genetic effects. The spurious associations due to stratification can be avoided by explicitly modeling the sub-populations. Genome-wide marker information can be used to uncover the underlying population structure. This can be done by unsupervised clustering using for instance the STRUCTURE programs (Pritchard *et al.* 2000), by means of principal components using for instance the EIGENSTRAT programs (Price *et al.* 2006), or – increasingly – by modeling a random polygenic effect with covariance structure proportionate to genome-wide kinship estimated from genotype data (f.i. Price *et al.* 2010).

An elegant and effective approach to avoid spurious association of passenger variants that are not closely linked to causative variants is to simultaneously test for linkage and association. In humans, this is typically achieved by analyzing parent-offspring trios and performing a "***transmission disequilibrium test***" (TDT) (f.i. Ewens & Spielman 2003). In the case of a binary trait such as a disease, the TDT tests whether a specific variant is over-transmitted by heterozygous parents to affected offspring. This will only be the case if the analyzed variant is associated (at the population level) and "linked" (at the familial level) with a causative variant (or itself a causative variant). Because of their specific structure, domestic animal populations offer ample opportunity to simultaneously extract linkage and association information. LD- and linkage information can be merged to estimate identity-by-descent (IBD) probabilities for all pairs of chromosomes in the dataset and these can be used to test whether a chromosome region is associated and in linkage with variants influencing the trait of interest (f.i. Meuwissen *et al.* 2002; Druet & Georges 2010).

Stratification will lead to an overall inflation of the test statistic for association. The occurrence of residual stratification effects can therefore be evaluated by examining the distribution of the test statistic (for all or a selection of variants scattered throughout the genome) using – for instance - a quantile-quantile (QQ) plot. A shift towards lower p-values suggests stratification. To control the level of false positives, the thresholds to declare significance can be decreased accordingly in a procedure referred to as "***genomic control***" (f.i. Devlin & Roeder 1999). It should be noted that the shift towards lower p-values resulting from population stratification will in some cases be due to actually causative polygenic variants. Some people have therefore rightfully argued that this procedure "*throws the baby with the bathwater*".

Compensating for incomplete genotype information. Thus far, we have made the presumptuous assumption that the entire genome sequence would be available for all individuals in the dataset. While this may become reality in the future, it is not yet the case. This hasn't stopped geneticists from engaging very actively in genome-wide association studies (GWAS) in many organisms, including men and domestic animals. This was made possible because of the pervasive linkage disequilibrium across the genome in these organisms. In most Out-of-Africa human populations, a panel of passenger SNPs with a density of ~ 1 SNP per 5Kb captures ~80% of common causative variants wit $r^2 \geq 0.8$ (f.i. The International HapMap Consortium 2007; Bhangale *et al.* 2008). It has become customary in human genetics to perform GWAS using SNP panels comprising between 300,000 and > 1 million SNP variants, which can be cost-effectively interrogated using commercially available micro-arrays. As LD extends over longer genomic regions in domestic animals than in human, GWAS are typically performed using panels interrogating an order-of-magnitude less markers than in human, i.e. from 50,000 to 750,000 (Goddard & Hayes 2009). The use of such panels should thus allow the identification of regions of the genome encompassing common causative variants. However, identification of the actual causative variants requires subsequent targeted "fine-mapping" efforts.

In some instances, the per-SNP information content has been optimized by taking advantage of prior knowledge about the LD structure of the genome (The International HapMap Consortium 2005, 2007, 2010). The human HapMap project genotyped 270 individuals representing three major ethnic groups for ~3.1 million common SNPs. Examination of the LD patterns between closely linked SNPs revealed a step-wise rather than gradual decrease in LD with distance. LD was found to be high within ~50Kb segments of the genome referred to as ***haplotype blocks***, which are separated from each other by recombination hotspots causing abrupt drops in LD. The typical haplotype block comprises 5 to 10 common haplotypes accounting for the majority of the chromosomes observed in the population. Rather than selecting SNPs at random, some manufacturers of SNP genotyping arrays selected panels of "***tagging SNP***" that tag as many common haplotypes as possible. Similar strategies, albeit with less resolving power, were applied to develop some of the SNP genotyping arrays used in domestic animals (f.i. Matukumalli *et al.* 2009).

So far, the vast majority of GWAS studies have used single SNP association tests. Thus, SNP genotype frequencies were compared between cases and controls or the effect of SNP genotype on the quantitative trait of interest was evaluated. The power to detect the effect on phenotype of a unseen causative variant by means of an interrogated "marker" SNP is strongly dependent on the degree of LD between the marker SNP and the causative variant. Assuming that a sample size of *n* would be needed to detect the effect of the causative variant if it were directly interrogated, a sample size of n/r^2 will be needed to detect the same effect via a marker SNP in LD of r^2 with the causative variant (as nr^2 corresponds to the Pearson test statistic for independence (f.i. Balding 2006). The detection power thus drops rapidly with decreasing LD between interrogated and causative SNPs. One way to overcome this is to perform association analyses using haplotypes, i.e. combination of adjacent SNPs. Ideally this requires "phasing" of the genotype data, i.e. sorting the alleles by parental origin. This is most reliably accomplished using genotype information from the parents. However, even in the absence of parental information (which is the most common scenario), the most likely linkage phase can be estimated with some degree of accuracy for strings of SNPs in LD. The hope of the ***haplotype-based approaches*** is that one of the haplotypes will be in higher LD with the causative variants than anyone of the composite SNPs considered individually, hence increasing the association signal. Many haplotype-based approaches use "windows". These can be sliding window with fixed number of SNPs (Lin *et al.* 2004). Alternatively the boundaries of the windows can be set such as to coincide with the limits of known haplotype blocks. Throughout this thesis, we use a Hidden-Markov-Model based approach that obviates the need for windows (f.i. Druet & Georges 2010).

An alternative approach to extract more LD information from the incomplete set of genotyped SNPs that is being extensively used in human genetics is "***genotype imputation***". Imputation corresponds to the *in silico* prediction of an individual's genotype for variants that have not been genotyped experimentally (Marchini & Howie 2010). In simplified terms, this is done by identifying individuals in a very densely genotyped reference population (f.i. the HapMap or full-sequenced 1,000 Genomes Project populations) that regionally carry the same haplotypes as the "individual to impute". The dense genotypes of the corresponding haplotypes are then projected from the reference population to the study population. This can be accomplished for the entire genome. GWAS are subsequently conducted one SNP at the time using genuinely genotyped SNPs as well as (the often more numerous) imputed SNP. All variants cannot be imputed with equal accuracy: low frequency variants or variants located in recombination hotspots are typically more difficult to impute. Such variants are penalized in the association studies, which complicates comparison between SNPs. Imputation has also been essential to merge datasets genotyped with different arrays in common meta-analyses.

Accounting for multiple testing. In nearly all instances, the objective of GWAS is to pinpoint chromosome regions that are thought to encompass truly causative variants. Therefore, one has to define a threshold for the test statistic above which to reject the null hypothesis of absence of association. This threshold has to account for the fact that a genome-scan implies the realization of many tests. Using a nominal threshold corresponding to a type-I error rate of 5% (for a single test) would thus generate ~5,000 false positive associations when testing 100,000 "independent" (not in LD) SNPs, even in the absence of a single true genetic effect. The traditional way to deal with this multiple testing issue is to adapt the threshold for the number of independent tests performed using either a ***Bonferroni or related Sidak correction***. This requires the determination of the number of independent test performed, which can be achieved using a variety of approaches often exploiting permutation testing. In human genetics, the recommended threshold for GWAS corresponds to a nominal p-value of 10^{-8}, implying the realization of 5 million independent tests (Hirschhorn & Daly 2005). Thresholds applied in animal genetics are typically somewhat more lenient as the number of tested SNPs is considerably lower and LD assumed to be more pronounced. In addition to imposing these very significant significance thresholds, good practice guidelines demand confirmation of the significant hits in an independent data set to warrant publication in the best journals.

An alternative approach, rather than considering individual p-values independently, exploits information from the distribution of p-values across all tests performed. If all tests correspond to true null hypotheses, the distribution of p-values is expected to be uniform, i.e. 5% of tests will have p-values between 0 and 5%, 5% will have p-value between 5 and 10%, etc. An excess of test with low p-values suggest the occurrence of true alternative hypotheses amongst the tests performed. Assume that 20% of the tests have a p-value between 0 and 5%, this implies that ~

three out of four of these tests are true alternative hypotheses. If we select these tests as "positive discoveries", we can therefore expect a ***false discovery rate (FDR)*** of one in four. The distribution of p-values across all tests performed is typically examined using QQ-plots. The exact FDR for individual tests can easily be computed using standard theory. The FDR approach will typically be more efficient at identifying true alternative hypotheses if these represent a large enough proportion of realized tests (Storey & Tibshirani 2003).

Rare variants. It was recognized from the onset that the SNP panels used for GWAS were best suited to tag common causative variants. Population genetic arguments supported the notion that common complex diseases would involve common risk variants, i.e. the so-called ***Common Disease Common Variant Hypothesis (CDCV)*** (Reich & Lander 2001). However, as surmised by some very early on (Pritchard 2001), it has become increasingly apparent that low frequency and rare risk variants also contribute to inherited risk and the heritability of quantitative traits in human populations. As a matter of fact low frequency and rare variants appear to have larger effects than common variants resulting in stronger purifying selection (which reduces their frequency) (Gibson 2012). The contribution of rare variants to the heritability of agriculturally important traits in livestock remains largely unknown.

Detecting rare causative variants poses specific challenges. Rare variants are typically poorly tagged by interrogated SNPs, certainly if analyzed one-by-one. It is likely that GWAS will soon be conducted with exome-wide followed by genome-wide resequence data rather than SNP genotype data, which should alleviate this detection issue. However, as rare variants are – by definition – *rare*, performing an association test remains difficult as too few individuals carry the variant to allow for the realization of a meaningful test. This limitation becomes obvious for "singletons", i.e. variants that are only observed once in the studied data set. A singleton observed in a case confers an infinitely large relative risk that will, however, never be significant: it's p-value is 0.5.

One way to include rare variants in association studies is to analyze them in "aggregate" rather than individually. The first such a family of approaches are the ***"burden tests"***. The underlying premise is that variants disrupting the function of the causative gene will be enriched in individuals with extreme phenotypes. In case-control studies, cases can be considered as extremes. For quantitative phenotypes one can select individuals in the tails of the distribution.

Given our present limited understanding of molecular biology, the only variants that can confidently be predicted to be "disruptive" are those affecting the protein sequence, including nonsense, frame-shift, splice-site, and – to a lesser extend – missense variants. Performing a burden test therefore typically consist in (i) sequencing one or more (or all) genes of interest in extreme individuals, (ii) identifying low frequency and rare variants that are predicted to disrupt gene function, and (iii) comparing the cumulative frequency of the corresponding variants between opposite extremes (f.i. cases vs (super-)controls) (Bansal *et al.* 2010). This approach has been applied to several common diseases and has unexpectedly revealed as many cases of enrichment of rare risk variants in cases, as of enrichment of rare protective variants in controls (Nejentsev *et al.* 2009; Momozawa *et al.* 2011; Rivas *et al.* 2011).

One of the limitations of the burden tests is that they assume that all disruptive variants in given gene affect the phenotype in the same manner: either all of them increase the phenotype (f.i. disease risk), or all of them decrease it. The C-alpha test has been developed to overcome this limitation for low frequency variants (it is not applicable to singletons). It looks for an aggregated overdispersion of the distribution of disruptive variants (in a given gene) between f.i. cases and controls, i.e. the fact that some of the variants tend to preferentially cluster in cases, while others tend to preferentially cluster in controls (Neale *et al.* 2011).

Identifying causative genes

Identifying causative variants influencing diseases and agronomically important traits is certainly one of the major objective of GWAS, yet identifying the genes which these variants perturb, i.e. the causative genes, is certainly equally if not more important. Indeed, it is this knowledge that paves the way to improved treatment regimes in medicine.

If the causative variants are coding variants, i.e. they change the amino-acid sequence of the gene, the identity of the target causative gene leaves little doubt. There is growing evidence however that a substantial proportion of causative variants are regulatory variants, affecting the expression profile of target genes rather than the structure of their product. As cis-acting regulatory elements can be hundreds of thousands and even millions of base-pairs away from the genes they regulate, the identification of the causative genes perturbed by regulatory variants remains a major challenge.

An extremely elegant genetic test of gene causality has been developed and applied in the model organisms *D. melanogaster* and *S. cerevisiae*: the ***reciprocal hemizygosity test*** (Long *et al.* 1996; Steinmetz *et al.* 2002). The typical scenario in which this test becomes useful in model organism, is following the mapping of a Quantitative Trait Locus (QTL) to a specific location in an intercross population. Assuming that the confidence interval for the QTL contains a number of genes, the question becomes which one or several of the genes in the interval underlie the QTL effect. To respond to that question pairs of reciprocal F1 hemizygotes are generated for all positional candidate genes. This can be done increasingly effectively by homologous recombination. Thus for each gene, a pair of hemizygotes is produced by knocking out the allele coming from either the "A" or the "B" strain. If the examined gene is not involved in the QTL effect, the reciprocal hemizygotes will be functionally equivalent: their phenotype will not differ significantly (yet may differ from the original F1 individuals). If, on the contrary, the gene is the causal gene, one reciprocal hemizygote will be functionally "Q-", while the other will be "-q" (in which Q and q represent the alternate alleles at the QTL). This will cause the reciprocal hemizygotes to differ phenotypically.

Generating series of reciprocal hemizygotes is obviously an arduous task in organisms other than yeast. A less demanding variation of the reciprocal hemizygosity test is "***quantitative complementation***" (Mackay 2001; Georges 2007). In this approach, one generates one knock-out per positional candidate gene, which is subsequently mated to animals from the mapping population in order to generate "A-" and "B-" animals. The premise of the test is that the contrast between the phenotype of these alternative hemizygous lines will be larger than the "A+" versus "B+" contrast if the candidate gene is causative, and not otherwise. Quantitative complementation has been applied in *D. melanogaster* and a couple of times in the mouse (Yalcin *et al.* 2004). Its application in outbred populations including human and domestic animals may appear impossible. Yet, naturally occurring null alleles for a substantial number of genes segregate in these populations at sometimes appreciable frequencies. It may thus, at least in theory, be possible to identify individuals with "C-" and "c-" genotype (where C and c would be alternative alleles for a causative variant) and compare their phenotypes. This approach has been proposed and applied once in cattle (Karim *et al.* 2011).

The other formal test for gene causality - which is easier to apply in outbred populations - is the previously described “***burden test***”. Imagine that a locus influencing a trait of interest has been identified by GWAS and that it encompasses n genes. The positional candidate genes can be deeply sequenced in cohorts of extreme individuals (f.i. cases and controls). The demonstration of a differential “burden” of rare disruptive mutations in either cases or controls for one of the candidate genes would unambiguously identify the causative gene(s). One of the main difficulties with the burden test is to reliably identify disruptive mutations. Contaminating the collection of candidate disruptive mutation with neutral ones rapidly undermines statistical power. It has recently been suggested that it might be better to avoid missense variants (and only use stop-gains, frameshift and essential splice site variants) in the burden test, as well as to incorporate information about de novo mutations (requiring the analysis of parent-offspring trios) (He *et al.* 2013).

While elegantly simple in principle, the burden test requires sequencing of very large cohorts to achieve adequate power. This power is rapidly eroded by multiple testing if one intends to study many positional candidate genes. This is probably one of the main reasons why application of the burden test at genome-wide level by exome sequencing has not yet yielded the results that were hoped for. It is thus advisable to carefully preselect the candidate genes to subject to a burden test. One way to do this is to first apply the burden test to genes mapping to GWAS-identified loci. One can additionally increase the prior probability of success by applying ***network analysis***. Gene networks that are shown to be significantly overpopulated within such sets of GWAS-identified positional candidates are likely to be enriched in genuine causative gene (Raychaudhuri *et al.* 2009; Rossin *et al.* 2011). An alternative strategy to prioritize positional candidates is to use ***eQTL information*** if available. Thus, an eQTL association signal for a positional candidate gene in a phenotype-relevant cell type that would resemble the association signal for the phenotype would be a strong candidate gene to subject to a burden test (Montgomery & Dermitzakis 2011).

It remains unclear whether the burden test is applicable to all types of causative genes. Many gene products fulfill distinct functions in multiple tissues and at different stages of development, driven by tissue/timing-specific enhancers. Disruptive coding variants, which form the basis of the burden test, are affecting all these functions without discrimination. Could it be that specific phenotypes are not associated with a differential burden of disruptive mutations in the coding sequence, but rather in tissue/timing-specific regulatory elements?

Missing heritability

Hundreds of GWAS have been conducted in humans, domestic animals and plants, for a broad range of medically and agriculturally important phenotypes. Risk loci and QTL have been identified for nearly all examined traits amounting to thousands of hits. Yet a systematic finding is that the identified loci typically only account for a small fraction of the heritability of the studied trait. This recurrent observation has raised the issue of the “missing heritability” and what its underlying causes might be (Manolio *et al.* 2009). The factors contributing to the missing heritability are most likely multiple. We will herein briefly survey the contributing factors that are most commonly invoked.

One possible contributor to the missing heritability that has not received a lot of attention is the possibility that the identified loci explain more of the trait variance than what has been assumed. The variance explained is typically computed by assuming that the lead SNP is the only causative variant in the locus. Recent fine-mapping efforts strongly suggest that – at least in humans and rat– most loci harbor multiple common causative variants (Michael *et al.* 2013 and unpublished observations), i.e. that ***allelic heterogeneity*** is the rule rather than the exception. Accurately computing the variance explained by the locus should account for this complexity and this will nearly certainly increase the variance explained.

The claim of missing heritability, i.e. the fact that the identified risk variants only explain part of the trait heritability, assumes that the heritability is estimated accurately. Trait heritabilities are classically estimated from “epidemiological” data. One may rightfully question the accuracy of this approach. Especially in humans, genetic and environmental resemblance are often confounded. It has even been argued that monozygotic twins might be treated more uniformly than dizygotic twins, hence possibly leading to overestimated heritabilities from twin data. More recently, it has been suggested that ***heritability estimates could be inflated*** if epistatic effects were contributing to trait heritability (Zuk *et al.* 2012). It is worthwhile noting in this regard that Peter Visscher and colleagues recently proposed to use genome-wide genotype data within families to provide unbiased estimates of at least the narrow sense h^2 (Visscher *et al.* 2006). The method is based on the estimation of the correlation between the phenotypic resemblance between sibs and the estimated fraction of their genomes inherited identical-by-descent. While being very insightful, the approach unfortunately requires a very large number of sibs to yield accurate estimates.

A third possible contributing factor to the missing heritability is the ***incomplete genome coverage*** of the utilized SNP panels. This seems less and less likely to be a major contributing factor as SNP panels continue to improve. Yet, it is certain that some part of the genome remain poorly tagged. This is probably the case for variants located within recombination hotspots as well as for variants mapping to segmental duplications. CNV also remain difficult to interrogate and are ignored in most GWAS. Although data suggest that most common CNV are satisfactorily tagged by flanking SNPs (McCarroll *et al.* 2008), part of the missing heritability might still be hiding in these difficult to interrogate parts of the genome.

One of the most thoroughly scrutinized hypotheses to account for the missing heritability is the potential ***importance of low frequency and rare causal variants***. The SNP panels that have been used predominantly for GWAS until recently primarily include common variants that are by definition (at least if used one at a time) only suitable to tag common causal variants. The need to be able to better study contribution of low frequency rare variants to disease heritability is one of the major drivers of the 1,000 Genomes Project (Nielsen 2010). SNP panels interrogating low frequency coding variants have been designed based on the ensuing information and are presently being used in GWAS with very large case-control cohorts. Genotype imputation is now routinely used to genotype “GWAS-sed” cohorts *in silico* for low frequency variants detected in the 1,000 Genomes Project. Moreover, a growing number of GWAS are presently being conducted using exome-sequencing, waiting for genome-wide resequencing to become an affordable norm. A growing number of studies have performed targeted resequencing of candidate genes (f.i. positional candidate genes from GWAS; cfr. above) to evaluate the contribution of rare variants. At present, the main message in human genetics appears to be that rare risk variants indeed do exist, that their effects indeed appear to be larger than those of common variants, yet that they only explain a very small fraction of the genetic variance and hence missing heritability (Momozawa *et al.* 2011; Rivas *et al.* 2011).

Another hypothesis that is receiving growing attention is the ***"polygenic" or "quasi-infinitesimal" hypothesis***. According to this view many complex traits, in medicine and agriculture, would be influenced by a very large number of genetic variants (i.e. thousands to tens of thousands) with individually very small effects. The variance accounted for by most such variants would be too small to have been detected by GWAS so far. There is growing experimental support for this polygenic hypothesis. In human genetics, the first wave of GWAS were typically conducted with of the order of 1,000 cases and 1,000 controls and typically yielded at most a handful for genome-wide significant loci. Merging datas*et al.*lowed for a second generation of meta-analyses, which typically generated additional loci at a rate that was strongly related to the size of the analyzed cohorts. Power calculations clearly indicated that many loci were still missed (i.e. loci were detected with effects size that had only limited probability to be detected). Merging ***meta-analyses*** in even large datasets indeed continued to reveal additional loci. This trend has been very clear for height (Lango Allen *et al.* 2010; Berndt *et al.* 2013) and – more recently – for schizophrenia (Ripke *et al.* 2013). Additional evidence in support of the polygenic hypothesis came from the application of "***genomic selection***" and related methods to animal data first, and – more recently – human data. As previously mentioned, in domestic animals trait heritabilities are typically estimated using the mixed "individual animal model" (Lynch & Walsh 1998). This linear model includes a random polygenic (additive) effect proper to each animal. The covariance between individual animal effects are assumed to correspond to $2Qs_A^2$, where Q is the kinship coefficient for the considered pair of individuals and s_A^2 the additive genetic variance. Kinship coefficients are classically computed from genealogical records, allowing estimation of s_A^2 and hence the heritability. If the heritability is known, the individual animal effects (corresponding to their breeding values) can be estimated as Best Linear Unbiased Predictors (BLUPs). More recently it has become possible to estimate kinship coefficients of pairs of individuals using genome-wide SNP genotypes obtained using SNP genotyping arrays that are now available for all major domestic animal species. Not surprisingly, estimates of s_A^2 obtained with the corresponding variance structure are very similar to those obtained on the basis of genealogical records, leading some scientists to claim that "*There is no missing heritability* *problem in animal genetics!*". Individual breeding values can be estimated using the corresponding covariance structure using so-called "***GBLUP***". Yang *et al.* (2010) applied this GBLUP model to human height data and showed that the polygenic effect captured ~50% of the phenotypic variance, while QTL mapped by association in the same data set would explain less than 10%. They suggested that a better SNP panel might have allowed them to explain as much as ~80% of the phenotypic variance. The same models are increasingly been applied to common complex diseases providing growing support that the polygenic hypothesis may accounts for a substantial fraction of the missing heritability (Cross-Disorder Group of the Psychiatric Genomics Consortium 2013). GBLUP assumes that all segments of the genome account for the same proportion of the genetic variance. It is obvious that this assumption will not be valid in most circumstances. To better account for possible heterogeneity in explained variance, more sophisticated ***Bayesian models*** are being developed that assume various prior distributions of QTL effects. Application of these models results - at least for some traits – in capturing a higher proportion of the trait variance and making more accurate predictions of individual "breeding values" (Hayes *et al.* 2010). It is noteworthy that the same models are increasingly be used to estimate the effects of all genetic variants simultaneously, and hence hopefully improve the distinction between causative and passenger variants. In less than five years, genomic selection has revolutionized animal breeding and has become the method of choice to identify elite breeding stock (Goddard & Hayes 2009). It seems reasonable to speculate that the same methodologies may contribute to the development of novel diagnostic approaches in human medicine (de los Campos *et al.* 2010).

A commonly proposed source of missing heritability is ***epistasis***. The idea is that the effects of some genetic variants are dependent of the genotype at one or more other variants. Properly modeling such epistatic interactions might explain a higher proportion of the trait variance than by considering the marginal effects only. Recent experiments conducted in experimental crosses of yeast showed that gene-by-gene interactions explained between 0 and 50% of the heritability and that detectable pairwise interactions explained from 0 to 71% of this epistatic variance (Bloom *et al.* 2013). So far, the search for gene-by-gene interactions has been less successful in outbred populations for reasons that remain only partially understood (Hill *et al.* 2008).

Another suggested cause of missing heritability are ***parent-of-origin effects***. The mammalian genome is known to harbor ~150 genes that are subject to parental imprinting (http://www.geneimprint.com/site/genes-by-species). For these genes, and although being diploid, the organism only uses one allele (in all or some tissues): the padumnal allele for halve of the imprinted genes, the madumnal allele for the other halve. Kong *et al.* (2009) reanalyzed GWAS data (for three common complex diseases) in the vicinity of known imprinted genes assuming an “imprinting model”. They indeed observed 10 associations with parent-of-origin effects. More sophisticated parent-of-origin effects are known to exist. The best understood is probably the callipyge phenotype, which is only expressed in heterozygous animals inheriting the *CLPG* mutation from their sire ($+^{Mat}/CLPG^{Pat}$), a mode of inheritance referred to as “polar overdominance” (Cockett *et al.* 1996). QTL mapping experiments performed in purpose-generated F3 mouse crosses indicated that polar overdominance and related parent-of-origin effects might be more common than generally recognized, and not limited to genomic regions harboring known imprinted genes (Lawson *et al.* 2013). It is important to recognize however that “mismodelling” may generate what appears to be imprinting effects but are in fact statistical artifacts. Unaccounted for maternal effects are one such source of pseudo-imprinting (Hager *et al.* 2008), while the erroneous assumption of fixation of alternate QTL alleles in the parental lines used to generate line-crosses are another. The latter has been the cause of a flurry of erroneous reports of imprinted QTL in livestock (De Koning *et al.* 2002; Sandor & Georges 2008).

De novo mutations have been shown to underlie a significant proportion of cases of autism and possibly other complex diseases (Girard *et al.* 2011; Kong *et al.* 2012). More recently, searching for highly disruptive *de novo* mutations in cases has been proposed as a strategy to identify causative genes. However, cases involving *de novo* mutations are typically sporadic. Even if these would account for a substantial proportion of the disease incidence, it is hard to imagine how such cases would contribute to disease heritability. A disease that would entirely be due to *de novo* mutation would have a heritability of zero, unless the mutation process by itself was heritable.

Finally, is it possible that the missing heritability reveals one or more important yet unsuspected novel molecular mechanisms of inheritance? In *C. elegans*, RNA-mediated interference (RNAi) has been shown to mediate multigenerational epigenetic inheritance (Buckley *et al.* 2012). The non-Mendelian inheritance pattern of a coat color variant involving a specific mutation in the murine cKit gene suggest that similar mechanisms might operate in mammals (Rassoulzadegan *et al.* 2006). However, more work is needed to evaluate the generalities of what still remain oddities, at least in mammals.

Uses of genomic markers in livestock management

Parentage control. Tracing the ancestry has been part of animal breeding for a long time, particularly since the introduction of breed creation and the accompanying registration in herd- and studbooks. Monitoring the accuracy of ancestry recording using genetic markers has become routine practice in many domestic animal species (f.i. Werner *et al.* 2004). It started with the use of blood group antigens (with subsequent addition of biochemical polymorphisms), and has evolved into the systematic use of panels of microsatellite markers – which are still in use in most countries under the auspices of the International Society of Animal Genetics (ISAG)(http://www.isag.us). As the use of genomic selection is increasing, a growing number of animals are being genotyped with genome-wide SNP arrays. It appears likely therefore that SNPs will soon phase microsatellites out. The possibility to "impute" microsatellite genotypes from SNP data should facilitate this transition.

Tracability. Consumers are paying increasing attention to the certification of the origins of the food products they eat. In several countries, including Belgium, biological samples are being stored to allow retrospective tracing of the origins of meat products (f.i. Dalvit *et al.* 2007). This is presently achieved using microsatellite markers. In this thesis, we have developed a method that can be viewed as a tracing method operating on sample mixtures (Blard *et al.* 2012). The aim of the method is to quantify the number of somatic cell counts (an indicator of mastitis) in the milk of all cows in a farm, just by genotyping a sample of milk from the farm's tank (hence containing a mixture of milk from all its cows). We demonstrate that by confronting the SNP allele frequencies in the tank's milk with the known SNP genotypes of the cows in the farm it is possible to determine what proportion of the somatic cell counts present in the tank was contributed by each individual cow, and hence determine her individual somatic cell counts.

This approach could contribute to improved ways of monitoring subclinical mastitis, as well as of monitoring illegal delivery of milk from infected cows to the milk factory.

Biodiversity. Intensive agricultural practices have considerably reduced the number of varieties and breeds that are being used. The ensuing reduction in genetic variability has rightfully been seen as a threat to future food security, in addition to the major cultural loss it means (f.i. Taberlet *et al.* 2008). Efforts have therefore been initiated to survey domestic biodiversity worldwide before some of the breeds disappear (f.i. Scherf 2001). Breed compilations are meant to inform future conservation efforts. One way to quantitatively characterize domestic biodiversity is to use genetic distance as unifying metric. This can be done using the available collections of genetic markers, including SNP panels. Conservation efforts can then focus on a set of breeds selected to maximize retention of genetic diversity.

Genetic defects. It has recently been shown that humans carry of the order of 100 loss-of-function variants (MacArthur *et al.* 2012). A handful of these appear to be highly deleterious recessives and to cause either embryonic death or severe developmental anomalies upon homozygosity. Homozygosity for such highly deleterious variants is uncommon in humans in the absence of consanguinity, as most of these are rare. In domestic animal populations, the widespread use of elite sires - including by artificial insemination (AI) – may cause highly deleterious variants to rapidly reach moderate to high frequencies. As avoidance of consanguineous matings is less rigorous than in humans, domestic animal populations are recurrently affected by outburst of genetic defects. This in turn may have caused more effective purifying selection against deleterious variants, which may explain their lower frequency in domestic animals than in human despite a higher nucleotide diversity (than in human) observed at neutral sites (Wanbo Li & Carole Charlier, unpublished observations).

With the development of genome-wide SNP arrays in livestock it has recently become possible to locate the genes responsible for such defects in a matter of weeks if not days (Charlier *et al.* 2008). Several examples demonstrating the efficacy of this process are included in this thesis (Dupuis *et al.* 2011; Sartelet *et al.* 2012). As a matter of fact, mapping causative loci has probably become the most cost-effective way to demonstrate the inherited nature of a newly reported anomaly. Going from the locus to the culprit gene and variant remains more challenging. Yet, the possibility to rapidly obtain the complete sequence of the locus using either a targeted or whole genome approach has - in most cases - considerably accelerated this step as well (f.i. Charlier *et al.* 2012).

Using a test interrogating the causative variant, or even neutral variants constituting a haplotype that is in strong linkage disequilibrium with the not yet identified causative variant, it is now possible to effectively manage recessive defects in livestock population. This is typically done by avoiding the selection of animals which carry common deleterious variants, as elite sires. More care should probably be taken to avoid the creation of genetic bottlenecks by eliminating too many sire lines, which could only exacerbate the issue. Moreover, too drastic selection against deleterious recessives may affect genetic progress for other traits, thereby *in fine* causing an economic loss. There is a need to optimize breeding schemes incorporating this new source of information.

It has become apparent that part of these deleterious recessives cause embryonic lethality (Charlier *et al.* 2012; Kumar Kadri *et al.* 2013; Sonstegard *et al.* 2013). It has therefore been speculated that the drop in fertility that has been observed in several livestock species might be due in part to an increase in embryonic loss due to homozygosity for embryonic lethal alleles. That this is indeed the case is substantiated by the observation – at least in cattle - of highly significant depletions in homozygosity for specific haplotypes (VanRaden *et al.* 2011; Fritz *et al.* 2013). As embryonic loss is a trait that is difficult to record, it may be more effective to apply a genotype-driven, reverse genetic approach to this problem. Along those lines, the Unit of Animal Genomics has mined genome-wide sequence data from ~50-500 animals of a number of cattle breeds for highly disruptive variants that should be enriched in embryonic lethals. Subsequent genotyping of large cohorts indeed demonstrates a significant absence of homozygotes as well as a direct effect on fertility, indicating that the corresponding variants indeed are embryonically lethal (Charlier *et al.* in preparation).

Genomic selection. One of the earliest drivers of “gene mapping” in domestic animals was the perspective to apply more effective “marker assisted selection” (Kashi *et al.* 1990) based on mapped Quantitative Trait Loci (QTL). Unfortunately, the limited variance explained by the identified QTL and the poor mapping resolution initially achieved with low-density microsatellite maps, made that proposition unattractive and ineffective. In 2001, Meuwissen *et al.* (2001) proposed a revolutionary concept, dubbed “Genomic Selection”, that would (i) take advantage of across-family linkage disequilibrium rather than within-family linkage (hence perform association mapping), (ii) take advantage of – at that time still hypothetical – genome-wide panels of cost-effectively genotyped SNP, and (iii) forgo the use of stringent significance thresholds, to estimate breeding values based on genome-wide marker information. The authors presented a number of approaches including GBLUP and various Bayesian models that assumed a non-uniform distribution of gene effects. The characterization of millions of SNPs and the concomitant development of cost-effective genotyping platforms in the years to follow, allowed the evaluation of the proposed method. It rapidly proved to offer opportunities to outcompete the ongoing breeding practices, including sire progeny-testing in cattle. In a matter of 24 months, genomic selection has been integrated in breeding programs worldwide, first in cattle and now – increasingly - in other livestock species as well. The superiority of genomic over conventional selection probably stems mainly from (i) the possibility to differentiate the breeding values of full-sibs on the basis of Mendelian segregation at a stage where conventional selection can't, and (ii) the higher accuracies of prediction obtained with genomic selection for low heritability traits even in the absence of Mendelian segregation (f.i. the information on fertility in dairy cattle obtained by genomic selection equates to the information of a progeny-group counting more than 150 daughters). Moreover, if large effects contribute to the genetic variance for a trait of interest (f.i. *DGAT1* variants to milk fat yield), Bayesian models provide a better fit to the data than the infinitesimal animal model, which increases the accuracy of prediction.

Increasing numbers of animals are presently being genotyped as part of genomic selection programs. As genotyping costs are still high for agricultural applications, schemes have been developed to minimize these costs. At present, this is typically done by exploiting the close familial relationships that exist within breeding stock. If a sire and dam have been genotyped with a high density SNP array, it is a waste of resources to genotype their offspring at equivalent density. Low-density SNP arrays combined with linkage approaches are sufficient to accurately predict the genotype of the offspring for all SNPs genotyped in the parents by a process, which in fact equates to "imputation". The preferred scheme today is pyramidal. It consists in genotyping the animals at the top of the breeding pyramid at the highest density. As a matter of fact, this tier of animals will rapidly be sequenced genome-wide. The 1,000 Bulls project (http://www.1000bullgenomes.com) aims at accomplishing this for cattle as a worldwide community effort. The next tier of animals in the pyramid would be genotyped at medium density, while the bottom tier would be genotyped at low density. Imputation would be used to project the whole-sequence information on the entire population. One publication in this thesis deals with methods of imputation in this context (Blard *et al.* 2012).

Genomic selection is a genuine advance in animal breeding. It is clear however that their remains room for improvement. Biases are observed in the breeding value estimates and accuracies still need to improve. One avenue towards the latter goal is to increase variant density (hence the pyramidal schemes described above) to include causative variants in the collection or at least increase the linkage disequilibrium between passenger and causative variants. However, increasing variant density increases the problem of over-fitting in statistical modeling. One possibly way to overcome this issue is to prioritize variants based on their predicted effect on gene function, i.e. estimate the probability that a given variant is causative rather than passenger. Efforts are underway to generate “Encode-like” data (Gerstein *et al.* 2012; Sanyal *et al.* 2012; Thurman *et al.* 2012; Ball 2013; http://encodeproject.org/ENCODE/) for livestock to better enable this distinction.

Objectives

The objectives of the present thesis were:

1. To develop a method to effectively map loci influencing binary and quantitative traits in outbred populations.

2. To apply the corresponding method to the analysis of economically important traits in livestock.

3. To develop approaches to more effectively use marker information in livestock production.

Part 1. Development and characterization of a haplotype-based method for association mapping of complex traits

Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification

Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification

Z. Zhang, F. Guillaume, A. Sartelet, C. Charlier, M. Georges, F. Farnir and T. Druet

Abstract

Motivation: In many situations, genome-wide association studies are performed in populations presenting stratification. Mixed models including a kinship matrix accounting for genetic relatedness among individuals have been shown to correct for population and/or family structure. Here we extend this methodology to generalized linear mixed models which properly model data under various distributions. In addition we perform association with ancestral haplotypes inferred using a hidden Markov model.

Results: The method was shown to properly account for stratification under various simulated scenario presenting population and/or family structure. Use of ancestral haplotypes resulted in higher power than SNPs on simulated datasets. Application to real data demonstrates the usefulness of the developed model. Full analysis of a dataset with 4600 individuals and 500 000 SNPs was performed in 2 h 36min and required 2.28 Gb of RAM.

Bioinformatics 28 (2012) 2467-2473

Introduction

Genome-wide association studies (GWASs) identify genetic variants (e.g. SNPs, CNV or indels) affecting traits of interest such as those related to human health or of agronomical importance. With the development of high-throughput genotyping and next-generation sequencing, these studies have been particularly successful. Hundreds of loci associated with diseases were detected through GWAS (e.g. Donnelly, 2008). Association studies proved equally valuable in other organisms such as Arabidopsis thaliana (Aranzana *et al.* 2005), mice (Threadgill *et al.* 2002), dog, crops (Malosetti *et al.* 2007; Yu *et al.* 2006) or livestock species.

Although very effective, genetic association studies still face a number of potential pitfalls. One major problem in GWAS comes from the spurious associations that may occur as a result of relatedness between individuals (e.g. familial relationships or population structure). Another issue is that, especially for complex traits, non-genetic factors (e.g. sex, age, etc.) may have profound impact on the scrutinized phenotype, raising the need for proper modeling of these effects.

An appealing solution to these problems is to use a mixed-model framework. Indeed, this methodology makes it possible to include covariates in the model and to account for the average genomic relatedness among individuals (population or family structure). Such models have been used for many years for QTL mapping especially in animal breeding (George *et al.* 2000). Recent studies (Kang *et al.* 2008; Malosetti *et al.* 2007; Yu *et al.* 2006; Zhao *et al.* 2007) have demonstrated that inclusion of such effects in mixed-models properly corrects for stratification and that the use of mixed models to control for stratification resulted in fewer false positives and/or higher power than other techniques such as genomic control (Devlin and Roeder, 1999), structured association (Pritchard *et al.* 2000) or principal components analysis (Price *et al.* 2006). In addition, mixed-models were able to capture the multiple levels of population structure and genetic relatedness. All these features make mixed-models a very promising tool to perform association analyses while controlling for relatedness structure.

Linear mixed-models (LMMs) assume that traits are normally distributed. Use of generalized linear mixed models (GLMMs) allows extension of the mixed-model approach to other types of traits, such as binary traits for example. With these models, a linear function of different covariates including polygenic and local genomic effects is used to describe the expected value of the observed phenotype through a so-called link function. Tzeng and Zhang (2007) developed a variance-components score test for association studies which can be used in the GLMM framework.

Analyses can be performed using single SNP or haplotypes of multiple SNPs. Haplotypes are specific combinations of alleles on the same chromosome. They extract more information on the relation between DNA variation and phenotypes than single SNPs and may present higher correlation with underlying mutations [depending on the marker density and the linkage disequilibrium (LD) pattern in the population]. Furthermore, haplotype tests can model allelic heterogeneity or find several (interacting) mutations at different tightly linked sites. However, the power of haplotype-based tests is potentially reduced due to the extra degrees of freedom needed in these analyses (e.g. Su *et al.* 2008; Tzeng and Zhang 2007). Different strategies to minimize this problem have been proposed in the literature, relying mainly on grouping haplotypes based on similarity (e.g. Blott *et al.* 2003; Durrant *et al.* 2004; Druet *et al.* 2008; Seltman *et al.* 2003). Various clustering algorithms are available. Those relying on sliding window approaches are not optimal as the optimal window size varies from one region to another (Browning, 2008). With the localized haplotype clustering method (Browning and Browning 2007), clusters of haplotypes are parsimoniously selected. This model allows for greater flexibility because haplotype lengths and the number of haplotypes are variable. Browning (2008) stated that this model is conceptually similar to the clusters of the Scheet and Stephens (2006) model. For each position along the genome, this later model assigns haplotypes to a predetermined number of ancestral haplotypes present several generations ago from which all haplotypes within a cluster are assumed to have descended. Each haplotype can be associated to a cluster for a different length making the model flexible. In addition, the model can group haplotypes with small difference (missing genotypes, genotyping errors or recent mutations). Su*et al.* (2008) proposed to use these ancestral haplotypes in association testing, whereas we suggested to use them for QTL fine-mapping and genomic selection (Druet and Georges 2010). In the same study, we showed that these clusters group haplotypes having a recent common ancestor (with short time to coalescence) and high identity-by-descent (IBD) probabilities [as estimated with the method of Meuwissen and Goddard (2001)].It was as efficient as methods using these IBD probabilities to cluster haplotypes (e.g. Druet *et al.* 2008). The use of these ancestral haplotypes proved already efficient for QTL fine-mapping (Karim *et al.* 2011) and genomic selection (de Roos *et al.* 2011).

In the present study, we develop a haplotyped-based method for association mapping relying on GLMM accounting for stratification and other covariates affecting the modeled trait.

Methods

The proposed method relies on GLMM. To account for stratification and/or polygenic background, the model includes a vector of random polygenic effects (e.g. Yu *et al.* 2006) in addition to the random haplotype effects:

$$\boldsymbol{\eta}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Zu}+\mathbf{Wh}$$

where $\boldsymbol{\eta}$ is a vector of n linear predictors (with n equal to the number of observed phenotypes), $\boldsymbol{\beta}$ is a vector of fixed effects (such as the overall mean in the present study), **u** is a vector of random n' polygenic effects (with n' equal to the number of individuals for which genomic information is available – typically, n' = n), **h** is a vector of p random ancestral haplotype effects, (with p equal to the chosen number of ancestral haplotypes (Druet and Georges, 2010)). **X**, **Z** and **W** are incidence matrices relating respectively fixed effects, polygenic and ancestral haplotypes effects to observations. The variance of the random polygenic effects **G** is 2**K** V_G (Yu *et al.* 2006) where **K** is a relative kinship matrix obtained from the marker data (see below) and V_G is the genetic variance. The variance of the random ancestral haplotype effects **T** is equal to **I** V_H where V_H is the "haplotypic" variance. Haplotypes effects are assumed independently identically (and normally) distributed.

The linear predictors are transformed to the observed scale (for example, disease status) through the inverse link function (e.g. McCullagh and Nelder, 1989). In our analyses, the logit link function was used to model binomial data such as disease traits (or affection status). The probability for individual i to be affected by the disease h(η_i) is therefore obtained through the inverse of the logit function:

$$\mu_i = h(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

Coding 0 a healthy individual and 1 an affected one, this probability is also the expected value for the trait. The solutions of the GLMM were obtained using an iterative procedure based on the

Laplacian approximation of the likelihood (McCullagh and Nelder, 1989; Breslow and Crayton, 1993). Indeed, the GLMM equations can be approximated by the following mixed model equations (**MME**) (e.g., Kachman 2000):

$$\begin{bmatrix} X^{*\prime}R^{-1}X^{*} & X^{*\prime}R^{-1}Z^{*} & X^{*\prime}R^{-1}W^{*} \\ Z^{*\prime}R^{-1}X^{*} & Z^{*\prime}R^{-1}Z^{*}+G^{-1} & Z^{*\prime}R^{-1}W^{*} \\ W^{*\prime}R^{-1}X^{*} & W^{*\prime}R^{-1}Z^{*} & W^{*\prime}R^{-1}W^{*}+T^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \\ \hat{h} \end{bmatrix} = \begin{bmatrix} X^{*\prime}R^{-1}y^{*} \\ Z^{*\prime}R^{-1}y^{*} \\ W^{*\prime}R^{-1}y^{*} \end{bmatrix}$$

where $X^*= XH$, $Z^*= ZH$, $W^*= WH$, $\mathbf{H} = \frac{\delta\mu}{\delta\eta}$, **R** is the residual variance matrix and $\mathbf{y}^*=\mathbf{y}-\boldsymbol{\mu}+\mathbf{H}\boldsymbol{\eta}$.

In the case of the logit function (and one record per individual), **H** and **R** are diagonal matrices with $\mu_i(1-\mu_i)$ on the diagonal (e.g., Kachman 2000). Therefore, an iterative procedure must be used. First, starting values for **β**, **u** and **h** are used to compute **μ** and to build the approximate MME. These are then solved to obtain new estimates of **β**, **u** and **h**. The process is repeated until convergence.

The relative kinship matrix **K** was estimated based on similarity scores as in Eding and Meuwissen (2001) and Hayes and Goddard (2008):

$$S_{xy,l} = 0.25\ [I_{11}+ I_{12}+ I_{21}+ I_{22}]$$

where $S_{xy,l}$ is the similarity score between individuals x and y at locus l and I_{ij} is an indicator variable equal to 1 if allele i on locus l in the first individual and allele j on the same locus in the second individual are identical, otherwise it is 0. In our analyses, we replaced SNP alleles by ancestral haplotypes (similar to multi-allelic markers). The relationships are then based on closer founders. S_{xy}, averaged over the whole genome is then used as an estimator of the kinship relationship f_{xy} as in Lynch (1988) or Eding and Meuwissen (2001):

$$\hat{f}_{xy} = \frac{S_{xy-s}}{1-s}$$

where s is the minimal value of S_{xy} in the matrix (Hayes and Goddard 2008). Zhao *et al.* (2007) and Kang *et al.* (2007) concluded that use of similarity score to construct relationship matrices was as efficient as more complex methods and avoided problems of non-positive definite matrices. We tested different methods to construct relationship matrices (based on SNP or haplotypes) but these had little impact on estimation of polygenic effects and even less on residuals.

Associations are tested for every marker position along the genome by a significance test of σ_h^2=0.Explicit evaluation of the likelihoods in GLMM is cumbersome, making application of likelihood ratio tests (LRTs) challenging. Therefore we used the score tests as proposed by

Verbeke and Molenberghs (2003). Schaid *et al.* (2002) and Tzeng and Zhang (2007) used score tests in haplotype-based association studies with binary traits. The score tests are based on the value of the first derivative of the log-likelihood under the null hypothesis (i.e. the variance of the haplotypes is null). A significant positive first derivative with respect to the variance component indicates that the maximum-likelihood estimator of the haplotypes variance is significantly different from zero.

Tzeng and Zhang (2007) derived a test statistic T based on the score tests for haplotype-based models for GLMM. In the case of the logit function, the test statistic is equal to:

$$T=0.5(\mathbf{y-\mu})'\mathbf{WW}'(\mathbf{y-\mu})=0.5(\mathbf{W'}(\mathbf{y-\mu}))'* \mathbf{W'}(\mathbf{y-\mu})$$

where $\mathbf{y-\mu}$ is a vector of residuals (observations corrected for estimated fixed and random effects) obtained from a GLMM under the null hypothesis (no haplotype effect) where $\sigma_h^2=0$. Since T relies on estimation of residuals from a model without haplotype effects, the procedure is similar to the two-step procedure proposed in Aulchenko *et al.* (2007). Therefore it has the same advantages: the mixed models must be solved only once to obtain the residuals, which considerably speeds up computations, and since residuals are corrected for stratification, they are free from familial correlations and the data become exchangeable (Aulchenko *et al.* 2007) which means that permutation techniques may be applied.

Tzeng and Zhang (2007) demonstrated that the distribution of the T test statistic under the null hypothesis could be approximated using a gamma distribution. We perform 1,000 permutations of the residuals to estimate the mean and the variance of the gamma distribution (or the shape and scale parameters). Parameters of the gamma distribution are estimated for each tested position (marker) because the distribution is influenced by the structure of the incidence matrix relating haplotypes to residuals, which is potentially position specific. We will refer to this strategy as 'gamma approximation'. In addition, empirical P-values can be computed by repeatedly permuting the phenotypes (residuals) among the individuals (referred to as permutation hereafter).

A data set of 3547 genotyped Holstein, Jersey or crossbred bulls (see Karim *et al.* 2011) was used to simulate case/controls studies. Individuals were genotyped for the Illumina Bovine SNP50 SNP chip (Illumina, San Diego, CA). After data edition, 37,647 SNPs were conserved on the 29 autosomes. DualPHASE from the PHASEBOOK package (Druet and Georges, 2010) was used to infer haplotypes from the genotyped individuals and to assign them to K ancestral haplotype clusters (K was set equal to 5, 10 or 20).

To simulate different stratification scenarios including breed (i.e. structured populations) and polygenic (i.e. familial relationships) effects, and mimicking causal variant (SNP with significant impact) effects, the following model was used:

$$\boldsymbol{\eta}=\mathbf{X_1}\,\mu+\mathbf{X_2}\beta+\mathbf{Z_1}\mathbf{u}+\mathbf{Z_2}\mathrm{v}$$

where μ is the mean effect, $\mathbf{X_1}$ is a vector of "1", β is the breed effect, $\mathbf{X_2}$ is a vector containing the percentage of Holstein blood (ranging from 0 to 1), $\mathbf{Z_1}$ is a matrix (n x 1000) containing the number of alleles "1" of a set of 1000 SNPs affecting the phenotype, **u** is a vector containing the allelic substitution effects of 1000 SNPs used to simulate a polygenic effect, $\mathbf{Z_2}$ is a vector containing the number of alleles "1" for the SNP of interest and v is the allelic substitution effect for that SNP. The breed effect was equal to 0, 0.2, 0.4 or 0.7 according to the scenario (corresponding to odd ratios (**OR**) equal to 1.0, 1.22, 1.5 and 2.0, respectively), the individual SNP effects (1000 SNPs) were drawn from a gamma distribution (shape = 0.4 and scale = 1) (Calus or Meuwissen). The variance of the polygenic effects (the sum of 1000 individual SNP effects) was then rescaled to 0 or 0.16. Finally, the studied SNP effects were equal to 0.5 and 0.8 (corresponding to OR equal to 1.65 and 2.22, respectively).

The phenotype of each individual was then sampled from a binary distribution with mean equal to:

$$\mathrm{p}_i=\frac{e^{\eta_i}}{1+e^{\eta_i}}$$

The overall mean effect was set to obtain an incidence of the disease of 27 % in the population. Finally, 500 cases and 500 controls were randomly sampled from the 3547 genotyped individuals. Simulations were repeated for 1,000 different SNPs chosen as potential causal variants. One thousand SNPs were selected as potential major variants and were removed from the dataset prior to phasing. A total of 10 000 and 100 000 simulations were performed per scenario to estimate power and to compute QQ-plots, respectively.

Results

Simulated data

The maximum LD (measured by r^2) between each of the 1000major variants and the remaining SNPs (hereafter called marker SNPs) or with the 5, 10 or 20 ancestral haplotypes was estimated within a 2Mb interval (1Mb on each side of the major variant). The maximum r^2 was on average equal to 0.40 with marker SNPs and to 0.51 and 0.72 with 10 or 20 ancestral haplotypes. However, in some cases, SNPs can still present higher LD. Indeed, using marker SNPs or K = 20 haplotypes, 7.5 % of the major variants were captured with an r^2 of 1.0 whereas this value dropped to 0% with 5 or 10 haplotypes. From here on, K = 20 for the remainder of the simulation study.

To test whether our model correctly accounted for stratification, type-I errors were estimated by testing the model under H0 (the major variant had no effect: v was set equal to 0) and QQ-plots were generated. Simulations were performed including a breed effect of 0.4, but no polygenic effect. Four models were fitted to the generated dataset: using haplotypes or marker SNPs, and including or not a polygenic effect accounting for stratification. In Figure 1, models without polygenic effect clearly present an excess of small P-values. After inclusion of the polygenic effect, the regression slopes of the QQ-plot were below one, indicating that stratification was correctly accounted for but that tests are slightly too conservative.

In Table 1, regression coefficients of QQ-plots obtained with the four fitted models applied to different simulated scenarios are presented. In all cases where stratification was simulated, models without polygenic effect showed excess of small P-values, particularly with haplotypes which tend to capture more stratification effects. The inclusion of a polygenic effect resulted in regression coefficients below 1.0, even when both breed and polygenic effects were simulated. In these simulations, two methods were used to estimate P-values, namely the permutation test and the gamma approximation test, with both yielding approximately the same regression slopes after correction for stratification.

Table 2 compares the power of models with marker SNPs or haplotypes for different OR and frequencies of the major variant (permutations were used to estimate P-values). In GWAS, due to multiple testing, low P-values must be achieved but the number of simulations allowed us only to estimate the power at α= 0.001.The marker with the strongest association is not always the closest to the major variant. Therefore, power was tested in a 2Mb window centered on the major variant, spanning ~ 30 SNPs. To correct for the resulting multiple testing (and correlation among successive tests along the region), chromosomes were randomly shuffled across individuals 10 000 times. For each permutation, the best P-value was stored and the association test was declared significant at P < 0.001 if one of the P-values was lower than 9990 of the best P-values obtained by permutation. Major variants, with resulting OR equal to 1.65 and 2.22, accounted only for a small fraction of the variation

(r^2 between the SNP and the binary trait below 0.04 and 0.08 according to the SNP effect). Due to incomplete LD with causative SNPs, haplotypes and genotyped SNPs captured an even smaller fraction of that variance. Therefore the power was low (tests are already corrected for ~ 30 repeated correlated tests by the permutation procedure explained above), particularly when minor allele frequency (MAF) was below 0.2 and for small OR (1.65). For larger SNP effects, power increased, particularly when using ancestral haplotypes in the model which proved better in the present simulations for MAF above 0.10.

The power was also compared for different levels and types of structures (Table 3). In all cases, use of ancestral haplotypes resulted in higher power than for single SNPs and power decreased in datasets presenting structure (particularly for haplotypes and in presence of polygenes).

Real data

Data from the Belgian Blue cattle breed heredo-surveillance platform were used to test the method on real datasets. Four phenotypes were analyzed: 3 monogenic recessive diseases [gingival hamartoma (33 cases), arthrogryposis (13 cases) and prolonged gestation (25 cases)] and color-sidedness which is monogenic dominant (8 cases). The causative variants are known for hamartoma (Sartelet *et al.* in preparation) and color-sideness (Durkin *et al.* 2012) whereas for arthrogryposis and prolonged gestation, diagnostic tests have been developed based on markers in LD with the causative variants. In addition to cases, genotypes from 300 controls were available. Individuals were genotyped for a custom made 50K bovine chip described in Charlier *et al.* (2008). After deleting markers with a call rate below 0.90 or having a MAF below 0.05, 41 878 SNPs mapping to autosomal chromosomes were used in the study. Haplotypes were reconstructed using DualPHASE (Druet and Georges, 2010) with 10 ancestral haplotypes (we reduced the number of ancestral haplotypes to 10 since the number of individuals is much smaller than in the simulation study). Association studies were also performed with EMMAX (Kang *et al.* 2010) which performs single point (SNPs) association studies with LMM that account for stratification (through inclusion of a kinship matrix). After Bonferroni correction for ~ 50 000 tests, genome-wide significance was set at10^{-6}.

For the 3 monogenic recessive diseases P-values (estimated with the gamma approximation) below 10^{-40} were obtained (Manhattan plots are available in Supplementary Figures) in regions in which almost all cases were homozygous for a specific ancestral haplotype whereas almost none of the controls was homozygous for that haplotype, suggesting high LD between this ancestral haplotype and the causative variant. The identified regions were in agreement with the previous findings. The Manhattan plot for color-sidedness is presented in Figure 2a. The lowest P-value is below 10^{-10} and the corresponding position is located at 0.7Mb from a CNV causing the phenotype, a copy of a chromosomal segment on BTA6 encompassing the KIT locus which translocated to BTA29 (Durkin *et al.* 2012). All color-sided individuals carry at least one copy of the same ancestral haplotype (it has dominant behavior). Some controls also carry the haplotype but the phenotype is not observed. Indeed, the phenotype has an incomplete penetrance since it cannot be observed on individuals with completely white coats (homozygous genotypes for a frequent common codominant mutation at the roan locus results in white coats [Charlier *et al.* 1996]).

Associations performed with EMMAX are presented in Supplementary Figures for monogenic recessive diseases and Figure 2b for color-sideness. This software relies on LMM and assumes that the traits are normally distributed. As other LMM packages, it can still be applied on binary traits and perform well as shown in Supplementary Figures where several SNP in the region surrounding the causative SNP were highly significant and no other SNP reached such levels of significance. However, some SNPs reach genome-wide significance in non-causative regions (e.g. association study for arthrogryposis).

For color-sideness, use of EMMAX resulted in a Manhattan plot where the causative region is non-significant and difficult to identify (other regions of the genome show higher significance). These examples illustrate that considering binary trait as normally distributed and using SNP as covariates can result in situations where the region harboring the causative mutation is difficult to identify. In such situations, extension to GLMM and use of ancestral haplotypes resulted in associations where many positions in the region of interest have high level of significance, clearly above the remainder of the genome and with less non-causative regions reaching genome-wide significance.

Computational performance

To test the computational efficiency of the developped software, we ran an implementation compiled with Intel Fortran using openMP and MKL libraries on a dataset with a total of 4600 individuals genotyped for 500 000 SNPs. The analysis was performed on a Intel Xeon E5520 processor at 2.27GHz using four threads. The full analysis lasted 2 h 36min and required a total of 2.28Gb of memory.

Discussion

Our simulation studies showed that taking into account genomic relationships among individuals through inclusion of a polygenic effect in GLMM accounted for stratification, as previously observed with LMM (e.g. Malosetti *et al.* 2007; Yu *et al.* 2006; Zhao *et al.* 2007). It was also observed that corrections were effective in structured populations (breed effect) and/or when family structure was present (polygenic effect). Most designs in model organisms, plants or animal species present a high level of stratification because either several populations are used in the study or the individuals are closely related, making these robust corrections essential. In addition to correcting for stratification, the GLMM framework offers additional flexibility because it allows for a better modeling of the phenotypes through the inclusion of additional covariates (e.g. sex, age, etc.) and a consequently better association study where all known nuisance factors have been corrected for. Such a possible nuisance factor could be a measure of the population structure, as suggested in some studies (e.g. Price *et al.* 2006; Pritchard *et al.* 2000). The effect of adding this correction concurrently with the genomic relatedness structure might be necessary in some populations albeit not always (Kang *et al.* 2008; Zhao *et al.* 2007). Another point stressing flexibility of LMM is that association can easily be performed with either SNPs or (ancestral) haplotypes. Extension from LMM to GLMM is important for traits that are not normally distributed. However, in many situations such as large balanced case/control studies where variants are not very rare and have low or moderate effects, LMM perform well with binary traits. GLMM are recommended for strong deviations from normality, when cases or controls are rare within a cell (covariates of the model such as fixed, SNP or haplotype effects). Such situations occur more often in smaller designs with few cases (or controls) and with rare variants (or haplotypes). Such designs are still common in animal or plant species and our applications to real data illustrate that in such cases, use of a GLMM with a logit link function results in cleaner association than LMM.

Slopes of QQ-plots indicated that statistical tests were too conservative resulting in a loss of power. Similar deflation has been described in Amin *et al.* (2007) and is due to the fact that polygenic effects (used for correction) and SNP or haplotype effects are correlated (e.g. SNP or haplotypes are used to estimate K). This correlation is stronger with haplotype which are therefore more affected by the over-correction. One solution would be to correct for deflation with genomic control as in Amin *et al.* (2007) or to perform a test modeling simultaneously polygenic and haplotype effects at most interesting positions. Without such corrections, the proposed model may be more conservative than some other methods relying on LMM.

Although computational efficiency was not the main goal of the present study, the developed method presents interesting computational features. First, GLMM must be solved only once (with the SNP or haplotype variance set to 0). Solving the GLMM equations and inferring the variance components are potentially time consuming. With likelihood ratio tests, variances are inferred for a model including polygenic and SNP (or haplotype) random effects for each tested position. With the present method, the test score is obtained from a simple statistic based on records corrected for effects of the model under the null hypothesis (without SNP or haplotypic effect).Performing this test is much faster than inferring variance in the full model. This is very similar to the approach used in GRAMMAR (Aulchenko *et al.* 2007). In addition, since the data are corrected for family and population structure it becomes exchangeable and permutations can be performed freely, which was not the case for the raw data. Thanks to permutations, empirical P-values corrected for multiple testing can easily be obtained. Still, with permutations it would be time consuming to obtain small P-values as typically needed in GWAS studies; in that case, approximation of the test score distribution with a gamma distribution (with scale and shift parameters obtained empirically through 1000 permutations) seems to perform well.

As in previous studies (Druet and Georges, 2010; Su *et al.* 2008), assigning chromosomes to ancestral haplotypes resulted in high LD between haplotype groups and underlying mutations. In the present study, the LD was much higher than when using SNP for a cattle population and with ~

50 000 SNPs covering the genome. The picture might change with different marker densities or in other populations. The method for clustering haplotypes has also been successfully applied to the fine-mapping of a QTL affecting bovine stature (Karim *et al.* 2011) in a crossbred population. In that study, the association between the ancestral haplotypes and the later candidate causative variants was almost perfect (2 misclassified haplotypes out of 1490). More recently, our method allowed to fine-map a mutation causing dwarfism in cattle which was always associated to the same ancestral haplotype (Sartelet *et al.* 2012). Other examples in the present study illustrate the high LD between the ancestral haplotypes obtained with DualPHASE (Druet and Georges, 2010) and ungenotyped variants. The association should be better for more recent variants which rapidly increased in frequency due to selection. In that case, the length of the haplotype associated to the variant would be longer than for random variants (as those used in the simulation study), making it easier to identify the haplotype. Ancestral haplotypes can be associated with different types of variants including SNP, multiple alleles, several-linked SNP (a small haplotype), insertions/deletions and duplications. In the real data example on color-sidedness, ancestral haplotypes presented high LD with a trans CNV and other examples of association between ancestral haplotypes and deletions or duplications (either in cis or trans position) can be found in Durkin *et al.* (2012). Note, when large reference populations genotyped at high density (or sequenced) are available, imputation followed by single point association would probably result in higher power than use of ancestral haplotypes (if there is only one causal variant and if SNPs in high LD with this variant are genotyped in the reference panel). However, such reference populations are only available in a few species. While providing high LD with underlying variants, the use of ancestral haplotypes also controls the number of haplotype groups to be used in the study, which is important to maintain statistical power. This method proves also flexible since there is no need to define arbitrary windows and since haplotype origin can change at any position along the chromosome. For instance, recombinant haplotypes do not create additional haplotypes groups: they are simply potentially assigned to different groups on each side of the cross-over position. Finally, haplotypes with small differences due to genotyping errors or new non-causative mutations can still be grouped together whereas with less flexible methods, new haplotype groups would be defined for each difference, resulting in a loss of power. For the same reason, missing genotypes are easily handled. Our score test framework can easily be

applied with other methods for clustering haplotypes, even those modeling a correlation between haplotypic effects. Our method does not rely on a particular biological model but identifies ancestral haplotypes significantly associated with the disease. Therefore it can be applied to monogenic recessive diseases, dominant diseases, phenotypes with complete or incomplete penetrance, oligenic or polygenic diseases or complex traits. It is also robust to misclassified samples which will only reduce slightly the power since there are no strong assumptions such as sharing of an IBD segment in all cases. For instance, the method was used to fine-map a variant causing dwarfism in Belgian Blue cattle (Sartelet *et al.* 2012) which was the cause only for a subset of cases (14 out of 33). Still, the method detected with high significance (P-value <10E-11) the region harboring the causative mutation. Further veterinary examination revealed that dwarfs could be classified into different categories and that the 14 cases corresponded to a specific sub-group. Even without that knowledge, the variant was identified, stressing the robustness of the method. The example of color-sidedness also illustrates that with only a few cases (8), a dominant gene (cases carry only one haplotype) and incomplete penetrance (the phenotype is not observed on white animals) the method still identifies with high significance the region harboring the causative variant.

Acknowledgments

We acknowledge LIC for providing data. Tom Druet and Carole Charlier are, respectively, Research Associate and Senior Research Associate of the Fonds de la Recherche Scientifique —FNRS. The authors thank the GIGA Bioinformatics platform and SEGI (University of Liege) for providing computing resources.

Funding: This work was funded by grants from the Walloon Ministry of Agriculture (Rilouke) and the European Union (PIGENDEF).

Conflict of Interest: None declared.

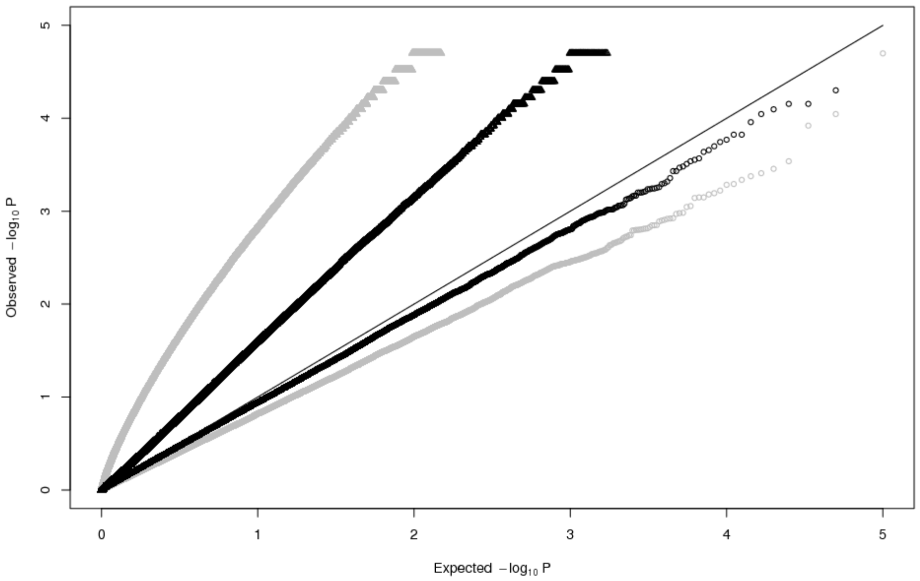


Fig. 1. QQ-plots obtained with marker SNPs (black) or ancestral haplotypes (gray) with (circles) or without (triangles) polygenic terms accounting for stratification included in the model.

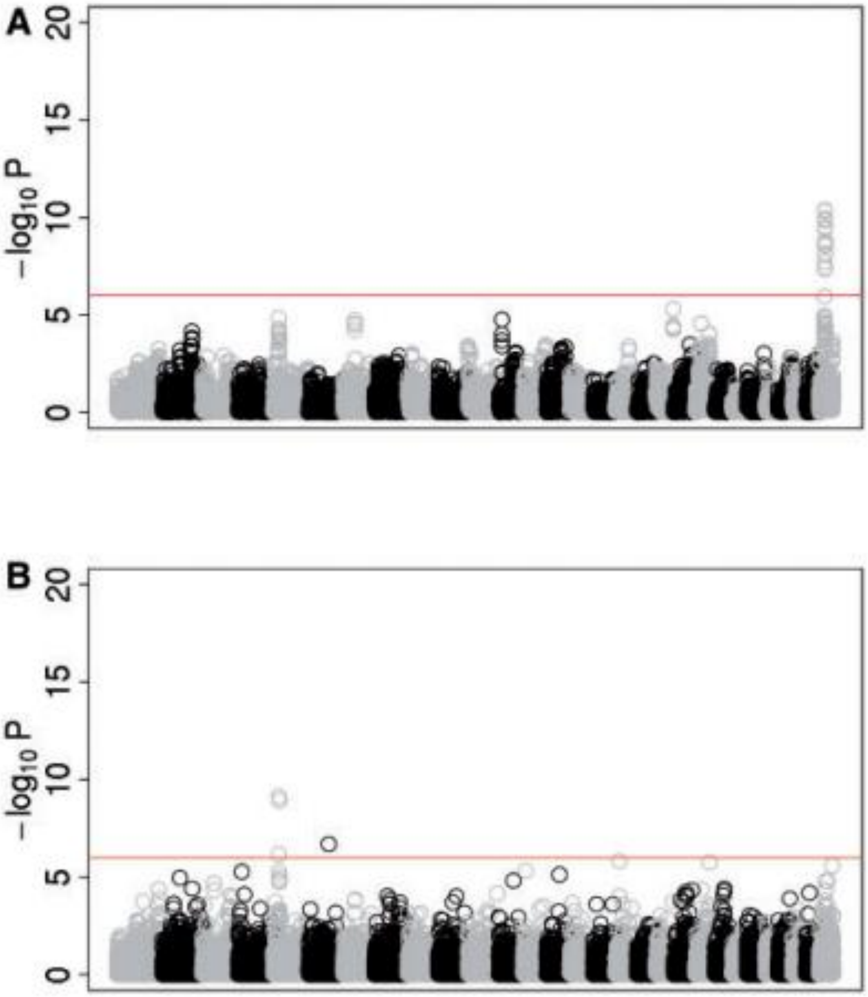


Fig. 2. Manhattan plot for association study for color-sideness with (A) GLMM score tests using ancestral haplotypes and (B) EMMAX

Table 1. Regression coefficients of QQ-plots obtained with four fitted models on four simulated designs [P-values computed with permutation test or with a gamma approximation test (in parenthesis)]

Simulated breed effect	Simulated polygenic variance	Fitted model			
		without polygenic effect		with polygenic effect	
		SNP[a]	Anc. Hap.[a]	SNP	Anc. Hap.
0	0	1.03 (1.01)	0.99 (1.02)	0.97 (0.96)	0.86 (0.87)
0.4	0	1.59 (1.57)	2.54 (3.06)	0.94 (0.94)	0.82 (0.83)
0	0.16	1.39 (1.36)	2.02 (2.22)	0.95 (0.95)	0.86 (0.86)
0.2	0.16	1.65 (1.63)	3.29 (3.89)	0.94 (0.94)	0.82 (0.83)

[a] Association is performed either with SNP or ancestral haplotypes.

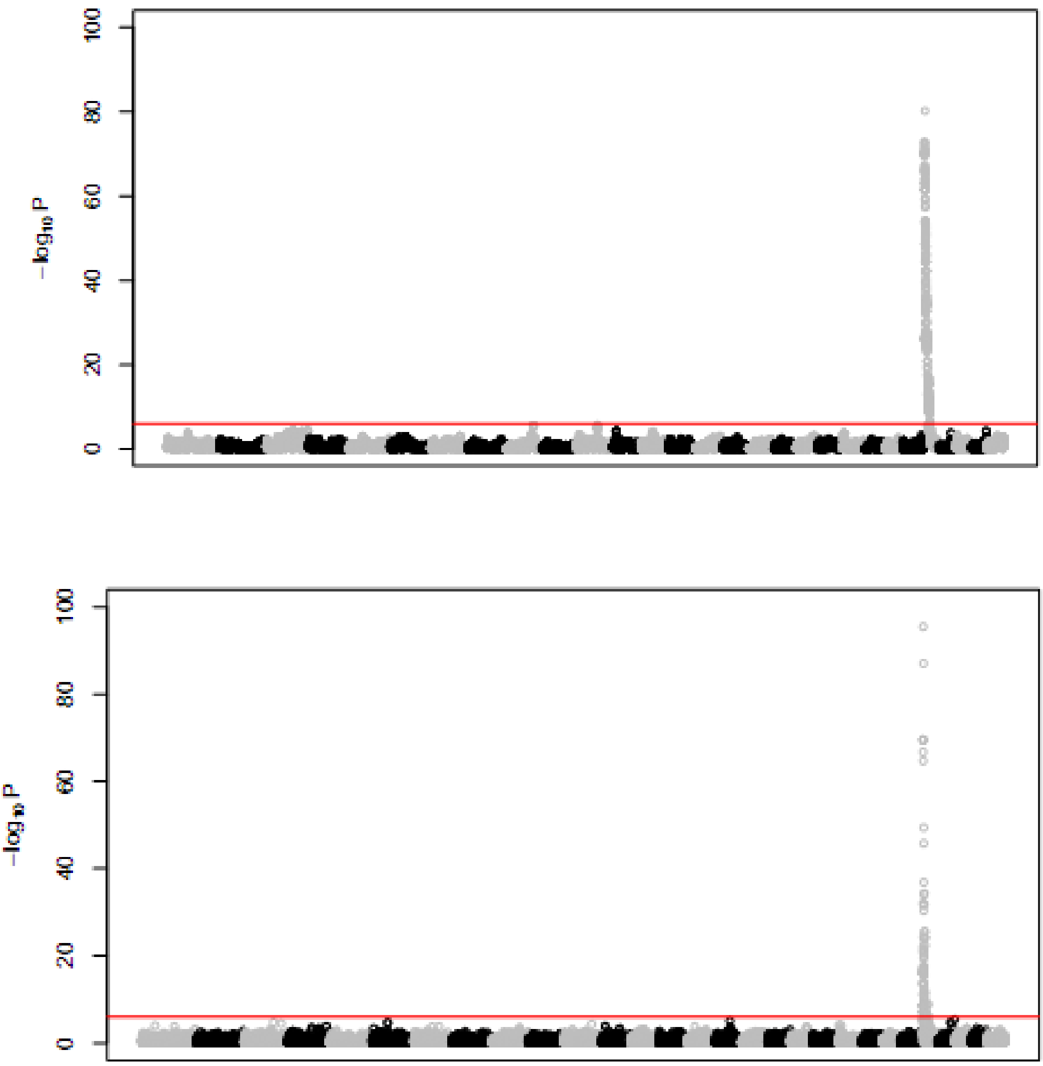
Table 2. Variant detection power (α= 0.001) with SNP or ancestral haplotypes in a design without structure (no breed and polygenic effects) for different minor allelic frequency (MAF) classes of the causal SNP

MAF Class		OR of variant = 1.65			OR of variant =2.2	
	r^{2a}	SNP	Anc. Hap	r^2	SNP	Anc. Hap
[0,0.1]	0.008	0.003	0.002	0.018	0.020	0.016
(0.1,0.2]	0.017	0.013	0.030	0.040	0.072	0.194
(0.2,0.3]	0.024	0.033	0.071	0.057	0.129	0.421
(0.3,0.4]	0.030	0.050	0.113	0.071	0.136	0.520
(0.4,0.5]	0.032	0.053	0.152	0.076	0.150	0.622

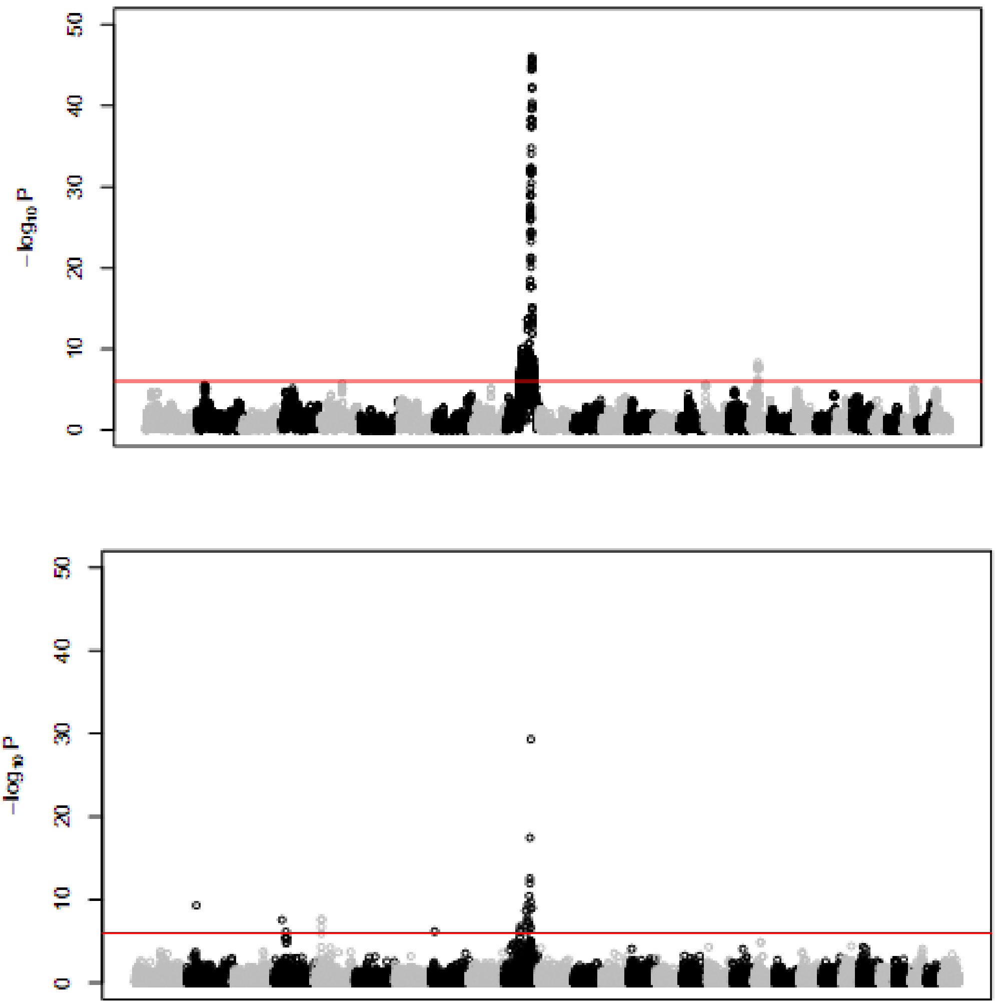
[a] r^2 between causative variant and observed phenotype.

Table 3. Power of association mapping (α= 0.001) with SNP and haplotypes in different designs with stratification (statistics are provided across all MAFs)

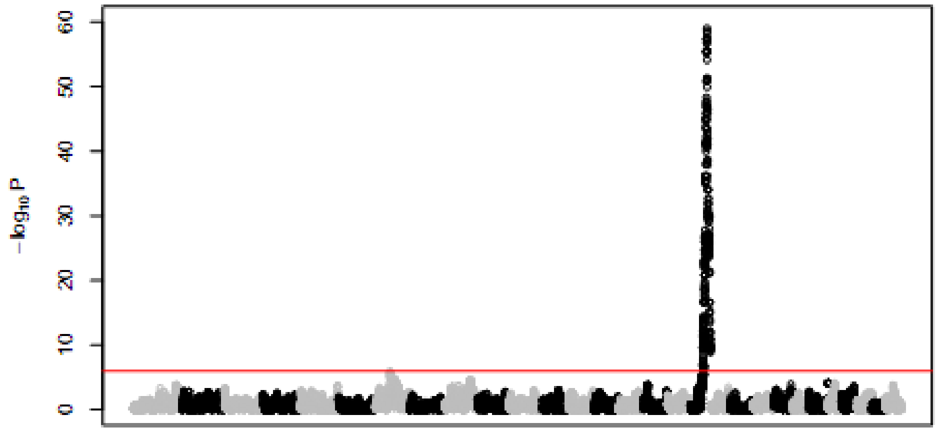
Simulated	Simulated	OR of variant = 1.65		OR of variant =2.2	
breed effect	polygenic variance	SNP	Anc. Hap.	SNP	Anc. Hap.
0	0	0.023	0.078	0.079	0.373
0	0.16	0.025	0.057	0.077	0.322
0.4	0	0.027	0.067	0.077	0.359
0.7	0	0.028	0.071	0.076	0.365
0.2	0.16	0.023	0.053	0.074	0.322

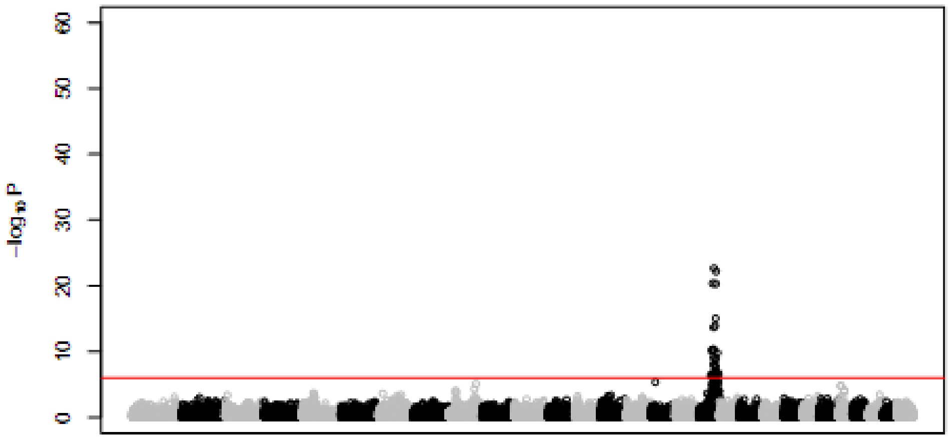


Supporting Figure 1: Manhattan plot for association study for gingival hamartoma with GLMM score tests using ancestral haplotypes and EMMAX (from top to bottom).



Supporting Figure 2: Manhattan plot for association study for arthrogryposis with GLMM score tests using ancestral haplotypes and EMMAX (from top to bottom).





Supporting Figure 3: Manhattan plot for association study for prolonged gestation with GLMM score tests using ancestral haplotypes and EMMAX (from top to bottom)

Part 2: Application of a haplotype-based method for association mapping of complex traits

2.1 Genome-Wide Association Study Reveals Constant and Specific Loci for Hematological Traits at Three Time Stages in a White Duroc x Erhualian F2 Resource Population

2.2 Results of a haplotype-based GWAS for recurrent laryngeal neuropathy in the horse

2.3 A splice site variant in the bovine RNF11 gene compromises growth and regulation of the inflammatory response

2.4 Detection of copy number variants in the horse genome and examination of their association with recurrent laryngeal neuropathy

2.5 Serial translocation by means of circular intermediates underlies colour sidedness in cattle

Genome-wide Association Study Reveals Constant and Specific Loci for Hematological Traits at Three Time Stages in a White Duroc × Erhualian F2 Resource Population

Z. Zhang, Y. Hong., J. Gao, S. Xiao, J. Ma, W. Zhang, J. Ren, L. Huang

Abstract

Hematological traits are important indicators of immune function and have been commonly examined as biomarkers of disease and disease severity in humans. Pig is an ideal biomedical model for human diseases due to its high degree of similarity with human physiological characteristics. Here, we conducted genome-wide association studies (GWAS) for 18 hematological traits at three growth stages (days 18, 46 and 240) in a White Duroc × Erhualian F_2 intercross. In total, we identified 38 genome-wide significant regions containing 185 genome-wide significant SNPs by single-marker GWAS or LONG-GWAS. The significant regions are distributed on pig chromosomes (SSC) 1, 4, 5, 7, 8, 10, 11, 12, 13, 17, and 18, and most of significant SNPs reside on SSC7 and SSC8. Of the 38 significant regions, 7 showed constant effects on hematological traits across the whole life stages, and 6 regions have time-specific effects on the measured traits at early or late stages. The most prominent locus was the genomic region between 32.36 and 84.49 Mb on SSC8 that is associated with multiple erythroid traits. The *KIT* gene in this region appears to be a promising candidate gene. The findings improve our understanding of the genetic architecture of hematological traits in pigs. Further investigations are warranted to characterize the responsible gene(s) and causal variant(s) especially for the major loci on SSC7 and SSC8.

PLoS One 8 (2013) e63665

Introduction

In the immune system, hematological traits include three components: leukocytes (white blood cells, WBCs), erythrocytes (red blood cells, RBCs) and platelets. All of these components represent important parameters of immune capacity of individuals (TULLIS 1952). Hematological related cells in the peripheral blood execute a range of functions including the transport of oxygen, innate and adaptive immunity, vessel wall surveillance, homeostasis and wound repair. As blood incessantly flows within the circulatory system around organs and tissues, it can reflect any slightly abnormal changes in the body rapidly by testing the changes of cells number and (or) cells volume. Deviations outside normal ranges for these parameters are indicative of different kinds of disorders including cancer and cardiovascular, metabolic, infectious and immune diseases (Soranzo *et al.* 2009). Measurements of erythrocytes within the blood are becoming a routine examination to uncover various hematological related disorders.

The count and volume of cellular elements in circulating blood are highly heritable and vary considerably among individuals (Edfors-Lilja *et al.* 1994; Evans *et al.* 1999; Garner *et al.* 2000). In human, genome-wide association studies (GWAS) have identified > 60 loci associated with hematological parameters in individuals of European ancestry, Japanese population, and African Americans (Soranzo *et al.* 2009; Ganesh *et al.* 2009; Kamatani *et al.* 2009, 2010; Meisinger *et al.* 2009; Nalls *et al.* 2008; Uda *et al.* 2008). However, these polymorphisms explain only a small fraction of the genetic variance in hematological traits. This is so called “missing heritability” (Yang *et al.* 2010). Well-designed study in animal model is an efficient way to identify additional genetic factors contributing to complex phenotypic variance. The domestic pig is a large-animal model for human genetic diseases due to its high degree of similarity with human physiological characteristics. Identification of responsible genes and causal variants for hematological traits in pigs would benefit researches on human medicine.

So far, 239 quantitative trait loci (QTL) for swine hematological traits have been reported by linkage mapping in the AnimalQTLdb database (Hu *et al.* 2007), but the confidence intervals of these QTL are generally large (>20 cM) and harbor hundreds of functional genes, thereby hampering the characterization of plausible candidate genes. Compared to traditional QTL mapping strategies, GWAS based on high-density markers is a more powerful tool to identify genomic regions for phenotypic traits. To our knowledge, only two very recent studies have reported the GWAS for hematological parameters in pigs (Luo *et al.* 2012; Wang *et al.* 2013). The two studies identified 10 and 62 genome-wide significant loci for hematological traits. However, only one locus for RDW on pig chromosome (SSC) 12 was consistently detected in the two studies, implying the complexity and heterogeneity of hematological traits.

In our previous studies, we conducted a whole genome linkage mapping in a White Duroc × Erhualian F_2 resource population using 183 microsatellite markers, and identified a number of QTL affecting hematological traits measured at 3 growth stages (Yang *et al.* 2009; Zou *et al.* 2008). To fine map the identified QTL and uncover new genetic variants associated with hematological traits, we herein performed GWAS on the F_2 resource population using the PorcineSNP60 Genotyping BeadChip technology (Illumina, USA). The experimental data are available upon the readers' request.

Material and Methods

Ethics statement

All the procedures involving animals are in compliance with the care and use guidelines of experimental animals established by the Ministry of Agriculture of China. The ethics committee of Jiangxi Agricultural University specifically approved this study.

Animals and phenotypic measurements

A detail description of the White Duroc × Erhualian F_2 resource population and phenotype recording have been presented in our previous publications (Yang *et al.* 2009; Zou *et al.* 2008; Guo *et al.* 2009). Briefly, the three-generation resource population comprising 1912 F_2 individuals was developed by crossing 2 White Duroc boars and 17 Erhualian sows. All animals were kept under a consistent standard pigpen and were fed with same diet at the experimental farm of Jiangxi Agricultural University. Eighteen hematological parameters were measured for 1449 individuals at three age stages: days 18, 46 and 240. Blood samples of 5 ml were collected from each animal and were directly injected into eppendorf tubes containing 30 ul of 20% EDTA in polybutadiene-styrene. A standard set of hematological data were recorded using a CD1700 whole blood analyzer (Abbott, USA) at the First Affiliated Hospital of NanChang University, China. The 18 hematological parameters include 7 baseline erythroid traits (hematocrit (HCT), hemoglobin (HGB), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobina concentration (MCHC), mean corpuscular volume (MCV), red blood cell count (RBC), and red blood cell volume distribution width (RDW)), 7 leukocyte traits (granulocyte count (GRAN), granulocyte count percentage (GRAR), monocyte count (MON), monocyte count percentage (MONR), lymphocyte count (LYM), lymphocyte count percentage (LYMA), and white blood cell count (WBC)), and 4 platelet traits (plateletcrit (PCT), platelet distribution width (PDW), platelet count (PLT), and mean platelet volume (MPV)).

Genotyping and quality control

Genomic DNA was extracted from ear tissues using a standard phenol/chloroform method. All DNA samples were qualified and standardized into a final concentration of 20 ng/ul. A total of 1020 individuals in the F_2 pedigree were genotyped for the Porcine SNP60 Beadchips on an iScan System (Illumina, USA) following the manufacturer's protocol. Quality control was executed to exclude SNPs with call rate < 95%, minor allele frequency (MAF) < 5%, severely Hardy Weinberg disequilibrium (*P* < 10E-5) and Mendelian inconsistency rate > 10%. Moreover, individuals with missing genotypes >10% or Mendellian errors >5% were discarded from the data set.

Statistical analyses

Single-marker GWAS: The allelic difference of each SNP in phenotypic traits was tested using a general linear mixed model (Breslow and Clayton 1993; Bradbury *et al.* 2007; Yu *et al.* 2005). The model included a random polygenic effect, and the variance-covariance matrix was proportionate to genome-wide identity-by-state (Hayes and Goddard 2008). The formula of the model in mathematic is: $Y = Xb + \mathrm{S}\alpha + Zu + e$, where Y is the vector of phenotypes, b is the estimator of fixed effects including sex and batch, α is the SNP substitution effect and u is the random additive genetic effect following multinomial distribution u ~ N(0, $\mathbf{G}\sigma_\alpha^2$), in here **G** is the genomic relationship matrix that was constructed based on SNP markers as described in Eding *et al.* (2001) , and σ_α^2 is the polygenetic additive variance. X, Z are the incidence matrices for b and u, S is incidence vector for α , e is a vector of residual errors with a distribution of N (0, $\mathbf{I}\sigma_e^2$). All single-marker GWAS were conducted by GenABEL packages (Aulchenko *et al.* 2007). The genome-wide significant threshold was determined by bonferroni correction, which was defined as

0.05/N, where N is number of tested SNPs. In our study, the number of studied SNPs is 39622 and the corresponding genome-wide threshold is 1.26e-6.

GWAS of time serials data: As all experimental individuals were recorded for hematological traits at three time stages (days 18, 46 and 240), we assumed that measurements at different stages in the same individual would be more correlated than those obtained from different individuals. To conceptualize this assumption, phenotypic records on the three age stages were analyzed together using a mixed effect approach to distinct the correlations within and/or among individuals (Harville 1997; Laird and Ware 1982). The model was similar to the above-mentioned single-marker GWAS model except that the phenotypic variance was partitioned to five parts rather than four parts: variance explained by SNPs, by fixed factors such as sex and batch, by polygenic effects, and by the time stage and by residual errors. The longitudinal GWAS were performed by LONG-GWAS to adjust the variance and covariance structure among the three age stages (Furlotte and Eyheramendy 2012).

Haplotype-based association studies: A haplotype-based association study was also performed to identify genomic regions associated with the tested hematological traits (Druet and Farnir 2011). Haplotypes corresponding to a predetermined number (K = 20) of hidden haplotype states was conducted with a hidden Markov model via PHASEBOOK (Druet and Georges 2010). Association between phenotypes and the hidden haplotypes was detected under a generalized linear mixed framework that corrected population stratification by fitting a random polygenic effect. The mathematic formula of the mixed model was the same as the single-marker analysis, except that S was incidence matrices of hidden haplotype states rather than SNP genotypes and that the estimated haplotype effects were set as random effects.

Linkage disequilibrium (LD) analysis: LD extents were estimated for significant SNPs using HAPLOVIEW (Barrett *et al.* 2005). The LD blocks were determined according to the four-gamete rule to pinpoint plausible candidate genes for hematological traits.

Results

Phenotype statistics and SNP characteristics

Descriptive statistics of the measured traits in the current experimental population are presented in **Table 1**. Of the 7 baseline erythroid traits, 4 parameters including HCT, HGB, MCHC and RBC increased with age and one measurement (RDW) declined with age, while MCV and MCH first decreased and subsequently increased. Of the 7 leukocyte traits, GRAN and GRAR increased while LYMA decreased with age. The tendency of the other leukocyte traits varied irregularly. No consistent variation pattern was observed for platelet traits.

After quality control, 3077 SNPs with call rate <90%, 15711 SNPs with MAF < 0.05, 6502 SNPs showing Hardy Weinberg disequilibrium (*P* < 10E-5) and 208 markers exhibiting Mendelian inconsistency were excluded for further analyses. All individuals are qualified samples. A final set of 39622 SNPs on 1020 individuals were explored for the subsequent GWAS.

Summary of significant loci identified by GWAS

For GWAS on F_2 populations, a cluster of significant SNPs would be typically detected at a significant locus due to large LD extent in such populations. If the distance between two genome-wide significant SNPs was small than 10 Mb, the two SNPs were treated as the same locus otherwise they were considered two independent loci in this study. According to this criterion, 13 genomic regions on 11 autosomes were strongly associated with blood cell parameters by single-marker GWAS or LONG-GWAS (**Table 2 & Table S1**). The 13 regions harbor 185 significant SNPs and 91 annotated genes (data not shown). Of the 185 SNPs, 119 SNPs corresponding to 63 genes were evidenced by both association strategies. In the single-marker GWAS, 147 significant SNPs representing 78 genes were identified in comparison with 158 associated SNPs from 76 genes revealed by LONG-GWAS.

Loci for erythrocyte traits

In total, we found 140 SNPs and 118 SNPs significantly associated with 7 erythrocyte traits by single-marker GWAS and LONG-GWAS, respectively. The corresponding Manhattan plots are shown in **Figure 1** and **Figure S1**.

Single-marker GWAS for erythrocyte traits: For HCT and HGB, we identified 3 genome-wide significant loci on SSC7. The loci cover the region from 33 to 56 Mb and had the effect exclusively on the measurements at day 240. The lead SNPs within the regions are identical for the two traits. No genome-wide significant SNP was found for these traits at days 18 and 46. For MCH, MCV and RBC, all significant SNPs were detected on SCC8 which appeared to harbor two independent associated regions (32.36 - 50.1 Mb and 66.03 - 84.49 Mb). The regions showed constant effect on MCV across the three age stages whereas exhibited time-specific influence on MCH and RBC as the association disappeared on early-stage samples at day 46. Notably, a cluster of SNPs (>100) on this chromosome exhibited signals of strong association with MCV and MCH at day 240 with the top SNP (ss131369293) at 50096834 bp. For MCHC, only ss131260759 at 4248936 bp on SSC4 achieved the genome-wide significant level for the trait measured at day 46.

LONG-GWAS for erythrocyte traits: A total of 118 SNPs within 10 genomic regions showed strong association with erythrocyte-related traits by LONG-GWAS. The results confirmed the findings of single-marker GWAS for HCT on SSC7, and for MCV, MCH and RBC on SSC8. It indicates that these loci consistently regulate red blood cells at different stages. Moreover, we uncovered two novel loci for HGB on SSC1 and SSC12 with the lead SNPs at 65994430 (ss120021119) and 29107229 (ss131459230) bp on the two chromosomes, respectively. One new locus was identified for HCT at 85640695 bp on SSC11. Only one significant locus for MCHC was found at 4168738 bp on SSC10, which differed from the result from the single-marker GWAS.

Loci for white blood cell counts

Only 6 genome-wide significant SNPs on 4 autosomes were identified for leukocyte traits by single-marker GWAS (**Figure 2**). Two SNPs within the *RAB44* gene on SSC7 were associated with WBC at day 240. Two significant SNPs (ss131368550 and ss131364780) for GRAR at days 18 and 46 were found at different positions (40.78 Mb and 26.31 Mb) on SSC8. Moreover, one SNP (ss131544979) at 2971217 bp on SSC17 was associated with LYM at the age of 18 days. A single SNP (ss478935524) at the position of 5860648 bp on SSC18 was associated with LYMA by LONG-GWAS.

Loci for platelet traits

Analysis of platelet traits revealed two significant loci on SSC13 and SSC5 by single-marker GWAS and 39 associated loci by LONG-GWAS (Figure 2). In simple GWAS results, ss131296370 at 95971272 bp on SSC5 and ss107854351 at 14848132 bp on SSC13 were associated with PLT and PDW at day 46 with *P*-values of 1.14 E-8 and 3.94 E-7 respectively. For LONG-GWAS, two significant loci for PDW were detected on SSC8. The lead SNPs at the two loci were SS131368505 at 40852645 bp and ss131371056 at 75662581 bp with a distance of 34.81 Mb, implying at least 2 loci controlling PDW on chromosome 8.

Discussion

GWAS versus QTL mapping

We have previously performed genome scans on the F_2 population using 183 microsatellites. We detected 46 genome-wide significant QTL for baseline erythroid traits, 8 for leucocyte-related traits and 6 for platelet-related traits. These QTL are distributed on SSC1, 2, 7, 8, 10, 12, 13, 15 and X (Yang *et al.* 2009; Zou *et al.* 2008). In the present study, we did not test the association of SNPs on chromosome X as the currently available GWAS statistical models are hard to handle the random inactivation situation on the sexual chromosome. By using single-marker GWAS or LONG-GWAS on the 60K SNP data, we confirmed the previously identified QTL on SSC7 and SSC8 accounting for 46% of the total detected QTLs and uncovered two additional genome-wide loci for HCT and HGB on SSC11 and SSC12, respectively. QTL mapping studies in F_2 populations were generally conducted by comparing the phenotypic difference between F_2 individuals inheriting different alleles from the founder breeds under the assumption that QTL alleles were alternatively fixed in each founder breed of the F_2 intercross. The advantage of this mapping strategy is that we can anchor genomic regions affecting phenotypic traits using sparse markers. However, it could result in false negative signals If QTL alleles are segregating within founder lines. Moreover, the confidence intervals of most QTL were larger than 20 cM. In our QTL mapping study, the smallest intervals were 3 and 4 cM for MCV and MCH at day 18 on SSC8, respectively (Yang *et al.* 2009). In comparison, GWAS test the average phenotypic difference grouped by alternative alleles of high density markers and without any assumption; thereby it could identify significant signals even if QTL is not fixed in founder breeds. Moreover, GWAS could narrow down the confidence interval of QTL to small genomic regions. In the current study, the confidence interval for MCV at day 18 on SSC8 were 0.80 Mb based on the 1.5 Lods drop rule (Dupuis and Siegmund 1999). Only a handful of genes exist in such small regions.

QTL replication with other studies

Until now, there were only two very recent GWAS papers describing significant genomic loci for hematological parameters in pigs (Luo *et al.* 2012; Wang *et al.* 2013). Luo *et al.* (2012) detected 62 genome-wide significant and 3 chromosome-wide significant SNPs associated with erythrocyte traits by performing GWAS on a Large White × Chinese Min F_2 intercross. All significant SNPs were found on SSC7 and SSC8 except one SNP associated with RBC on SSC1 and two SNPs for RDW on SSC12. Our results confirmed all of these findings. Wang *et al.* (2013) identified 111 SNPs including 10 genome-wide significant SNPs and 101 chromosome-wide significant SNPs for 15 hematological traits in 2 Western breeds and one Chinese synthetic breed. The 111 SNPs are distributed on all autosomes except for SSC7, SSC8 and SSC18. However, none of these SNPs were replicated in this study. The reasons for the inconsistence could be different genetic background of experimental populations in the two studies, the complex genetic basis of hematological related cells, and different trait recording methods. Wang *et al.* (2013) measured hematological traits on pigs at day 35 after immunized with classical swine fever vaccine at day 21.

Time constant and specific QTL

The single-marker GWAS revealed that the significant locus on SSC8 was consistently associated with MCV measured at days 18, 46, and 240, suggesting that a common variant regulates MCV at the whole life stage. The constant effect of this locus on MCV was further confirmed by LONG-GWAS that treated MCV data at the three ages together and obtained the same finding as the single-marker GWAS. In contrast, ss131544979 at 2971217 bp on SSC17 showed time specific effect on LYM at day 18. SNPs within two different regions on SSC8 were associated with GRAR at days 18 and 46, respectively, indicating that distinct genes are involved in development stages of granulocyte cells. Time specific loci were also evidenced for PLT and PDW on SSC5 and SSC13 as the association signals were observed only from the data at day 46. A high proportion of SNPs on SSC7 and SSC8 were identified for erythroid traits and leukocytes traits at day 240. Notably, the SNPs on SSC7 for HCT, HGB and WBC had a significant effect only at day 240 (**Table 2**) and therefore can be viewed as a late-acting QTL. It should be noted that all significant SNPs on SSC7 were not located in the SLA region (24.7 Mb - 29.8 Mb), which was response for immune system and suspected to a range of diseases.

LONG-GWAS analysis

Currently, standard GWAS (e.g. GenABEL) only utilized one time point for each individual. If QTL constantly control the traits during the whole life process, it is reasonable to assume that jointly analysis of data at all-time points may be more powerful than the single-time-point approach. We thus used LONG-GWAS that utilized multiple phenotype measurements for each individual as proposed by Furlotte *et al.* (2012). We not only replicated the linkage mapping results for hematological trait, but also identified five new QTL affecting HGB, MCHC, HCT and LYMA on chromosomes 1, 12, 10, 11, and 18 respectively. One disadvantage of LONG-GWAS is that the significant signal may be overwhelmed by putting all time point's data together if QTL effects vary during time stage. In this study, 5 time-specific expressed QTL for MCH, MCHC, PLT, PWD and LYM at early stage from 18 to 46 days were identified by the singer-marker GWAS, while theses loci were not detected by LONG-GWAS.

Haplotype analysis for single SNP associated with measured phenotypes

In the standard GWAS, a prominent locus is usually featured by a lead SNP and a cluster of surrounding significant SNPs within a genomic region especially in the F_2 pedigrees, in which high LD extents are expected. However, only one genome-wide significant SNP was associated with GRAR at days 18 and 46 (**Table S2**). None of suggestive SNP was detected in the neighboring region of the top SNP. To test if the signal was false positive result or real association, we conducted a haplotype based GWAS for this trait (**Figure S2**). We showed that dozens of genome-wide SNPs in a large interval were uncovered for GRAR at day 46, and the position of the top SNP was exactly the same to the lead SNP in the single-marker GWAS. For GRAR at day 18, the most significant SNPs were moved to another position (86.8 Mb), but the second top marker was identical to the top SNP identified in the single-marker GWAS. The findings support the reliability of the single significant marker for GRAR.

Plausible candidate genes at the significant locus on SSC8

In the present study, the most interesting finding is the major locus for multiple hematological traits on SSC8. More than half of the detected regions (14 regions) were located on SSC8 that were associated with 9 hematological traits. Hundreds of significant SNPs in a large single region of more than 45 Mb on SSC8 were identified for MCH and MCV at day 240. In contrast, two segments in the ~45 Mb region appeared to be independently associated with the two measurements at day 18 as no significant SNP was found in the inter-segment region (**Table 2**). To investigate whether one or two loci affect MCV and MCH on this chromosome, we re-analyzed MCH and MCV data conditional on the allelic effect of the lead SNPs (**Figure S3**). For MCH at day 18, after controlling for the effect of ss131369293, the second significant region disappeared. Also, no SNP showed association with MCH and MCV at day 240 after correcting for the effect of the lead SNP (ss131369293). Moreover, the complete LD (*D*' = 1) was observed for the two top SNPs (ss131369009 at 44.93 Mb and ss107827400 at 66.03 Mb) for MCV at day 240 in the two intervals within the ~45 Mb region on SSC8. These observations support one major locus for the tested hematological traits on this chromosome. However, we can not rule out the possibility that two neighboring genes contribute to the phenotypic traits as the conditional analysis can not distinguish the effects of two adjacent loci due to LD.

We noticed that SNPs associated with MCH, MCV, RBC and GRAR at the early stage mainly reside in the region of 36.31 to 50.10 Mb. The stem cell growth factor receptor (*KIT*) gene around 43.55 Mb on this chromosome stands out as a compelling candidate gene as it is essential for the development of hematopoietic stem cells and has been highly expressed in hematopoietic cells (Escribano *et al.* 1998; Sakurai *et al.*i 1996). Several mutations in *KIT* have significant influence in RBC in the mouse (Jackson *et al.* 2006). Johansson *et al.* (2006) showed strong association of *KIT* variants with erythroid traits in piglets. Moreover, Fésüs *et al.* (2005) reported the mild effect of *KIT* on hematological parameters in adult pigs. Our observation of the *KIT* region associated with hematological traits reinforces the assumption that the *KIT* gene has a significant effect on peripheral blood cell measures in pigs (Cho *et al.* 2011).

Conclusion

In conclusion, a total of 185 genome-wide significant SNPs corresponding to 91 genes were identified for 18 hematological traits at the three growth ages in the White Duroc × Erhualian F_2 intercross. These loci confirmed the previously identified QTL and showed both time constant and specific effects on the measured traits. Of these findings, the most prominent one was the genomic region between 32.36 and 84.49 Mb on SSC8 that is associated with multiple erythroid traits. The *KIT* gene on this chromosome appears to be a promising candidate gene. The findings improve our understanding of the genetic architecture of hematological traits in pigs. Further investigations are warranted to characterize the responsible gene(s) and causal variant(s) especially for the major loci on SSC7 and SSC8.

Table 1. Descriptive statistics of 18 hematological traits at three growth stages in the F_2 resource population

Trait	**Abbreviation**	**Value (No.)**		
		Day 18	**Day 46**	**Day 240**
Hematocrit (%)	HCT	0.30 ± 0.07 (1447)	0.30 ± 0.07 (1010)	0.41 ± 0.05 (1010)
Hemoglobin (g/l)	HGB	95.55 ± 22.83 (1444)	100.67 ± 18.71 (1025)	137.35 ± 16.08 (1025)
Mean corpuscular hemoglobin (pg)	MCH	19.34 ± 3.21 (1445)	18.24 ± 4.34 (1009)	19.28 ± 1.49 (1009)
Mean corpuscular hemoglobin content (g/l)	MCHC	319.79 ± 30.86 (1442)	341.41 ± 51.29 (1009)	337.13 ± 15.91 (1009)
Mean corpuscular volume (fl)	MCV	60.66 ± 9.64 (1446)	52.70 ± 8.13 (1019)	57.25 ± 4.29 (1019)
Red blood cell count (10^{12})	RBC	4.92 ± 0.94 (1447)	5.67 ± 1.12 (1010)	7.14 ± 0.87 (1010)
Red cell distribution width (%)	RDW	25.45 ± 4.65 (1405)	24.10 ± 6.34 (1002)	18.94 ± 3.39 (1002)
Granulocyte count (10^{9})	GRAN	1.31 ± 1.35 (1433)	2.96 ± 2.91 (797)	7.47 ± 4.31 (797)
Granulocyte count percentage(%)	GRAR	11.89 ± 10.21 (1433)	15.77 ± 13.28 (910)	41.71 ± 20.69 (910)
Lymphocyte count (10^{9})	LYM	9.28 ± 5.43 (1433)	13.76 ± 5.4 (1024)	7.85 ± 3.74 (1024)
Lymphocyte count percentage(%)	LYMA	79.29 ± 13.79 (1448)	75.06 ± 16.39 (1025)	46.29 ± 17.85 (1025)
Monocyte count (10^{9})	MON	0.21 ± 0.24 (550)	0.40 ± 0.32 (754)	0.22 ± 0.23 (754)
Monocyte count percentage(%)	MONR	1.96 ± 2.11 (551)	2.19 ± 1.68 (754)	1.30 ± 1.35 (754)
White blood cell count (10^{9})	WBC	11.54 ± 6.03 (1449)	18.45 ± 5.97 (1024)	17.02 ± 4.39 (1024)
Mean platelet volume (fl)	MPV	7.95 ± 1.67 (666)	8.30 ± 1.94 (760)	7.93 ± 1.25 (760)
Plateletcrit (%)	PCT	0.44 ± 0.23 (549)	0.75 ± 1.08 (753)	0.23 ± 0.1 (753)
Platelet distribution width (%)	PDW	15.53 ± 2.44 (664)	14.43 ± 1.85 (761)	15.49 ± 0.94 (761)
Platelet count (10^{9})	PLT	559.26 ± 219.65 (1386)	590.21 ± 271.91 (1009)	295.67 ± 117.75 (1009)

Values are shown in mean ± standard deviation; the numbers of recorded individuals are given in parentheses. Description of 7 erythroid traits in ~1420 animals genotyped for 183 microsatellite markers has been shown in Zou *et al.* (2008).

Table 2. Genome-wide significant loci associated with hematological traits by the single-marker analysis

Trait [1]	**Top SNP**	**Chr** [2]	**Pos (bp)** [3]	***P*-Value**	**Num** [4]	**Interval (Mb)** [5]	**Nearest gene** [6]
HCT240	ss107806758	7	35177641	8.98E-10	3	33.18 - 35.18	*SPDEF*
HCT240	ss131349087	7	45420438	1.69E-08	1	45.42	*CDC5L*
HCT240	ss131354973	7	56230077	5.90E-08	2	54.81 - 56.23	*EFTUD1*
HGB240	ss107806758	7	35177641	8.97E-11	8	33.18 - 35.25	*SPDEF*
HGB240	ss131349087	7	45420438	1.36E-07	1	45.42	*CDC5L*
HGB240	ss131354973	7	56230077	2.91E-07	1	56.23	*EFTUD1*
MCH18	ss131369009	8	44927836	5.03E-07	5	39.15 - 50.1	*TLL1*
MCH18	ss131083163	8	76480145	1.43E-07	2	76.48 - 77.78	*SHROOM3*
MCH240	ss131369293	8	50096834	1.17E-19	122	31.53 - 79.81	*PPID*
MCHC46	ss131260759	4	42489363	8.57E-07	1	42.49 - 42.49	*PGCP*
MCV18	ss131369293	8	50096834	2.66E-10	6	36.97 - 50.1	*PPID*
MCV18	ss131083163	8	76480145	2.77E-09	13	66.03 - 79.51	*SHROOM3*
MCV46	ss478938668	8	42150857	6.81E-07	1	42.15	*KIT*
MCV46	ss131076611	8	76669199	1.26E-06	1	76.67	*SHROOM3*
MCV240	ss131369293	8	50096834	1.08E-17	103	34.39 - 79	*PPID*
RBC18	ss131094241	8	49881116	1.40E-09	9	32.36 - 49.88	*RXFP1*
RBC18	ss107827400	8	66027033	3.12E-08	7	66.03 - 84.49	*TECRL*
RBC240	ss131369293	8	50096834	7.42E-11	14	34.39 - 50.1	*PPID*
RBC240	ss107906810	8	72567179	4.31E-08	5	66.03 - 74.51	*U6*
GRAR18	ss131368550	8	40777538	3.58E-07	1	40.78	*OCIAD1*
GRAR46	ss131364780	8	26310331	7.61E-07	1	26.31 - 26.31	*RPS18*
LYM18	ss131544979	17	2971217	2.64E-07	1	2.97	*ssc-mir-383*
WBC240	ss131344940	7	37288793	1.93E-07	2	37.23 - 37.29	*RAB44*
PDW46	ss107854351	13	14848132	3.94E-07	1	14.85	*LRRC3B*
PLT46	ss131296370	5	95971272	1.14E-08	1	95.97	DCN

[1] Abbreviations of hematological traits are given in Table 1. e.g. HCT240 is hematocrit at 240 days.

[2] Chromosomal locations of top SNPs.

[3] Positions of the top SNPs according to *Sus scrofa* Build 10.2 genome assembly.

[4] The number of genome-wide significant SNPs for each hematological trait

[5] The associated interval was defined as the region in which the distance between any two neighboring genome-wide significant SNPs was less than 10 Mb.

[6] Annotated genes nearest to the top SNPs

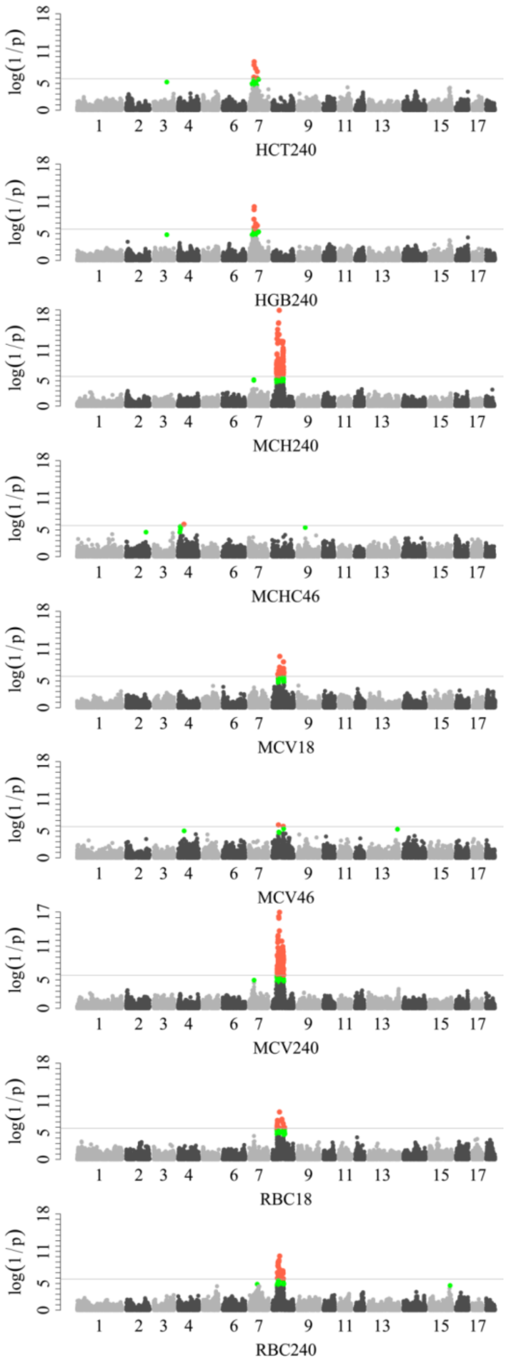


Figure 1. Manhattan plots for the single-marker analysis of erythrocyte traits. $\log_{10}(1/p)$ values are shown for all SNPs that passed quality control. The dotted line denotes the Bonferroni-corrected genome-wide significant threshold. SNPs surpassing the genome-wide threshold are highlighted in pink and SNPs reaching the suggestive threshold in green. HCT240: hematocrit at 240 days; HGB240: hemoglobin at 240 days; MCH240: mean corpuscular hemoglobin at 240 days; MCHC46: mean corpuscular hemoglobin content at 46 days; MCV18, MCV46, MCV240: mean corpuscular volume at 18, 46 and 240 days; RBC18 and RBC240: red blood cell count and at 240 days.

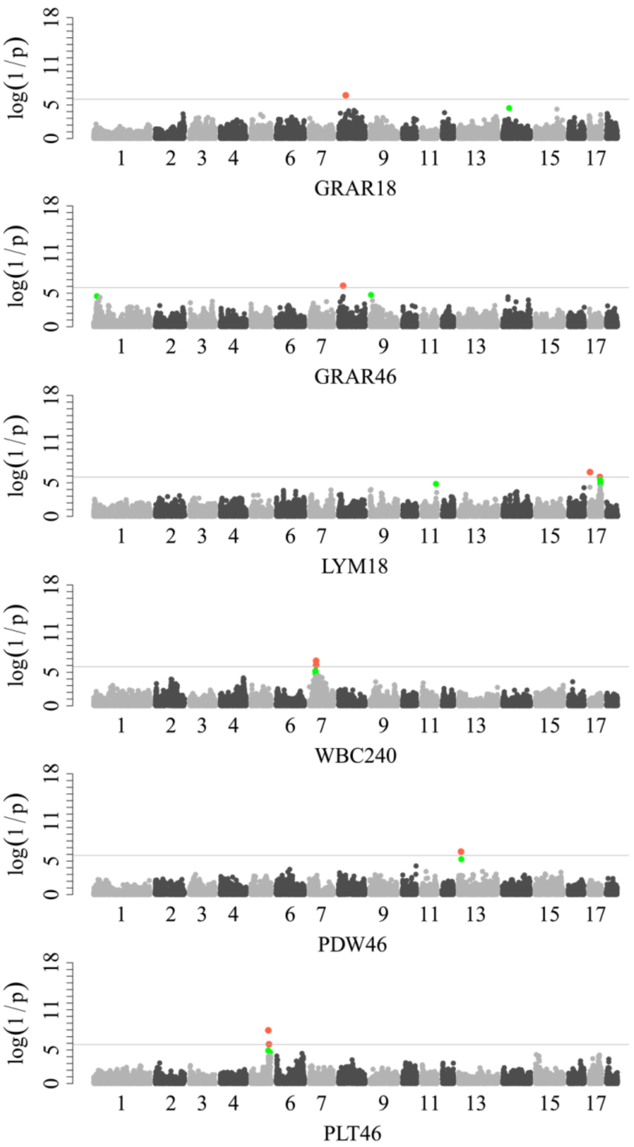


Figure 2. **Manhattan plots for the single-marker analysis of white blood cell and platelet traits.** log10(1/p) values are shown for all SNPs that passed quality control. The dotted line denotes the Bonferroni-corrected genome-wide significant threshold. SNPs surpassing the genome-wide threshold are highlighted in pink and SNPs reaching the suggestive threshold in green. GRAR18 and GRAR46: granulocyte count percentage at 18 and 46 days; LYM18: lymphocyte count at 18 days; WBC240: white blood cell count at 240 days; PDW46: platelet distribution width at 46 days; PLT46: plateletcrit at 46 days.

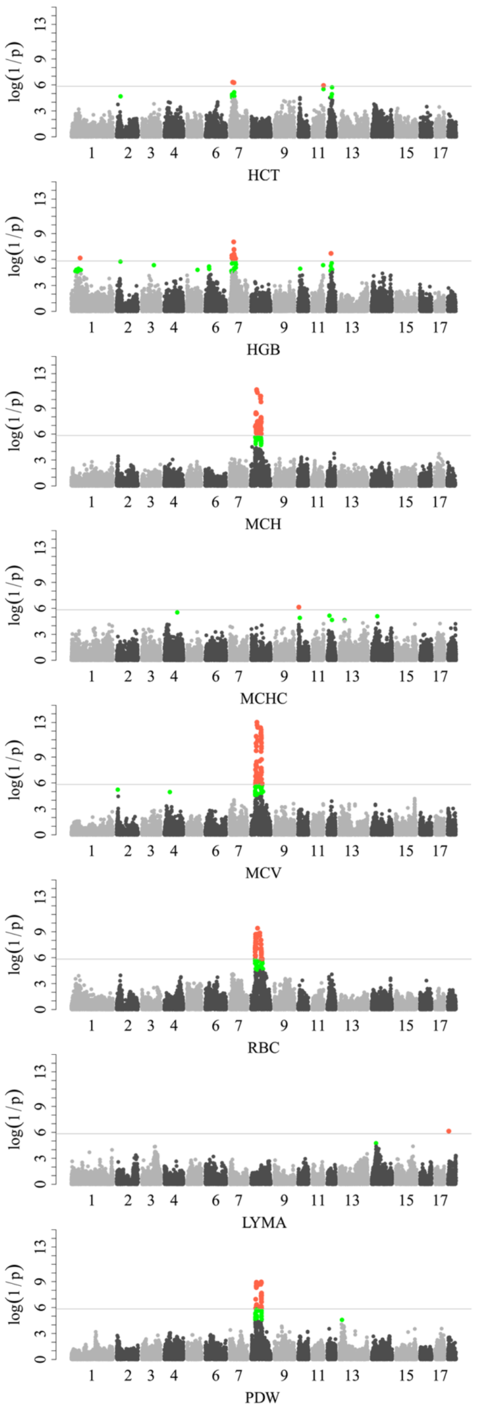


Figure S1. **Manhattan plots for the LONG-GWAS analysis of hematological traits**. $\log_{10}(1/p)$ values are shown for all SNPs that passed quality control. The dotted line denotes the Bonferroni-corrected genome-wide significant threshold. SNPs surpassing the genome-wide threshold are highlighted in pink and SNPs reaching the suggestive threshold in green. HCT: hematocrit; HGB: hemoglobin; MCH: mean corpuscular hemoglobin; MCHC: mean corpuscular hemoglobin content; MCV: mean corpuscular volume; RBC: red blood cell; LYMA: lymphocyte count percentage; PDW: platelet distribution width.

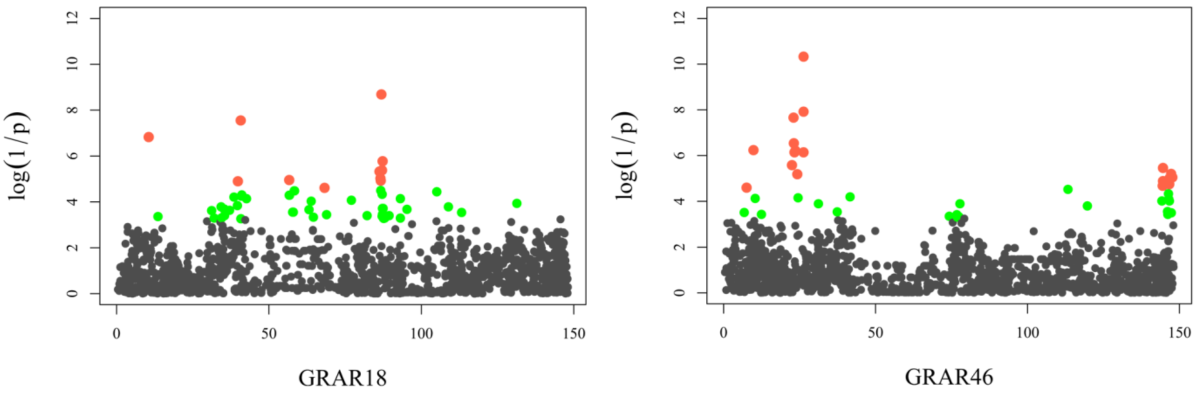


Figure S2. Manhattan plots for the hidden haplotypes analysis of GRAR at 18 and 46 days on SSC8 where only one SNP was associated with each trait in the single-marker analysis. SNPs surpassing the genome-wide threshold are highlighted in pink and SNPs reaching the suggestive threshold in green.

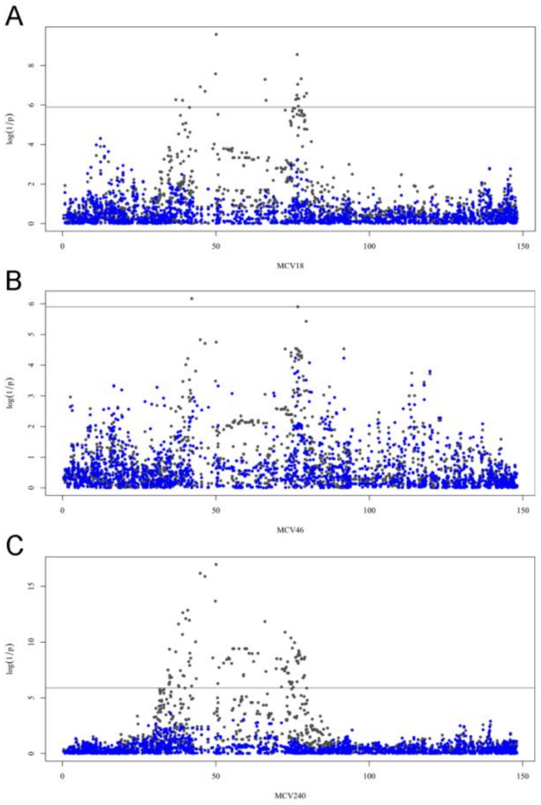


Figure S3. **Manhattans plots for conditional GWAS of MCV at 18 (A), 46 (B) and 240 (C) days.** Grey and blue dots denote the results for SNPs before and after controlling for the top SNP (ss131369293) at 50.10 Mb on SSC8, respectively. Grey lines represent the genome-wide significant threshold.

Table S1. Genome-wide significant SNPs associated with hematological traits by LONG-GWAS.

Trait [1]	Top SNP	Chr [2]	Pos (bp) [3]	*P*-Value	Num_SNP[4]	Interval (Mb) [5]	Nearest Gene [6]
HCT	ss131338218	7	21815831	4.85E-07	2	21.82 - 31.03	*SLC17A4*
HCT	ss131455151	11	85640695	1.17E-06	1	85.64 - 85.64	*TUBGCP3*
HGB	ss120021119	1	65994430	6.71E-07	1	65.99 - 65.99	*7SK*
HGB	ss131341609	7	31027719	9.32E-09	15	17.05 - 45.42	*TINAG*
HGB	ss131459230	12	29107229	1.94E-07	2	29.02 - 29.11	*CA10*
MCH	ss131369009	8	44927836	7.01E-12	95	34.9 - 79.19	*TLL1*
MCHC	ss131567944	10	4168738	7.36E-07	1	4.17 - 4.17	*FAM5C*
MCV	ss131369009	8	44927836	9.28E-14	95	34.39 - 84.49	*TLL1*
RBC	ss131094241	8	49881116	4.00E-10	29	31.09 - 50.1	*RXFP1*
RBC	ss478935224	8	66349700	1.37E-09	27	66.03 - 85.12	*TECRL*
LYMA	ss478935524	18	5860648	7.13E-07	1	5.86 - 5.86	*GALNTL5*
PDW	ss131368505	8	40852645	1.18E-09	13	34.39 - 50.1	*OCIAD1*
PDW	ss131371056	8	75662581	1.07E-09	26	66.03 - 79	*PPEF2*

[1] Abbreviations of hematological traits are given in Table 1. e.g. HCT240 is hematocrit at 240 days.

[2] Chromosomal locations of top SNPs.

[3] Positions of the top SNPs according to *Sus scrofa* Build 10.2 genome assembly.

[4] The number of genome-wide significant SNPs for each hematological trait

[5] The associated interval was defined as the region in which the distance between any two neighboring genome-wide significant SNPs was less than 10 Mb.

[6] Annotated genes nearest to the top SNPs

Table S2. Simple statistic results for GRAR at 18 and 46 days classified by the genotypes of the top SNPs.

	Mean ± standard deviation (Number of individuals)		
Genotype	11	12	22
GRAR18	3.67 ± 1.90 (13)	8.32 ± 7.80 (280)	10.92 ± 9.48 (478)
GRAR46	10.57 ± 2.12 (3)	24.93 ± 15.07 (127)	13.55 ± 12.56 (646)

Results of a haplotype-based GWAS for Recurrent Laryngeal Neuropathy in the horse

M. C. Dupuis, Z. Zhang, T. Druet, C. Charlier, P. Lekeux and M. Georges

Abstract

Recurrent laryngeal neuropathy (RLN) is a major upper airway disease of horses, which causes abnormal respiratory noise during exercise and can impair performance. Aetio-pathogenesis remains unclear but genetic factors have been suspected for many decades. The objective of this study was to identify risk loci associated with RLN. Horses with extreme phenotypes were carefully selected. The Illumina Equine SNP50 Beadchip was used to genotype 234 cases (196 Warmbloods, 20 Trotters, 14 Thoroughbreds and 4 Draft Horses), 228 breed-matched controls and 69 parents. Statistical analyses included quantification of population structure, single marker and haplotype-based association studies, and family-based analyses. Population stratification was corrected by modeling a random polygenic background effect with covariance structure estimated from genome-wide SNP data. Two genome-wide suggestive loci were identified respectively on chromosomes 21 ($p = 1.62x10^{-6}$) and 31 ($p = 1.69x10^{-5}$), using haplotype-based association studies in Warmbloods. Both signals were driven by the enrichment of a “protective” haplotype in controls compared to cases.

Mamm. Genome 22 (2011) 613-620

Introduction

Recurrent Laryngeal Neuropathy (RLN), also known as idiopathic laryngeal hemiplegia, is the most common obstructive upper airway disorder of the horse. It is usually detected at two to three years of age, at the onset of training, as RLN causes abnormal inspiratory sounds during effort (“roaring” or “whistling”) and can impair performance (Cahill and Goulden 1987; Dixon *et al.* 2001). At endoscopic examination, the disease manifests itself as a reduced abduction of the left arytenoid cartilage. It is thought to be due to laryngeal muscular dysfunction caused by degeneration of the recurrent laryngeal nerves. Affected animals often undergo surgery to improve performance and the condition therefore has important welfare and economic implications (Dixon *et al.* 2009; Robinson 2004).

Estimates of RLN prevalence vary depending on breed and criteria used to define the phenotype: between 1 to 8.3% in Thoroughbreds (Brown *et al.* 2005; Garrett *et al.* 2010; Lane *et al.* 1987; Sweeney *et al.* 1991), 24 to 42% in draft horses (Archer *et al.* 1989; Brakenhoff *et al.* 2006), and as high as 50% in adult Warmbloods (including animals with slight paresis; (Ohnesorge *et al.* 1993). It seems to occur at higher frequency in taller horses and in males (Beard and Haynes 1993; Goulden and Anderson 1981a). Contrary to inherited myelinopathies or axonopathies in humans or dogs, RLN is not classified as a polyneuropathy but as a bilateral mononeuropathy (Hahn *et al.* 2008). Indeed, horses affected by RLN do not show clinical signs of polyneuropathy (e.g. limb muscle weakness, sensory symptoms, and megaoesophagus) and typical histological lesions have only been reported in the recurrent laryngeal nerves, especially the left one which is the longest nerve in the horse. These lesions include signs of degeneration (i.e. loss of large myelinated fibers, most severely distally) and regeneration (Cahill and Goulden 1986a, b, c; Duncan *et al.* 1978).

In two retrospective studies, etiologic factors of laryngeal paresis or paralysis (such as irritant perivascular injection, trauma, guttural pouch mycosis, hepatic encephalopathy, poisoning) were identified in 6-11% of affected horses (Dixon *et al.* 2001; Goulden and Anderson 1981b). For the remaining 89-94%, which were qualified to be RNL cases, the etiology was undetermined. Genetic factors have been suspected for many decades (Marti and Ohnesorge 2002). Indeed, sseveral studies have reported a higher incidence of RLN in offspring of affected than unaffected stallions (Ohnesorge *et al.* 1993; Poncet *et al.* 1989). The heritability has been estimated at 0.2 in Thoroughbreds (Ibi *et al.* 2003), 0.4 in Clydesdale (Barakzai 2009) and 0.6 in German saddle horses (Ohnesorge *et al.* 1993).

The aim of our study was to carry out a genome-wide association study (GWAS) using the Equine SNP50 Beadchip from Illumina to identify risk loci associated with RLN in horses.

Materials and Methods

Animals and phenotypes

Clinical data were collected between 2008 and 2010 at the CIRALE (French center of imaging and research on equine locomotor disorders) with the collaboration of seventeen veterinary clinics located in France, Belgium, Germany and Switzerland. For each horse, we recorded breed, sex, date of birth, pedigree and medical history. Throat and neck were examined very thoroughly to exclude laryngeal paralysis secondary to for instance jugular phlebitis. The laryngeal mobility was assessed under laryngoscopy, without sedation, for at least one minute per nostril. Abductory and adductory muscles were stimulated by inducing swallowing and transient hyperventilation by nasal occlusion. All videoendoscopies were recorded. Horses with laryngeal grade ≥ III.2 (Havemeyer workshop grading system (Robinson 2004); full abduction of the arytenoid cartilage never achieved) were considered as cases. Horses with laryngeal grade I (synchronous and symmetrical movements) and ≥ 3 years were used as controls, as it has been shown that laryngeal grade may evolve from normal to abnormal before the age of 3 (Anderson *et al.* 1997). Horses with intermediate grades were excluded from the study because of reported unpredictable disease progression (Dixon *et al.* 2002), low intra-observer repeatability (Ducharme *et al.* 1991; Perkins *et al.* 2009), and low correlation with histopathology (Collins *et al.* 2009; Piercy *et al.* 2009). Blood samples were collected on EDTA for all examined horses and stored at -20 ℃.

We collected samples from a total of 234 cases (196 Warmbloods (W), 20 Trotters (TR), 14 Thoroughbreds (TH) and 4 draft horses) and 228 breed-matched controls. Thirty-four cases and 17 controls belonged to six large paternal half-sib W families. The remaining 200 cases and 211 controls were descendent of 360 sires and 408 dams. Samples were available for 43 (unphenotyped) sires and 26 (unphenotyped) dams (Table 1). Male horses were overrepresented in both cases (75%) and controls (64%).

Genotyping and quality control

Genomic DNA was extracted from 350 µl of blood using the MagAttract DNA Blood Midi M48 Kit (Qiagen) and was quantified using the Picogreen assay (Invitrogen). Genotyping was performed using the Equine SNP50 Beadchip (Illumina) including 54,602 evenly distributed SNPs on the 31 equine autosomes and X chromosome (average distance between markers: 43kb), and standard protocols recommended by the manufacturer.

Average and minimum call rate per individual was 99.68% and 96% respectively, allowing us to keep all individuals. Gender, replicates and parental relationships were checked with the Genome Studio software (Illumina). Minor allele frequency (MAF) was calculated for each SNP in each breed. 7,556 SNPs were excluded based on genotyping rate (call freq < 90%; 329 SNPs), minor allele frequency (MAF < 5%; 6986 SNPs), and Hardy-Weinberg equilibrium in Warmbloods ($p < 10^{-3}$; 433 SNPs).

Statistical analyses

Population structure. The PLINK toolset (version 1.07, http://pngu.mgh.harvard.edu/~purcell/plink/) was used to quantify population stratification based on pairwise identity-by-state (IBS) distances (Purcell *et al.* 2007).

Single marker association studies. Genotype frequencies at 47,046 SNPs were compared between cases and controls using a standard association test implemented with PLINK. Quantile-quantile plots (QQ Plots) were generated to detect inflation of statistics due to population stratification. Association studies accounting for stratification were conducted on autosomal SNPs with the GenABEL toolset (available at http://mga.bionet.nsc.ru/~yurii/ABEL/GenABEL/) that models estimated relatedness of individuals based on genome-wide SNP data (Aulchenko *et al.* 2007). SNP-specific test statistics (chi-squared values) obtained within each of the three analysed breed (W, TR and TH) were summed to obtain an across-breed statistic and p-value (chi-squared statistics with degrees of freedom equal to the number of breeds). Draft horses were not included in the analyses as there were too few of them.

Haplotype-based association studies. We reconstructed autosomal haplotypes using the Phasebook software package, that exploits population (linkage disequilibrium) and familial information (Mendelian segregation and linkage) in a Hidden Markow Model setting (Druet and Georges 2010). We tested three different numbers of hidden states (10, 15, 20). Population frequencies of the hidden haplotype states were then compared between cases and controls using a generalized mixed model, which includes a random polygenic effect for which the variance-covariance matrix is proportionate to genome-wide identity-by-state. Data from the three breeds were combined by summing the SNP-specific score test values (which have a gamma distribution) to obtain an across-breed statistic and p-value (gamma statistic with the mean equal to the sum of the means and the variance equal to the sum of the variances).

Within-family analyses. One of the six larger paternal half-sib pedigrees was unusual in that the frequency of RLN in male and female offspring of the sire was reportedly close to 50%, suggesting autosomal dominant inheritance. Samples were available from nine affected and seven non-affected offspring. As no DNA was available for the stallion, precluding straightforward linkage analysis, we compared allele and haplotype frequencies between half-sib cases and controls using Fisher's exact test.

Significance thresholds. Nominal p-values corresponding to genome-wide thresholds to declare associations as being significant (one such signal expected per 20 genome scans by chance; $p_{genome\text{-}wide} = 0.05$) or suggestive (one such signal expected per genome scan by chance; $p_{genome\text{-}wide} = 0.63$) (Lander and Kruglyak 1995) were determined by applying a conservative Bonferroni correction for the realization of 47,046 tests (corresponding to the number of usable SNPs), yielding respective thresholds of 1.09×10^{-6} (significant, log(1/p)=5.96) and 2.11×10^{-5} (suggestive, log(1/p)=4.67).

Results

Informativeness of the utilized SNP panel. The frequency distributions of MAF are represented in figure 1 for W, TR and TH. Generally speaking, MAF of polymorphic SNPs (MAF > 0) were uniformly distributed in the three populations, and their average heterozygosities comparable (W: 0.325, TR: 0.323, TH: 0.346). However, the proportion of polymorphic SNPs differed considerably between breeds (W: 0.97, TR: 0.91 TH: 0.87). The higher proportion of polymorphic SNPs in W remained when accounting for differences in sample size (Suppl. Fig. 1).

The power of association studies is directly proportionate to the level of linkage disequilibrium (LD) between causative variants and SNP markers measured using r^2. To assess the level of genome coverage provided by the utilized SNP panel in the studied populations, we treated each one of the 47,046 markers in turn as pseudo-causative variant and identified the SNP with highest r^2 among the remaining ones, as well as the distance between them. As estimates of r^2 are known to be inflated by the quantity 1/*n* (where *n* is the size of the utilized sample), calculations were performed using a random sample of same size (*n*=37) for the three populations. Figure 2 shows the cumulative frequency distribution of r^2_{max} in W, TH and TR. r^2_{max} values were highest in TH (mean r^2_{max} 0.82), followed by TR (0.74) and W (0.69). As r^2 is a measure of effective population size, the observed ranking was as expected, reflecting tight regulation of reproduction in TH (closed studbook) when compared to TR (partially open studbook) and W (admixed breeds). Of note, r^2_{max} values were inferior to 0.5 for 37% of SNPs in W (providing the largest case-control cohort for GWAS), indicating suboptimal LD-coverage of the marker panel for single-point analyses, in at least some regions.

Between- and within-population structure. Population structure was quantified using genome-wide pairwise identity-by-state (IBS) distances calculated by PLINK. Supplementary Figure 2 shows the first four principal components of the multidimensional scaling (MDS) plot of all analyzed animals. In general, TH, TR and Draft horses form tight, well-separated clusters, while W are considerably more scattered and located between the other clusters with whom they partially overlap (particularly with TH). This is as expected given W origin: a group of horse types bred for equestrian sports, descending from heavier agricultural types upgraded by "hotblood" influence, mainly TH and sometimes TR.

Figure 3 shows the MDS-plot for the W, sorted in cases and controls. At first glance, spatial distribution of cases and controls seemed very comparable. Nevertheless, examination of the average between group (case-case, control-control, case-control) IBS metrics using PLINK revealed a significant ($p < 0.04$) excess resemblance within controls, leading to a significant stratification ($p < 0.0001$) between cases and controls. This is likely related to the fact that controls came from a smaller number (148) of stables than cases (216).

Population-based GWAS. QQ-plots of the p-values obtained using standard single-SNP association studies conducted with PLINK confirmed the stratification (λ_{GC}=1.36) anticipated from the analysis of the IBS metrics (Suppl. Fig. 3). Rather than controlling Type I error by applying genomic control (Devlin and Roeder 1999; Devlin *et al.* 2001), thereby reducing detection power (Balding 2006), we modeled a random polygenic background effect in restricted maximum likelihood (REML)-type analyses. This was done using GenABEL (Aulchenko *et al.* 2007) for single-SNP analyses, and Phasebook (Druet and Georges 2010) for haplotype-based analyses. The effect on the distribution of p-values is reflected in the corresponding QQ-plots (Supplementary Fig. 4 and Fig. 4A). While considerably improved, the distribution still showed a modest shift towards low p-values. We therefore applied an additional genomic control-type correction such that p-values < 10^{-3} would lie on the expected diagonal (Fig. 4A).

When using the single-marker model accounting for stratification, a single SNP on chromosome 21 yielded suggestive evidence of association when analyzing W alone, while no SNP exceeded this threshold when analyzing the three breeds jointly (Supplementary Fig. 4). When performing haplotype-based analysis (K=20) in the W population, we obtained suggestive associations on chromosomes 21 (log(1/p)=5.79) and 31 (log(1/p)=4.77) (Fig. 4A). Analysis of hidden haplotype frequencies in cases and controls revealed that in both instances (chromosomes 21 and 31) the signal was due to the enrichment of a specific haplotype in controls (Fig. 4B and 4C). For both chromosomes, the "protective" haplotypes had a frequency of ~0.35 in controls versus ~0.20 in cases. The two loci were also detected when modeling smaller number of Hidden States albeit with lower log(1/p) values (chr 21: 4.99 (K=10), 4.57 (K=15); chr 31: 1.81 (K=10), 4.52 (K=15)). When adding information from TH and TR in an across-breed analysis, the signal on chromosome 21 exceeded genome-wide significance (log(1/p)=6.03), while the signal on chromosome 31 remained suggestive yet increasing in significance (log(1/p)=4.94) (Supplementary Fig. 5). However, as no obvious peaks were detectable at the corresponding chromosome 21 and 31 positions when analyzing TH and TR separately, we consider this increase in significance with caution and consider both loci as suggestive.

Removing p-values corresponding to marker positions within 2 Mb from the most strongly associated positions on chromosomes 21 and 31 from the W QQ-plot reveals two closely linked SNPs on chromosome 18 departing from the null distribution indicating that this might be a genuine association signal (Supplementary Fig. 6).

Family-based analysis. As previously mentioned, ~50% of the offspring of one of the sires of the six available half-sib families were reportedly affected with RLN, compatible with the transmission of an autosomal dominant mutation for which the stallion would be heterozygous. The stallion himself was reportedly affected (no video endoscopy available). To test this hypothesis despite the fact that DNA could not be obtained for the sire, we compared SNP allele frequencies in his nine affected and seven unaffected offspring using Fisher's exact test. The outcome of this analysis is shown in Fig. 5. Log(1/p-values) reaching 4.09 were obtained on chromosome 1 between positions 50 and 58Mb. Haplotype analysis of the corresponding region pinpointed the two supposedly paternal haplotypes of which one was transmitted to seven of the nine cases and only one of the seven controls. The nominal statistical significance of this transmission disequilibrium equals 0.02. Assuming that the genome segregates from parent to offspring as ~300 independent units (M. Georges, unpublished), this p-value is obviously not significant after correction for multiple testing.

Discussion

We herein describe the results of a GWAS conducted with the EquineSNP50 array to detect risk loci for RLN in the horse. We report two genome-wide suggestive loci, respectively on chromosomes 21 and 31. Establishing the bona fide nature of both risk loci will require confirmation in independent cohorts (Chanock *et al.* 2007), which we hope will result from the publication of our results.

The signal for both loci was clearly enhanced by performing haplotype-based analyses. This is most likely due to the fact that haplotypes have the potential to be in higher LD with the causative variants than individual SNPs, especially when using medium density SNP panels.

When performing an across-breed analysis (W+TH+TR), the signal on chromosome 21 reached genome-wide significance. This was accomplished by summing, at each marker position, the gamma statistics obtained separately in each breed. By doing so, we search for congruent evidence for association across breeds, yet without imposing any constraints on haplotype effects or sign in the different populations, i.e. we don't force the same haplotypes to be associated with RLN in the different populations. As a matter of fact, phasing and haplotype clustering was performed separately for each breed. However, we consider this results with caution as no clear signal was observed when considering the TH and TR populations alone.

Unexpectedly, in W both the chromosome 21 and 31 signals were largely driven by the enrichment of a “protective” haplotype in controls compared to cases. A priori, one would predict that the RLN defect result from one or more (mildly) deleterious mutations enriched in cases. Note that protective variants have been identified for complex diseases in human, including Crohn’s disease (Duerr *et al.* 2006; Momozawa *et al.* 2011). Relative protection conferred by the identified haplotypes was 1.77 (± SE 0.11) for loci on chr 21 and 1.91 (± SE 0.13) for the one on chr 31. These are fairly large effect when compared to relative risks typically reported for complex traits in humans (Altshuler *et al.* 2008). This could be due to their inflation as a result of the “winners curse”, or – if genuine – provide additional evidence that complex traits in domestic animal populations are influenced by larger effects which are less effectively selected against as a result of the reduced effective population size (Goddard and Hayes 2009).

Analysis of intergroup IBS metrics as well as QQ-plots of the p-values obtained when performing standard association test reveals statistically significant levels of population stratification, expected to generate spurious associations. We therefore conducted our association studies in a mixed model framework including a random polygenic background. Covariances between individual animal effects were proportionate to the genome-wide IBS computed using either SNP genotypes (GenABEL) or the hidden haplotype states (Phasebook). As expected, including these random polygenic effects had a drastic “buffering” effect on the distribution of p-values approaching the distribution expected under the null hypothesis, with the exception of the lowest p-values corresponding primarily to chromosome 21 and 31 signals. Nevertheless, the QQ-plots still revealed a trend towards p-values lower than expected under the null hypothesis. This could either be due to the fact that the polygenic effect does not completely control for stratification, or that predisposition to RLN is influenced by a very large number of loci with individually small effects, i.e. that RLN has quasi-infinitesimal architecture akin to human height (Visscher 2008). However, to avoid type I errors in declaring detection of loci influencing RLN, we conservatively applied an additional genomic control-type correction to the p-value such that the bulk of corrected p-values would follow the null distribution.

To the best of our knowledge, no associations have been reported in any other mammal between neuropathies and the orthologous regions of the two identified loci. Candidate genes that could be implicated in the pathological process of peripheral neuropathies were searched within 1 Mb of the associated SNPs on chromosomes 1, 21 and 31. The pathology of RLN is characterised by a proximal to distal loss of large myelinated fibres. Histological lesions of recurrent laryngeal nerves suggest that RLN may be a primary axonopathy: collapsed myelin sheath without axis cylinder, increased myelin sheath thickness, regenerating Schwann cell membrane clusters, axonal debris. In addition to axonal degeneration, there is also evidence of chronic myelin damages: Büngner's band, onion bulbs, variation in internodal length, myelin digestion chambers (Cahill and Goulden 1986a, b, c; Duncan *et al.* 1978). Some of these characteristics are common with several forms of Charcot-Marie-Tooth disease (CMT), which represents a group of clinically and genetically heterogeneous inherited neuropathies affecting humans. More than 30 loci and 20 causative genes are known to be associated with CMT (Barisic *et al.* 2008; Irobi *et al.* 2004). None of these genes were found in the regions of interest detected by association analyses, except for one on chromosome 1. In the homologous region on human chromosome 10, the *early growth response 2 (EGR2* also known as *KROX20)* was identified. This gene is part of a multigene family encoding zinc-finger proteins implicated in myelination of the peripheral nervous system (Warner *et al.* 1998). Allelic variants of *EGR2* have been implicated in hypomyelinating neuropathy, Charcot-Marie-Tooth disease type 1D, Dejerine-Sotas neuropathy. We sequenced the open reading frame of *EGR2* in cases and controls belonging to the family of interest, but failed to identify any DNA sequence variant of interest (results not shown).

Genes associated with CMT are involved in many different functions: neuronal structure maintenance, axonal transport, nerve signal transduction, RNA processing, housekeeping functions (Barisic *et al.* 2008; Irobi *et al.* 2004). Several genes in the identified regions may participate in these cellular functions without having previously implicated in disease pathogenesis (for example *USE1* involved in endosome-lysosome transport, *MAP1S* a microtubule associated protein, *MSS1* that may play a role in mitochondrial tRNA modification...). However, stronger association evidence is required to justify their molecular analysis.

In conclusion, our study indicates that predisposition to RLN in the horse, despite having high heritability, is unlikely to be determined by one or a small number of genes with major effect. However, this hypothesis cannot be totally excluded as the LD coverage provided by the utilized SNP panel was not optimal. Moreover, a major locus with allelic heterogeneity may also have escaped our scan despite the use of a haplotype-based method. Nevertheless, it seems more likely that RLN has a complex determinism involving a large number of loci, of which the candidates on chromosome 21 and 31 may have the largest effects. If the genuine nature of these loci is confirmed in independent cohorts, identifying the causative genes may increase our understanding of RLN pathogenesis and possible suggest novel therapeutic opportunities. Assuming that RLN is very polygenic, predictive diagnosis and selection may be more effective using a genomic selection type of approach (Georges 2007; Meuwissen *et al.* 2001). However, the latter approach would require genotyping of a much larger training cohort.

Acknowledgments

We are grateful to veterinarians for phenotypic data and sample collection (Drs Aebisher, Antys, Bayssat, Berthelot, Betsch, Boening, Bussy, Ciantar, Colin, Corde, Delarue, Foursin, Geffroy, Goupil, Lefere, Palmers, Pechayre, Picandet, Plainfossé, Scicluna, Simon, Stockwell, Tessier, Van Erck), to Julien de Curraize for his help with horses and data management, to Naima Ahariz, Nadine Cambisano, Sarah Géron and the GIGA genotranscriptomic core facility for their assistance with genotyping, to Wambo Li and Cecile Lam for their assistance with sequencing. This work was funded by grants from the National Research Agency, the Fond Eperon, the French Institute for the Horse and Equitation and the Regional Council of Basse-Normandie.

Table 1: Cohorts characteristics.

A total of 531 horses from four breeds (Warmbloods, Trotters, Thoroughbreds and Draft horses) were genotyped.

	Cases	Controls	Sires	Dams	TOTAL
Warmbloods	196	188	17	17	418
Trotters	20	18	22	5	65
Thoroughbreds	14	16	3	4	37
Draft horses	4	6	1	0	11
TOTAL	234	228	43	26	531

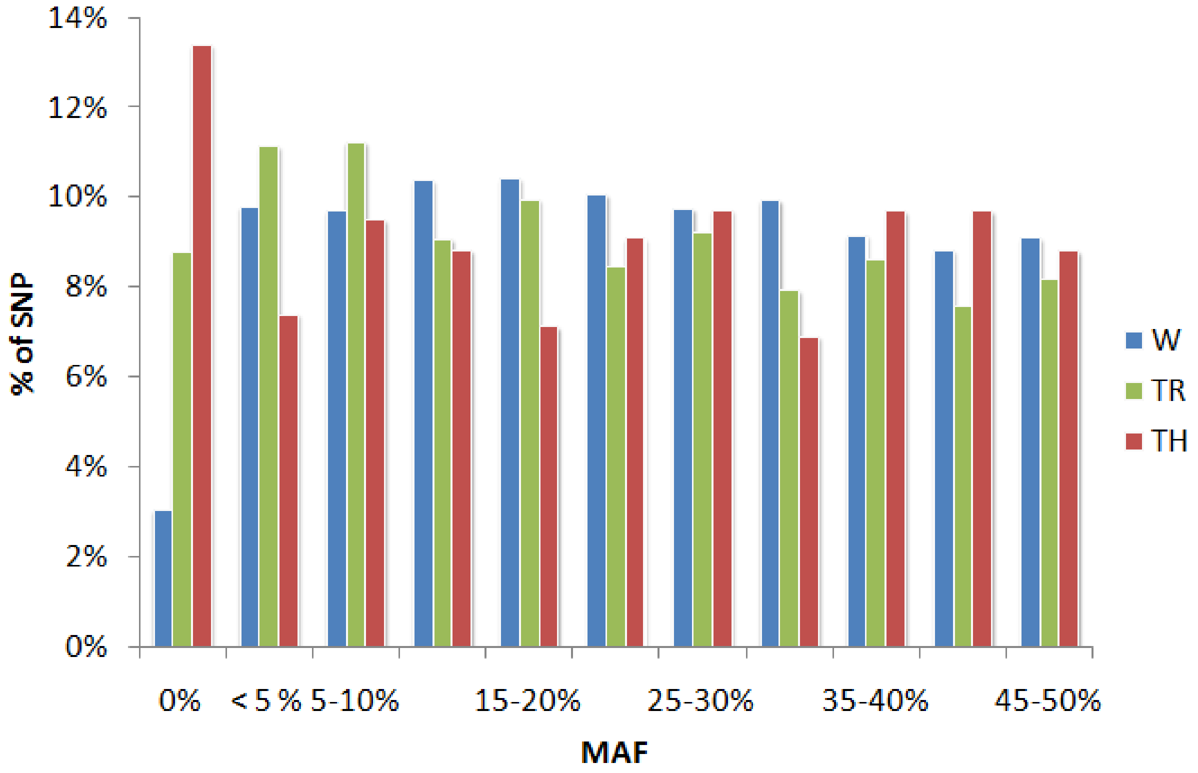


Figure 1: Frequency distributions of minor allele frequencies (MAF) calculated for all markers available on the Equine SNP50 array in W (n=418), TR (n=65) and TH (n=37).

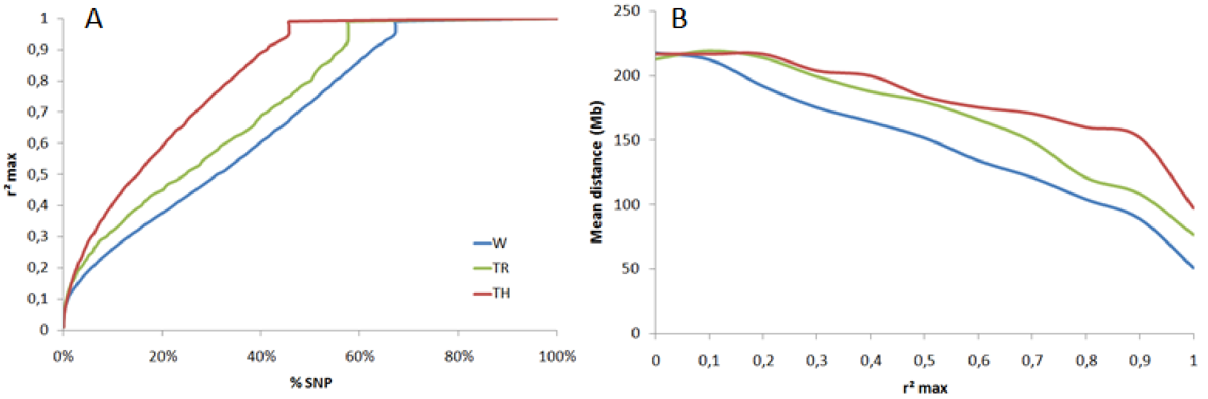


Figure 2: A. Cumulative frequency distributions of highest r^2 values for each of the 47,046 SNPs used in GWAS. **B.** Mean distances (Mb) between SNPs of highest r^2values. Calculations were performed with the same sample size (n=37) in the three breeds.

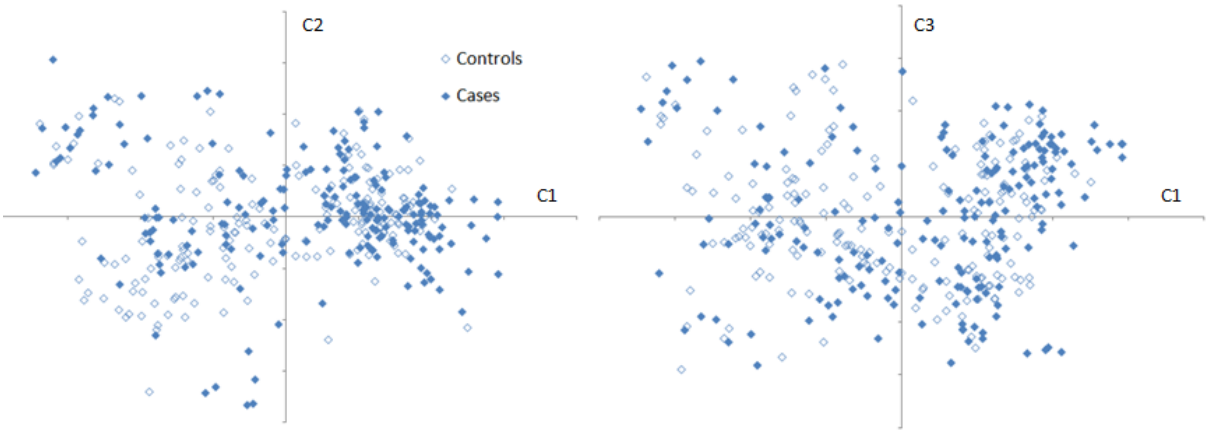


Figure 3: First four principal components of a Multidimensional Scaling (MDS) analysis (based on pairwise identity-by-state (IBS) distances) of W sorted in cases and controls.

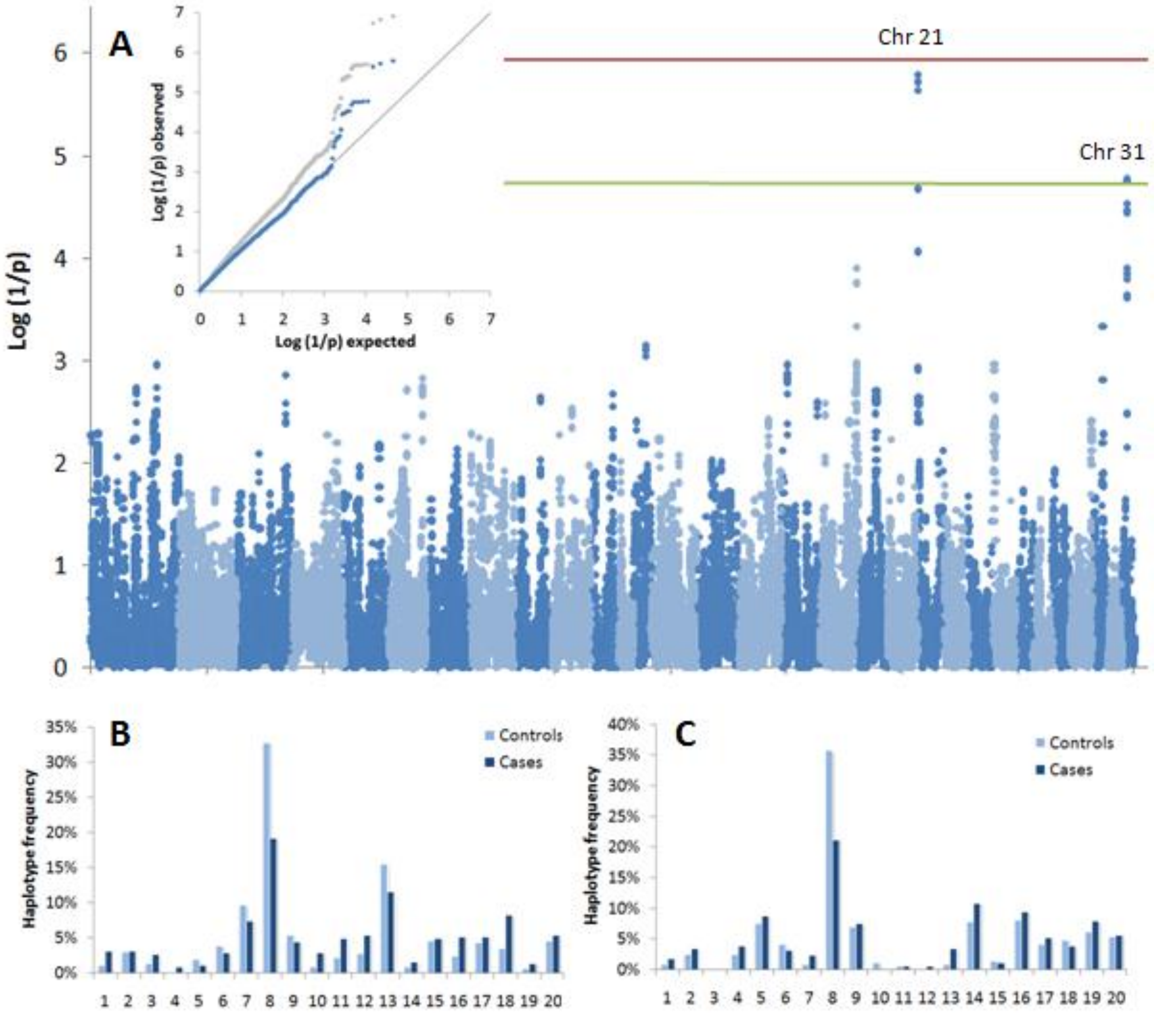


Figure 4: (A) Results of a GWAS for RLN. Dots marker the nominal significance (log(1/p)) obtained at each SNP positions using an haplotype-based (20 Hidden Haplotype States) analysis in W correcting for stratification by means of a random polygenic effect (pair-wise covariance proportionate to SNP-based genome-wide IBS), implemented with Phasebook. Alternating blue colors mark the chromosome limits. The green and red horizontal lines correspond to the genome-wide (Bonferroni corrected) suggestive and significant thresholds, respectively. The inset shows the corresponding QQ plot, before (grey) and after (blue) application of genomic control. (B) Frequency of the 20 Hidden Haplotype States in cases (dark blue) and controls (light blue) for map position 2,203,650 on chr 21. (C) Same as (B) for map position 9,863,278 on chr 31.

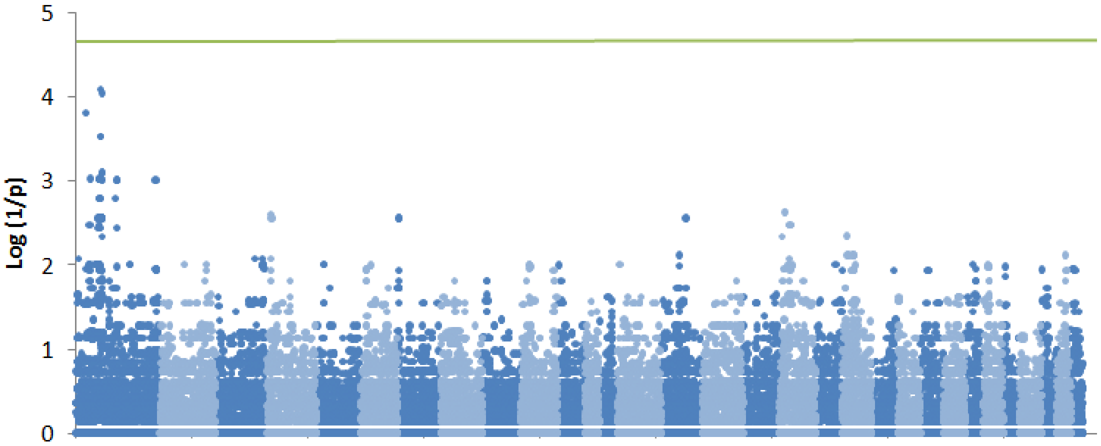
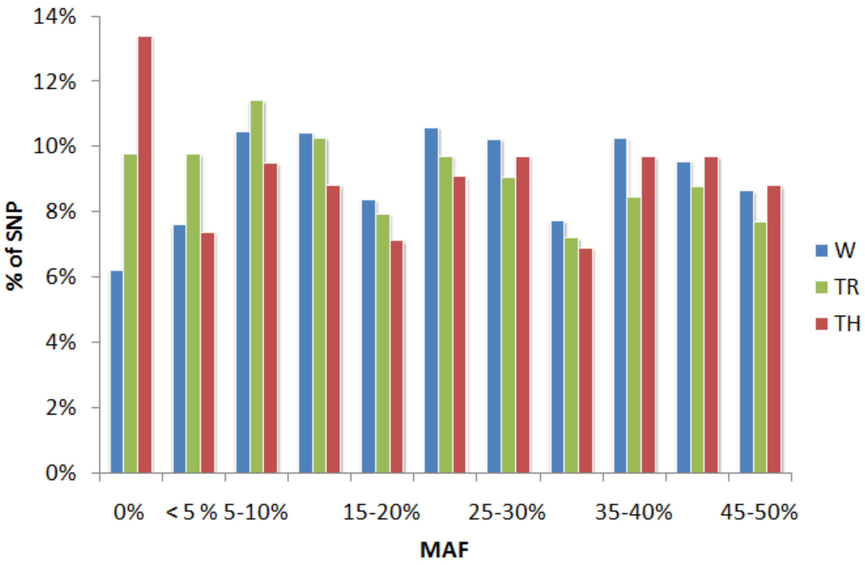
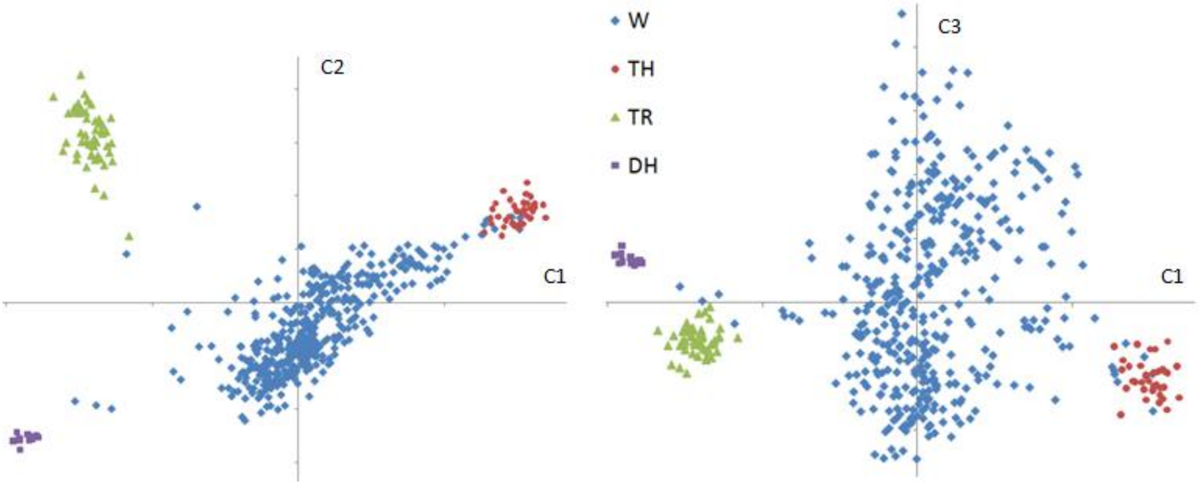


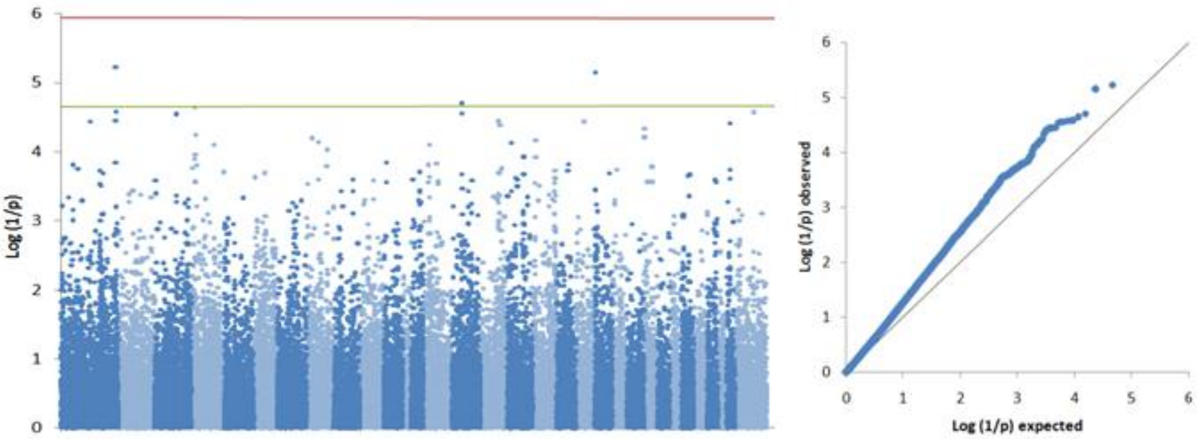
Figure 5: Log(1/p) values (computed using Fisher's exact test) of the differences in SNP allele frequencies between nine affected and seven non-affected half-sibs, offspring of an affected stallion with 50% RLN incidence in his descendants. Log(1/p-values) > 3 (reaching 4.09) were obtained on chromosome 1 between positions 50 and 58Mb. The green horizontal line corresponds to the genome-wide (Bonferroni-corrected) suggestive threshold.



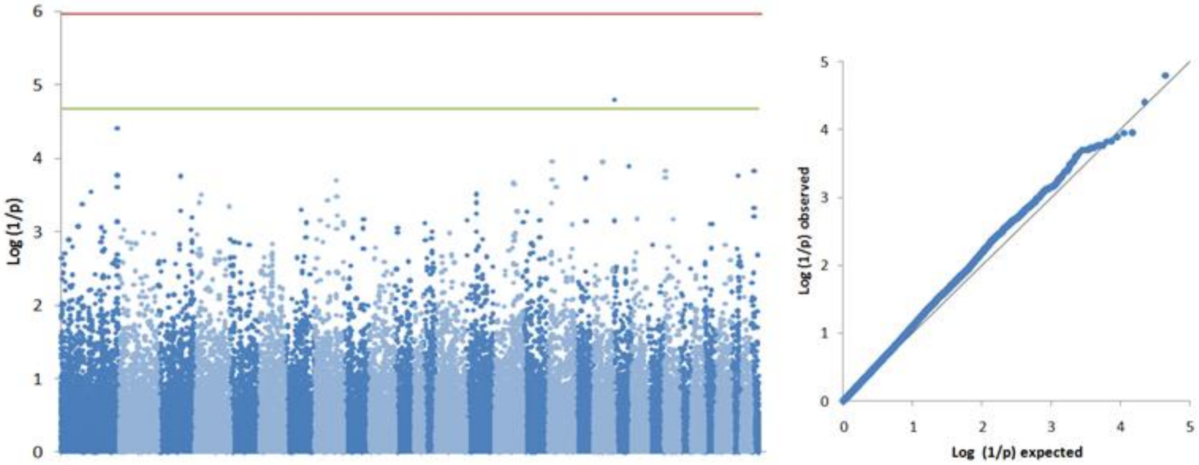
Supplementary figure 1: Frequency distributions of minor allele frequencies (MAF) computed in samples of the same size (n=37) in three breeds.



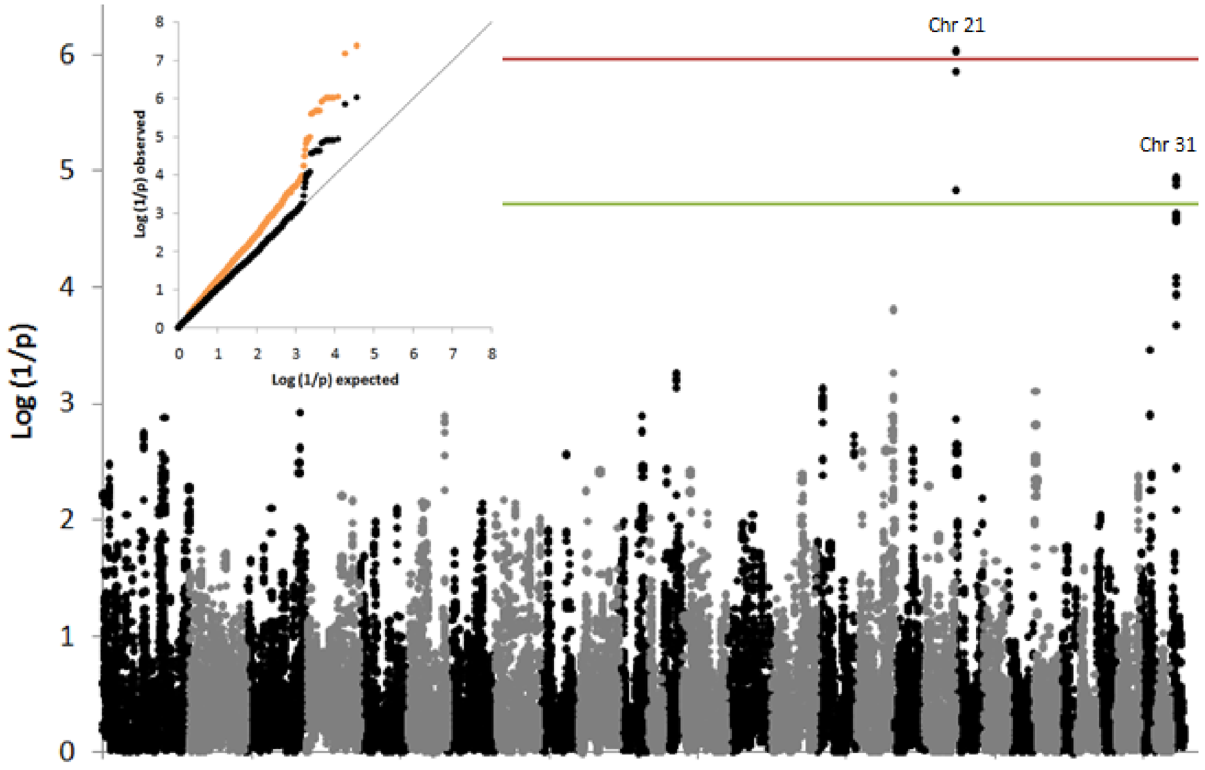
Supplementary figure 2: First four principal components of Multidimensional Scaling (MDS) analysis of all available horses on the basis of pairwise identity-by-state (IBS) distances.



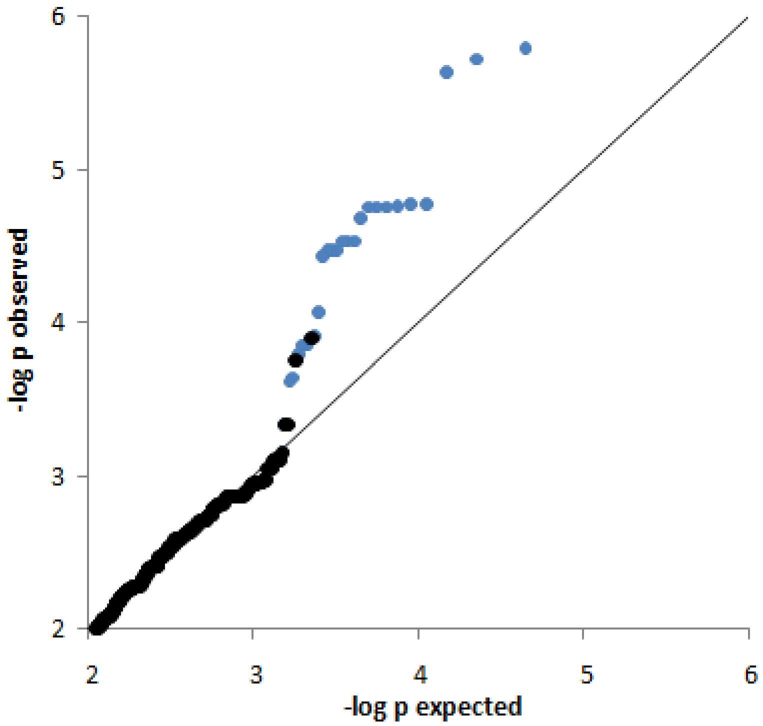
Supplementary figure 3: Log quantile-quantile (QQ) and Manhattan plots of the p-values obtained using standard single-SNP association studies implemented with PLINK in W. Deviation across the entire distribution towards high log(1/p-values) suggests stratification (λ_{GC}=1.36).



Supplementary figure 4: Log quantile-quantile (QQ) and Manhattan plots of the p-values obtained using single-SNP association studies in W corrected for stratification by modeling a random polygenic background effect in a REML type analysis implemented with GenABEL (estimated relatedness of individuals based on genome-wide SNP data). The correction improved the distribution of p-values which lie close to the expected diagonal.



Supplementary figure 5: Log quantile-quantile (QQ) and Manhattan plots of the p-values obtained using an across-breeds (W+TR+TH) haplotype-based association studies correcting for stratification by means of a random polygenic effect (pair-wise covariance proportionate to SNP-based genome-wide IBS), implemented with Phasebook. Alternating black and grey colors mark the chromosome limits. The green and red horizontal lines correspond to the genome-wide (Bonferroni corrected) suggestive and significant thresholds, respectively. The inset shows the corresponding QQ plot, before (orange) and after (dark) application of genomic control.



Supplementary Figure 6: Log quantile-quantile (QQ) plot before (blue dots) and after (black dots) removing p-values corresponding to marker positions within 2 Mb from the most strongly associated SNPs on chromosomes 21 and 31.

A Splice Site Variant in the Bovine RNF11 Gene Compromises Growth and Regulation of the Inflammatory Response

A. Sartelet, T. Druet, C. Michaux, C. Fasquelle, S. Géron, N. Tamma, Z. Zhang, W. Coppieters, M. Georges and C. Charlier

Abstract

We report association mapping of a locus on bovine chromosome 3 that underlies a Mendelian form of stunted growth in Belgian Blue Cattle (BBC). By resequencing positional candidates, we identify the causative c124-2A>G splice variant in intron 1 of the *RNF11* gene, for which all affected animals are homozygous. We make the remarkable observation that 26% of healthy Belgian Blue animals carry the corresponding variant. We demonstrate in a prospective study design that approximately one third of homozygous mutants die prematurely with major inflammatory lesions, hence explaining the rarity of growth-stunted animals despite the high frequency of carriers. We provide preliminary evidence that heterozygous advantage for an as of yet unidentified phenotype may have caused a selective sweep accounting for the high frequency of the *RNF11* c124-2A>G mutation in Belgian Blue Cattle.

PLoS Genet 8 (2012) e1002581

Author Summary

Recessive defects in livestock are common, and this is considered to result from the contraction of the effective population size that accompanies intense selection for desired traits, especially when relying heavily on artificial insemination (as males may concomitantly have a very large number of offspring). The costs of recessive defects are assumed to correspond to the loss of the affected animals. By performing a molecular genetic analysis of stunted growth in Belgian Blue Cattle (BBC), we highlight (i) that the economic impact of recessive defects may outweigh the only loss of affected animals and (ii) that some genetic defects are common for reasons other than inbreeding. We first demonstrate that a splice site variant in the RING finger protein 11 (*RNF11*) gene accounts for ∼40% of cases of stunted growth in BBC. We then show that a large proportion of animals that are homozygous for the corresponding *RNF11* mutation die at a young age due to compromised resistance to pathogens. We finally demonstrate that carriers of the mutation benefit from a selective advantage of unidentified origin that accounts for its high frequency in BBC.

Introduction

Growth is one of the economically most important phenotypes in livestock production. While genetic variants with large effects on stature account for part of the between-breed variation (Karim *et al.* 2011), within-breed variation is likely to be highly multifactorial and polygenic. Accordingly, quantitative trait loci (QTL) influencing growth are reported on all autosomes in the cattle QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index).

The BBC breed is a beef breed that is famous for its "double-muscling" phenotype caused in part by a disruptive 11-bp deletion in the myostatin (*MSTN*) gene (Grobet *et al.* 1997). As in other breeds, growth performances are paramount in BBC as they control duration of the fattening period and final carcass weight, hence directly determining profit.

In recent years, an increasing number of young animals with growth retardation as primary symptoms were reported to our heredosurveillance platform. We established this platform in 2005 to rapidly detect genetic defects emerging in the BBC, identify the culprit genes and mutations, and develop diagnostic tests to limit their negative impact (Charlier *et al.* 2008). Animals with growth retardation underwent a standard protocol including a genome-wide association study (GWAS) to identify putative causative loci. We herein report the mapping of a locus accounting for ∼40% of growth-retardation cases, and identify the causative loss-of-function mutation in the RING finger protein 11 (*RNF11*) gene. Moreover, we perform a prospective study that indicates that as much as one third of homozygous mutants die from infection before six months of age. We finally present evidence that carriers of the mutation might benefit from a selective advantage that may account for its unexpectedly high frequency (~13%) in the BBC population.

Results

A major growth-stunting locus maps to BTA3

Between 2008 and 2011, we collected blood samples and epidemiological data from 147 BBC individuals, aged between 3 months and 3 years old, with pronounced (~15% reduction in stature when compared to contemporaries) yet proportionate growth retardation as primary distinctive feature. We initially genotyped 33 of these with a custom-designed 50 K medium-density bovine SNP array [3]. None of these animals would be homozygous or compound heterozygote for the previously identified c.2904-2905delAG (Fasquelle *et al.* 2009) and c.1906T>C (Sartelet *et al.* 2011) MRC2 mutations causing Crooked Tail Syndrome and known to affect stature. Using the genotypes of the corresponding SNPs (yet obtained with a distinct, high-density bovine SNP array) from 275 healthy sires as control, we performed a GWAS using an approach based on hidden haplotype states with a generalized mixed model accounting for stratification (Zhang *et al.* 2012). A genome-wide significant signal was obtained on BTA3 driven by haplotype state 17, observed at a frequency of 52% in cases versus 12% in controls (Figure 1A). Fourteen of the 33 cases (42%) were homozygous for the corresponding haplotype, causing a significant deviation from Hardy-Weinberg expectations in cases (expected: 27%, p<0.002), hence suggesting recessivity.

Retrospective phenotypic analysis of the 14 homozygotes revealed shared features: proportionate growth retardation appearing around 5–6 months of age (not observed at birth), normal muscular development, close forehand, long and thin neck, hairy, long and thin head (Figure 2). Pedigree analysis indicated that the 14 individuals traced back to *Galopeur des Hayons* (a once popular BBC sire) on sire and dam side.

A splice site mutation in the RNF11 gene is the likely causative mutation

Direct examination of the SNP genotypes of the 14 cases homozygous for hidden state 17 revealed a 3.3 Mb (100,727,788–104,017,608 - Btau 4.0) segment of autozygosity (Figure 1B). It encompassed 19 annotated genes of which none was an obvious candidate (Figure 1C). We thus undertook the systematic re-sequencing of all open reading frames (ORF) and intron-exon boundaries. During this process (and after completion of 14/19 genes), we identified an A to G transition (c124-2A>G) mutating the intron 1 acceptor splice site of the *RNF11* gene (Figure 1D). *RNF11* encodes a highly conserved, ubiquitously expressed protein with 154 amino-acids (Azmi and Seth 2005), recently recognized as a subunit of the A20 ubiquitin-editing complex regulating NF-κβ signaling (Shembade *et al.* 2009). We developed a 5′-exonuclease assay and genotyped (i) the case-control cohort used for GWAS (33 cases, 275 controls), (ii) a diversity panel encompassing 141 animals from eleven breeds other than BBC, (iii) 549 additional normal adult BBC animals, and (iv) *Galopeur des Hayons.* The c124-2A>G variant appeared in near perfect linkage disequilibrium (D′=1; r2=0.984) with haplotype state 17 in the case-control cohort. It was not present in non-BBC animals. It had an allelic frequency of 13% amongst the 824 genotyped healthy adult BBC animals, yet without a single animal being homozygous *GG* ($p<0.01$ under Hardy-Weinberg equilibrium). *Galopeur* was indeed confirmed to be carrier of the c124-2A>G mutation.

The effect of the c124-2A>G mutation on *RNF11* transcripts was examined by RT-PCR using RNA extracted from skeletal muscle, spleen, mesenteric lymph node, thymus, lung, trachea of one *GG* and one *AA* animal. Using two primers located respectively in exon 1 and 3 and RNA from wild-type *AA* animals, we obtained a unique 360-bp RT-PCR product in all examined tissues, and showed by sequencing that it encompassed the expected exon 2 sequence (data not shown). The same experiment performed with RNA from a homozygous mutant *GG* animal yielded (i) a major product of ~190 bp, and (ii) a minor product of ~360 bp (Figure 3A). The major product was shown by sequencing to correspond to a transcript skipping exon 2. The minor product missed the first seven base pairs of exon 2, and resulted from the activation of a cryptic splice site in exon 2. RT-PCR conducted with primers located respectively in exon 1 and 2 confirmed the existence of transcripts containing exon 2 in homozygous mutants (Figure 3B). Both forms are expected to cause a frameshift, appending 29 (major product) and 14 (minor product) illegitimate residues to a severely truncated (41/154 amino-acids) *RNF11* protein missing the ubiquitin interaction and RING-finger domains. The transcript corresponding to the minor form is expected to undergo non-sense mediated RNA decay (NMRD) (Chang & Wilkinson 2007), due to the occurrence of a stop codon in exon 2 of three. NMRD is not expected to affect the transcript corresponding to the major form as the corresponding open reading frame terminates in exon 3 of three. We compared the levels of *RNF11* transcript in mesenteric lymph node and spleen of a wild-type *AA* and a mutant *GG* animals, using quantitative RT-PCR with primer sets targeting the second (outside of the 7-bp deletion) and third *RNF11* exons, respectively, as well as three internal control genes. In spleen, we observed a 1.1-fold reduction ($p=0.4$) in the amount of exon 3 containing transcripts, and a 11-fold reduction ($p<0.005$) in exon 2 containing transcripts. Assuming NMRD of the minor but not of the major product, this allows us to estimate (i) that ~80% of the *RNF11* pre-mRNAs skip exon 2, while ~20% use the exon 2 cryptic splice site, and (ii) that 55% of exon 2 retaining transcripts are being degraded by NMRD. The same analysis conducted in lymph node reveals a ~2-fold reduction ($p<0.05$) in exon 3 containing transcripts, and ~37-fold reduction ($p<0.0005$) in exon 2 containing transcripts, corresponding to (i) ~44% of *RNF11* pre-mRNAs skipping exon 2 and ~56% using the exon 2 cryptic splice site, and (ii) ~95% of exon 2 retaining transcripts being degraded by NMRD (Supporting Information S1).

Taken together, our findings strongly support the causality of the c124A>G *RNF11* mutation in determining stunted growth in homozygous *GG* animals.

Increased juvenile mortality accounts for incongruent carrier frequency and disease incidence

The ~26% carrier frequency amongst healthy individuals is incompatible with the number of reported cases of stunted growth. As an example, ~6% of offspring of known carrier bulls should be affected, and such high figures were never recorded. We reasoned that this lower than expected incidence of cases might reflect elimination of mutant animals either before or after birth. Embryonic mortality of homozygous mutant fetuses has been reported for deficiency in uridine monophosphate synthetase (DUMPS) (Shanks and Robinson 1989), Complex Vertebral Malformation (CVM) (Thomsen *et al.* 2006; Malher and Philipot 2006) and Brachyspina (BS) (Charlier *et al.* submitted for publication).

To test these hypotheses we first examined field data and tested the effect of sire carrier status on (i) “non return (in oestrus) rate” of inseminated cows between 28 and 280 days after insemination, and (ii) rate of mortality, morbidity and culling of offspring between birth and 14 months of age (Hanset and Boonen 1994). Non-return rates tended to be slightly decreased when cows were inseminated with semen from carrier sires (i.e. reproductive failure increased), but the effect was not significant ($p=0.66$). Mortality, morbidity and culling tended to be increased in offspring of carrier sires, but this effect was not significant either ($p=0.89$) (Supporting Information S1).

As analysis of field data did not provide conclusive results, we performed a prospective study. We identified 105 carrier dams in 22 farms that were pregnant following insemination with semen from known carrier sires. We followed the ensuing 105 calves up to 12 months after birth. The responsible veterinarian (AS) and the breeders were not aware of the calves' *RNF11* genotype until completion of the study. Genotypic proportions at birth did not deviate significantly from Mendelian expectations (AA: 26

(=24.8%); AG: 56 (=53.3%); GG: 23 (=21.9%); p=0.72). All calves looked normal, and there was no significant effect of *RNF11* genotype on weight or height at birth. However, one year after birth, 10 calves had died and eight had been culled for health-related reasons. Strikingly, all but one of these were homozygous mutant GG, while one was *AG* (p<0.0005) (Figure 4A). While the *AG* animal was euthanized with a limb fracture, the nine deceased *GG* animals died with severe inflammation (primarily pneumonia) (Supporting Information S1). The c124-2A>G genotype had a highly significant (p≤0.001) effect on post-natal growth. Indeed, all surviving *GG* animals exhibiting stunted development after 6 months (Figure 4B). A *contrario*, the growth pattern of *AG* and *AA* animals was indistinguishable.

Taken together, our data indicate that as much as one third of homozygous *GG* calves die with major inflammation, while all remaining calves exhibit stunted growth and are hence systematically culled prematurely.

Selective advantage of heterozygotes may underlie the high carrier incidence

The 26% carrier frequency amongst healthy BBC animals is puzzling given the observed purifying selection against *GG* animals. This suggests that heterozygotes might benefit from a selective advantage that would maintain the G allele at high frequency in the population. Such balanced polymorphism has been demonstrated for MRC2 loss-of-function mutations causing Crooked Tail Syndrome in homozygotes, yet increased muscle mass in carriers (Fasquelle *et al.* 2009; Sartelet *et al.* 2011).

To test this hypothesis, we first used field data and examined the effect of *RNF11* c124-2A>G sire carrier status on own and progeny performances for recorded traits including size, muscularity, type and general appearance (Hanset and Boonen 1994). We obtained conflicting results: carrier status appeared to negatively affect the perceived quality of sire, yet improve the quality of its offspring (Supporting Information S1).

As an alternative approach to test for a putative selective advantage benefitting carriers, we evaluated whether the incidence of carriers amongst active AI sires was compatible with Mendelian (0.5:0.5) inheritance of a neutral mutation from the founder bull *Galopeur*. Assuming that the c124-2A>G mutation improves zootechnical performances in heterozygotes, carriers should be over-represented amongst AI sires related to *Galopeur*. Two hundred and six of the 262 BBC AI sires born between 2003 and 2007 were related to *Galopeur* and 58 (=28%) of these proved to carry the *RNF11* c124-2A>G mutation. Using gene dropping in the known genealogies, we computed the probability that 58 or more descendants would be carrier in the absence of selection (no systematic transmission distortion). This probability was 0.0002, 0.0006 and 0.01 assuming a frequency of 0, 0.01 and 0.05 for the c124-2A>G mutation outside the *Galopeur* lineage (Figure 5A). These results suggest that the c124-2A>G mutation indeed underwent a recent selective sweep in the BBC population, although the phenotype that is being selected remains unclear. That 58/206 descendents of *Galopeur* carry the c124-2A>G mutation is best explained by assuming that the mutation has ~10% excess probability (i.e. 60%) to be transmitted by a carrier parent to an AI sire or one of its ancestors (Figure 5B).

Homozygosity at the *RNF11* c124-2A>G mutation accounted for 14 of the first 33 analyzed cases (i.e. 42%), raising the question of what caused stunted growth in the others. To address this, we genotyped the remaining 114 cases for the c124-2A>G mutation. In agreement with genotypic proportions in the first 33 cases, 47/114 (41%) were homozygous and 23/114 (20%) heterozygous. Therefore, carrier frequency amongst non c124-2A>G homozygous cases was 34% (29/86), which does not differ significantly (p=0.10) from the frequency of c124-2A>G carriers in the control cohort (211/829=26%). This suggests that the c124-2A>G mutation is the only common *RNF11* mutation involved in stunted growth in BBC.

To identify putative other loci involved in stunted growth, we genotyped the remaining 67 non c124-2A>G homozygous cases with a medium density 50 K SNP array (Illumina), and rescanned the genome as described before using only non c124-2A>G homozygous cases (86) and the same control cohort (275). As expected, there was no evidence for a residual effect of the *RNF11* locus. Neither was there any genome-wide significant evidence for other loci on any one of the 29 autosomes (Supporting Information S1).

Discussion

We herein demonstrate that a loss-of-function mutation in the *RNF11* gene affects normal growth and disease resistance in calves. This is the first report of a phenotypic effect associated with *RNF11* mutations in any organism, including human and mouse (Shembade *et al.* 2009).

We postulate that the increased disease susceptibility of homozygous c124-2A>G calves is related to the demonstrated role of *RNF11* in feedback down-regulation of NF-κB by the A20 complex (Shembade *et al.* 2009). Indeed, the nine c124-2A>G homozygous calves that underwent necropsy were affected by extensive inflammation of the respiratory tract (eight) or by polyarthritis (one). Of note, A20 knock-out mice die prematurely from multi-organ inflammation (Lee *et al.* 2000). The fact that only ~1/3 of homozygous mutant calves died prematurely is compatible with a defect in the control or resolution of inflammation. External factors, including pathogens, may trigger an intendedly salutary innate and/or adaptive response, that evolves in pathogenic non-resolving inflammation (Nathan & Ding 2010).

The effects on growth may be secondary to hidden episodes of uncontrolled inflammation, as proposed for A20- and ITCH-deficient mice and human (Lee *et al.* 2000; Schembade *et al.* 2008; Lohr *et al.* 2010). However, the fact that several of the surviving homozygous c124-2A>G calves appeared perfectly healthy upon clinical examination, suggest that growth retardation might be directly related to alternative functions of *RNF11* as modulator of growth factor receptor signaling (particularly TGF-β and EGFR signaling) and transcriptional regulation (Azmi & Seth 2005). It is also noteworthy, that *RNF11* has been found to be highly expressed in bone cells during osteogenesis (Gao *et al.* 2005).

Calf mortality is an economically important trait. It is generally considered highly complex and multifactorial, and its heritability is always very low. It is thus difficult to improve using conventional selection strategies. We herein demonstrate that genomic approaches may help dissect such complex phenotypes in sub-components including some with simple Mendelian determinism amenable to effective “marker assisted selection”. The situation uncovered in this work is reminiscent of bovine leukocyte deficiency (BLAD) in Holstein-Friesian (Shuster *et al.* 1992), an immune deficiency resulting from CD18 deficiency and causing increased susceptibility to infection in young calves (Nagahata 2004).

We provide suggestive evidence that the high incidence of the *RNF11* c124-2A>G mutation in BBC is not only due to drift, but may be due to the superiority of heterozygotes for unidentified selection criteria. Such a situation would be reminiscent of previously described pleiotropic effects on conformation of mutations in the gene encoding the calcium release channel (CRC) in pigs (causing malignant hyperthermia and porcine stress syndrome in homozygotes) (Fujii *et al.* 1991) and in the MRC2 gene in cattle (causing Crooked Tail Syndrome in homozygotes) (Fasquelle *et al.* 2009; Sartelet *et al.* 2011). These examples illustrate some of the issues resulting from the selection of animals with extreme performances.

Materials and Methods

Ethics statement

Blood samples were collected from sires, cows and calves, by trained veterinarians following standard procedures and relevant national guidelines.

Genotyping

Genomic DNA of cases was extracted from 350 µl of blood using the MagAttract DNA Blood Midi M48 Kit (Qiagen). Genomic DNA of controls was extracted from frozen semen using the MagAttract Mini M48 Kit (Qiagen). The 33 cases of the initial genome scan were genotyped using a custom-made 50 K SNP array (Charlier *et al.* 2008). The 67 cases of the second scan (excluding *RNF11* c124-2A>G homozygotes) were genotyped with the BovineSNP50 v2 DNA analysis BeadChip (Illumina). The 275 control sires were genotyped with the BovineHD BeadChip (Illumina). SNP genotyping was conducted using standard procedures at the GIGA genomics core facility.

Genome-wide haplotype-based association studies

Phasing of the SNP genotypes and assignment of the haplotypes to a predetermined number of hidden haplotype states was conducted with PHASEBOOK (Druet & Georges 2010). Hidden haplotype state-based association analysis was conducted using GLASCOW (Zhang *et al.* 2012). GLASCOW uses generalized linear models and fits a random hidden haplotype state effect as well as a random polygenic effect to correct for population stratification. Locus-specific p-values were determined from 1,000 permutations assuming a gamma distribution of the used score test (Zhang *et al.* 2012). We applied a conservative Bonferonni correction assuming 50,000 independent tests to determine the genome-wide significance thresholds.

Mutation scanning

Coding exons of positional candidate genes were amplified from genomic DNA of a homozygous case and a healthy control using standard procedures. The primers used for the *RNF11* gene are listed in the Supporting Information S1. PCR products were directly sequenced using the Big Dye terminator cycle sequencing kit (Applied Biosystem, Foster City, CA). Electrophoresis of purified sequencing reactions was performed on an ABI PRISM 3730 DNA analyzer (PE Applied Biosystems, Forster City, CA). Multiple sequence traces from affected and wild-type animals were aligned and compared using the Phred/Phrap/Consed package (www.genome.washington.edu).

5′ exonuclease diagnostic assay of the c124-2A>G RNF11 mutation

A 5′exonuclease assay was developed to genotype the c124-2A>G *RNF11* mutation, using 5′-AGG AAG AAA CAA AAG GAA AAC ATT ACC TAG A-3′ and 5′-TGT TGG ATG ATA GAC CGG AAC TG-3′ as PCR primers, and 5′-ACT TGT TCC TAA ATT TT-3′ (wild type A allele) and 5′-TTG TTC CCA AAT

TTT-3′ (mutant G allele) as probes (Taqman, Applied Biosystems, Fosters City, CA). Reactions were carried out on an ABI7900HT instrument (Applied Biosystems, Fosters City, CA) using standard procedures.

RT–PCR and cDNA sequencing

Total RNA from *RNF11* c124-2A>G *AA* and *GG* animals was extracted from lung, lymph nodes, spleen, skeletal muscle, thymus and trachea using standard procedures (Trizol, Invitrogen). After DNase-treatment (Turbo DNA-free, Ambion), cDNA was synthesized using the SuperScript III First-Strand Synthesis SuperMix (Invitrogen). A cDNA segment was amplified using two *RNF11* specific primers sets: one encompassing exon 2 with primers located in exon 1 and exon 3 (E1–E3) and one encompassing the exon1-exon2 boundary (E1–E2) (Supporting Information S1). PCR products were separated by electrophoresis on a 2% agarose gel containing 0.0001% of SYBR Safe DNA gel stain (Invitrogen) at 100 volts during 40 min and size was evaluated with SmartLadder 200 lanes (Eurogentec). The PCR products were directly sequenced as described above.

Real-time quantitative RT–PCR

Total RNA from *RNF11* c124-2A>G *AA* and *GG* animals was extracted from lymph node, spleen as described above. After DNase-treatment (Turbo DNA-free, Ambion), 500 ng of total RNA was reverse transcribed in a final volume of 20 µl using SuperScript III First-Strand Synthesis SuperMix (Invitrogen). PCR reactions were performed in a final volume of 10 µl containing 4 µl of 5-fold diluted cDNA (corresponding to 100 ng of starting total RNA), 1X of ABsolute Blue QPCR SYBRE Green ROX Mix 2X (Thermo Fischer Scientific), 0.3 µM forward and reverse primers and nuclease free water. PCR reactions were performed on an ABI7900HT instrument (Applied Biosystems, Forster City, CA) under the following conditions: 10 min at 95 ℃ followed by 40 cycles at 95 ℃ for 15 sec and 60 ℃ for 1 min. Two primers sets were used to test *RNF11* expression and three genes were included as candidate endogenous controls: (1) Beta-Actin (ACTB), (2) Ribosomal Protein Large P0 (RPLP0), (3) Tyr-3- & Trp-5-Monooxygenase Activation Protein Zeta (YWHAZ). The corresponding primer sequences are given in Supporting Information S1. A standard curve with a five point two-fold dilution series (total RNA=100, 200, 400, 800 and 1600 ng from lymph node and spleen from a *AA* wild-type individual) for each *RNF11* primer set was used to determine the amplification efficiency. All sample/gene combinations were analyzed in triplicate. ACTB and YWHAZ genes were selected as endogenous controls using geNorm

(Vandesompele *et al.* 2002). Normalized relative *RNF11* expression, for exon 2- and exon 3-containing transcripts, in the lymph node and the spleen of a wild-type *AA* and a mutant *GG* animal accounting for primer efficiency were computed using the qbaseplus software package (Biogazelle) (Hellemans *et al.* 2007).

Estimating the effect of carrier status for the RNF11 c124-2A>G mutation on agronomically important traits measured in the field

The effect of the sire's *RNF11* c124-2A>G genotype on non-return rate (NRR) of its mates was estimated using a mixed model including sire's *RNF11* genotype (fixed), year and month at insemination (fixed), mate's herd (random), individual animal effect of the offspring (random) and error. NRR are computed from the AI information collected by inseminators working with the Association Wallonne de l'Elevage (AWE; http://www.awenet.be/) at seven time-points after AI. The analysis was performed on 479,674 cows mated to 340 AI sires.

The effect of the sire's *RNF11* c124-2A>G genotype on the rate of mortality, morbidity and culling of its offspring was estimated using a mixed model including sire's *RNF11* genotype (fixed), calf's gender (fixed), year and month of calf's birth (fixed), mate's parity (fixed), calf's in utero position (fixed; forward or backward), calf's herd (random), individual animal effect of the calf (random), and error. The corresponding phenotypes are collected by AWE technicians visiting farms, for (i) newborn calves, and (ii) calves having reached the age of 14 months since last visit. The number of records for newborn offspring was 317,350 from 332 AI sires, and for 14 month-old offspring was 126,098 from 288 AI sires.

The effect of the sire's *RNF11* c124-2A>G genotype on its own zootechnical performances was estimated using a mixed model including sire's *RNF11* genotype (fixed), sire's MRC2 genotype (fixed) (Fasquelle *et al.* 2009; Sartelet *et al.* 2011), year and month at scoring (fixed), sire's body condition at scoring (fixed), sire's age at scoring (quadratic regression), individual animal effect for the sires (random) and error (Lynch and Wash 1997). Zootechnical performances of AI sires are recorded between 15 and 56 months of age as 22 linear scores (0–50 score) that are summarized as indexes evaluating size, muscularity, meaty type and general appearance (Hanset and Boonen 1994). Three hundred and eleven sires were used in this analysis.

The effect of the sire's *RNF11* c124-2A>G genotype on the zootechnical performances if its offspring was estimated using a mixed model including sire's *RNF11* genotype (fixed), sire's MRC2 genotype (fixed) [(Fasquelle *et al.* 2009; Sartelet *et al.* 2011), offspring's gender (fixed), year and month at scoring (fixed), offspring's body condition at scoring (fixed), offspring's age at scoring (quadratic regression), offspring's herd (random), individual animal effect for the offspring (random) and error (Lynch and Walsh 1997). The first data set corresponded to the same five global scores (cfr. sire's own performances) measured on 92,475 36-month-old daughters of 306 sires by AWE technicians. The second data set corresponded to weight (Kg), size (cm) and conformation (1–9 score) measured on 95,045 14-month-old offspring of 315 sires.

Covariances between random individual animal effects were assumed to be proportionate to twice the kinship coefficient computed from known genealogies. Variance components and fixed effects were computed using MTDFREML (Boldman *et al.* 1995).

Acknowledgments

We are grateful to the GIGA Genomics Platform for their technical assistance, to the breeders for their collaboration, and to the Walloon Breeding Association (AWE) and the Belgian Blue Beef Herd-Book (HBBBB) for pedigree, performance, and progeny test data.

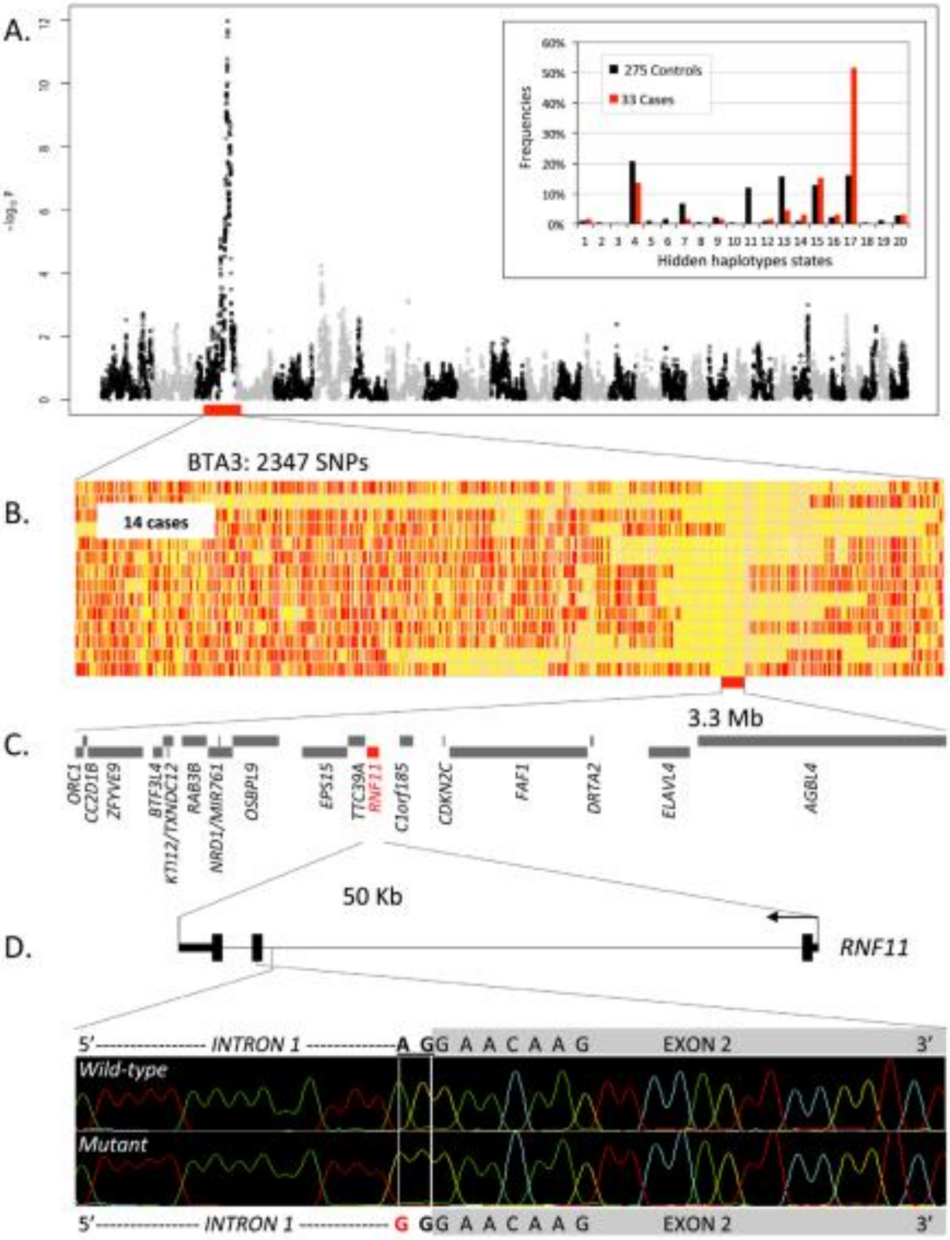


Figure 1. Genome-wide haplotype-based association mapping of a growth stunting locus on BTA 3. (A) Manhattan plot for the haplotype based genome-wide association study for stunted growth using a model with 20 ancestral haplotypes. Alternating colors (black and grey dots)mark the limits between autosomes. Inset: frequency of the 20 hidden haplotype states in the 33 cases (red) and the 275 controls (black) at position BTA3:103,391,968 bp. (B) Genotypes of the 14 cases homozygous for hidden haplotype state 17 for 2,347 BTA3 SNPs. Homozygous genotypes are shown in orange or yellow and heterozygous genotypes in red. The limit of the homozygous haplotype shared by the 14 cases is highlighted in red. (C) Gene content of the 3.3 Mb shared interval (19 genes). (D) *RNF11* gene model, and representation of the *RNF11* c124-2A>G splice site variant.

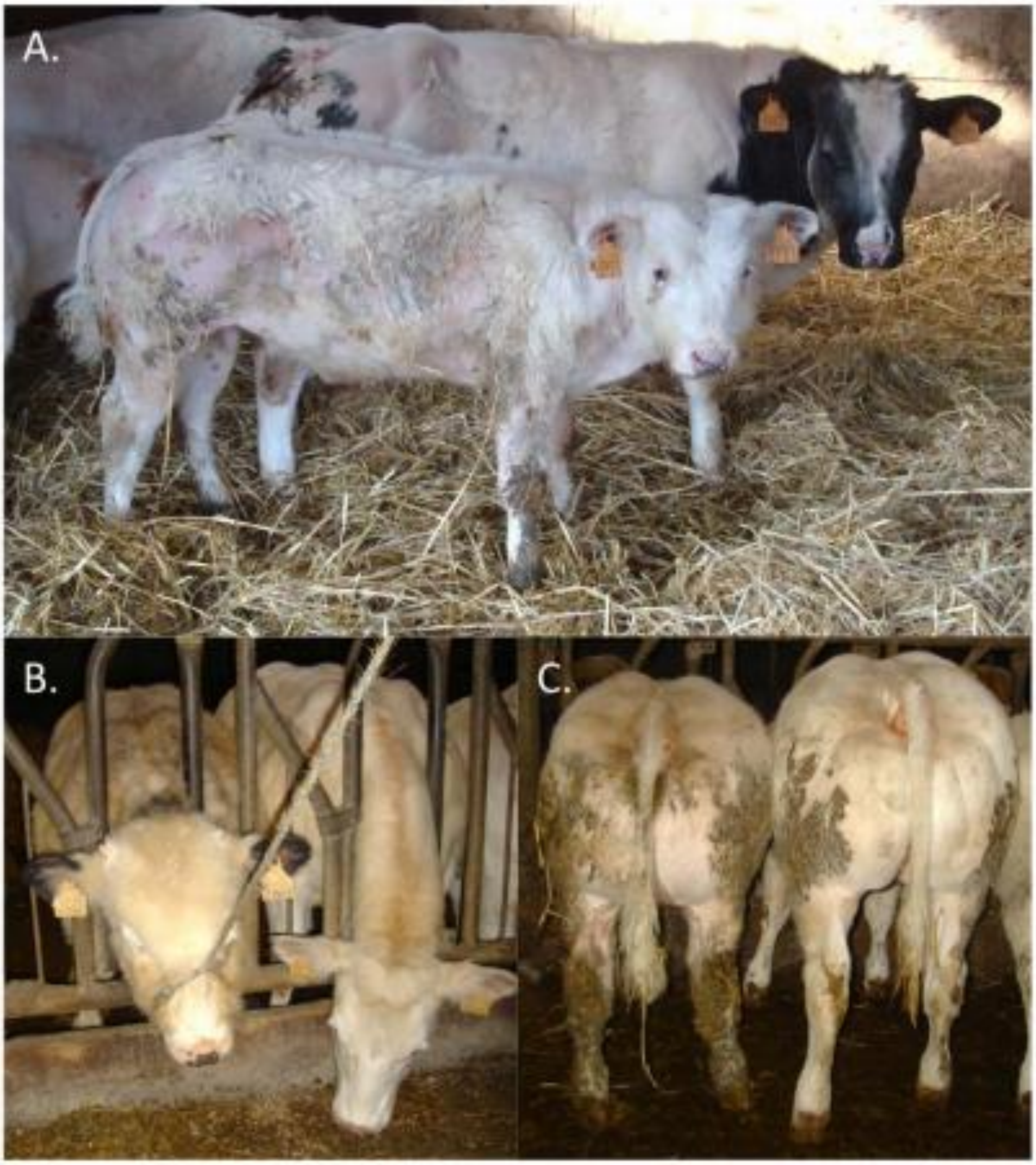


Figure 2. Features of animals homozygous for the *RNF11* c124-2A>G mutation. Affected (front) and control (back) calves of same age, illustrating the proportionate growth retardation, close forehand, and hairy head masking a narrow skull (A). Illustration of the hairy head (B), and normal muscle development (C) of animals homozygous for the *RNF11* c124-2A>G variant.

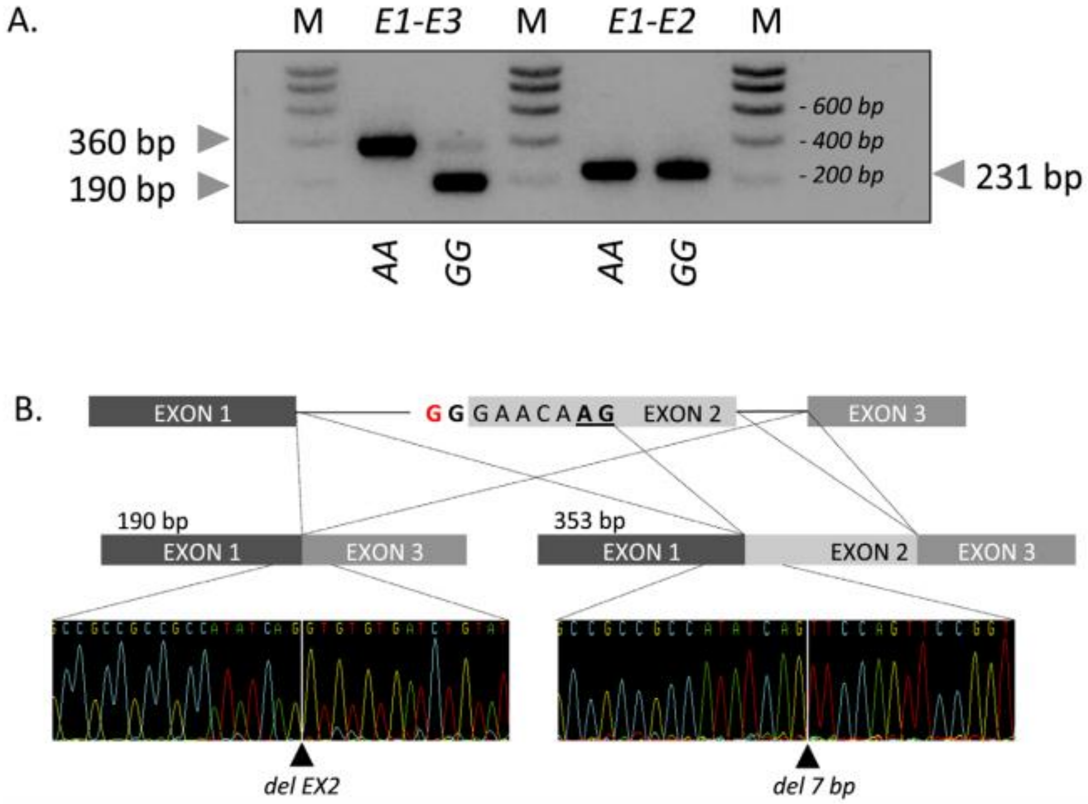


Figure 3. Effect of the c124-2A>G splice site variant on *RNF11* transcripts. (A) Gel electrophoresis of RT-PCR products obtained from mesenteric lymph node from homozygous wild-type (*AA*) and mutant (*GG*) animals using primer sets located respectively in exon 1 and 3 (E1–E3) and exon 1 and 2 (E1–E2). M: molecular weight marker. (B) Sequence analysis and structure of the 190-bp and 360-bp RT-PCR products obtained from an affected (*GG*) animal.

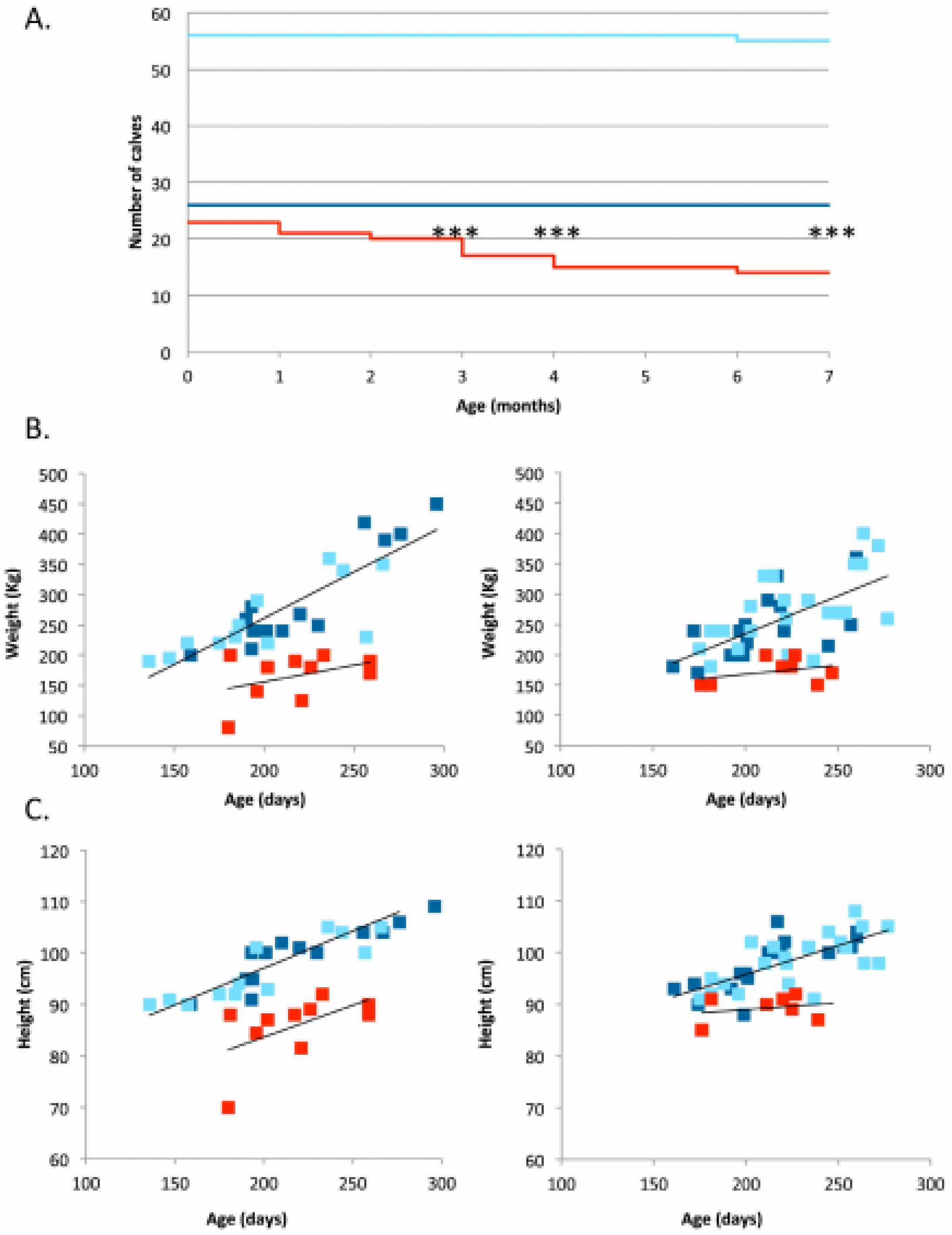


Figure 4. Survival and growth of 105 calves born from matings between carrier sires and dams. (A) Survival (from birth to 7 months of age) of calves sorted by c124-2A>G genotype (red: *GG*, dark blue: *AA*, light blue: *AG*) (***: p<0.001). (B) Weight (estimated from heart girth length) and (C) height at withers (from birth to 7 months of age) of calves sorted by c124-2A>G genotype (red: *GG*, dark blue: *AA*, light blue: *AG*). Regression lines (black) were fitted separately for affected and non-affected animals.

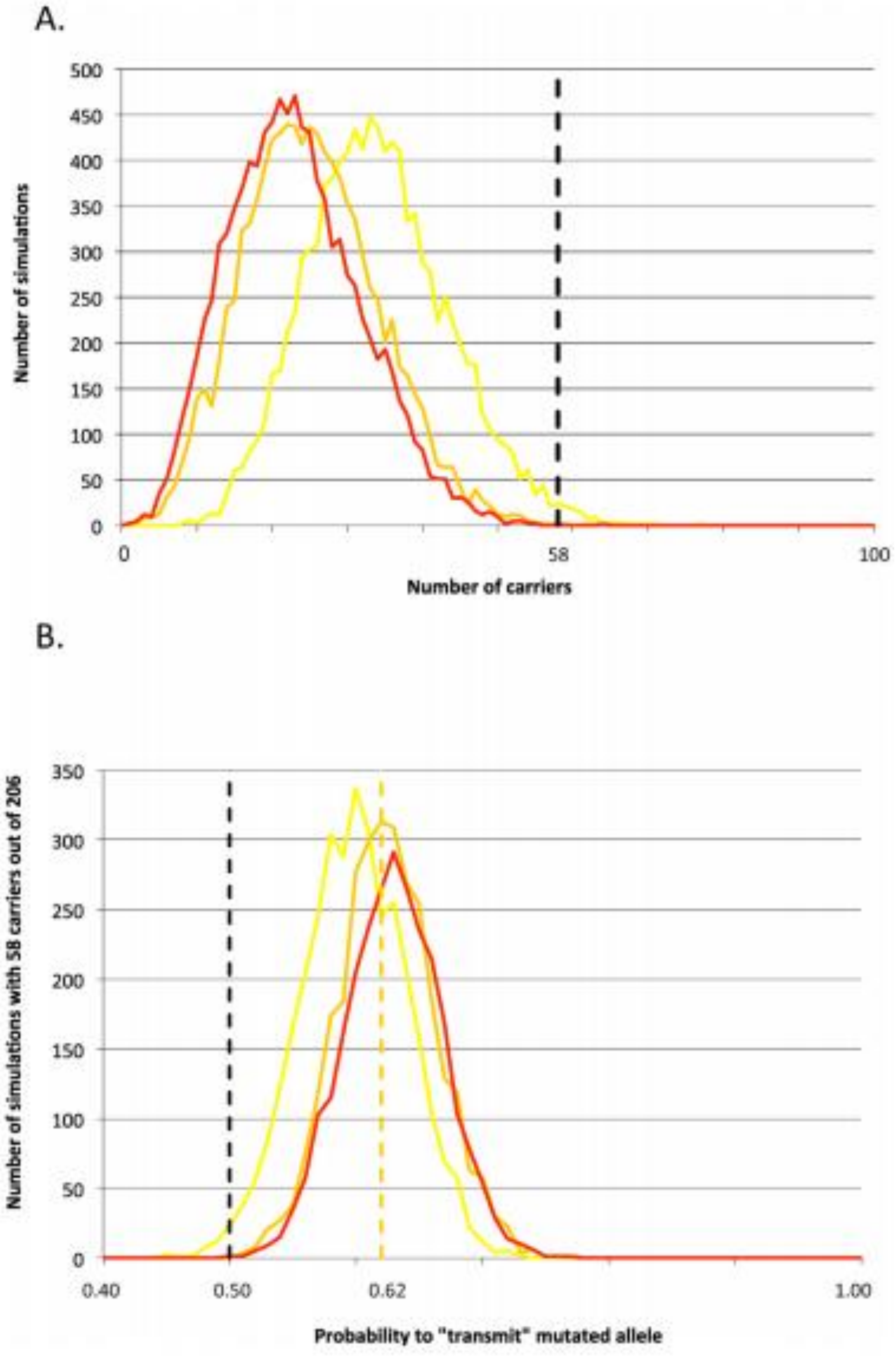


Figure 5. Signature of selection. (A) Frequency distribution (number of simulations out of 10,000) of the number of sires tracing back to the *Galopeur* founder (total: 206) that are expected to carry the c124-2A>G mutation assuming that it segregates in the corresponding pedigree according to Mendelian expectations, and that the frequency of c124-2A>G outside the Gallopeur lineage is 0% (red), 1% (orange), or 5%

(yellow). The dotted vertical marks the actual number of carrier sires (58) amongst the 206 descendants of *Galopeur*. (B) Distribution of the number of simulations (out of 10,000) yielding 58 carriers out of 206 descendants of *Galopeur* (Y-axis), as a function of the rate of transmission of the mutation from heterozygous carriers (X-axis). Three curves are given corresponding to frequencies of the mutation outside of the *Galopeur*'s lineage of 0% (red), 1% (orange), and 5% (yellow). The dotted orange vertical line corresponds to a transmission rate of 62%, maximizing the number of simulations yielding 58 carriers for a mutation frequency (outside of the *Galopeur*'s lineage) of 1%.

Detection of copy number variants in the horse genome and examination of their association with recurrent laryngeal neuropathy

M. Dupuis, Z. Zhang, K. Durkin, C. Charlier, P. Lekeux and M. Georges

Abstract

We used the data from a recently performed genome-wide association study using the Illumina Equine SNP50 beadchip for the detection of copy number variants (CNVs) and examined their association with recurrent laryngeal neuropathy (RLN), an important equine upper airway disease compromising performance. A total of 2797 CNVs were detected for 477 horses, covering 229 kb and seven SNPs on average. Overlapping CNVs were merged to define 478 CNV regions (CNVRs). CNVRs, particularly deletions, were shown to be significantly depleted in genes. Fifty-two of the 67 common CNVRs (frequency ≥ 1%) were validated by association mapping, Mendelian inheritance, and/or Mendelian inconsistencies. None of the 67 common CNVRs were significantly associated with RLN when accounting for multiple testing. However, a duplication on chromosome 10 was detected in 10 cases (representing three breeds) and two unphenotyped parents but in none of the controls. The duplication was embedded in an 8-Mb haplotype shared across breeds.

Anim. Genet. 44 (2013) 206-208

Results

Recurrent laryngeal neuropathy (RLN) is the most common upper airway pathology in the horse (Robinson 2004). Estimates of narrow-sense heritability range between 20% and 60% depending on the population (Marti and Ohnesorge 2002; Ibi *et al.* 2003; Barakzai 2009). We recently conducted a genome-wide association study with 234 cases and 228 breed-matched controls, using the Illumina Equine SNP50 array (Dupuis *et al.* 2011). We identified two genomic regions with suggestive evidence for association of chromosomes 21 and 31. However, both loci jointly explained at most 8% of inherited predisposition.

To further exploit the generated data set, we used Penncnv to call copy number variants (CNVs) from the available SNP genotypes (Wang *et al.* 2007; see Materials and methods in Appendix S1). After exclusion of horses with more than 10 CNVs (given the frequency distribution of CNV per horse, Fig. S1), we detected 2797 CNVs for 477 horses (average CNVs/sample = 5.9), including 10.7% homozygous and 2.6% heterozygous deletions, and 55.1% heterozygous and 1.6% homozygous duplications. On average, individual CNVs spanned 229 kb and seven SNPs. Overlapping duplications and deletions were merged separately to define 478 CNV regions (CNVRs) (Tables S1 & S2). Overlap between duplication and deletion CNVRs was observed 30 times. The majority of CNVs were rare: 86% were observed in four or fewer horses (i.e., <1%), of which 67% were singletons. The chromosome distribution of the CNVRs is shown in Fig. S2. Thirty-one of the 478 CNVRs jointly contained olfactory receptor genes (average, 15 genes; range, 1–94). In addition, the 238 duplication CNVRs overlapped 273 genes, and the 240 deletion CNVRs overlapped 174 genes. When compared to a random sample of non-overlapping genome segments, deletion CNVRs were very significantly depleted in gene content ($P < 0.001$), whereas duplication CNVRs tended to be depleted albeit non-significantly ($P = 0.07$; Fig. S3). This is thought to reflect purifying selection against CNV-dependent gene loss and supports the genuine nature of a large proportion of predicted CNVRs (Conrad *et al.* 2010). In addition to olfactory receptors, CNVRs-encompassed taste receptors, members of the cytochrome P450 family and solute carrier family, T cell receptors, and immunoglobulin lambda-like polypeptide, known to be subject to CNVs in other species as well (Feuk *et al.* 2006; Redon *et al.* 2006).

We performed several tests to evaluate the specificity and sensitivity of the CNV calling procedure (see Materials and methods in Appendix S1). These analyses were limited to the 67 common CNVRs (≥1%; Table S2). First, we searched for an association between the CNV copy number and the genotype of SNPs located in 2-Mb flanking windows. The test statistic obtained with the cognate CNVR was compared with that obtained for the same windows with the 66 other common CNVRs. For 33/67 CNVRs, the test statistic obtained with the cognate CNVR was the highest of the list. For 12 additional ones, the test statistic had a rank of three or two, corresponding to a nominal P value <5% (Fig. S4). Thus, the genuine nature of ~67% of common CNVRs is supported by evidence of cis-association. We further extended the association studies to the entire genome to detect possible trans-associations (Durkin *et al.* 2012). Log(1/P) values exceeded a conservative threshold of 8 for two duplication CNVRs without evidence of cis-association (CNVR62, 66) and one CNVR with clear cis-association (CNVR43). Surprisingly, three deletion CNVRs with significant cis-association (CNVR4, 22, 45) also yielded a trans-signal exceeding the threshold of 8.

We then searched for Mendelian inheritance of both duplication and deletion CNVRs. We used eight available trios (sire–dam–offspring) to verify whether a CNVR in an offspring was detected in at least one of the parents and used 74 offspring of 52 single parents to verify whether CNVRs detected in parents were transmitted to offspring. This allowed us to confirm three additional CNVRs (CNVR33, CNVR37, and CNVR44), for which no significant association was obtained. Twenty-five of the 48 CNVRs confirmed by association were additionally validated by Mendelian inheritance.

We finally checked for Mendelian inconsistencies within deletion CNVRs. Two additional CNVRs (CNVR46, CNVR61) were validated using this approach. Among the 22 deletion CNVRs confirmed by association, 17 exhibited an excess of Mendelian inconsistencies.

Taken together, we provide genetic validation evidence for ~80% of the common CNVRs in our data set.

To obtain an estimate of the sensitivity of CNV detection, we examined the proportion of offspring inheriting a CNVR from a carrier parent. The analysis was limited to validate CNVRs. Overall, 35% of offspring from a carrier parent also carried the CNVR, whereas 50% was expected. Thus, the overall sensitivity of detection of common CNVs was estimated at 70%.

We then searched for association between RLN and CNVR genotype. To that end, we compared the proportion of cases and controls with copy number differing from two using Fisher's exact test. We considered only the 67 common CNVRs in this analysis. The significance threshold was consequently set (Bonferonni correction) at 0.00075. None of the 67 CNVRs exceeded this threshold. However, the most significant signal was obtained with CNVR35, which was observed in nine cases (three warmbloods, three trotters, three draft horses), one unphenotyped parent and none of the controls, yielding a nominal P value of 0.0036. CNVR35 corresponds to a 62-kb duplication that was validated on the basis of a cis-association. Examination of the SNP genotypes indicated that the duplicated CNVR35 allele is embedded in a remarkably long (8 Mb) haplotype shared among the three breeds. We identified two animals that carried the corresponding haplotype and visual examination of their log R ratio and B allele frequencies indicated that indeed both were carrying the CNVR allele with the duplication (Fig. S5). One was an unphenotyped sire of one of the affected draft horse offspring (shown by Penncnv to carry the duplication), whereas the other was an additional affected warmblood. Including this additional case in the analysis yields a nominal association P value of 0.0018. The corresponding genomic region is not known to behave as a recombination cold spot. Thus, the conservation of such a long haplotype across three different breeds is intriguing. One possible explanation is that the duplication resides in an 8-Mb inversion, precluding recombination with wild-type haplotypes. It is tempting to speculate that the inversion rather than the CNV might underlie the association with RLN.

Acknowledgements

We are grateful to veterinarians for phenotypic data and sample collection, to the GIGA genotranscriptomic core facility for their assistance with genotyping, and to INRA for their collaboration on the GENEQUIN research program. This work was funded by grants from the Agence Nationale de la Recherche (ANR), the Fond Eperon, the Institut Français du Cheval et de l'Equitation (IFCE), and the Région Basse-Normandie (France).

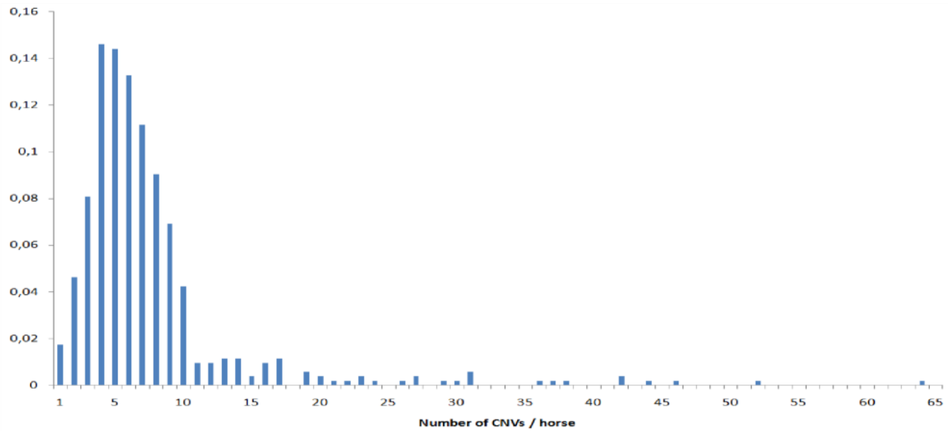


Figure S1: Frequency distribution of the number of CNVs per horse detected with the PennCNV software on 520 individuals.

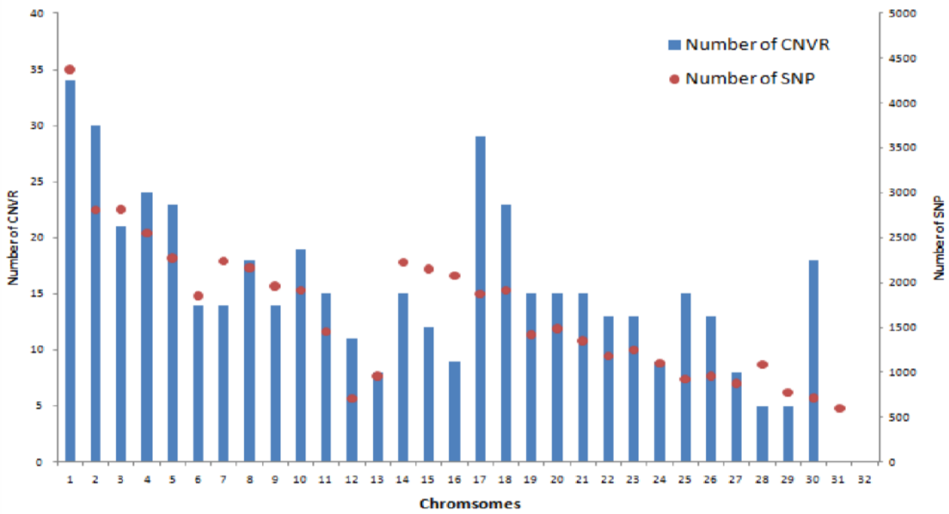


Figure S2: Chromosomal distribution of the CNVRs compared to number of SNPs per chromosome.

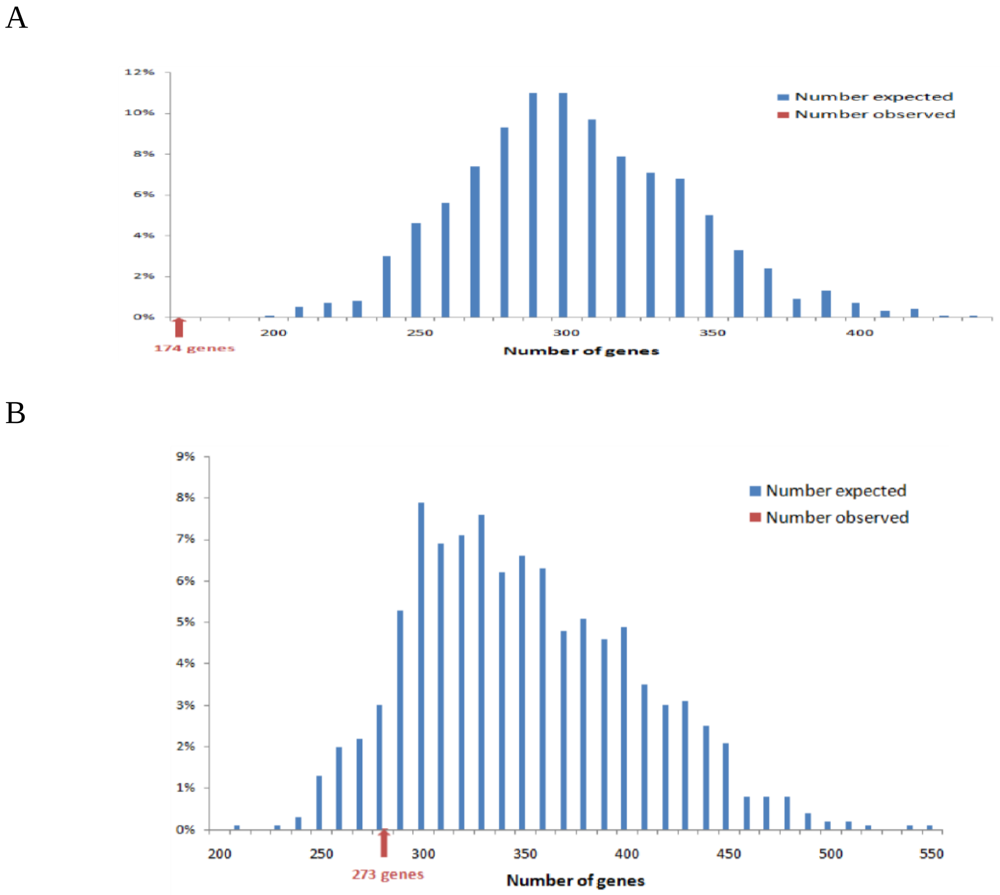


Figure S3: Frequency distributions of the number of genes observed in a random sample (1000 permutations) of non-overlapping random segments (matched in term of size and number) for deletions (A) and duplications (B).

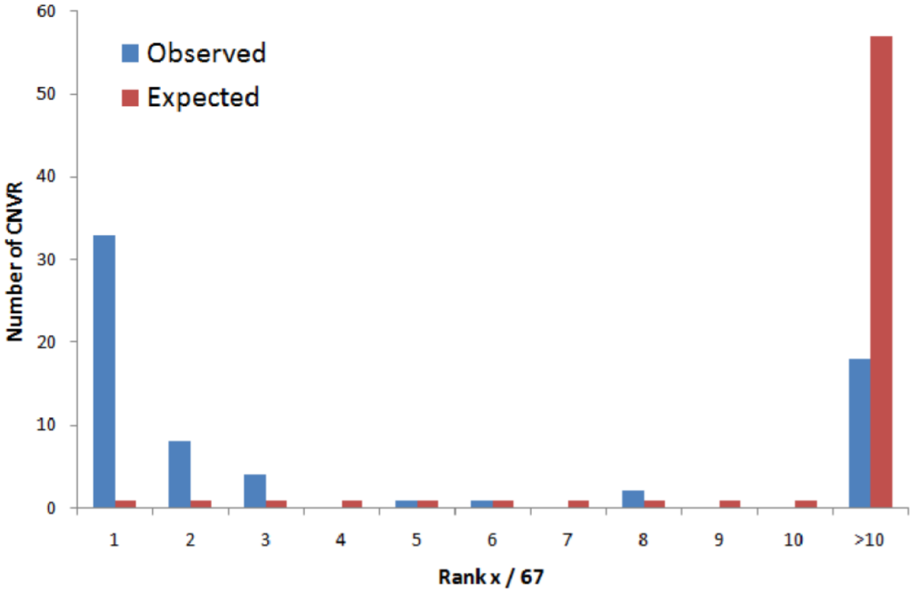
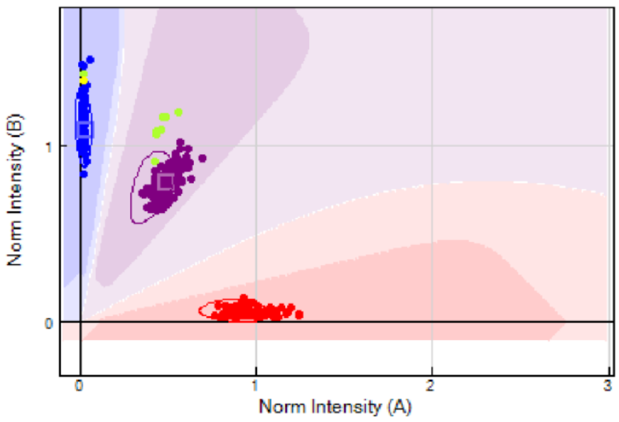
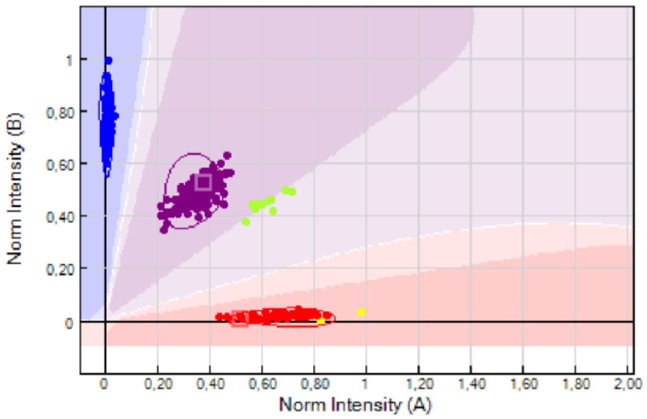


Figure S4: Evidence of association between copy number variants and the genotype of SNPs located in 2Mb flanking windows revealed by the rank of the test statistic obtained with the cognate CNVR (blue) compared with that obtained for the same windows with the other common 66 CNVR (red).

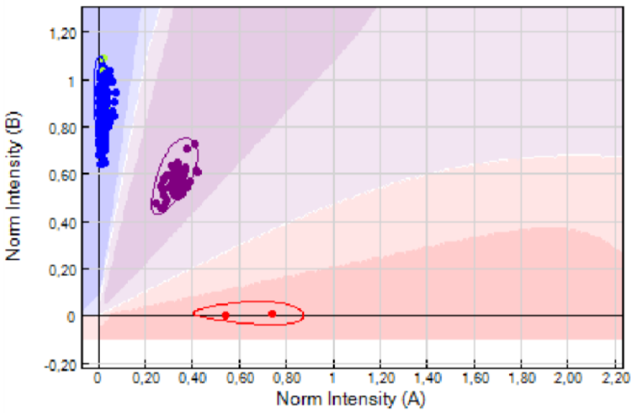
BIEC2-112446



BIEC2-112447



BIEC2-112455



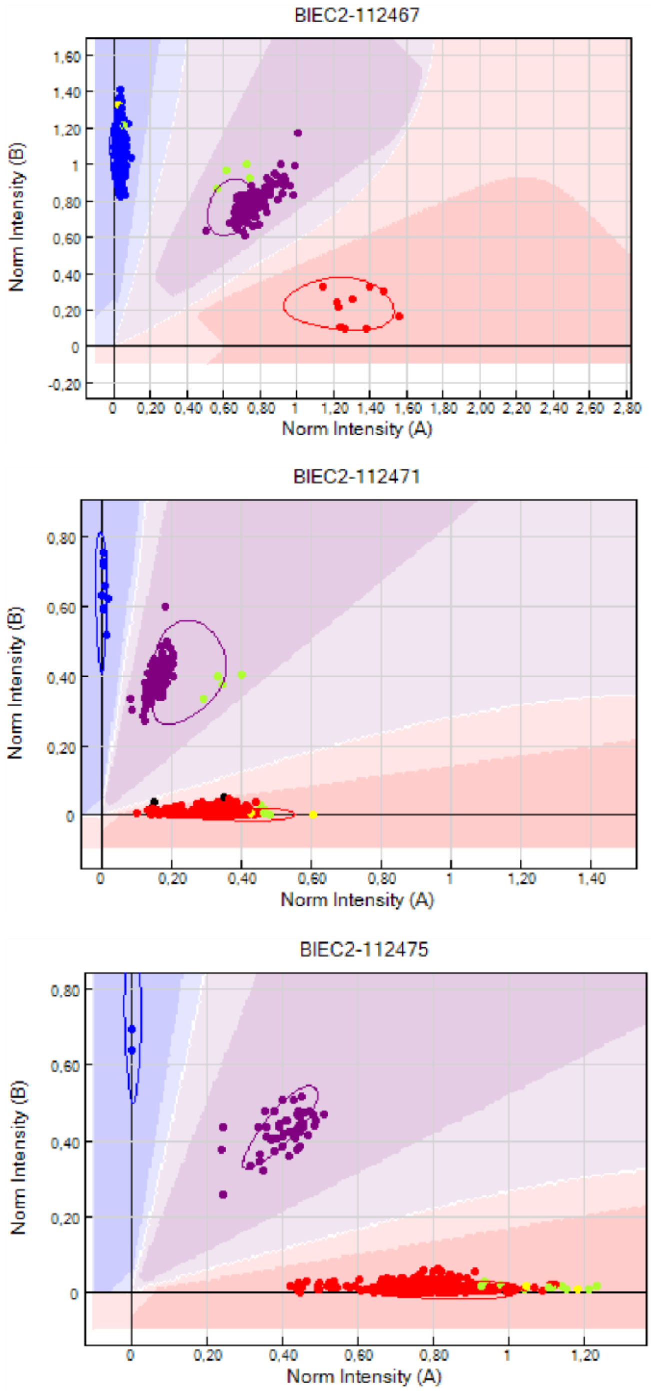


Figure S5: Visualization with illumina Genome Studio software of fluorescence intensities of 6 SNPs (from BIEC2-112446 to BIEC2-112475) contained in a duplication (CNVR35) detected in twelve horses (10 by PennCNV, in green, and 2 by visual examination, in yellow).

Table S1. Details of the 478 CNVR: chromosome and position in bp, number of horses (Nb) detected with the CNVR and copy-number （CN） (homozygous deletions 0, heterozygous deletion 1, heterozygous duplications 3 or homozygous duplications 4)

Chr	Start	End	Nb	CN	Chr	Start	End	Nb	CN
1	2404732	2428759	1	3	13	5704290	5788444	1	3
1	3377373	3430737	1	3	13	7000988	7010566	1	3
1	6988046	7101467	1	3	13	7359845	7378465	14	3
1	16485051	16675322	1	1	13	8023293	8336831	7	3
1	25680190	25688764	1	3	13	13961628	13968758	4	1
1	28700170	28750510	1	1	13	22580668	22763058	1	3
1	38333182	38425494	3	3	13	24691212	24715679	1	3
1	49442370	49612915	1	1	14	1296596	1392437	2	3
1	54099833	54226216	1	1	14	3819727	3867523	1	3
1	54099833	54226216	3	3	14	17756907	17826590	1	1
1	93174703	93315929	1	3	14	18057043	18077886	1	1
1	93604112	93623884	57	3	14	29605818	29613676	1	3
1	102790715	102924399	2	1	14	41858449	41859074	1	3
1	109437334	109473816	26	0 or 1	14	42327252	42341408	1	3
1	109437334	109473816	1	3	14	45136536	45192152	1	3
1	111524212	111565898	1	1	14	52997893	53417443	40	3
1	121724591	121759945	1	1	14	59460837	59463001	3	3
1	127066527	127072827	1	1	14	59794863	59856137	1	1
1	127066527	127072827	1	3	14	64933907	64951664	3	1
1	127890314	127937605	1	3	14	67970273	67973544	1	1
1	136452610	136587104	1	3	14	68068476	68095557	2	3
1	138375254	138425445	1	3	14	92832873	92874359	4	3
1	150538517	150589441	1	1	15	7204308	7215311	1	3
1	154902949	155593582	1	1	15	13360186	13546091	31	3
1	155487276	155656642	78	0 or 1	15	21073741	21139258	4	1
1	155795029	156870455	81	0 or 1	15	36124731	36215145	1	1
1	158969159	159109725	9	1	15	47772108	47776161	1	1
1	158969159	159109725	20	3 or 4	15	47772108	47776161	2	3
1	165920447	166094027	1	3	15	48199824	48278494	16	1
1	166388258	166511143	1	1	15	57107237	57203204	10	1
1	169080745	169106171	1	1	15	57107237	57124675	2	3
1	178553782	178573079	1	1	15	58977518	58988902	1	3
1	178798269	178815370	1	1	15	59421724	59448498	1	4
1	178798269	179550475	3	3	15	65519182	65526115	1	3
2	11066626	11138697	1	1	16	2716637	2787127	11	3
2	15112913	15284362	1	1	16	15852883	15869728	1	1
2	23718743	23738529	1	3	16	17790340	17885080	1	3

2	33516505	33523263	1	1	16	38714465	38751894	1	1
2	47257405	47320158	2	0	16	58842993	58952082	1	3
2	48153515	48240185	1	1	16	63556037	63561798	1	1
2	51877913	51881552	3	3	16	76887717	76938027	1	3
2	58352438	58411027	1	3	16	78418181	78701360	1	3
2	58518774	58529951	1	1	16	80679310	80681397	1	3
2	71776762	71852657	3	1	17	2877987	3010511	1	3
2	76470475	76600049	1	1	17	6214434	6412483	1	3
2	78026093	78027107	1	1	17	7219757	7291227	3	3
2	83746531	83778839	1	1	17	15505612	15561039	1	1
2	85856142	85856257	1	3	17	15505612	15512365	1	1
2	90673871	90752478	1	3	17	20197651	20198683	2	3
2	96664019	96702010	1	1	17	32444026	32635391	1	3
2	96664019	96702010	1	3	17	32635391	32770316	1	3
2	97687916	97689054	2	1	17	32828288	32828619	1	1
2	97917014	98286668	1	1	17	35323872	35577301	1	3
2	101905815	101930407	1	3	17	36644251	36688452	1	3
2	103758278	103782300	1	3	17	36846001	36977325	3	3
2	106062109	106064917	7	0 or 1	17	38001228	38063886	1	1
2	106917495	106967946	1	3	17	40886200	41219289	1	3
2	107853583	107936731	5	3	17	43477906	43563179	1	3
2	108615709	108630836	1	1	17	44277770	44330951	1	1
2	110360397	110411873	1	1	17	48562840	48603600	1	1
2	110487412	110551367	2	1	17	48562840	48603600	1	3
2	111089904	111104849	1	3	17	49886727	50011096	2	1
3	21596958	21636056	1	3	17	51747591	51884825	1	1
3	35305705	35321182	1	1	17	53466267	53539910	1	1
3	41567820	41635136	8	3	17	55492569	55582048	1	1
3	41621380	41635136	1	3	17	55690076	55690662	1	1
3	42828373	42864867	1	3	17	57337550	57338181	4	1
3	47915872	47970675	1	1	17	57962518	58094803	1	1
3	51525184	51575785	1	3	17	58048425	58048581	1	1
3	57169047	57214428	1	1	17	60468135	60479962	1	3
3	65932710	66015792	11	1	17	62254110	62526107	3	3
3	65705932	66065643	50	3	17	65459354	65462839	1	1
3	66838815	66956851	1	1	18	3286734	3440128	2	3
3	67191980	67212427	1	1	18	9608605	9810377	2	3
3	67197969	67248561	3	3	18	9608605	9746042	1	3
3	70627351	70640054	1	3	18	10104665	10224646	1	3
3	70845155	70967718	1	1	18	10812787	10881369	1	3
3	73033604	73156602	8	3	18	13135454	13194121	1	3
3	81610006	81656738	1	1	18	16627010	16713910	1	3
3	81758548	81778280	1	3	18	24529189	24674961	1	1
3	90966188	90971491	1	1	18	25534223	25595772	1	1

3	100081272	100245562	1	1	18	25803363	25806480	1	1
3	103598907	103691288	1	1	18	28863825	28964267	2	3
3	104737609	104821920	1	3	18	36449291	36449543	1	1
3	114181081	114298463	1	3	18	37572919	37625592	1	1
4	2309850	2634210	1	1	18	40187377	40342808	1	1
4	9061371	9121425	2	3	18	46009970	46010221	1	1
4	9121425	9323693	1	3	18	49404497	49408946	1	3
4	9310410	9500292	53	0 or 1	18	51078308	51113066	1	1
4	16830516	16836115	1	3	18	60165814	60166198	1	1
4	22227867	22288213	1	1	18	60519874	60601238	1	1
4	22737039	22769640	1	1	18	70473204	70562048	1	3
4	23011397	23158551	1	1	18	75780975	75879559	27	3
4	27642526	27651073	1	1	18	75864062	75879559	3	3
4	29536635	29844870	1	1	18	76611180	76632666	1	1
4	32764495	32768967	1	1	19	6860	503825	1	3
4	33736922	33747940	1	1	19	1116053	1669010	1	3
4	37013872	37014111	1	1	19	1346278	1396027	1	3
4	38547233	38625012	1	1	19	3679642	3815186	1	1
4	52194368	52612016	1	3	19	5012965	5014569	2	1
4	70612792	70683966	1	1	19	5924490	5959965	1	3
4	72343609	72349365	13	3	19	8396001	8473556	1	1
4	81521531	81535530	2	1	19	32598913	32638781	46	0 or 1
4	81577925	81616460	1	1	19	32598913	32621031	5	3
4	82490516	82508616	1	1	19	38547055	38646893	1	1
4	85453445	85457390	3	1	19	43032770	43092034	1	1
4	96922930	97192480	86	3 or 4	19	46577737	46676246	1	1
4	98289000	98481816	1	1	19	51656828	51766306	1	1
4	102853989	102936414	1	3	19	51883240	51924633	1	3
5	15890795	15914852	1	1	19	54601785	54640517	1	3
5	27624425	27686502	1	1	20	17395157	17406109	1	1
5	28782503	28814723	1	3	20	22807631	22870569	1	1
5	37840041	37916448	5	0	20	22837509	22906369	1	1
5	37994476	37998277	1	1	20	24257326	24485375	1	3
5	44138490	44261644	9	3	20	26371568	26564881	9	3
5	45358974	45514810	1	3	20	28355086	28731640	4	1
5	45442281	45551283	1	1	20	31961012	32367245	19	0 or 1
5	47239959	47259141	1	1	20	32059082	32250493	9	3
5	47359920	47399241	47	3	20	37245113	37303121	2	3
5	53365624	53382340	1	3	20	40725082	40812984	1	3
5	56110634	56177960	1	3	20	43520722	43544976	3	3
5	60139794	60239913	1	1	20	45309182	45347364	3	1
5	63904057	63920127	1	1	20	48306105	48355432	5	1
5	69413874	69429487	1	1	20	50391131	50435199	1	3
5	69413874	69429487	1	3	20	50435199	50483331	2	3

5	72522420	72674477	1	1	21	976423	1008072	3	3
5	73582972	73620324	1	3	21	2768320	2859633	1	1
5	77417050	77435333	1	1	21	8453609	8487022	1	1
5	77417050	77685828	17	3	21	18811395	18814925	5	3
5	88243192	88349479	27	3	21	18960857	19035089	1	1
5	89599857	89784932	1	1	21	23358420	23383825	1	1
5	99595549	99631010	2	3 or 4	21	31684806	31841466	1	1
6	286052	337057	1	3	21	34531228	34550625	1	1
6	8932759	8991955	1	3	21	35167053	35205963	1	3
6	20120126	20157740	10	3	21	36692422	36697404	1	1
6	23555139	23576576	1	3	21	37245803	37287979	1	1
6	26104234	26118925	19	0 or 1	21	37245803	37287979	1	3
6	26102028	26126581	1	3	21	40039074	40217182	1	1
6	28525354	28582837	1	3	21	40983019	42820265	2	1
6	37380955	37550849	1	1	21	42207465	42211062	1	1
6	38268897	38278874	3	3	22	2493448	2554255	1	1
6	38756173	38783638	1	1	22	5268548	5383622	1	4
6	39643333	39648248	1	1	22	12929717	12980228	2	1
6	45467796	45552702	2	1	22	15434589	15483680	2	1
6	71956823	72607543	71	0 or 1	22	20254977	20311433	1	1
6	73072905	73412971	2	1	22	20922073	20926117	2	3
7	8577492	8584515	1	1	22	23459207	23496489	1	3
7	9160976	9413168	6	1	22	23909436	24034489	1	3
7	16006458	16050482	1	1	22	27752979	27827795	1	3
7	30834310	30995260	1	1	22	36435729	36545238	9	3
7	31406445	31529855	10	1	22	36732680	36920538	20	0 or 1
7	52328145	52498211	1	1	22	36850262	36920538	4	3
7	52610482	52677786	39	3	22	37835325	37868430	1	3
7	55286808	55351233	1	1	23	41804	275355	2	3
7	55286808	55410254	2	3	23	4729518	4739287	1	3
7	73083306	73197149	3	1	23	5855795	6056873	2	1
7	73371661	73666395	78	0 or 1	23	6284614	6425213	1	3
7	73504993	73666395	9	3	23	7331191	8512483	9	3
7	74362741	74418075	3	3	23	8512483	8734195	2	3
7	92601495	92604434	1	1	23	16302756	16397289	1	3
8	1722878	2213144	1	1	23	17021853	17100602	2	1
8	1910659	2115738	3	3	23	18031161	18042294	1	1
8	2115370	2213144	3	1	23	29305683	29334199	1	1
8	2897484	2996427	2	3	23	30217610	30259948	1	1
8	3608369	3821757	54	3	23	37630447	37630804	1	1
8	4183178	4430473	2	1	23	47584956	47585275	1	1
8	4391896	4646812	43	0 or 1	24	7904451	7908956	1	1
8	5712244	5716266	2	3	24	13702903	13726613	1	1
8	16028660	16028807	1	3	24	13910912	13988805	1	1

8	16071303	16171233	2	3	24	27656238	28204834	2	3
8	16085130	16171233	1	1	24	28564090	28682383	1	1
8	24030822	24188533	1	3	24	32416012	32628728	20	3
8	24528856	24575032	1	3	24	35847276	35875578	1	1
8	28260573	28355849	2	3	24	38370785	38375555	35	3
8	41810441	41869404	1	1	24	41033593	41035832	2	3
8	81506440	81550319	1	3	25	2996811	3028209	1	3
8	82849217	82877751	1	1	25	16489352	16560013	1	1
8	82877751	82982694	2	3	25	17593725	17703603	1	3
8	87181631	87279745	2	3	25	17955032	17979476	18	0 or 1
9	20586158	20588207	1	0	25	17955032	18141741	21	3
9	21946854	22030940	1	1	25	17703603	17979476	1	3
9	29889627	29892897	10	0 or 1	25	18481781	18574374	1	3
9	50803693	50808048	4	0 or 1	25	18860537	18971635	1	3
9	56052920	56125503	2	1	25	25249663	25309442	2	3
9	56385984	56567910	1	1	25	26318531	26942120	99	0 or 1
9	57755597	58055411	17	0 or 1	25	29618659	29684315	28	3
9	70971456	71020023	2	1	25	37673569	37759471	1	1
9	70960794	71020023	108	3	25	37673569	37780451	1	3
9	72114145	72307507	3	3	25	38084663	38546385	9	3
9	72247701	72307507	1	1	25	38457363	38546385	1	3
9	74224040	74488164	1	3	26	52232	357957	1	1
9	81398267	81402542	3	3	26	2942000	2972786	1	1
9	82977765	83060823	1	1	26	6633318	6643399	1	1
10	674485	1141923	10	3	26	13391944	13474118	1	1
10	3909288	4005578	2	3	26	17252368	17325508	1	3
10	13004730	13070456	2	3	26	18158669	18163088	2	1
10	13917500	14021914	2	3	26	18834339	19052616	1	1
10	16554166	16568462	2	3	26	21312489	21381972	2	3
10	21301438	21335645	1	3	26	25045181	25045560	1	1
10	26215698	26249034	1	3	26	26883163	26924886	1	1
10	30764949	30827421	10	3	26	28446287	28500869	1	1
10	39119202	39541740	1	1	26	33521081	33536968	4	3
10	48605291	48704891	1	1	26	37233957	37329333	4	3
10	48866033	48869952	1	1	27	41769	258452	1	3
10	52948258	52949299	1	1	27	1749770	1797283	1	1
10	53808057	53837679	1	1	27	5750034	5868146	17	3
10	53830445	53837679	1	3	27	6539738	6597118	1	3
10	54831255	54848777	1	3	27	6990066	7033521	1	3
10	65304239	65372917	1	1	27	13198799	13303733	1	3
10	69091035	69125778	3	1	27	17601828	17746986	3	3
10	70245511	70506856	1	1	27	26139989	26381437	1	1
10	77814178	77851298	1	1	28	14158917	14230291	1	1
11	635764	780294	6	3	28	16674286	16683495	1	1

11	1926549	1926646	1	3	28	35535261	35924600	1	3
11	2045893	2153152	1	3	29	477117	631698	2	1
11	2517091	2524317	2	3	29	399673	946361	1	1
11	3176125	3435909	1	3	29	282613	631698	24	3
11	10225315	10418099	1	3	29	5666076	5733601	1	1
11	23950817	23992162	5	3	29	8565963	8611891	1	1
11	26158220	26328579	1	3	29	16200670	16335779	1	1
11	33269366	33337791	1	3	29	29212867	29320084	2	3
11	36069308	36078928	1	1	30	11367620	11430634	1	3
11	38181763	38192091	1	3	30	13446880	13565090	1	3
11	54640169	54812394	1	1	30	15052775	15104826	1	3
11	54812394	54929094	1	1	30	17054598	17084652	1	1
11	59722467	59737161	1	1	30	17054598	17084652	1	3
11	60402199	60405936	1	3	30	19565060	19826443	1	1
12	4475059	4585884	1	1	30	20499484	20500654	1	1
12	12524489	13488187	32	0	30	24398152	24603312	1	1
12	13945011	14777981	114	0 or 1	30	30032173	30051326	1	3
12	14124730	14172327	1	1	31	4829031	5085857	1	3
12	14587232	14777981	2	1	31	8225583	8227054	1	3
12	12193543	14913468	240	3 or 4	31	9688131	9865005	1	3
12	15362879	16525629	2	3	31	11290101	11304886	1	1
12	20138808	20154683	3	3	31	11852511	11903912	1	1
12	21568982	21582407	1	3	31	12363898	12778909	1	3
12	28033991	28066331	2	3	31	17905592	18089361	1	3
12	30065284	30158615	3	3	31	17951839	17955728	1	1
13	5704290	5788444	1	1	31	22891972	22950057	1	3

Table S2. Characteristics of the 67 common CNVRs: position (Chr: chromosome, Start: start position in bp, End: end position in bp), copy-number (Dupl: duplication, Delet: deletion), total number of horses (Nb), number of horses in each breed (W: Warmblood, TR: Trotter, TH: Thoroughbred, DH: Draft Horse), results of haplotype-based mapping (localisation in cis and/or trans, –log *P* value in cis, rank of the test statistic for association mapping in cis, Chr in trans: localization of the signal for significant result for association mapping in trans, –log P values for these signals in trans), familial analyses (P+/O+PP (trio): number of parents with the CNV compared to the number of offspring with the CNV in trios, O+/OP+: number of offspring with the CNV amongst the offspring from parents with the CNV, Mendelian inconsistencies: 0 duplications, + enrichment for deletions, **-** no enrichment), genes content and association with RLN (Number of cases / controls / parents, total number of horses, *P* **-**values of the Fisher test).

CNV	Chr	Start	End	CN	Nb	W	TR	TH	DH	Localisation	-LogP in cis	Rank	Chr in trans	-LogP in trans	P+/O+PP (trio)	O+/OP+	Mendel. Incon.	Genes function	Case	Control	Parent	Total	FISHER
CNV01	1	93604112	93623884	Dupl	**57**	45	6	3	3	cis	4.1	2			0	5/21	0	MESP2, AP3S2	29	21	7	**57**	0.2955
CNV02	1	109437334	109473816	Delet	**26**	18	6	2	0	cis	7.4	1			1/2	1/4	+	No	14	8	4	**26**	0.2749
CNV03	1	155487276	155656642	Delet	**78**	69	4	3	2	cis	16.0	1			0	2/8	-	Olfactory receptors	29	43	6	**78**	0.0531
CNV04	1	155795029	156870455	Delet	**81**	64	7	7	3	cis and trans	12.5	2	Chr 13	12.6	2/2	15/25	+	Olfactory receptors	34	37	10	**81**	0.605
CNV05	1	158969159	159109725	Delet	**9**	9	0	0	0	no	1.5	16			0	0/2	-	CCNG1 cyclin family	2	6	1	**9**	0.1692
CNV06	1	158969159	159109725	Dupl	**20**	14	4	2	0	cis	9.8	3			1/1	3/6	0	CCNG1 cyclin family	8	8	4	**20**	0.9999
CNV07	2	106062109	106064917	Delet	**7**	6	0	1	0	cis	13.5	1			0	0/1	-	No	4	2	1	**7**	0.6855
CNV08	2	107853583	107936731	Dupl	**5**	4	0	1	0	no	1.4	22			0	0/0	0	No	1	4	0	**5**	0.209
CNV09	3	41567820	41635136	Dupl	**8**	8	0	0	0	cis	5.4	1			0	1/1	0	No	6	1	1	**8**	0.1223
CNV10	3	65932710	66015792	Delet	**11**	9	1	1	0	cis	3.1	2			0/2	0/2	+	UGT, glucuronosyltransferase	6	3	2	**11**	0.5036
CNV11	3	65705932	66065643	Dupl	**50**	44	3	0	3	cis	11.7	1			0	0/3	0	UGT	21	26	3	**50**	0.4396
CNV12	3	73033604	73156602	Dupl	**8**	8	0	0	0	cis	3.8	2			0	0/0	0	No	2	5	1	**8**	0.2777
CNV13	4	9310410	9500292	Delet	**53**	50	2	0	1	cis	16.0	1			0/1	4/9	+	T receptor gamma	14	31	8	**53**	0.0068
CNV14	4	72343609	72349365	Dupl	**13**	13	0	0	0	no	1.1	23			0	0/0	0	No	4	9	0	**13**	0.1667

Part 2

CNV15	4	96922930	97192480	Dupl	**86**	63	9	13	1	cis	15.2	1			1/4	3/14	0	Olfactory receptors	37	38	11	**86**	0.8
CNV16	5	37840041	37916448	Delet	**5**	4	1	0	0	cis	14.2	1			0	0/3	+	APCS, CRP, Olf recept	3	1	1	**5**	0.6234
CNV17	5	44138490	44261644	Dupl	**9**	8	0	1	0	cis	6.1	2			0	0/2	0	S100 genes	4	3	2	**9**	0.9999
CNV18	5	47359920	47399241	Dupl	**47**	39	4	4	0	cis	5.7	1			0	3/18	0	Fragment of IgG	19	25	3	**47**	0.3397
CNV19	5	77417050	77685828	Dupl	**17**	14	1	2	0	cis	4.5	2			0	0/1	0	CLCA, ODFL	12	4	1	**17**	0.0719
CNV20	5	88243192	88349479	Dupl	**27**	22	0	5	0	cis	16.0	1			0	0/3	0	No	16	8	3	**27**	0.1412
CNV21	6	20120126	20157740	Dupl	**10**	9	0	1	0	cis	6.7	1			0	0/2	0	INPP5D	3	5	2	**10**	0.4976
CNV22	6	26104234	26118925	Delet	**19**	16	1	2	0	cis and trans	8.9	1	Chr 8	8.9	0	1/3	-	AQP12B	8	8	3	**19**	0.9999
CNV23	6	71956823	72607543	Delet	**71**	61	9	0	1	cis	11.9	1			0	3/12	+	Olfactory receptors	40	24	7	**71**	0.0568
CNV24	7	9160976	9413168	Delet	**6**	6	0	0	0	no	0.9	52			0	0/0	-	no	1	5	0	**6**	0.1168
CNV25	7	31406445	31529855	Delet	**10**	10	0	0	0	cis	7.3	1			0	0/0	-	Olfactory receptors	6	4	0	**10**	0.7513
CNV26	7	52610482	52677786	Dupl	**39**	30	9	0	0	cis	2.9	3			1/4	2/7	0	Olfactory rec + Zn finger	20	12	7	**39**	0.1996
CNV27	7	73371661	73666395	Delet	**78**	70	3	4	1	cis	16.0	1			0/1	6/15	+	Olfactory receptors	40	31	7	**78**	0.3626
CNV28	7	73504993	73666395	Dupl	**9**	8	1	0	0	no	3.3	5			0	0/1	0	Olfactory receptors	2	6	1	**9**	0.1692
CNV29	8	3608369	3821757	Dupl	**54**	36	8	10	0	cis	9.5	1			0	3/7	0	TOP,Ig lambda-like	18	29	7	**54**	0.0881
CNV30	8	4391896	4646812	Delet	**43**	34	2	0	7	cis	6.4	2			0	4/19	+	Ig lambda like pptide	18	22	3	**43**	0.5073
CNV31	9	29889627	29892897	Delet	**10**	6	4	0	0	cis	12.8	1			1/2	1/2	+	No	7	3	0	**10**	0.3386
CNV32	9	57755597	58055411	Delet	**17**	12	1	4	0	cis	5.2	1			0/2	0/2	+	No	8	7	2	**17**	0.9999
CNV33	9	70960794	71020023	Dupl	**108**	82	15	8	3	no	1.1	30			1/5	4/13	0	Gsdmc	49	46	13	**108**	0.9075
CNV34	10	674485	1141923	Dupl	**10**	10	0	0	0	no	2.5	6			0	0/1	0	UQCRFS	6	3	1	**10**	0.5036
CNV35	10	30764949	30827421	Dupl	**10**	4	3	0	3	cis	3.9	1			0	0/1	0	No	9	0	1	**10**	0.0036
CNV36	11	635764	780294	Dupl	**6**	6	0	0	0	cis	2.1	3			0	0/0	0	FOXK, NARF, HEXDC	2	4	0	**6**	0.4428
CNV37	11	23950817	23992162	Dupl	**5**	1	4	0	0	no	0.8	35			0	1/2	0	OSBP	2	1	2	**5**	0.9999
CNV38	12	12193543	14913468	Dupl	**240**	180	38	13	9	cis	16.0	1			5/5	31/47	0	Olfactory receptors	109	103	28	**240**	0.846
CNV39	12	12524489	13488187	Delet	**32**	27	0	5	0	cis	16.0	1			0	2/9	+	Olfactory receptors	14	14	4	**32**	0.9999
CNV40	12	13945011	14777981	Delet	**114**	103	0	11	0	cis	16.0	1			0	9/17	+	Olfactory receptors	52	50	12	**114**	0.9999

CNV41	13	7359845	7378465	Dupl	**14**	13	1	0	0	no	2.4	8			0	0/0	0	Cyto P450	10	4	0	**14**	0.1731
CNV42	13	8023293	8336831	Dupl	**7**	7	0	0	0	no	1.7	21			0	0/0	0	PILRA, MCM, TAF	3	4	0	**7**	0.7202
CNV43	14	52997893	53417443	Dupl	**40**	28	5	7	0	cis and trans	7.7	1	Chr 8	10.8	0	2/6	0	No	18	16	6	**40**	0.8591
CNV44	15	13360186	13546091	Dupl	**31**	30	0	1	0	no	0.8	48			0	1/2	0	MAL	14	16	1	**31**	0.7066
CNV45	15	48199824	48278494	Delet	**16**	16	0	0	0	cis and trans	13.5	1	Chr 19	13.1	0	0/0	-	No	12	4	0	**16**	0.0719
CNV46	15	57107237	57203204	Delet	**10**	10	0	0	0	no	2.3	8			0	0/0	+	No	4	6	0	**10**	0.5379
CNV47	16	2716637	2787127	Dupl	**11**	11	0	0	0	no	2.1	12			0	0/1	0	WNT7	7	3	1	**11**	0.3386
CNV48	18	75780975	75879559	Dupl	**27**	22	3	0	2	cis	2.9	3			0	0/4	0	Aox	14	11	2	**27**	0.682
CNV49	19	32598913	32638781	Delet	**46**	34	7	3	2	cis	13.3	1			1/2	7/27	+	DLG	23	17	6	**46**	0.4092
CNV50	19	32598913	32621031	Dupl	**5**	4	1	0	0	no	1.8	16			0	0/0	0	DLG	1	4	0	**5**	0.209
CNV51	20	26371568	26564881	Dupl	**9**	6	2	1	0	no	0.8	45			0	0/4	0	Olfactory receptors, Zn finger P	3	3	3	**9**	0.9999
CNV52	20	31961012	32367245	Delet	**19**	14	2	1	2	cis	7.6	1			0/1	0/1	+	TSPAN17	11	8	0	**19**	0.6408
CNV53	20	32059082	32250493	Dupl	**9**	6	3	0	0	no	1.3	19			0/1	0/3	0	TSPAN17	5	1	3	**9**	0.2155
CNV54	20	48306105	48355432	Delet	**5**	4	1	0	0	cis	8.7	1			0	0/0	+	No	3	2	0	**5**	0.999
CNV55	21	18811395	18814925	Dupl	**5**	5	0	0	0	cis	4.6	1			0	0/0	0	No	1	4	0	**5**	0.209
CNV56	22	36435729	36545238	Dupl	**9**	8	0	0	1	no	1.4	22			0/1	0/1	0	No	4	4	1	**9**	0.9999
CNV57	22	36732680	36920538	Delet	**20**	18	1	1	0	cis	7.1	1			0	0/2	+	No	12	6	2	**20**	0.2289
CNV58	23	7331191	8512483	Dupl	**9**	8	0	1	0	no	1.3	15			0	0/0	0	Solute carrier family, phosphatase	2	7	0	**9**	0.1006
CNV59	24	32416012	32628728	Dupl	**20**	20	0	0	0	cis	16.0	1			0	0/0	0	TTC8	8	12	0	**20**	0.3652
CNV60	24	38370785	38375555	Dupl	**35**	30	0	5	0	cis	11.5	1			0	1/2	0	No	15	18	2	**35**	0.5883
CNV61	25	17955032	17979476	Delet	**18**	16	1	1	0	no	1.5	20			0	0/2	+	Zn finger prot	5	11	2	**18**	0.1305
CNV62	25	17955032	18141741	Dupl	**21**	15	5	1	0	trans	2.4	11	Chr 22	8.2	0	1/5	0	Zn finger prot, solute carrier family	14	4	3	**21**	0.0279
CNV63	25	26318531	26942120	Delet	**99**	80	15	1	3	cis	16.0	1			0/1	6/10	+	Recepteur olfactif	46	44	9	**99**	0.9999

Part 2

CNV64	25	29618659	29684315	Dupl	**28**	27	1	0	0	cis	16.0	1			0	0/0	0	PreBcell leukemia PBX3	1 4	14	0	**28**	0.9999
CNV65	25	38084663	38546385	Dupl	**9**	9	0	0	0	cis	2.9	2			0	0/0	0	K-cl channel, phosphatase, ubiquitin ligas	e 5	4	0	**9**	0.9999
CNV66	27	5750034	5868146	Dupl	**17**	15	1	0	1	trans	1.4	26	Chr 2	9.4	0	0/2	0	ADAM32	8	7	2	**17**	0.9999
CNV67	29	282613	631698	Dupl	**24**	24	0	0	0	trans	0.9	29	Chr 30	16.0	0	1/10	0	Olfactory receptors	1 1	8	5	**24**	0.6408

Appendix S1

Materials and Methods

Cohorts, phenotypes and genotypes

We used a previously described RLN case-control cohort including 234 cases

(laryngeal grade ≥ III.2), 228 breed-matched controls (laryngeal grade I and ≥ 3 years), and 58 unphenotyped parents, representing four breeds (419 Warmbloods (W), 53 Trotters (TR), 37 Thoroughbreds (TH) and 11 draft horses (DH), Dupuis *et al.* 2011). All 520 horses were genotyped using the Illumina Equine SNP50 Beadchip and standard protocols recommended by the manufacturer. Raw data were analyzed with the Illumina GenomeStudio 2010.2 software. Average and minimum call rate per individual were 99.7% and 96%, respectively. Gonosomal SNPs were ignored for CNV analysis.

Identification of CNVs

SNP-specific (Log R ratio) and allele-specific (B allele frequency) fluorescence intensities were obtained from the Illumina GenomeStudio software and analyzed with PennCNV. PennCNV uses a Hidden Markov Model (HMM) incorporating Log R ratio, B allele frequency, inter-SNP distance and allelic frequency in the population for the detection of CNV (Wang *et al.* 2007). Overlapping CNVs were merged to define copy-number variant regions (CNVR).

Characterization of gene content

Analysis of gene content was based on the Ensembl gene Predictions (Ensemble 62, Hubbard *et al.* 2002).

Association between SNP and CNVR genotype

Association between SNP and CNVR genotype was examined for the 67 common

CNVR detected in at least 5 horses. The analysis was restricted to the W population. 7622 SNPs (within CNVR plus adjacent pairs) were first excluded from the data set.

Genotypes of the remaining 44,441 SNPs were phased using Phasebook (Druet and Georges, 2010) which assigns haplotypes to 20 hidden haplotype states. We then compared the frequency of haplotype states between cases and controls using ASSHAP (Druet *et al.* unpublished). ASSHAP uses a generalized mixed model including a random polygenic effect (to account for population stratification) for which the variance-covariance matrix is proportionate to genome-wide identity-by-state (computed on the basis of hidden haplotype states).

Results of haplotype-based mapping can show significant association signals in cis, due to an enrichment in one or several haplotypes in horses having the CNV, suggesting they had received the variant from a common ancestor. Signals in trans can also be detected for duplication and they can result from transposition of DNA from a chromosome to another or from genomic missassemblies. For example, one signal (CNV67) was probably due to an assembly error (the region is limited by two gaps, the orthology with other species is broken, and a human orthologous segment on chr 1 for the equine chr 29 flanks the orthologous segment of equine chr 30).

Mendelian inheritance of CNVR

We used 8 available trios (sire-dam-offspring) to verify whether a CNV in an offspring was detected in at least one of the parents, and 74 offspring of 52 single parents to verify whether

CNVRs detected in parents were transmitted to offspring.

Association of CNV with RLN

Association between CNVR genotype and RLN was done by comparing the number of cases and controls across the four breeds with CNVR copy number differing from two using Fisher’s exact test.

Serial translocation by means of circular intermediates underlies colour sidedness in cattle

K. Durkin, W. Coppieters, C. Dröemuller, N. Ahariz, N. Cambisano, T. Druet, C. Fasquelle, A. Haile, P. Horin, L. Huang, Y. Kamatani, L. Karim, M. Lathrop, S. Moser, K. Oldenbroek, S. Rieder, A. Sartelet, J. Sölkner, H. Stalhammar, D. Zelenika, Z. Zhang, T. Leeb, M. Georges and C. Charlier

Abstract

Colour sidedness is a dominantly inherited phenotype of cattle characterized by the polarization of pigmented sectors on the flanks, snout and ear tips (Olson 1999). It is also referred to as 'lineback' or 'witrik' (which means white back), as colour-sided animals typically display a white band along their spine. Colour sidedness is documented at least since the Middle Ages and is presently segregating in several cattle breeds around the globe, including in Belgian blue and brown Swiss (Olson 1999; Porter and Mason 2002).Here we report that colour sidedness is determined by a first allele on chromosome29 (Cs_{29}),which results from the translocation of a 492-kilobase chromosome 6 segment encompassing *KIT* to chromosome29, and a second allele on chromosome6 (Cs_6), derived from the first by repatriation of fused 575-kilobase chromosome 6 and 29 sequences to the *KIT* locus. We provide evidence that both translocation events involved circular intermediates. This is the first example, to our knowledge, of a phenotype determined by homologous yet non-syntenic alleles that result from a novel copy-number-variant-generating mechanism.

Nature 482 (2012) 81-84.

Results

To gain insights into the molecular basis of colour sidedness (Fig. 1) , we genotyped 21 colour-sided and 30 control Belgian blue animals with a custom-made 50K single nucleotide polymorphism (SNP) array (Charlier *et al.* 2008). As a result of segregation at the roan locus, Belgian blue animals are either black spotted ($r^{Bl}r^{Bl}$), blue spotted ($r^{Bl}r^{Wh}$) or white ($r^{Wh}r^{Wh}$) (Charlier *et al.* 1996). As white is epistatic to colour sidedness, we selected non-white control animals. We assumed autosomal dominant inheritance (Cs allele) and genetic homogeneity in Belgian blue, and thus scanned the genome of colour-sided animals for a shared haplotype (present in at least one copy) using the ASSDOM software (see Methods). This analysis yielded a single genome-wide significant signal ($P < 0.03$), mapping the Cs locus to bovine chromosome 29 (BTA29) (Fig. 2a and Supplementary Fig. 1). A sire transmitting the colour-sided phenotype to all its pigmented offspring was homozygous for the corresponding haplotype as expected. The shared haplotype spanned 1.9 Mb, and encompassed LUZP2 — not known to be involved in pigmentation — as the only gene. Sequencing the LUZP2 open reading frame (ORF) from colour-sided and control animals did not reveal any protein-sequence-altering variant (data not shown).

In an independent effort, we scanned the genome for copy number variants (CNVs) using a database of 50K SNP genotypes from .4,500 animals of different breeds and the PennCNV software (Wang *et al.* 2007). Intriguingly, this analysis revealed a private duplication encompassing 26 SNPs on chromosome 6 (BTA6), shared by the 21 colour-sided Belgian blue animals (Fig. 2b). We confirmed and refined the boundaries of this ~480 kb duplication by comparative genome hybridization (CGH) of genomic DNA of a Cs/Cs and 1/1 Belgian blue animal on a genome-wide Nimblegen tiling array (Fig. 2c). The corresponding CNV encompassed the *KIT* gene, known to be essential for melanocyte migration and survival (Yoshida *et al.* 2001), and to be associated with coat colour variation (Brooks *et al.* 2007; Marklund *et al.* 1998).

To reconcile these apparently discrepant results, we performed fluorescence in situ hybridization (FISH) with bacterial artificial chromosome (BAC) clones mapping respectively to the BTA29 association interval (labeled with a red fluorophore), and to the BTA6 CNV interval (labeled with a green fluorophore) on lymphocytes of a Cs/ + and a +/+ animal. Overlapping red and green signals on one of the BTA29 homologues of the colour-sided animal demonstrated that the CNV signal resulted from the translocation of the ~480kb BTA6 segment to BTA29 (Fig. 2d). Retrospective examination of the SNP genotypes indicated that on average 3.17 of the 26 SNPs mapping to the ~480 kb duplication could not be called in colour-sided animals, yet the remaining genotypes were sufficient to yield a suggestive association signal (genome-wide P< 0.37), reflecting the sharing of the extra *KIT* haplotype by colour-sided animals (Fig. 2a).

To define the translocation breakpoints, we generated mate-pair libraries from self-ligated, 5 kb DNA fragments of a homozygous Cs/Cs Belgian blue animal, and generated, 10 Gb of sequence on an Illumina GAIIx instrument. We expected two clusters of aberrant mate pairs spanning the left and the right breakpoint resulting from the insertion of an intact BTA6 fragment in a BTA29 segment. We refer to the intact BTA6 fragment as A–B–C–D–E, the BTA29 segment in which it inserts as a–b, and the resulting left and right breakpoints as a–A and E–b, respectively. However, we observed three aberrant mate-pair clusters corresponding respectively to a–D, E–A and C–b fusions (Supplementary Fig. 2). The corresponding topology is most parsimoniously explained by assuming that: (1) the translocated BTA6 fragment circularized (generating fusion E–A); (2) reopened in the C–D interval; and (3) integrated in the a–b BTA29 interval (generating fusions a–D and C–b; Fig. 3).

Using the genomic coordinates of the clustered mate pairs, we designed primer sets to amplify the three corresponding fusion points. Productive amplification was only achieved using genomic DNA from colour-sided Belgian blue animals, as expected (Supplementary Fig. 3A). We sequenced the corresponding PCR products to define the breakpoints at single-nucleotide resolution. The E–A fusion was characterized by 2bp micro-homology typical of non-homologous end joining (NHEJ) (Supplementary Fig. 4A), whereas the a–D and C–b fusions exhibited micro-duplications and micro-deletions reminiscent of replication dependent microhomology-mediated break-induced replication (MMBIR) (Supplementary Fig. 4B) (Hastings *et al.* 2009). All breakpoints mapped to intersperse non-homologous repeat elements (data not shown).

The dominance of the Cs allele is expected to reflect a gain of function resulting from dysregulated expression of the translocated *KIT* gene. To verify the transcriptional competence of the translocated *KIT* copy, we performed polymerase chain reaction with reverse transcription (RT–PCR) experiments using amplicons spanning two SNPs: the ss469414206 G-to-A transition in intron 1, and the ss469414207 C-to-T transition in intron 7. We extracted total RNA from pigmented and unpigmented skin sectors of a Cs/1 colour-sided Belgian blue animal with the GG genotype on BTA6 and the A genotype on BTA29 ($(GG)_6/A_{29}$) for ss469414206 and the $(CC)_6/T_{29}$ genotype for ss469414207. The resulting pre-mRNA-dependent RT–PCR products were directly sequenced and the ratio of T/C and A/G species was estimated using Peakpicker (Ge *et al.* 2005). T/A transcripts accounted for ~33% of the *KIT* output in both pigmented and unpigmented sectors, demonstrating the transcriptional potential of the translocated gene copy. Long-range RT–PCR analysis of the near complete *KIT* mRNA with primers located respectively in exon 1 and the 39 untranslated region (UTR) followed by amplicon sequencing did not show evidence for alternate transcripts in colour-sided animals (SupplementaryFig.5).

Linkage analysis performed in a brown Swiss pedigree segregating for colour sidedness with microsatellite markers targeting candidate genes (*KIT*, *KITL*, *MITF*, *EDNRB*, *ADAMTS20*), yielded a log of odds (lod) score of 6.9 maximizing in the immediate vicinity of the *KIT* locus (Supplementary Table 1). We genotyped four colour-sided (three Cs/Cs and one Cs/1), and five control brown Swiss animals using a 50K SNP array (Charlier *et al.* 2008). Colour-sided Cs/Cs brown Swiss animals indeed shared a 2.0 Mb autozygous haplotype encompassing the *KIT* locus. At first glance, these findings suggested a distinct determinism of colour sidedness in Belgian blue and brown Swiss.

We analyzed the corresponding SNP genotypes with PennCNV and performed CGH using genomic DNA from brown Swiss Cs/Cs versus +/+ animals. This revealed the duplication of a ~120-kb BTA6 segment nested in the Belgian blue duplication, but excluding the *KIT* gene. Intriguingly, it also revealed the duplication of a ~418-kb BTA 29 segment immediately flanking the Belgian blue insertion site and encompassing the last four of twelve LUZP2 exons (b–c; Fig. 4a, b). Moreover, fusion point C–b (but not a–C and D–A), specifying the Belgian blue Cs allele, could be amplified by PCR from genomic DNA of all examined brown Swiss colour-sided but not control animals (Supplementary Fig. 3). These findings established a clear link between the Belgian blue and brown Swiss Cs alleles.

We performed FISH analysis on lymphocytes of a Cs/+ brown Swiss animal using BTA6 and BTA29 BAC clones. Remarkably, we observed overlapping red and green signals on one of the BTA6 homologues, hence revealing the translocation of a BTA29 fragment on BTA6 (Fig. 4c). We generated mate-pair libraries from self-ligated ~5-kb and ~2-kb DNA fragments of a brown Swiss Cs/Cs animal and generated ~15 Gb of sequence on an Illumina GAIIx instrument (Supplementary Fig. 2). Analysis of the resulting sequence traces revealed two aberrant mate-pair clusters. The first corresponds to the Belgian blue C–β fusion point, previously detected by PCR. The second corresponds to a novelβ–γ fusion point (Supplementary Fig.2). The most parsimonious explanation accounting for all the data is that the brown Swiss Cs allele derives from the Belgian blue Cs allele by (1) excision of the B-C-β–γ fragment from the Belgian blue Cs allele on BTA29, (2) circularization, and (3) re-integration in the wild-type *KIT* locus by homologous recombination. This would result in a novel Cs allele mapping to the *KIT* locus and characterized by tandem duplicates of the B–C segment flanking the translocated β–γ BTA29 fragment.(A-B-**C**-β–γ-**B**-C-D-E;Fig.3)

Our model predicts that the C and B BTA6 segments immediately flanking β–γ carry the same haplotype as the Cs-allele of Belgian blue. To test this, we developed long-range PCR assays that would specifically amplify B, C, B and C segments from genomic DNA of a Cs/Cs brown Swiss animal, and – by sequencing the corresponding amplicons – determined the genotype of B and B and C and C for four SNP positions heterozygous in the high-throughput sequence reads of the brown Swiss Cs/Cs animal. We then determined the genotype of the Belgian blue Cs allele for the corresponding variants and showed that it matched the C and B segments of the brown Swiss Cs allele, in agreement with our model of homologous – recombination – dependent resolution of the circular intermediate (Fig. 3 and Supplementary Fig. 6).

Using the mate-pair genomic coordinates, we designed primer pairs to amplify the γ-**B** fusion point. As expected, productive amplification was only achieved using genomic DNA from colour-sided brown Swiss animals (Supplementary Fig. 3). We sequenced the corresponding PCR products to define the breakpoints at single-nucleotide resolution. The γ-**B** fusion presented hallmarks typical of microhomology mediated break induced replication (MMBIR) (Supplementary Fig. 4).

We obtained genomic DNA from colour-sided animals from seven additional cattle breeds and domestic yaks (Supplementary Fig. 7), which we tested by PCR for the presence of the two Belgian-blue-specific fusion points (E-A, α-D), the brown-Swiss-specific fusion point (γ-**B**), and the Belgian blue/brown Swiss shared fusion point (**C**-β). Colour-sided Dutch witrik and Ethiopian fogera animals were shown to carry the Belgian blue Cs allele (Cs_{29}), Austrian pustertaler sprinzen, Czech red-spotted cattle and French vosgienne the brown Swiss Cs allele (Cs_6), and Irish moiled, Swedish mountain and domestic yak carried both the Cs_{29} and Cs_6 alleles (Supplementary Fig. 3). We assume that Cs_{29} and Cs_6 alleles were introgressed in yak after domestication via well-documented hybridization of *Bos taurus* and *Bos grunniens*. These findings indicate that the Cs_{29} and Cs_6 alleles account for most if not all colour sidedness in cattle.

Analysis of colour sidedness has revealed a novel CNV-generating translocation mechanism involving circular intermediates. Whether this is a bovine idiosyncrasy or a more common mechanism remains to be determined. That some CNVs reflect translocation events is well established. As an example, 75 probable dispersed duplications in the human genome have been reported (Conrad *et al.* 2010), as well as at least four interchromosomal duplications (Liu *et al.* 2009, 2010). We ourselves performed genome-wide association mapping between the SNP and CNV genotype in human and cattle, and observed 21 and 4 putative ‘trans-associations’ (that is, non-syntenic CNV-defining SNPs and associated SNPs), respectively (Supplementary Material). Some of these dispersed duplications might involve circular intermediates. In support of the occurrence of other instances involving circular intermediates are recent observations of repeated translocation of five clustered ORFs in wine and bioethanol strains of *Saccharomyces cerevisiae*, apparently via resolution of circular intermediates (Borneman *et al.* 2011).The same mechanism may contribute to somatic mutations intumours. Indeed, episomes with theNUP214–ABL1fusiongene (observed in ,6% of T-cell acute lymphoblastic leukaemia) have been proposed to result from circularization and excision of a chromosome 9 segment bounded by the NUP214 and ABL1 genes, and to reintegrate ectopically by the same resolving mechanism proposed for colour sidedness in at least some patients (Graux *et al.* 2009). The repatriation of exogenous sequences (including exons) back to the original chromosomal location via circular ‘shuttling’ intermediates suggests that this mechanism might underlie a specific mode of exon shuffling.

Methods summary

A custom 50K SNP array (Charlier *et al.* 2008) was used to genotype 21 colour-sided and 30 control Belgian blue animals. The genome of colour-sided animals was scanned for a shared haplotype using the ASSDOM software. In the brown Swiss, microsatellites adjacent to five candidate genes were genotyped in three half-sibling families and two point linkage calculated using the Merlin software (Abecasis *et al.* 2002). CNVs were identified in the 50KSNParraydata using the PennCNV software (Wang *et al.* 2007). Array CGH was carried out on a custom 2.1 M oligonucleotide array (Roche-Nimblegen) with a non-colour-sided Belgian blue used as the reference in each hybridization. Metaphase spreads were generated from short-term lymphocyte cultures. BACs from the duplicated regions on BTA6 and BTA29 were identified using end sequences from the bovine RPCI42 BAC library. BACs were labelled with the appropriate fluorochrome by nick translation (Abbott Molecular), hybridized to the metaphase spreads and examined by fluorescent microscopy. Mate-pair libraries were generated using the Illumina mate-pair library kit v.2 for a Cs/Cs Belgian blue and a Cs/Cs brown Swiss animal. A paired-end library was also generated for a Cs/+ Belgian blue using the Illumina paired-end kit. Sequencing was carried out on an Illumina GAIIx instrument. PCR products spanning regions of interest were purified using QIAquick PCR purification kit (Qiagen) sequenced using Big Dye terminator cycle-sequencing kit v.3.1 (Applied Biosystems) and run on an ABI PRISM 3730 DNA analyser (Applied Biosystems). RNA was extracted form skin using the RNeasy fibrous tissue mini kit (Qiagen) and cDNA was synthesized using SuperScript II first-strand Synthesis SuperMix (Invitrogen). Long-range PCR was carried out using the Expand Long Template PCR System (Roche).

Methods

Association and linkage mapping of the colour-sided locus. SNP genotyping was conducted using custom-made 50K Infinium SNP arrays (Charlier *et al.* 2008) used according to the instructions of the manufacturer. Microsatellite genotyping was conducted as previously described (Drögemüller *et al.* 2009). Association analysis was conducted using ASSDOM. ASSDOM searches for chromosome segments devoid of SNPs for which cases have alternate homozygous genotype (say 11 versus 22), excluding the sharing of an identical-by-descent (single-copy) haplotype. Intervals bounded by such excluding SNPs receive a score corresponding to $\sum_{i=1}^{k} log(1-p_i^2)^n$, where p_i is the frequency of the allele missing among *n* cases, estimated in m controls. The genome-wide statistical significance of the ‘non-exclusion’ signal is determined by phenotype permutation of the disease status between the n cases and m controls. Two-point linkage analyses were conducted with Merlin (Abecasis *et al.* 2002).

Prediction of CNVs from SNP genotype data. The log R ratio signal intensity and B allele frequency from a custom 50K SNP array (Charlier *et al.* 2008) were obtained using Illumina BeadStudio software. PennCNV, a hidden Markov model based approach that takes into account signal intensity, allelic intensity ratio, distance between markers and allele frequency (Wang *et al.* 2007) was used to call CNVs. Regions of the genome that showed evidence of copy number change were inspected in greater detail. Plots of log R ratio and B allele frequency were examined in BeadStudio (Illumina) and the region checked in the University of California, Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu/).

Detection of CNVs by CGH. Array CGH was carried out on a custom 2.1 M oligonucleotide array (Roche-Nimblegen) based on the UMD3.0bovineassembly. The array contained 2,152,422 probes (50–75 mers) with a median spacing of 1,160bp. The reference animal used in hybridizations was a non-colour-sided Belgian blue. Genomic DNA labelling (Cy3 for sample and Cy5 reference), hybridization and washing were performed according to the manufacturer’s instructions and have been described elsewhere (Selzer *et al.* 2005). Slides were scanned using a GenePix 4000B 5mm microarray scanner (Axon Instruments). Images were processed using NimbleScan software (Roche-Nimblegen). Spatial correction was applied and data normalized[19], segmentation was performed using the DNACopy algorithm

(Olshen *et al.* 2004). The $\log_2$ ratios for each oligonucleotide were also examined visually for evidence of a change in copy number in regions of interest.

FISH. Peripheral blood was obtained from colour-sided and wild-type Belgian blue and brown Swiss animals. Pokeweed-stimulated lymphocyte cultures were established and chromosome spreads prepared for colour-sided and wild-type animals in each breed following standard cytogenetic procedures. End sequences from the RPCI42 bovine BAC library and the duplicated regions on BTA6 and BTA29 in the colour-sided Belgian blue and brown Swiss were downloaded from the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm. nih.gov/). BLAST was used to identify end sequences located in the duplicated regions. The BACs 160M9 and 156I13 overlap and cover the region chr6:72,566,605–72,817,995 (bosTau4), while the BACs 37P11 and 116G8 also overlap and cover the region chr29:20,772,406–21,035,251 (bosTau4). BAC clones were initially cultured at 37uC in 1 ml of 2YT media containing 30 ul ml^{-1} chloramphenicol. The cultures were plated on lysogeny broth (LB) agar plats with 30 ul ml^{-1} chloramphenicol to obtain single colonies, the identity of the BACs was confirmed by PCR using primers designed within the area encompassed by the respective BAC (Supplementary Table 2). The positive clones were then used to inoculate 100ml of 2YT media with 30 ul ml^{-1} chloramphenicol. Following 24 h at 37℃ with constant agitation, DNA was extracted using the Qiagen midiprep kit, following the manufacturer’s instructions. The DNA was labelled with the nick translation kit from Abbott Molecular, using the manufactures protocol. DNA from BACs 160M9 and 156I13 was mixed and labelled with spectrum green (Abbott Molecular) while 37P11 and 116G8 were mixed and labelled with spectrum orange (Abbott Molecular). Labelled DNA (100 ng) was combined with 1 ug of bovine Cot-1 DNA and 2 mg of bovine genomic DNA, precipitated then resuspended in 3 ul of purified water and 7 ul of hybridization buffer (Abbott Molecular). The separate probes were denatured at 73 ℃for 5 min and then combined on a slide containing metaphase spreads. These slides had been denatured in 70% formamide in 2 x SSC at 75℃ for 5 min followed by dehydration in ethanol. A coverslip was secured with rubber cement and the slide incubated overnight at 37 ℃ in a humidified chamber. Slides were then washed in 0.43 SSC/0.3% Tween-20 at 73℃ for 1–3 s followed by washing in 2 x SSC/0.1% Tween-20 at room temperature (~18℃) for 1–3 s and air dried. The slides were counterstained with DAPI II (Abbott Molecular) and visualized by fluorescent microscopy.

Next-generation sequencing of mate-pair and paired-end libraries. Mate-pair libraries with different insert sizes were generated using the Illumina mate-pair library kit, v.2. The manufacturer’s instructions were followed except for the step involving fragmentation of the circularized DNA, for which a bioruptor sonicator UCD-200 (Diagenode) was used. The 300 ml sample was placed in a 1.5 ml Eppendorf tube and sonicated for 8 min with the instrument set to high and a cycle of 30s on and 30s off. A 5-kb insert library was generated for a Cs/Cs Belgian blue animal and a 2-kb and 5-kb for a Cs/Cs brown Swiss animal. A paired-end library with a 400 bp insert size was also generated for a Cs/1Belgianblue animal using the Illumina paired-end kit, following the instructions of the manufacturer. The resulting libraries were quantified using Pico-Green (Quant-it, Invitrogen) and the Agilent 2100 Bioanalyzer High Sensitivity DNA kit (Agilent Technologies). Sequencing was carried out on an Illumina GAIIx instrument. Mapping of the 36bp from each end of the mate-pair libraries and the 110 bp from the ends of the paired-end library was performed using the BWA tool (Li and Durbin 2009). Breakpoints were identified by visually inspecting the mate pairs using the integrative genomics viewer (James *et al.* 2011) in the ~2-Mb region surrounding the Cs-specific duplications in Belgian blue and brown Swiss animals and looking for discordant mate pairs.

PCR amplification of translocation breakpoints. PCR primers were designed to span each of the breakpoints identified by mate-pair sequencing. The primers were tested on genomic DNA from colour-sided and wild-type animals. PCR products were visualized on a 2% agarose gel. Primers with amplification confined to colour-sided animals were purified using the QIAquick PCR purification kit (Qiagen), where multiple bands were observed the relevant band was excised and purified using the QIAquick gel extraction kit (Qiagen). The fragments were then sequenced using Big Dye terminator cycle-sequencing kit v.3.1 (Applied Biosystems) with the purified reaction run on a ABI PRISM 3730 DNA analyser (Applied Biosystems). Primers used are listed in Supplementary Table 2.

Analysis of Cs$_{29}$-derived *KIT* transcripts. To ensure that no mutations were present in the coding sequence of the Cs$_{29}$-specific *KIT* gene, primers were designed to amplify all the exons and the 39 UTR from genomic DNA (Supplementary Table 2). The resulting PCR products were sequenced as outlined above, which did not reveal any protein altering DNA sequence variant. Toexamine expression of *KIT* from the Cs $_{29}$ allele, a small biopsy of skin was removed from the back (white skin) and side (pigmented skin) of a colour-sided Belgian blue animal (the relevant ethical procedures were adhered to). The samples were immediately frozen in liquid nitrogen and stored at -80 ℃ untilRNA extraction. The tissue was homogenized using a Tissue Lyser (Qiagen), total RNA was extracted using the RNeasy fibrous tissue mini kit (Qiagen), following the manufacturer’s instructions. First-strand cDNA was synthesized using SuperScript II first-strand Synthesis SuperMix (Invitrogen) with a mixture of random hexamers and oligo (dT) primers. Genomic DNA was extracted from whole blood from the same animal using standard phenol-chloroform extraction. The intronic regions of the *KIT* gene were searched for SNPs using the mate-pair sequences produced from the Cs/Cs Belgian blue animal. Primers were designed around suitable SNPs and sequenced in the tissue donor. To establish the genotype of the Cs $_{29}$ allele, multiple colour-sided animals were sequenced. For SNP ss469414207(C/T) 40 colour-sided animals were sequenced. For all Cs/Cs animals, the C and T alleles had equal peak height, while in the remaining Cs/ + animals the T allele produced the smaller peak. Additionally seven non colour-sided Belgian blue animals were also sequenced and none possessed the T allele. For SNP ss469414206 (G/A), two Cs/Cs and three Cs/+ animals were sequenced and showed a pattern of peak heights consistent with the Cs $_{29}$ allele having the A genotype. To determine if the Cs$_{29}$ allele was expressed, the cDNA from the skin biopsies was amplified using primers spanning both SNPs and sequenced as outlined above. Moreover, and to ensure the integrity of Cs$_{29}^{-}$ derived *KIT* transcripts, primers located in the first exon and the 3'UTR (Supplementary Table 2) were used with the Expand Long Template PCR System (Roche) to amplify the near full-length *KIT* cDNA. The product was run on 1% gel and examined for evidence of alternative splicing.

Genotyping the duplicated B and C segments of the Cs$_6$ allele. Long-range PCR was carried out using the Expand Long Template PCR System (Roche). The genomic DNA used as template was extracted using QIAamp DNA Mini columns (Qiagen) following the manufacturer's instructions to produce high molecular weight DNA. For each reaction the following mix was prepared: 2 ul Buffer 1, 140uM dNTP, 120nM upstream and downstream primer, 0.3 ul enzyme mix and 100ng genomic DNA, final volume was 20ul. Extension time and the thermal profile recommended by the manufacturer was followed. Following 30 cycles 10 ul of the product was run on a 0.8% agarose gel to check for amplification, the remaining reaction mix was retained. The PCR primers spanning the relevant SNP where then used in a conventional PCR reaction using between 0.1 and 0.5 ul (amount determined by intensity of band) of the long-range product as template. When long-range PCR product was used as template the number of cycles was reduced to 20. Clean-up and sequencing was carried out as outlined above. Used primers are reported in Supplementary Table 2.

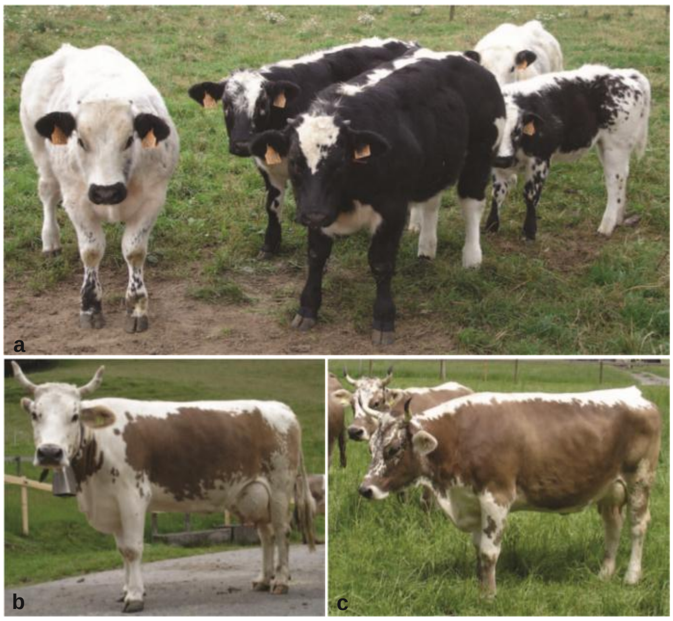


Figure 1 Colour-sided Belgian blue and brown Swiss animals. a, represented Belgian blue animals are heterozygous Cs/+. In addition to the effect of the Cs locus, which causes polarization of pigmented sectors to the flanks, ears and snout, Belgian blue animals exhibit considerable polygenic variation in the degree of white spotting. The colour-sided phenotype of the two nearly completely white animals (extensive degree of white spotting) is apparent from the pigmentation of ear tips and snout. **b, c,** In brown Swiss, which are generally devoid of white spotting, homozygous Cs/Cs (**b**) and heterozygous Cs/+ (**c**) animals differ by the extent of pigmentation, explaining why colour sidedness is also referred to as semi-dominant.

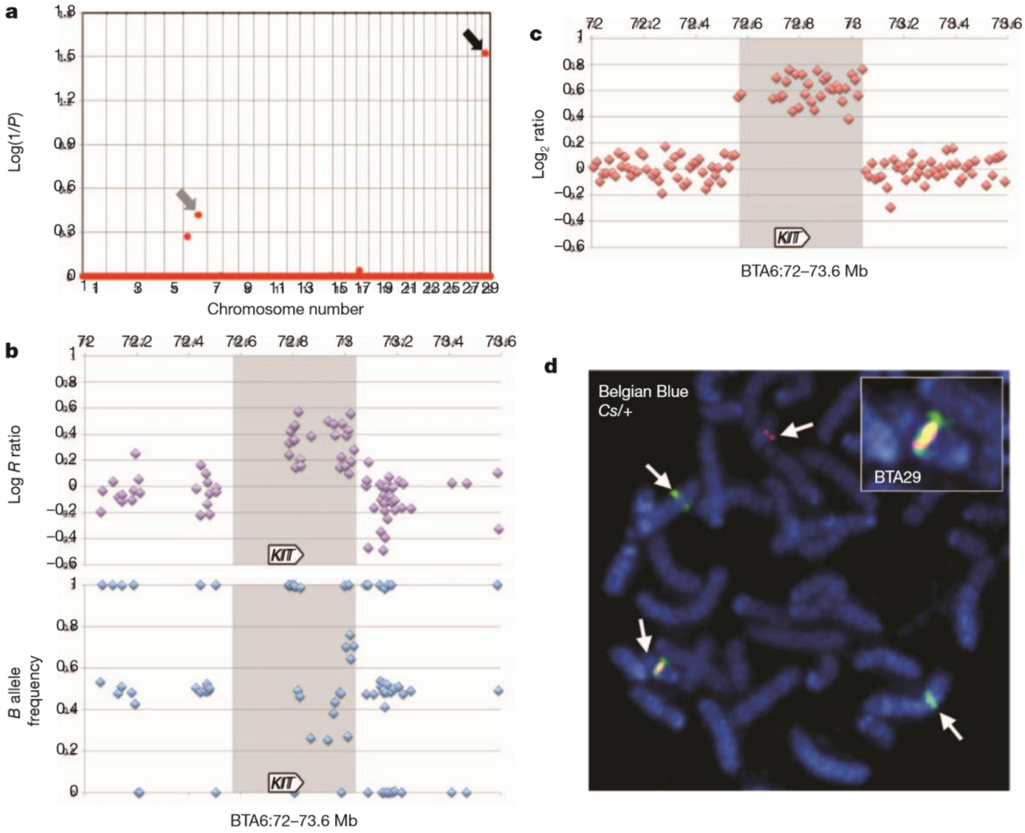


Figure 2 Identification and mapping of the colour-sided locus in the Belgian blue. a, Genome-wide association mapping using ASSDOM (see Methods), revealing a genome-wide significant association (P< 0.03) on BTA29 (black arrow) and suggestive association on BTA6 (grey arrow). Vertical lines separate chromosomes. **b**, Detection of a duplication on theBTA6 segment shared by 21 colour-sided Belgian blue animals using 50K SNP array data (Charlier *et al.* 2008) and PennCNV (Wang *et al.* 2007). **c**, Confirmation and boundary definition using CGH on a Nimblegen bovine tiling array. Shaded areas in **b** and **c** mark the boundaries of the CNV as defined by CGH. **d**, Demonstration by FISH of the translocation of a *KIT*-encompassing BTA6 segment onto BTA29 in a heterozygous Cs/+ Belgian blue animal. Magnification, X100.

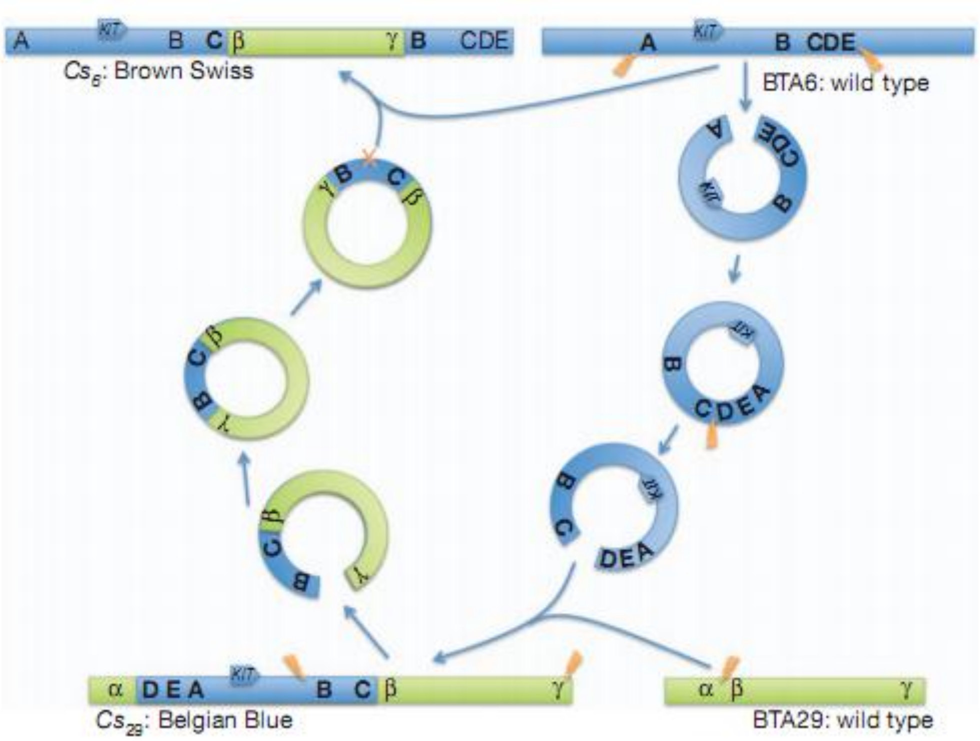


Figure 3 Model for the generation of the colour-sidedCs$_{29}$andCs$_{6}$alleles by serial translocation via circular shuttling intermediates.

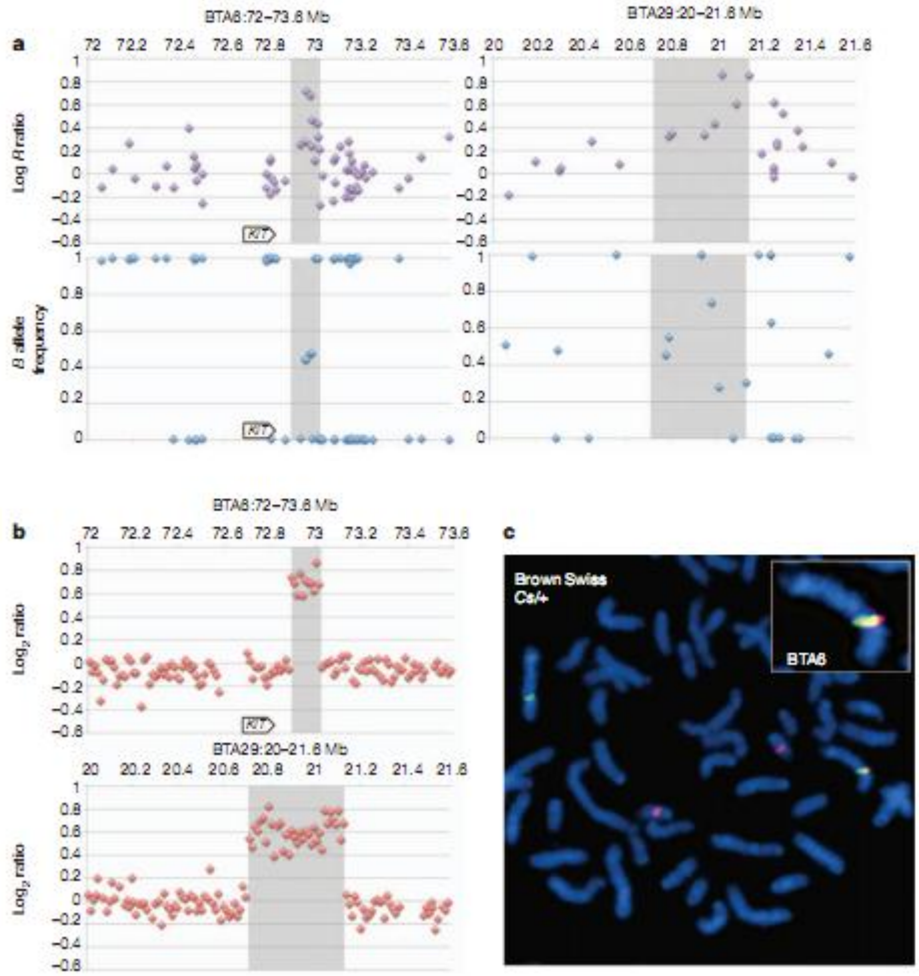
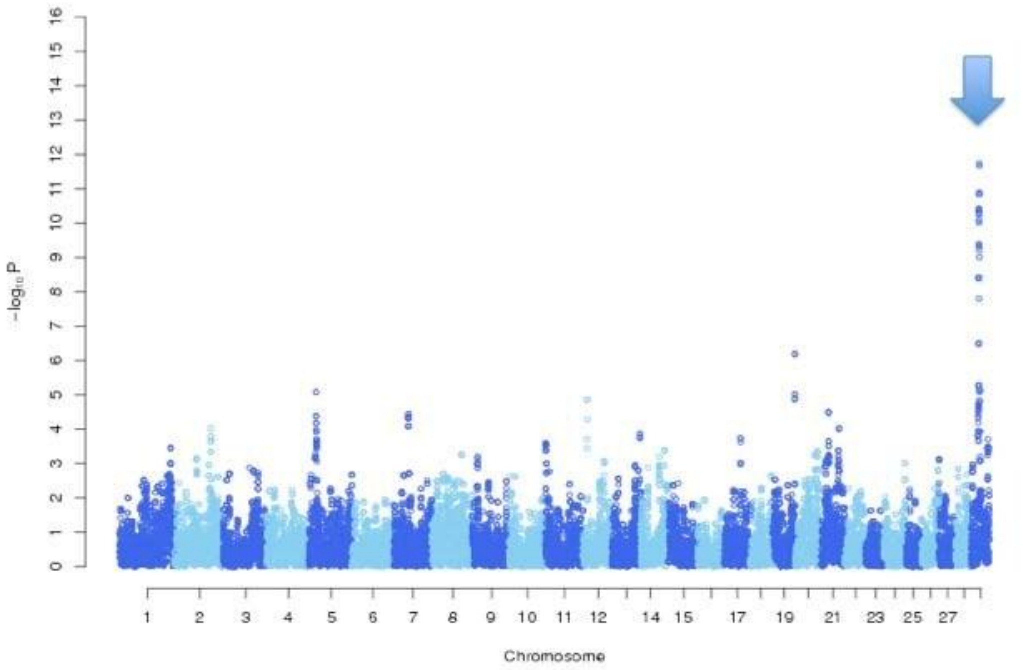


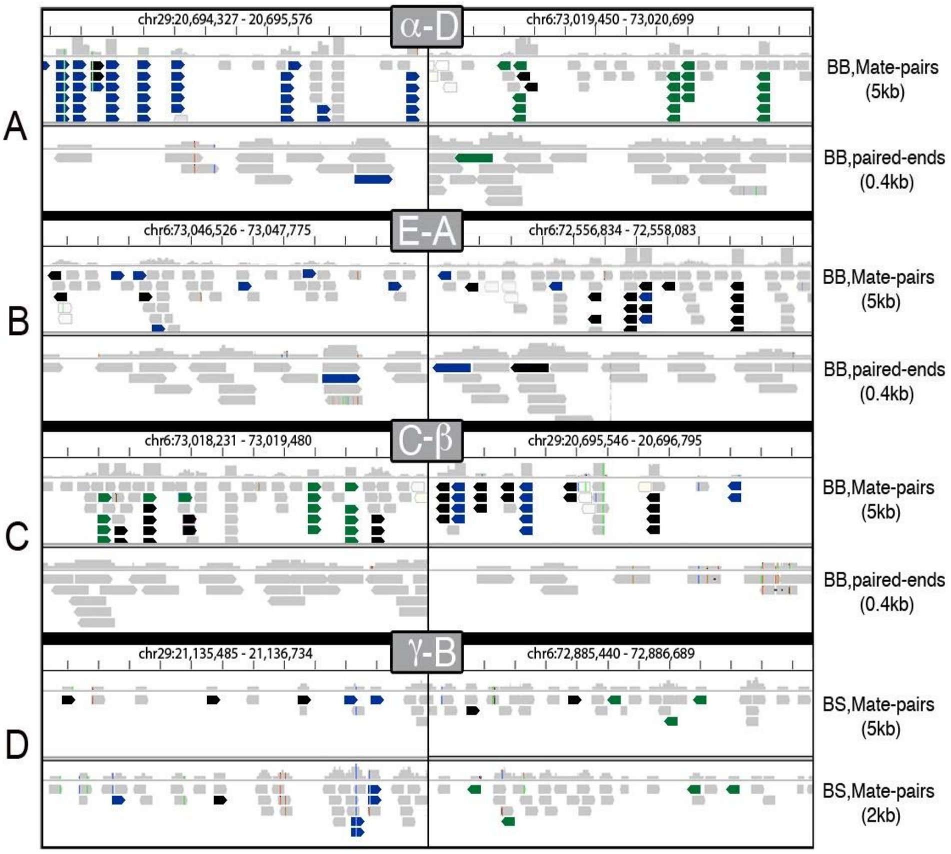
Figure 4 Identification and mapping of the colour-sided locus in brown Swiss animals. a, PennCNV evidence for BTA6 and BTA29 duplications in brown Swiss colour-sided animals. The slightly different distribution of BTA6 SNPs when compared to Fig. 2 is due to the use of a different version of the 50K SNP array (Olson 1999). **b**, Confirmation and boundary definition of the Cs$^-$ associated BTA6 and BTA29 duplication in brown Swiss animals using CGH on Nimblegen bovine tiling arrays. The shaded areas in **a** and **b** mark the boundaries of the CNV, as defined by CGH. **c**, Demonstration by FISH of the translocation of a BTA29 segment onto BTA6 in a heterozygous Cs/ + brown Swiss animal.



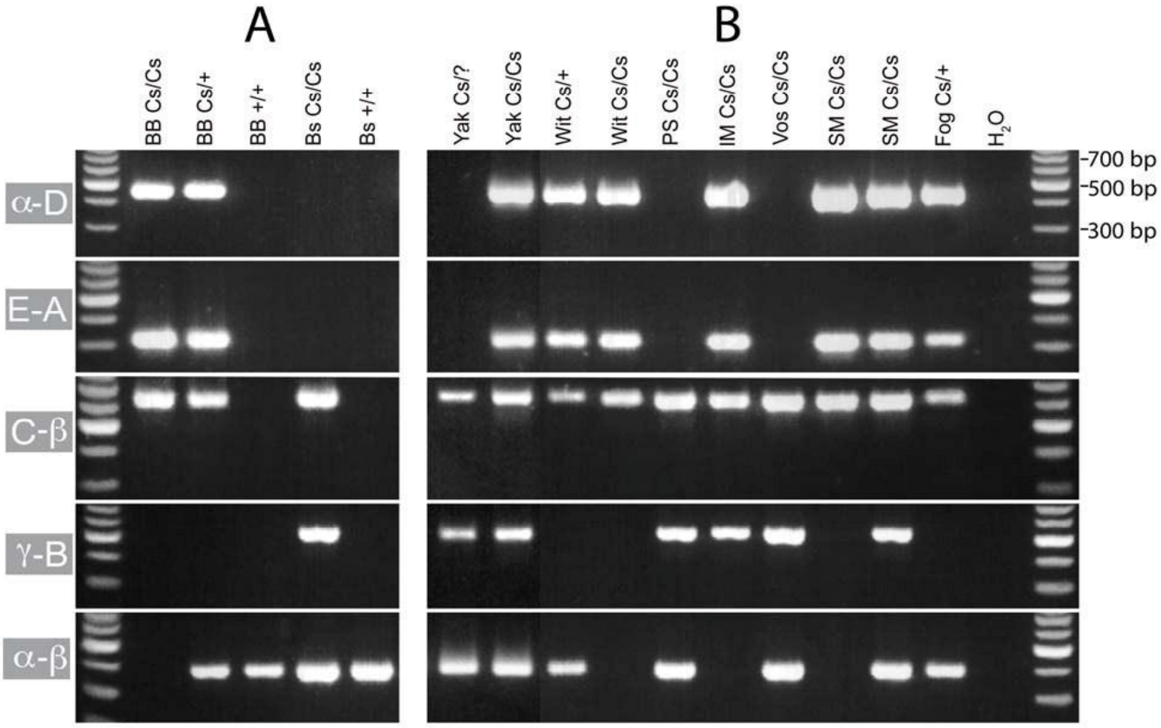
Supplementary Fig. 1:

Association mapping of the *Cs* locus conducted with Glascow (Druet *et al.* in preparation) in the Belgian Blue Cattle population using 21 color‐sided animals and 30 non‐white controls. SNP genotypes were first phased using the Hidden Markov Model described by Druet and Georges (2010) and implemented with Dualphase. Dualphase assigns haplotypes to a defined number of Hidden States corresponding to clusters of genealogically related chromosomes. Glascow then searches for association between haplotype states and a binary trait. It uses a Generalized Linear Mixed Model with the logit function as inverse link function to transform an underlying linear variable to a binary trait. The model includes a random haplotype effect (with zero covariance between the effects of distinct haplotypes) as well as a random polygenic effect (with variance‐covariance matrix proportionate to genomewide SNP identity‐by‐state) to correct for stratification. Following Verbeke & Molenberghs (2003) and Tzeng & Zhang (2007) , we evaluated the strength of association using a score test. Under the null hypothesis, the score test has an approximate gamma distribution. We empirically determined the mean and variance of the gamma distribution at each marker position from 1,000 permutations of the residuals (observations corrected for modeled effects).

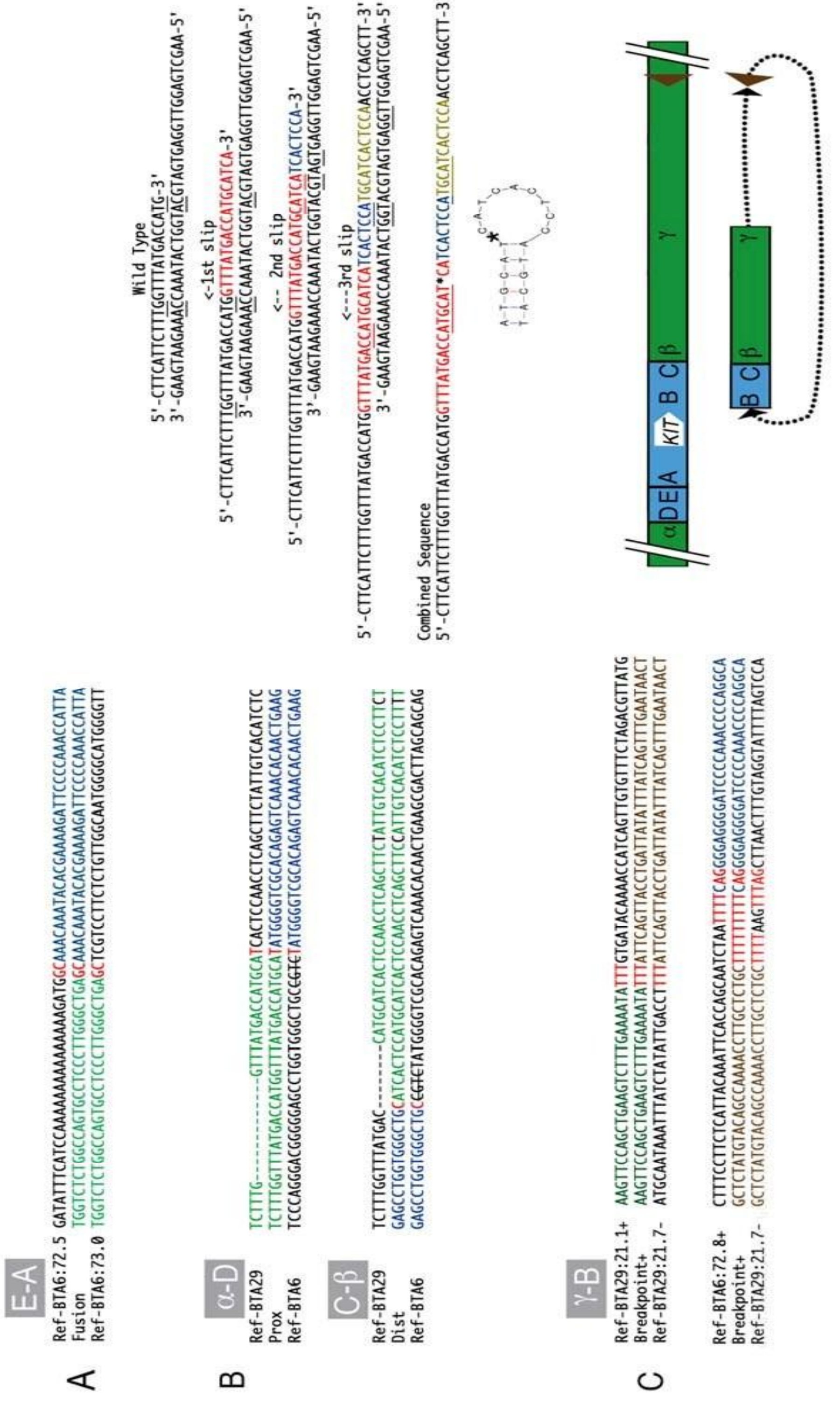
The graph shows the location scores (expressed as –log(p)) that were obtained using Glascow across the genome. The genome‐wide significant signal on chromosome 29 is highlighted by the arrow.


A
chr29:20,694,327 - 20,695,576
α-D
chr6:73,019,450 - 73,020,699
BB,Mate-pairs (5kb)
BB,paired-ends (0.4kb)
B
chr6:73,046,526 - 73,047,775
E-A
chr6:72,556,834 - 72,558,083
BB,Mate-pairs (5kb)
BB,paired-ends (0.4kb)
C
chr6:73,018,231 - 73,019,480
C-β
chr29:20,695,546 - 20,696,795
BB,Mate-pairs (5kb)
BB,paired-ends (0.4kb)
D
chr29:21,135,485 - 21,136,734
γ-B
chr6:72,885,440 - 72,886,689
BS,Mate-pairs (5kb)
BS,Mate-pairs (2kb)

Supplementary Fig. 2: Integrated Genome Viewer (IGV) (James *et al.* 2011) screen‐captures showing aligned mate‐pair (~36‐bp) or paired‐end (~110‐bp) reads around the four breakpoints associated with CS alleles (α‐D, A‐E, **C**-β and γ‐**B**; cfr. Fig. 3). Reads from aberrantly mapping pairs are color‐coded to indicate the chromosome to which the other pair maps (blue: = BTA6; green: BTA29; black: any other chromosome). The vertical line marks the position of the four breakpoints. Positions correspond to bosTau4 assembly.

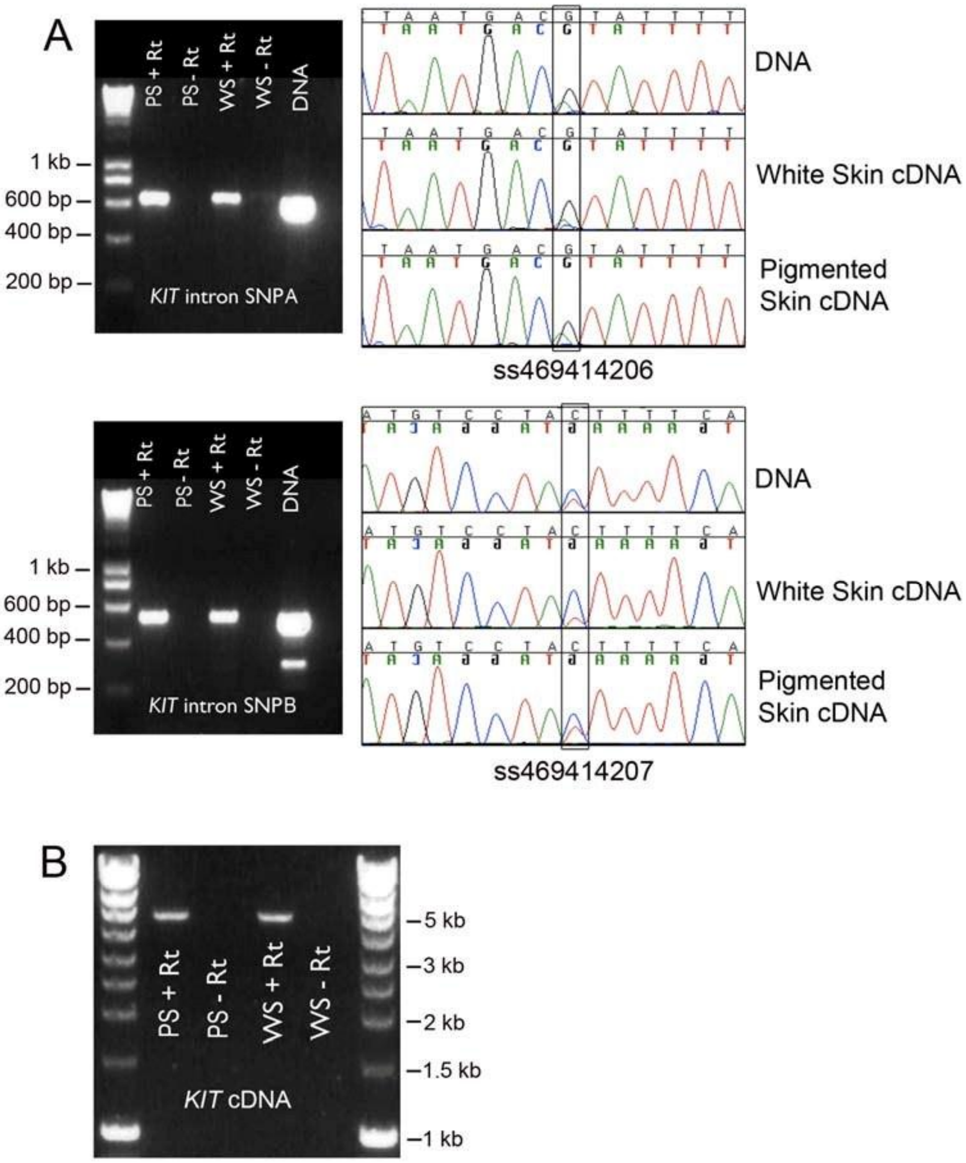


Supplementary Fig. 3: PCR amplification and agarose gel electrophoresis of amplicons spanning the breakpoints associated with color‐sidedness (α‐D, A‐E, C‐β, γ‐B and α‐β; cfr. Fig. 3). (**A**) BB: Belgian Blue; Bs: Brown Swiss; (**B**) Yak: domestic yak; Wit: Witrik; PS: Pustertaler Sprinzen; IM: Irish Moiled; Vos: Vosgienne; SM: Swedish Mountain; Fog: Fogera. Data from Czech Red Spotted Cattle not shown.

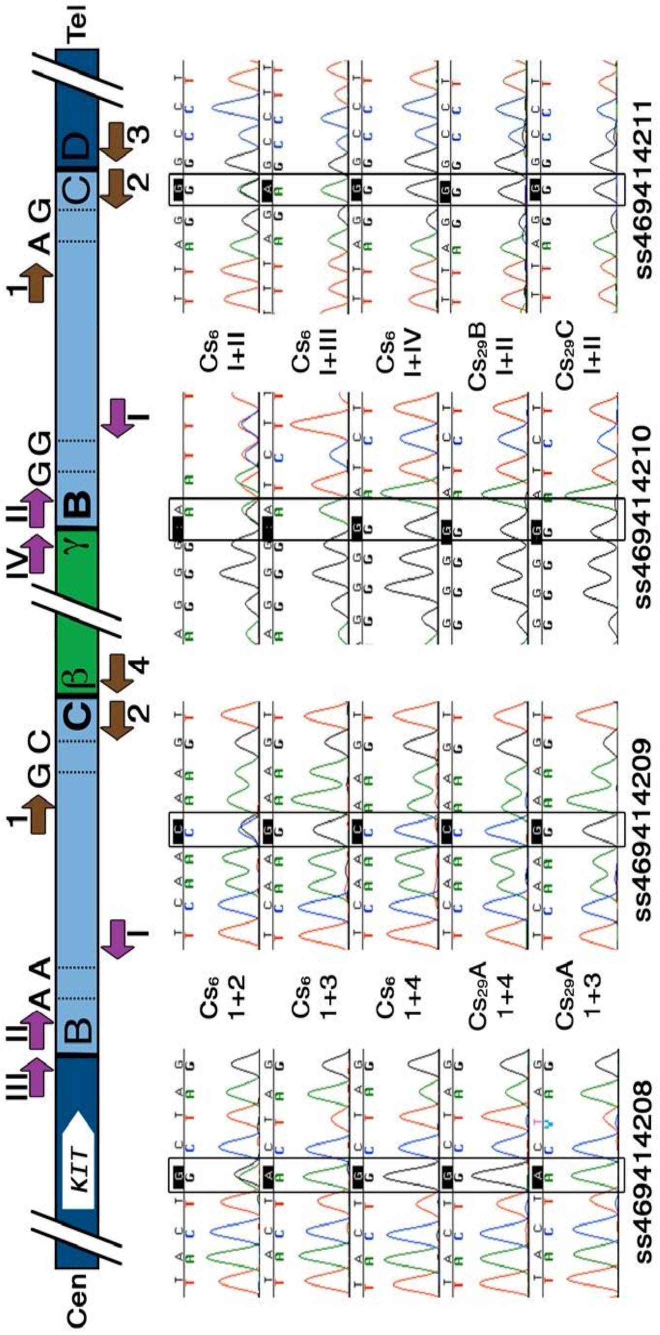

A
E-A
Ref-BTA6:72.5 GATATTTCATCCAAAAAAAAAAAAAAAAGATGGCAAACAAATACACGAAAAGATTCCCCAAACCATTA
Fusion TGGTCTCTGGCCAGTGCCTCCCTTGGGCTGAGCAAACAAATACACGAAAAGATTCCCCAAACCATTA
Ref-BTA6:73.0 TGGTCTCTGGCCAGTGCCTCCCTTGGGCTGAGCTCGTCCTTCTCTGTTGGCAATGGGGCATGGGGTT
B
α-D
Ref-BTA29 TCTTTG-------------GTTTATGACCATGCATCACTCCAACCTCAGCTTCTATTGTCACATCTC
Prox TCTTTGGTTTATGACCATGGTTTATGACCATGCATATGGGGTCGCACAGAGTCAAACACAACTGAAG
Ref-BTA6 TCCCAGGGACGGGGGAGCCTGGTGGGCTGC~~CGTC~~TATGGGGTCGCACAGAGTCAAACACAACTGAAG
C-β
Ref-BTA29 TCTTTGGTTTATGAC--------CATGCATCACTCCAACCTCAGCTTCTATTGTCACATCTCCTTCT
Dist GAGCCTGGTGGGCTGCATCACTCCATGCATCACTCCAACCTCAGCTTCCATTGTCACATCTCCTTTT
Ref-BTA6 GAGCCTGGTGGGCTGC~~CGTC~~TATGGGGTCGCACAGAGTCAAACACAACTGAAGCGACTTAGCAGCAG
C
γ-B
Ref-BTA29:21.1+ AAGTTCCAGCTGAAGTCTTTGAAAATATTTGTGATACAAAACCATCAGTTGTGTTTCTAGACGTTATG
Breakpoint+ AAGTTCCAGCTGAAGTCTTTGAAAATATTTATTCAGTTACCTGATTATATTTATCAGTTTGAATAACT
Ref-BTA29:21.7- ATGCAATAAATTTATCTATATTGACCTTTTATTCAGTTACCTGATTATATTTATCAGTTTGAATAACT
Ref-BTA6:72.8+ CTTTCCTTCTCATTACAAATTCACCAGCAATCTAATTTTCAGGGGAGGGGATCCCCAAACCCCAGGCA
Breakpoint+ GCTCTATGTACAGCCAAAACCTTGCTCTGCTTTTTTTTTCAGGGGAGGGGATCCCCAAACCCCAGGCA
Ref-BTA29:21.7- GCTCTATGTACAGCCAAAACCTTGCTCTGCTTTTAAGTTTAGCTTAACTTTGTAGGTATTTTAGTCCA
Wild Type
5'-CTTCATTCTTTGGTTTATGACCATG-3'
3'-GAAGTAAGAAACCAAATACTGGTACGTAGTGAGGTTGGAGTCGAA-5'
<-1st slip
5'-CTTCATTCTTTGGTTTATGACCATGGTTTATGACCATGCATCA-3'
3'-GAAGTAAGAAACCAAATACTGGTACGTAGTGAGGTTGGAGTCGAA-5'
<-- 2nd slip
5'-CTTCATTCTTTGGTTTATGACCATGGTTTATGACCATGCATCATCACTCCA-3'
3'-GAAGTAAGAAACCAAATACTGGTACGTAGTGAGGTTGGAGTCGAA-5'
<---3rd slip
5'-CTTCATTCTTTGGTTTATGACCATGGTTTATGACCATGCATCATCACTCCATGCATCACTCCAACCTCAGCTT-3'
3'-GAAGTAAGAAACCAAATACTGGTACGTAGTGAGGTTGGAGTCGAA-5'
Combined Sequence
5'-CTTCATTCTTTGGTTTATGACCATGGTTTATGACCATGCAT*CATCACTCCATGCATCACTCCAACCTCAGCTT-3'
A-T-G-C-A-T
T-A-C-G-T-A
C-A-T-C-A-C-C-T-C
*
α DE A KIT B C β γ
B C β γ

Supplementary Fig. 4: (A) Sequence of the E‐A fusion showing the GC micro‐homology at the two breakpoints, typical of replication‐independent non‐homologous end joining (NHEJ) (Hastings *et al.* 2009). **(B)** Sequences of the α‐D and C‐β fusion points showing microduplications and micro–deletions reminiscent of replication‐dependent microhomology‐mediated break‐induced replication (MMBIR)

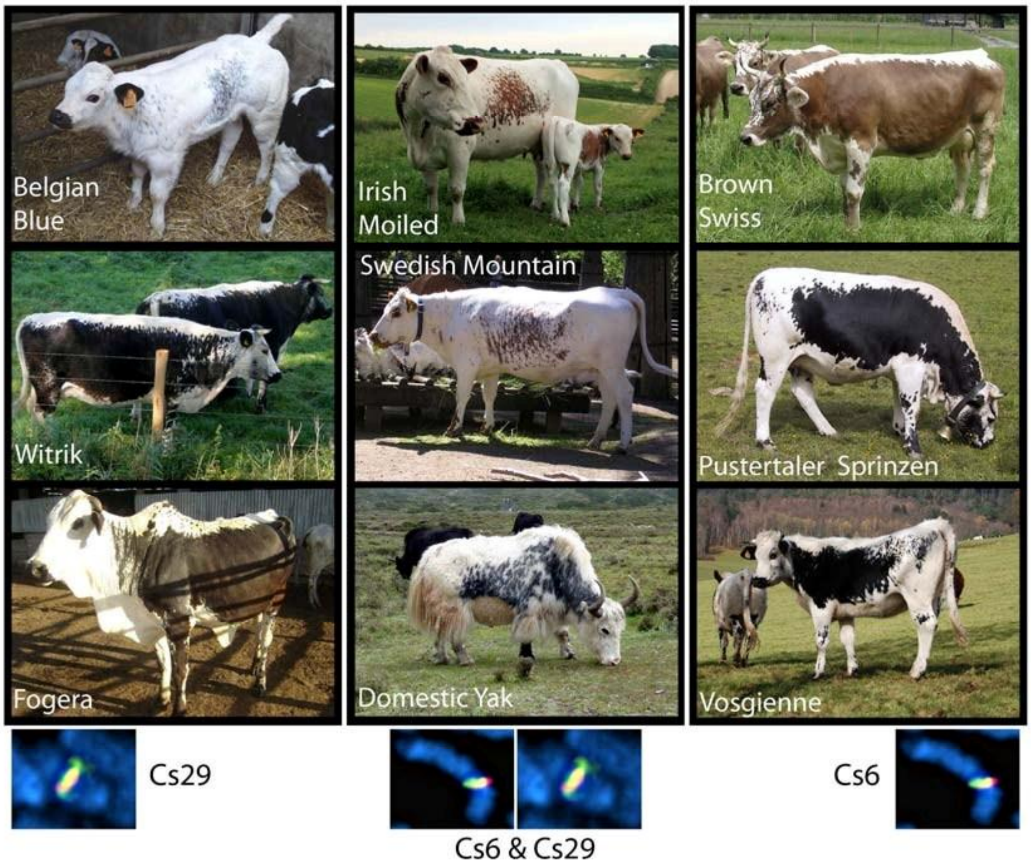
(Hastings *et al.* 2009). Examination of the BTA29 sequences flanking the BTA6 *Cs_{29}* insertion, suggests the occurrence of three slippage events that might have resulted from replication fork stalling and template switching (FoSTeS) (Hastings *et al.* 2009). The resulting sequence is characterized by an inverted palindromic 6‐bp repeat able to form a hairpin loop. The BTA6 sequences inserted at the base of the loop (marked by the asterisk). **(C)** Sequence of the γ‐B fusion showing signatures of replication‐dependent micro‐homology‐mediated break induced replication (MMBIR) including the micro‐insertion (between γ and B) of an inverted 75‐bp sequence originally found 644‐kb downstream of γ, flanked on either side by tracks of micro‐homology with the ends of B and γ respectively.



Supplementary Fig. 5: (A) RT‑PCR analysis of 673‑bp and 533‑bp *KIT* amplicons, encompassing the intron 1 “ss469414206” and intron 7 “ss469414207” SNPs, respectively. Direct sequencing of the RT‑PCR products obtained from RNA extracted from white (WS) and pigmented (PS) skin of a color‑sided animal with $(GG)_6/A_{29}$ (“rsA”) and $(CC)_6/T_{29}$ (“rsB”) genotype reveals BTA29‑derived T and A bearing mRNA molecules demonstrating the transcriptional potential of the *KIT* gene on the Cs_{29} allele. **(B)** Long range RT‑PCR analysis of the near complete *KIT* mRNA generates a product with expected size (4,527‑bp) using total RNA extracted from white (WS) and pigmented (PS) skin of the same color‑sided animal.



Supplementary Fig. 6: Determining the origin of BTA6 DNA flanking BTA29 DNA in Cs_6. Schematic representation of the Cs_6 allele: dark blue: non‐duplicated BTA6 segments (corresponding to A (left) and D‐E (right)); light blue: duplicated portions of BTA6 (B‐**C** and **B**‐C); green: translocated BTA29 segment (β‐γ). Primers 1 & 2 and I & II (amplifying segments C & **C** and **B** & B, respectively) were used to confirm four SNPs (“ss469414208”, “ss469414209”, “ss469414210” and “ss469414211”) identified in the Brown Swiss *Cs/Cs* next generation sequencing reads using conventional PCR followed by Sanger sequencing. Primers 1 and I were then paired (i) with primers 3 and III to amplify (under long range PCR conditions) the B and C segments flanking the non‐duplicated BTA6 sequences, and (ii) with primers 4 and IV to amplify the **B** and **C** segments flanking the BTA29‐derived sequences. Determining the genotype of the *Cs*$_{29}$ allele for segment **C** was achieved by sequencing amplicon 1‐4 obtained from genomic DNA of a Belgian Blue color‐sided animal. Determining the genotype of the *Cs*$_{29}$ allele for segment **B** was achieved by sequencing amplicon I‐II obtained from genomic DNA of multiple Belgian Blue color‐sided animals. The **B** and **C** segments flanking the BTA29 insert (β‐γ) in the *Cs*$_{6}$ allele were thereby shown to match the **B** and **C** segments of the *Cs*$_{29}$ allele, as predicted from the resolution of a circular intermediate by a non‐allelic homologous recombination event in the B‐C interval. Cs$_{6}$ = *Cs*$_{6}$/+ animal; Cs$_{29}$A, Cs$_{29}$B, CS$_{29}$C = *Cs*$_{29}$/+ Belgian Blue animals.



Supplementary Fig. 7: Genotypes of color‐sided cattle from eight breeds as well as color‐sided domestic yaks with regards to the Cs_{29} and Cs_{6} alleles. Both alleles were introgressed from cattle into domestic yak by hybridization. Czech Red Spotted Cattle not shown

Supplementary material

Association mapping of dispersed duplications in the human and bovine genome.

Methods:

Data. We took advantage of genotype data obtained (i) on 856 individuals of Northern European descent genotyped with the Human 660W‐Quad V1 DNA Analysis Beadchip array, (ii) 275 Belgian Blue animals genotyped with the Illumina Bovine 700K HD SNP array, and (iii) 191 Holstein‐Friesian animals genotyped with the Illumina Bovine 700K HD SNP array.

CNV detection: CNVs were detected using PennCNV (Want *et al.* 2007). Overlapping CNVs were merged into CNV‐regions (CNVR). Only CNVR encompassing only duplications (copy number 3 or 4) were retained for further analysis.

Phasing: Genotypes were phased using Dualphase (Druet and Georges 2010). As a result, all chromosomes in the data are assigned to a predetermined number (in this case 20) of Hidden Haplotype Clusters at each SNP position.

Association mapping: To perform association analysis between CNV genotype and SNP genotype, we considered all individuals with PennCNV assigned copy number of 3 or 4 as “cases” and all individuals with PennCNV assigned copy number of 2 as “controls”. The association analysis was conducted using Glascow (described under Suppl. Fig. 1). A random polygenic effect was added to the model for the cow data but not for the human data.

Significance thresholds. We generated QQ‐plots using all p‐values obtained with marker positions that were not syntenic with the SNPs involved in the corresponding CNVR. In both human and bovine, the QQ‐plot revealed a pronounced inflexion at p‐values of 10^{-9}, which was therefore selected as significance threshold (**A**).

Results:

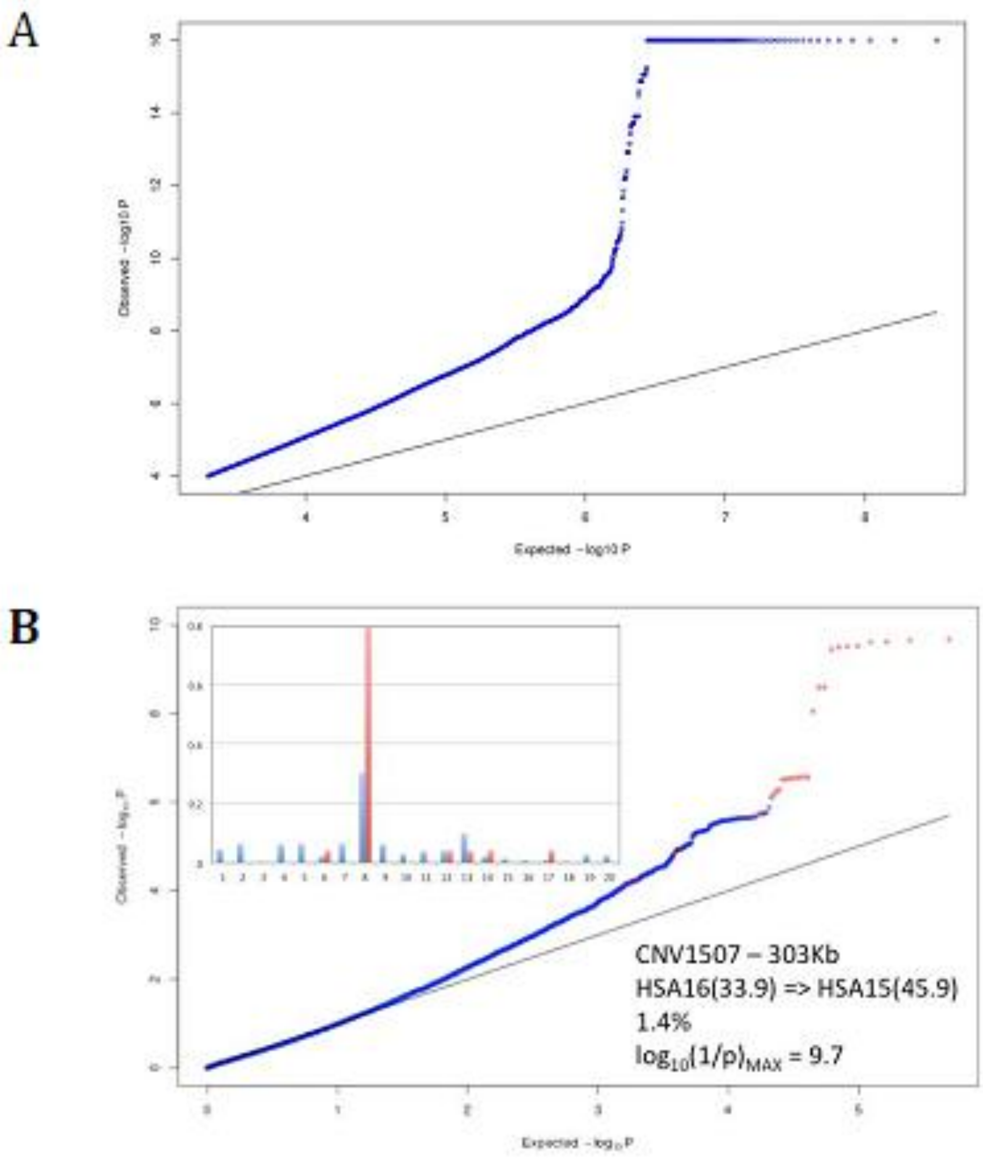
We observed 21 putative trans‐associations with p‐value < 10^{-9} in human and four in the cow. An example of such putative trans‐association is shown in panel (**B**). The main features of the corresponding putative dispersed duplications are summarized in table format.

Summary of putative dispersed duplications identified in human and bovine

	CNVR(egion)					Trans - association		
ID	**Chr**	**Start**	**End**	**Length**	**Frequ.[#]**	**Chr**	**Pos.**	**Log(1/p)**
				HUMAN				
1	1	20141	137979	117838	1.6%	17	11395390	10.8
78	1	142536304	142691538	155234	27.8%	4	52811970	16.0
339	3	133969263	133980309	11046	3.9%	12	56365699	10.8
1504	16	32472505	33801989	1329484	21.0%	8	87146973	16.0
383	4	939113	1092019	152906	42.5%	3	192311220	11.8
384	4	1237014	1245277	8263	1.7%	8	38433937	10.5
1190	12	751308	752432	1124	3.3%	3	192311220	10.4
1393	15	20201673	21259305	1057632	53.0%	2	69990290	11.0
1578	17	32058858	32066000	7142	6.6%	1	88973472	16.0
25	1	13374687	13488491	113804	2.0%	13	43798077	9.0
161	2	29335758	29338734	2976	2.3%	21	39227602	9.6
305	3	75605823	75615298	9475	4.7%	13	24007495	9.2
439	4	69371230	69502640	131410	3.5%	6	20487250	9.1
532	5	26219551	26238133	18582	25.7%	7	18935767	9.0
867	8	1947417	1951113	3696	3.5%	18	74779866	9.1
922	8	129763003	129779232	16229	2.2%	11	117322782	9.4
1382	14	104313212	104314910	1698	2.8%	12	56365699	9.6
1451	15	101791262	101792488	1226	2.1%	17	31536843	9.5
1482	16	16460524	16730728	270204	9.0%	9	95753319	9.6
1507	16	33872091	34191016	318925	1.4%	15	45945513	9.7
1654	18	44774027	44788908	14881	1.5%	1	218094151	9.8
				BOVINE				
360	18	27914135	28375996	461861	12.6%	12	809826	14.8
434	20	39350230	39388277	38047	15.7%	4	751020	16.0
971	8	51367642	51390418	22776	4.0%	17	12898336	14.4
1543	15	47513169	47522748	9579	3.6%	27	42483035	10.1

Frequ[#].: percentage of individuals with Penn - CNV copy number 3 or 4.

Assembly used, Bovine = UMD_3.1, Human = hg19



Supplementary Fig. 8: CNVR with all non‑syntenic SNP positions. A marked inflection was observed at ~ 9, justifying the choice of p‑values $\leq 10^{-9}$ as significance threshold. Very similar results were obtained with the bovine data (not shown). **(B)** Example of a putative dispersed duplication found by association in human. The main graph shows the QQ‑plot of all $-\log_{10}$p values obtained for the corresponding CNVR. Points in red correspond to all chromosome 15 SNPs, while the ponts in blue correspond to the SNP positions on all other chromosomes. The bar graph in the inset shows the frequency of the 20 Hidden Haplotype Clusters in "cases" (i.e. individuals with copy number 3 or 4) in red and in "controls" (i.e. individuals with copy number 2) in blue.

Supplementary table 1: Linkage analysis in Brown Swiss cattle

We selected two microsatellites each for five functional candidate genes from the cattle genome sequence (UMD 3.1 assembly). We genotyped these 10 microsatellites in three two-generation paternal half-sib families segregating for color-sidedness. The three families comprised a total of 37 informative meioses. We calculated twopoint linkage using the Merlin software assuming a fully penetrant dominant inheritance. The data clearly indicated that the KIT locus on BTA 6 is linked to colorsided in Brown Swiss cattle.

Gene	**BTA**	**Position (Mb)**	**Repeat motif**	**Alpha**	**Heterogeneity LOD (HLOD)**	**Two point LOD**
KITLG	5	18.567	$(AC)_{18}$	0	0	-∞
KITLG	5	19.360	$(AC)_{20}$	0.136	0.176	-∞
ADAMTS20[1)]	5	37.135	$(TA)_{27}$	n.d.	n.d.	n.d.
ADAMTS20	5	37.233	$(TA)_{21}$	0	0	-∞
KIT	6	71.789	$(TA)_{8}$ $(TG)_{20}$	1	6.894	6.894
KIT	6	71.844	$(TG)_{11}$ $(TA)_{15}$	0.819	3.306	3.236
EDNRB	12	53.296	$(GT)_{15}$	0	0	-∞
EDNRB	12	53.360	$(TG)_{19}$	0	0	-∞
MITF	22	31.750	$(CA)_{18}$	0	0	-1.548
MITF	22	31.794	$(GT)_{19}$	0	0	-0.836

[1)] This marker was not informative in the analyzed families.

Supplementary table 2: Primer sequences

	Product Name	**Forward (5'-3')**	**Reverse (5'-3')**	**Size bp**	**Region amplified bosTau4**
Primers to ID BACs	KIT_Ex1B	CTCGAAAGAACAGGGGTCAG	AGGAGCAGCAGAACGAAGAG	274	chr6:72,741,131-72,741,404
	BTA29_20.8	TTTACCCGGAAATCCACAAA	GGTTCAGAGCTAGGGTGTGC	379	chr29:20,899,945-20,900,323
Breakpoint primers	α-D	GGGGAGAATCTGTTTTCCTG	GGAAGGCCTTATTGCACACT	417	BTA29 to 6 breakpoint unique to BTA29
	A-E	GAAGCAACCCAGAGATGAGC	AAGGGAAGCCCATATGATGA	318	Fusion point on BTA6
	C-β	TCAACGAGGGACAACATGAA	CAATTGACCCCTCATTTTGG	606	Common breakpoint, found on BTA6 & BTA29

	γ-B	GCTGCAGAAAATGTTATTCCA	TCTTGAAGGGCCATAGCATC	525	BTA29 to 6 breakpoint unique to BTA6
	α-β	GGGGAGAATCTGTTTTCCTG	TAAAGTCGCCAGTGCAAGTG	394	Will not amplify in BTA29 Cs/Cs animals
Primers to check for SNPs in KIT coding sequence	KIT_Ex1	CTGGGCTCAGCCTTCTACC	TCCTGAAAGACTCGCAGCTC	928	chr6:72,740,730-72,741,657
	KIT_Ex2	GGAAACTTGACCCCGTTGTA	CATACCCGAAGCCACTATGC	743	chr6:72,779,376-72,780,118
	KIT_Ex3	CCGAAAGGCAACGTCTTAGAT	ATTTTGAGGCTGGGAGAACC	500	chr6:72,782,945-72,783,444
	KIT_Ex4	CATGGCTGAGGAAAAATGGT	GTGCTATGCAATGGGGAAAT	516	chr6:72,784,309-72,784,824
	KIT_Ex5	GCACTGCAGAGAATTTGGAA	TTGCTTTTGTGCTCTGGTTG	630	chr6:72,788,425-72,789,054
	KIT_Ex6	TCTTTCCGTTTCATTCTGCTG	AGCCCCAAACTTCCTTCTGT	533	chr6:72,793,394-72,793,926
	KIT_Ex7	GAGGCTGAACAGAGGACCAG	TCATGTGGTCAGCGAATTGT	624	chr6:72,795,451-72,796,074
	KIT_Ex8	GGAGCTTCAGCATCTTCACC	TCTACCTGCAGGCTGGAAAT	786	chr6:72,811,793-72,812,578
	KIT_Ex9	CCGATGCCTTCAGTTGATTT	GCCAGTGATGGAATGGACTT	342	chr6:72,813,937-72,814,278
	KIT_Ex10-11	TGGAGGTGAGAGGTGTTGTG	CTAAAGGCAATGCGATGTGA	459	chr6:72,815,316-72,815,774
	KIT_Ex12-13	CCACCACCACCATTTATTCC	CCATTTGGGTCAAAATCCTG	458	chr6:72,815,948-72,816,405
	KIT_Ex14	CTGACCCCTAATCAGGCAGA	GCCTTTCCCATGTTCCCTAT	862	chr6:72,817,391-72,818,232
	KIT_Ex15	ATAGCCTGCCTCTCACATGC	CAGTGACAACACCACCAAGG	562	chr6:72,819,627-72,820,188
	KIT_Ex16	TTCAGCACCTTCCTGTCCTT	TCAAGCGACACTCTGCATTC	378	chr6:72,820,095-72,820,472
	KIT_Ex17	GGCACCGAATGGTTTAAATG	TTCTCCTGCTGTGACCTTCA	567	chr6:72,821,253-72,821,819
	KIT_Ex18-19	TTGGATCTTTTGTGCTTCCA	GCGACCGAAATAACATTTGC	508	chr6:72,824,397-72,824,904
	KIT_Ex20	GTAAGGGCCCAGATGTCCTT	CCAAGAGAATGGAGGTCCTG	494	chr6:72,824,748-72,825,241
	KIT_Ex21	CATTCCAGCAGAAAAGCACA	TTTCCGCATCAAGGGATAAG	791	chr6:72,826,056-72,826,846
	KIT_3'UTR-A	GTCCTTCCCAAGGTTTCTCC	TGCTGAAAGCCAGGCTACTT	903	chr6:72,826,694-72,827,596
	KIT_3'UTR-B	TGATGCCGTTTGAAAAAGTG	GGAAAGGTGCGAGAGCATAG	712	chr6:72,827,333-72,828,044
	KIT_3'UTR-C	TGTTGTCTCGCAGGATTCAG	CATCTGGGAAACCTCACCTT	756	chr6:72,827,909-72,828,664
Expression primers	KIT_Intron_Rs1	AAGGAGGCGTTACCAGGTTT	TGCTGGAATGGAAATTAGGC	637	chr6:72,751,987-72,752,623
	KIT_Intron_Rs2	TTATTCCAGGTCCCTGCTTG	AGGGGCATCCTTGAGAGTTT	533	chr6:72,809,532-72,810,064
	KIT_Ex1-3'UTR	CTCTTCGTTCTGCTGCTCCT	GGAAAGGTGCGAGAGCATAG	4,527	KIT cDNA

Supplementary table 3: Primers to determine origin of BTA6 DNA flanking BTA29 DNA in Cs6

SupFig6 1+2	SNP group A	CGTGTGTGAGCATGCTAGGT	GACCAAAGACCCCACTTCCT	885	chr6:73,008,888-73,009,772
SupFig6 1+3	SNP group A 6-6*	CGTGTGTGAGCATGCTAGGT	CCGCCTCTCTCATTATCAGC	11,164	chr6:73,008,888-73,020,051
SupFig6 1+4	SNP group A 6-29	CGTGTGTGAGCATGCTAGGT	CAATTGACCCCTCATTTTGG	10,909	BTA6 DNA upstream of common breakpoint
SupFig6 II+I	SNP group B	GCAGTAGATGCCCCCATAGA	AGCAGAGGAAGAAGCAGCAG	792	chr6:72,896,736-72,897,527
SupFig6 III+I	SNP group B 6-6#	GCTAGCCAGATCCAGCAGTC	AGCAGAGGAAGAAGCAGCAG	11,792	chr6:72,885,736-72,897,527
SupFig6 IV+I	SNP group B 29-6	GCTGCAGAAAATGTTATTCCA	AGCAGAGGAAGAAGCAGCAG	11,687	BTA6 DNA downstream of BTA6 specific break

6-6* Primers span region C-D

6-6# Primers span region B and non duplicated part of BTA6

6-29, Primers span breakpoint C-β

29-6, Primers span breakpoint γ-B

SNP group A: ss469414208, ss469414209

SNP group B: ss469414210, ss469414211

Part 3: Taking advantage of genetic marker information in animal breeding

3.1 Marker imputation with low-density marker panels in Dutch Holstein cattle

3.2 Identifying cows with subclinical mastitis by bulk single nucleotide polymorphism genotyping of tank milk

Marker imputation with low-density marker panels in Dutch Holstein cattle

Z. Zhang and T. Druet

Abstract

The availability of high-density bovine genotyping arrays made implementation of genomic selection possible in dairy cattle. Development of low-density SNP panels will allow extending genomic selection to a larger portion of the population. Prediction of ungenotyped markers, called imputation, is a strategy that allows using the same low-density chips for all traits (and for different breeds). In the present study, we evaluated the accuracy of imputation with low-density genotyping arrays in the Dutch Holstein population. Five different sizes of genotyping arrays were tested, from 384 to 6,000 SNPs. According to marker density, overall allelic imputation error rate obtained with the program DAGPHASE, which relies on LD and linkage, ranged from 11.7 % to 2.0 % and from 10.7 % to 3.3 % with the program CHROMIBD which relies on linkage and the set of all genotyped ancestors. However, imputation efficiency was influenced by the relationship between low-density and high-density genotyped animals. Animals with both parents genotyped had particularly low imputation error rates, below 1 % with 1,500 SNPs or more. In summary, missing marker alleles can be predicted with 3 to 4 % errors with approximately 1 SNP / Mb (~3,000 markers). CHROMIBD proved more efficient than DAGPHASE only at lower marker densities or when several genotyped ancestors are available. Future studies are required to measure the impact of these imputation error rates on accuracy of genomic selection with low-density SNPs panels.

J Dairy Sci. 93 (2010) 5487-5494

Introduction

Genomic selection (Meuwissen *et al.* 2001) is now widely used in dairy cattle (de Roos *et al.* 2007; VanRaden, 2008; VanRaden *et al.* 2009). It relies on genotyping a livestock population for thousands of SNPs across the entire genome and to use these SNPs to estimate breeding values of any genotyped animal. Recently, Habier *et al.* (2009) proposed to genotype some animals on lower-density marker panels to make genomic selection more cost effective. Unlike strategies based on selection of subset of SNPs associated to selected traits, they suggested to use evenly spaced SNPs across the genome. Indeed, the first strategy, as implemented in Weigel *et al.* (2009), requires different set of SNPs for each trait while with the strategy of Habier *et al.* (2009), the same set of SNPs can be used for all the traits. This strategy is now considered by most dairy cattle breeding organizations that implemented genomic selection. Such a low-density chip with 3,000 SNPs has already been developed in the United States and Illumina plans to commercialize it in 2010.

However, the use of low-density marker panels requires methods for transferring information from individuals genotyped at higher density. Prediction of missing markers, or imputation, is one technique allowing the use of individuals genotyped on different marker panels. Imputation is already widely used in human genetics (Marchini *et al.* 2007; Marchini and Howie, 2008). However, in studies applied to human populations commonly used marker arrays are of much higher density (above 300 thousands SNPs) and imputation relies mostly on linkage disequilibrium (LD), whereas in livestock species, availability of extended pedigrees allows the use of linkage information for imputation. Druet and Georges (2010) and Druet and Farnir (personal communication) proposed two methods adapted for marker imputation with low-density SNPs panels in animal and plant breeding. The objective of the present study is to test these two methods with low-density SNPs on a Dutch Holstein dairy cattle data set.

Material and Methods

Data

A set of 4,734 dairy cattle Holstein individuals genotyped for 45,836 SNP markers on the CRV chip, a custom-made 60K Illumina panel described in Charlier *et al.* (2008), was used in this study. A more complete description of this data set can be found in Druet *et al.* (2010).

For testing imputation efficiency, the animals were assigned to two groups: reference individuals were genotyped on all the markers while the target animals were genotyped on lower density SNP panels (we erased markers not present on the low density panel).

Different sizes of reference populations were tested (500, 1,000, 1,500 and 2,000 reference individuals). Animals with higher number of genotyped descendants were preferentially included in the reference population (516 individuals had at least one genotyped descendent). Remaining reference individuals were chosen at random. Remaining animals (4,234, 3,734, 3,234 and 2,734) were considered as target individuals.

Creation of Low Density Marker Panels

To mimic low-density genotyping, different low-density panels were defined (see below) and genotypes of target animals were erased for the unselected markers. Five different sizes of marker panels were tested: 384, 768, 1,536, 3,000 and 6,000 SNPs. These sizes were chosen according to Illumina GoldenGate and iSelect BeadChip technologies which allow custom genotyping with 384 to 1,536 SNPs and with 3,000 to 60,000 SNPs, respectively. The markers were selected to obtain a compromise between uniform marker density and high minor allelic frequency (MAF) with the following method. Number of markers per chromosome was obtained by multiplying the desired marker density (total number of markers divided by the size (in Mb) of the genome) by the size of the chromosome in Mb. Each chromosome was divided in equal segments based on the desired number of markers. Then, the marker with the highest MAF was selected in the first segment. For subsequent segments, the marker with the highest score combining MAF and distance with the marker retained in the previous segment was selected:

$$\text{score}(i)=\text{MAF}(i)*[\text{ssize} - |\text{ssize-dist}(i)|]$$

where i is the indice of the tested marker, ssize is the size of each segment and dist(i) is the distance between the tested marker and the selected marker in the previous segment (Matukumalli *et al.* 2009).

Marker imputation method

Haplotype Reconstruction. Haplotypes of reference and target individuals were first partially reconstructed based on linkage and mendelian segregation rules with LinkPHASE (Druet and Georges, 2010). Then, haplotypes from reference individuals were fully reconstructed by using iteratively Beagle (Browning and Browning, 2007) with scale and shift parameters equal to 2.0 and 0.1 and DAGPHASE (Druet and Georges, 2010) as described in Druet and Georges (2010). In total, 20 such iterations were performed to estimate the directed acyclic graph (**DAG**) describing reference haplotypes. Linkage information was ignored for parents with 5 offspring or less because Druet and Georges (2010) observed that this procedure reduced haplotyping errors.

Missing Marker Imputation. Finally, a file containing completely reconstructed haplotypes for reference individuals and partially reconstructed haplotypes (obtained from LinkPHASE) for target individuals was used for marker imputation. Two different imputation programs were used: DAGPHASE (Druet and Georges, 2010) and CHROMIBD (Druet and Farnir, personal communication).

With DAGPHASE, linkage information is used when parents are genotyped: partial haplotypes of offspring are used to determine which haplotype was received from the parent and marker alleles of the transmitted haplotype are then used to fill-in missing markers in the progeny. For other haplotypes, LD information is used by inferring the path of these haplotypes in the DAG which summarizes all reference haplotypes. Probabilities of different paths are computed based on partial haplotypes (genotyped markers) with an hidden Markov model described in Druet and Georges (2010). Then, the marker alleles labeling the path in the DAG are used to impute missing marker alleles. The LD modeled in the DAG results from both recent and old ancestors. When the haplotype of a recent ancestor is present in the DAG, the partial haplotype will go through the same path as the haplotype of the ancestor over a long portion of the graph. Therefore, identification of the path will be precise and possible even with few genotyped markers. Portions of haplotypes transmitted from more distant ancestors are shorter and require therefore higher marker densities to be inferred precisely. With DAGPHASE, two outputs can be obtained: prediction of all the missing markers based on the most likely “hidden” chain (as in Druet and Georges, 2010) or prediction of posterior genotype probabilities based on the Baum-Welch algorithm as described in Druet *et al.* (2010).

CHROMIBD uses linkage and a hidden Markov model to estimate identity-by-descent probabilities (**IBD**) between a chromosome and parental chromosomes from genotyped ancestors. It relies only on the set of genotyped ancestors and LD is not used. Haplotypes of these ancestors and IBD probabilities are then used to predict posterior genotyped probabilities of missing markers (Druet and Farnir, personal communication).

Evaluation of Imputation Efficiency

When marker alleles are predicted, imputation efficiency was estimated by comparing imputed marker alleles and real marker alleles (this procedure slightly underestimates precision as real genotypes have some errors). The allelic imputation error rate per animal was then estimated as:

$$\frac{1}{2*N}\sum_{j=1}^{N}\left|O(n_{ij}^{1})-I(n_{ij}^{1})\right|$$

where N is the number of markers, $O(n_{ij}^{1})$ and $I(n_{ij}^{1})$ are the observed (real genotypes) and imputed number of allele "1" for individual i at marker j and 2*N is the total number of imputed alleles.

When posterior genotype probabilities were estimated, the expected number of "1" alleles was first computed as:

$$E(n_{ij}^{1})=2*P(G_{ij}=11)+1*P(G_{ij}=12)+0*P(G_{ij}=22)$$

where $E(n_{ij}^{1})$ is the expected number of allele "1" for individual i at marker j and is $P(G_{ij}=kl)$ the probability that individual i carries marker alleles k and l at marker j. The allelic imputation error rate per animal was finally estimated as:

$$\frac{1}{2*N}\sum_{j=1}^{N}\left|O(n_{ij}^{1})-E(n_{ij}^{1})\right|$$

Imputation efficiency was measured on the 2,734 individuals which were never reference individuals in order to compare results obtained with the same set of individuals across all designs.

Results and discussion

In this work, three factors affecting imputation efficiency were studied: density of markers in the small chip, relationship between target and reference individuals and number of reference individuals. The impact of each factor is dependent on the imputation method used (CHROMIBD or DAGPHASE).

Results are presented as allelic imputation error rates (proportion of incorrect imputed marker alleles). Genotype imputation error rates (proportion of incorrect imputed genotypes) are approximately equal to twice the allelic imputation error rates. With 384 markers, the ratio between error rates was ~1.90. This ratio consistently increased with number of markers and was ~1.98 with 6,000 markers (ratios were computed in designs with 1,000 and 2,000 reference individuals).

Influence of Marker Density

The main question when designing low density marker panels is what imputation efficiency can be achieved for a given number of SNPs on the small chip. Figure 1 presents mean imputation error rates obtained for small chips with 384 to 6,000 SNPs (with 1,000 reference individuals). With both DAGPHASE and CHROMIBD, the curve describing imputation efficiency as a function of marker density presented an hyperbolic pattern with steep decrease at low densities and moderate decrease with higher number of markers. Such a relationship was previously described by Druet *et al.* (2010). With DAGPHASE, allelic imputation error rates ranged from 11.7 % with 384 SNPs to 2.0 % with 6,000 SNPs. With CHROMIBD, relying only on linkage, the differences were smaller: from 10.7 % to 3.3 %. For comparison, imputing genotypes based on the most likely genotypes or randomly according to genotype frequencies resulted in respectively 20.0 % and 27.0 % allelic imputation error rates. With DAGPHASE (CHROMIBD), adding approximately 1,500 SNPs decreased errors rates by 6.9 % (5.4 %) when the initial number of markers was equal to 384 and by only 1.7 % (1.3 %) when having 3,000 markers instead of 1,536. Similarly, using 3,000 markers instead of 384 markers dramatically decreased error rates by 8.6 % (6.7 %) whereas adding 3,000 more markers, reduced errors only by 1.1 % (0.7 %). To obtain overall mean allelic imputation error rates around 5 %, chips with 1,536 SNPs or more should be used. Such error rates might still be to high to obtain accurate estimation of breeding values with genomic selection. Panels with 3,000 SNPs or more are necessary to get around 3 % error rate or less (with DAGPHASE). The relationship between imputation error rates and precision of genomic selection must be studied before defining optimal SNPs panels. Technical and economical considerations must be taken into account to decide whether larger chips should be used.

The results presented for different SNPs panels are overall results, averaged over all animals. However, imputation accuracy is highly dependent on the relationship between target and reference individuals.

Relationship between Target and Reference Individuals

Druet *et al.* (2010) and Druet and Farnir (personal communication) already showed that DAGPHASE and CHROMIBD result in higher imputation accuracy when the relationship between target and reference individuals is higher. This relationship was measured as a score equal to the expected proportion of the genome inherited from reference individuals (scores of 0.5, 0.75, 0.875, 0.9375 and 0.96875 corresponds to all male parents genotyped for 1, 2, 3, 4 and 5 generations, respectively.). Such behavior is expected since

CHROMIBD directly models IBD probabilities between the target and the reference individuals whereas DAGPHASE uses haplotypes of reference individuals to estimate the optimal DAG. If more ancestors of an individual were used to construct the DAG, there is more chance that the haplotype of that individual is present in the DAG, and over longer stretches (see methods).

Table 1 contains for all tested marker panels and different categories of relationships between target and reference individuals, the mean imputation error rates of DAGPHASE and CHROMIBD. The computed scores ranged from 0.0 to 1.0 with the median equal to 0.9375. 94.8 %, 84.6 % and 70.0 % of the target individuals had a score higher or equal to 0.75, 0.85 and 0.90. When almost no ancestors were genotyped (score below 0.50), errors rates ranged according to the marker density from 4.8 % to 20.2 % with DAGPHASE and from 22.3 % to 26.8 % with CHROMIBD. For higher scores, imputation error rates decreased, particularly with CHROMIBD. For target individuals with a score above 0.75, DAGPHASE resulted in imputation error rates below 3 and 4 % with 6,000 or 3,000 SNPs. With CHROMIBD, such error rates were obtained only for target individuals with scores above 0.90.

For animals with both parents in the reference group (score = 1), the mean imputation rate was clearly lower than for other animals, ranging from 3.20% to 0.30% with DAGPHASE and from 4.20% to 0.40 % with CHROMIBD. For these animals, haplotypes received from parents are inferred accurately by using linkage (for a single generation). For other animals, the maternal haplotype (and sometimes the paternal haplotype too) is more difficult to model. The imputation error rate per animal is the mean from imputation error rates from paternal and maternal haplotypes. Since sires are most often genotyped, we can assume that imputation efficiency for paternal haplotypes is approximately equal to the imputation efficiency observed for animals with both parents in the reference group (because all these haplotypes are described through linkage). Therefore, the imputation error rate for maternal haplotypes is probably higher (approximately equal to twice the error rate per animal minus the error rate observed when the score equals 1).

Influence of the number of individuals genotyped in the reference panel

Figure 2 describes the relationship between number of individuals genotyped on all markers and imputation efficiency with DAGPHASE or CHROMIBD for 5 different marker densities. With DAGPHASE, imputation efficiency constantly increased when number of reference individuals increased: from 12.2 % to 11.3 %, 8.8 % to 6.9 %, 5.9 % to 3.9 %, 4.1 % to 2.3 % and 2.7 % to 1.4 % with 384, 768, 1,536, 3,000 and

6,000 markers on the small chip, respectively. With CHROMIBD, almost no differences were observed when increasing size of the reference population. In comparison with the influence of the number of SNPs or of the relationship between target and reference individuals, the impact of the number of individuals genotyped in the reference panel was small.

Druet *et al.* (2010) studied the relationship between number of reference individuals and imputation efficiency at higher marker densities. The benefit was more pronounced when very few reference individuals (below 500) were available. Above 1,000 genotyped reference individuals, only small gains in efficiency were observed (by improving estimation of LD between markers). In many dairy cattle populations, large numbers of individuals are already genotyped on high-density marker panels. Therefore, the size of the reference population should already be large enough and not be a major factor influencing imputation accuracy using small chips for a portion of the population. The main benefit of increasing number of reference individuals will be obtained through increasing relationship between target and reference individuals (see above).

Comparison of Imputation Methods

The results presented in earlier sections highlighted differences between DAGPHASE and CHROMIBD. First, CHROMIBD is more efficient than DAGPHASE with few SNPs per panel whereas DAGPHASE is better than CHROMIBD when panels have 1,536 SNPs or more (Figure 1). Second, the improvement of imputation accuracy when relationship between target and reference animals increases is more pronounced with CHROMIBD. With DAGPHASE, the imputation efficiency improves only slightly when the score increases from 0.75 to 0.95 whereas with CHROMIBD, the benefit of having additional genotyped ancestors still results in large difference in imputation efficiency. For low scores, CHROMIBD is clearly less accurate than DAGPHASE. CHROMIBD results better for high scores. However, the score above which CHROMIBD achieves lower imputation error rates depends on the number of SNPs on the small panel: at low density (384 SNPs), CHROMIBD is better for animals with a score higher than 0.85 whereas at higher densities (6,000 SNPs), using CHROMIBD is more efficient for animals with a score higher than 0.95. For animals with both parents genotyped, imputation performed with DAGPHASE had the lowest error rates. Indeed, in that situation both methods use the same information (linkage) but DAGPHASE models both haplotypes jointly whereas CHROMIBD models them independently. Finally, CHROMIBD was barely affected by size of the reference population whereas having more reference individuals improved imputation with DAGPHASE.

All these observations can be explained by differences in the methods. DAGPHASE uses LD due to recent and old ancestors. LD is stronger at higher marker densities and increasing size of reference population improves estimation of LD in the DAG. At low marker densities, LD is relatively weak and, as a result, imputation efficiency decreases. CHROMIBD models the IBD process along a chromosome between at target chromosome and a set of parental chromosomes. It relies therefore on recent ancestors which transmitted relatively long chromosome fragments and can work at lower marker densities. However, the efficiency is strongly influenced by the number of ancestors which are genotyped.

As mentioned in the methods section, posterior genotypes probabilities can also be obtained with DAGPHASE. Use of this output resulted in slightly higher imputation error rates (results not shown). However, use of posterior genotype probabilities (as with CHROMIBD too) gives additional information than simply predicting likely genotypes. Indeed, the reliability of the imputation is known when using posterior probabilities and the user knows which marker alleles can be imputed with limited risk. It is for instance possible to call only marker alleles with posterior genotype probability above a certain threshold. In addition, these probabilities can be used to estimate the reliabilities of genomic predictions obtained with imputed markers. Finally, Weigel *et al.* (2010) mentioned that it may advantageous to use genotype probabilities in genomic selection rather than calling genotypes.

In agreement with Druet and Georges (2010), results obtained with DAGPHASE show that it is beneficial to use both linkage and LD information for imputation rather than relying solely on LD (e.g. with Beagle). In Druet and Georges (2010), the algorithm of Beagle was found to be more efficient for marker imputation than the algorithm of fastPHASE (Scheet and Stephens, 2006). We also tested HiddenPHASE relying on fastPHASE algorithm for description of LD (Druet and Georges, 2010) but imputation error rates and running times were higher (data not shown). Weigel *et al.* (2010) used fastPHASE (Scheet and Stephens, 2006) and IMPUTE (Marchini *et al.* 2007) to study the feasibility of imputation in Jersey cattle for different sizes of SNP panels for three chromosomes. Their imputation method relied purely on LD. When comparing results from the two studies, different factors must be considered: the relationship between target and reference individuals (the authors mention that most animals had a score above 0.75 without further detail), the reference panel (2,542 Jersey animals) and LD within breed. Weigel *et al.* (2010) report that previous studies indicated that LD may be higher in Jerseys than in Holstein (e.g., Villa-Angulo *et al.* 2009). With approximately 430, 870, 2,170, 4,340 and 8,680 SNPs, they obtained, with their best method, genotype imputation error rates between 25.7 and 31.1 %, 19.9 and 27.3 %, 9.5 and 11.4 %, 5.8 and 8.3 % and between 3.6 and 5.8 %, respectively. In our studies, genotyping error rates (with 2,000 reference individuals and the best method) were equal to 20.2, 13.0, 7.4, 4.3 and 2.7 % with 384, 768, 1,536, 3,000 and 6,000 SNPs, respectively. These results indicate that with fewer markers and reference individuals and probably lower LD, our method achieved higher imputation accuracy although it is difficult to compare results from different data sets. Higher accuracy is explained by the fact that the method accounts for linkage information. The study of Weigel *et al.* (2010) was indeed designed to obtain lower bounds of imputation accuracy because it relied on LD alone and ignored linkage information (Weigel *et al.* 2010). Use of Beagle algorithm (or CHROMIBD) might also achieve better imputation accuracy than fastPHASE. In addition, computational costs of methods used in our study were lower (see below). With fastPHASE and IMPUTE, Weigel *et al.* (2010) mentioned computational problems and the need to break chromosomes in several pieces.

Computational Requirements

Computation times and memory requirements were measured on a computer cluster with Intel Xeon "Harpertown" L5420 at 2.50 Ghz. Each chromosome was processed on its own CPU core. Computational requirements are presented for a design with 3,000 SNPs on the low density panel and 1,000 reference individuals. Estimation of parameters of the DAG and haplotype reconstruction of reference individuals required from 5 to 35 minutes per chromosome (Beagle was run with 2 Gb memory and DAGPHASE used less memory). For imputation with DAGPHASE, computation times per chromosome ranged from 2 minutes to 10 minutes and memory requirements per chromosome ranged from 275 to 620 Mb memory. With CHROMIBD, computational requirements were higher: from 4 minutes to 15 minutes and from 420 Mb to 1.16 Gb memory.

When new genotypes are available, repeating the imputation should improve accuracy of prediction of missing markers. However, repeating imputation for all animals has huge computational costs. To reduce these costs, the DAG should not be estimated at every imputation (unless the DAG has been initially estimated on a small sample) and a reference training set might be used for this purpose (not the whole genotyped population). In addition, haplotypes of reference individuals should be re-estimated only new available information (such as new progeny or parents genotypes) is expected to significantly improve this process. Finally, when imputing a set of newly genotyped animals, only haplotypes of these animals and of their genotyped parents are needed. Running the imputation on such a subset would result much faster and requires less memory. Imputation should be redone for individuals only if accuracy will significantly increase for them (e.g, if a parent of the target individual has been newly genotyped or if haplotype reconstruction of a parent has been improved thanks to new genotyped offsprings).

Conclusions

With both tested imputation methods, missing marker alleles can be predicted with 3 to 4 % errors with approximately 1 SNP / Mb (~3,000 markers). These figures can further be reduced by adding more markers. With DAGPHASE, these error rates can be achieved for animals with at least their sire and maternal grand-sire genotyped. With CHROMIBD, four generations (or more) of male ancestors are required to obtain error rates below 4 % (with 3,000 SNPs). Such situations are fortunately frequent in the Dutch Holstein population. When both parents are genotyped, imputation error rates were below 1 % with 1,500 SNPs or more. CHROMIBD is more efficient than DAGPHASE only at lower marker densities or when several genotyped ancestors are available.

Future studies are required to measure the impact of these imputation error rates on accuracy of genomic selection with low-density SNPs panels and to determine which animals should be genotyped on low density chips to optimize breeding programs. In addition to genomic selection, use of low-density genotypes might also be beneficial in F2 designs used for QTL mapping. F0 animals would then be genotyped at high density and small chips would be used for later generations. In that case, imputation would be efficient since both parents are genotyped. Such a procedure could speed up QTL mapping and keep costs relatively low.

Acknowledgements

Zhiyan Zhang has a PhD fellowship from the China scholarship council. Tom Druet is Research Associate from the F.R.S. – FNRS. The authors thank CRV for access to their data. This work was funded by grants of the Service Public de Wallonie and from the Communauté Française de Belgique (Biomod ARC). We acknowledge University of Liège (SEGI and GIGA bioinformatics platform) for the use of NIC3 and GIGA-grid supercomputers.

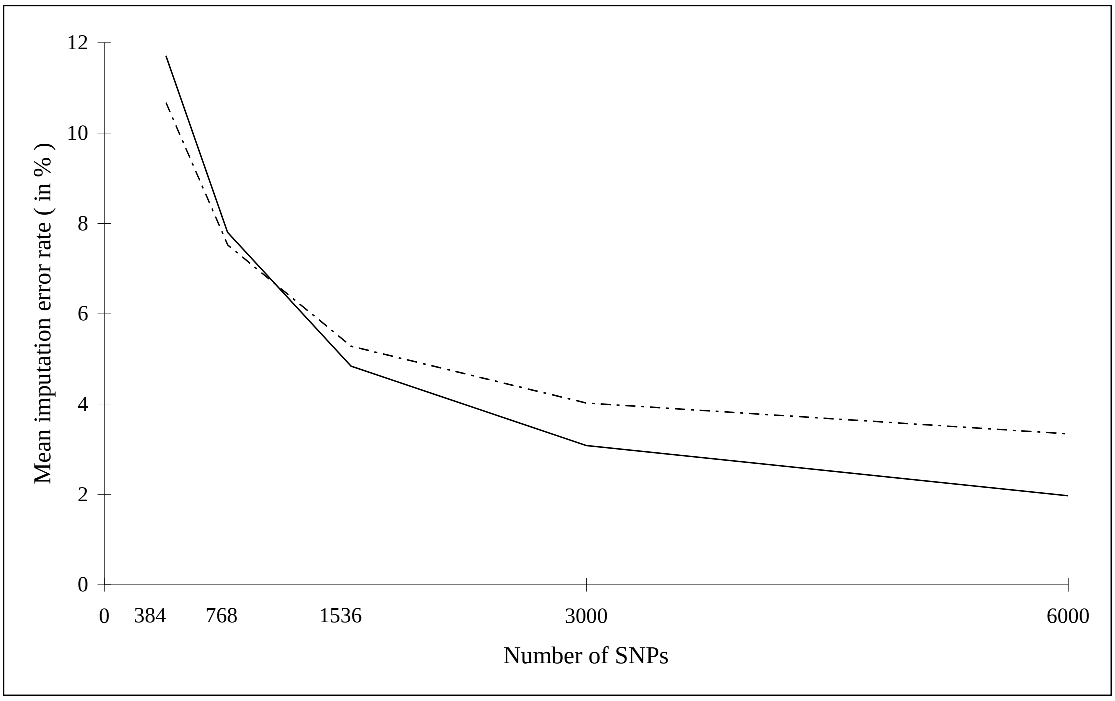


Figure 1. Mean allelic imputation error rate (in %) obtained for 1000 reference individuals imputed with DAGPHASE (solid line) and CHROMIBD (dashed line) for different sizes of low-density panels.

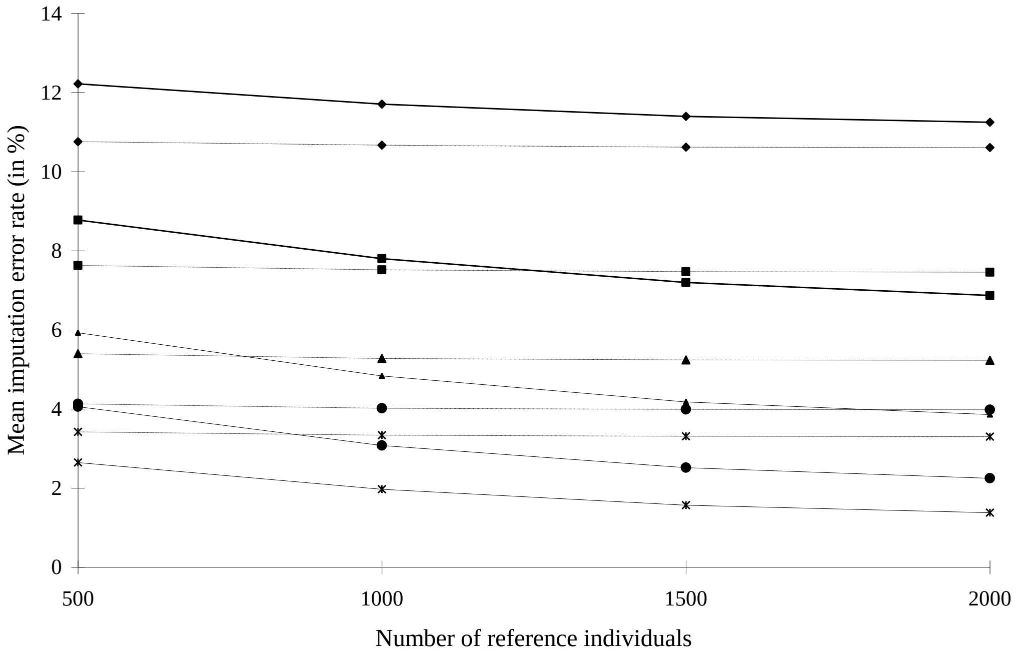


Figure 2. Mean allelic imputation error rate (in %) for low-density panel with 384 (◆), 768(■), 1536(▲), 3000(●) and 6000 (*) SNPs obtained with DAGPHASE (solid line) and CHROMIBD (dashed line) for different sizes of reference population.

Table 1. Allelic imputation error rates (in %) obtained with DAGPHASE and CHROMIBD for animals grouped by a score representing the expected proportion of the genome inherited from a reference individual (results are presented for different sizes of SNPs panels).

Relationship	**Number of**	**DAGPHASE**					**CHROMIBD**				
Score	**individuals**	384	768	1,536	3,000	6,000	384	768	1,536	3,000	6,000
<0.50	14	20.2	15.2	10.8	7.2	4.8	26.8	25.5	24.2	23.2	22.3
[0.50-0.75)	128	15.5	11.4	7.8	5.3	3.5	20.4	18.7	17.2	16.2	15.4
[0.75-0.80)	157	13.1	9.0	6.0	4.0	2.7	13.9	11.4	9.6	8.4	7.7
[0.80-0.85)	122	13.0	8.8	5.6	3.7	2.4	13.1	10.2	8.0	6.6	5.8
[0.85-0.90)	397	12.7	8.5	5.4	3.6	2.3	11.7	8.5	6.3	5.0	4.2
[0.90-0.95)	608	12.7	8.5	5.3	3.3	2.1	10.9	7.4	4.9	3.5	2.7
[0.95-1.00)	943	12.9	8.5	5.1	3.1	1.9	10.2	6.4	3.7	2.3	1.5
1	365	3.2	1.7	0.9	0.5	0.3	4.2	2.3	1.2	0.6	0.4

Identifying cows with sub-clinical mastitis by bulk SNP genotyping of tank milk

G. Blard, Z. Zhang, W. Coppieters and M. Georges

Abstract

Mastitis remains the most important health issue in dairy cattle. Improved methods to identify cows developing subclinical mastitis would benefit farmers. We herein describe a novel method to determine the somatic cell counts (SCC) of individual cows by bulk genotyping a sample of milk from the milk tank with panels of genome-wide single nucleotide polymorphisms (SNP). We developed a simple linear model to estimate the contribution of individual cows to the genomic DNA present in the tank milk from 1) the known genotypes of individual cows for the interrogated SNP and 2) the ratio of SNP alleles in the tank milk. Using simulations, we estimate that 3,000, 50,000, and 700,000 SNP are sufficient to accurately (R^2> 0.98) estimate individual SCC in tanks containing milk from 25, 100, and 500 cows, respectively. Using actual data, we demonstrate that the SCC of 21 cows can be estimated with a co-efficient of determination of 0.60 using approximately 9,000 SNP. The proposed method increases the value of the proposition of SNP genotyping individual cows for genomic selection purposes.

J Dairy Sci. 95 (2012) 4109-4113

Introduction

Mastitis is generally regarded as the most important health issue in dairy farming. Costs to the farmer result from treatment, decreased milk yield and value, and culling. In the Netherlands, the annual cost of mastitis to farmers has been estimated at ~80 Euros per cow per year (f.i. Hogeveen *et al.* 2011).

One often distinguishes clinical from subclinical mastitis. Mastitis is said to be clinical when resulting in either milk or udder anomalies that are manifest to the farmer. Subclinical mastitis is defined as the presence of microorganisms in combination with elevated somatic cell counts (SCC) in the milk. 200,000 cells/ml is often utilized as SCC cut-off value. Subclinical mastitis is costly on its own because of its negative impact on milk yield, which is decreasing log-linearly with SCC.

At present, the control of mastitis is primarily driven by the detection and treatment of clinical mastitis. More effective detection of subclinical mastitis would be a valuable addition to mastitis control. In most circumstances, SCC for individual cows are only measured periodically (typically every 4-6 weeks) precluding close monitoring of the udder health status of individual cows. Advanced versions of milking robots will precociously detect changes in conductivity or even SCC, yet automatic milking has been associated with a decrease rather than an improvement in udder health (f.i. Hovinen & Pyörälä, 2011). Thus novel approaches to monitor SCC of individual cows could be a benefit to the sector.

We herein propose a method that allows determination of SCC for individual milking cows (and hence detection of cows with subclinical mastitis) by genotyping a sample of milk from the farm’s tank for a panel of Single Nucleotide Polymorphisms (SNPs). At present, 3 favored bovine SNP panels are available, including approximately 3,000 (**3K**), 50,000 (**50K**) and 700,000 (**700K**) SNPs distributed across the genome. All of these are developed and commercialized by Illumina Inc. (http://www.illumina.com/). The proposed method assumes that individual genotypes for the same SNPs are available for all cows on the farm. As genomic selection is becoming common practice in dairy cattle, an increasing proportion of the cow population is being SNP genotyped. As an example, in the spring of 2011, more than 85,000 Holstein-Friesian animals had been genotyped with one of the SNP arrays mentioned above in the US alone. The majority of these were females, and their proportion was rapidly increasing with time. We anticipate that farms in which all cows will be SNP genotyped will be common place in the near future.

Materials and Methods

Method.

Assume a dairy farm with *n* cows. Assume that the volume of milk contributed by cow *i* to the milk tank is known (as it is often in reality) and equals $\boldsymbol{v_i}$. Assume that the somatic cell counts per liter of milk contributed to the tank by cow *i* is $\boldsymbol{c_i}$. Assume that all cows on the farm have been genotyped for an array interrogating *m* SNPs. Assume that $\boldsymbol{g_{ij}}$ is an indicator variable for the genotype of cow *i* for SNP *j* (taking a value of 1 for genotype 11, 0.5 for genotype 12, and 0 for genotype 22). Assume that a sample of DNA extracted from the milk tank has been genotyped with the same array, and that the estimated frequency of allele "1" of SNP *j* in the milk sample is f_i. The proportion of somatic cells contributed by cow *i* to the tank, $\boldsymbol{pc_i}$ can be estimated from a set of *m* linear equation of the form:

$$f_j = \sum_{i=1}^{n} pc_i \times g_{ij} + \varepsilon_j$$

by minimizing

$$SSE = \sum_{j=1}^{m} \varepsilon_j .$$

From the obtained $\boldsymbol{pc_i}$ values one can then determine the somatic cell counts per liter of cow *i* relative to the rest of the herd as

$$rc_i = pc_i / pv_i = c_i / \bar{c}$$

in which $\mathbf{pv}_i$ is the known proportion of the tank's volume contributed by cow *i*, and $\bar{c}$ is the average somatic cell count per liter in the herd. If $\bar{c}$ is known (it can be measured in the milk tank), the absolute somatic cell counts per liter for cow *i* can be computed from rc_i.

Simulated data

We simulated farms with 25, 100 and 500 cows. The cows' daily milk production (liter) was assumed normally distributed with mean 30 l and standard deviation 0.2. Somatic cell counts per liter were assumed to be proportional (x 8×10^6) to a chi-squared distribution (2df). We assumed the use of SNP arrays with 3,000, 50,000 or 700,000 SNPs corresponding closely to presently available commercial products. Cows were assumed to have genotypes for each SNP (in practice, missing values would be filled in by imputation). Estimates of SNP allele frequencies ($0 < f < 1$) in the tank milk were assumed normally distributed around the true frequency with standard error of 0.05.

Actual data

We collected a sample of milk from a tank containing known quantities of milk (mean: 29.5 liter; SD: 5.3 liter) contributed the same day by 20 cows that had been previously genotyped using either of two custom-made ~50K Illumina arrays (Charlier *et al.* 2008). Individual somatic cell counts were determined on the same day for each cow using a Fossomatic FC instrument (Foss, Hilerod, Denmark) DNA was extracted from the milk sample using standard procedures and genotyped with the USDA ~50K Illumina array (Matukumalli *et al.* 2009). Estimates of B-allele frequencies (corresponding to f_i in Method) were directly obtained from the BeadStudio software package (http://www.illumina.com/). We also had SNP genotypes for 20 cows from the same farm that did not contribute milk to the tank.

To calibrate the relationship between B-allele frequency as computed with BeadStudio, and actual B-allele proportion (in the Methods section), we took advantage of DNA samples available for 95 Dutch Holstein-Friesian samples that had previously been genotyped with the Illumina BovineSNP50 genotyping BeadChip (Matukumalli *et al.* 2009). The DNA concentrations were estimated as the average of 2 independent fluorometric measurements performed using PicoGreen (Life Technologies, Carlsbad, CA) according to the instructions of the manufacturer. We mixed equal volumes of DNA solutions from the 95 sires and genotyped the resulting DNA pool using the Illumina BovineSNP50 genotyping BeadChip. The relationship between 1) BeadStudio-derived B-allele frequency and 2) B-allele proportion computed from the known SNP genotypes and DNA concentrations of the sires is shown in Supplementary Figure 1 (available online at http://www.journalofdairyscience.org/) for 45,248 SNP with complete genotype data across the 95 sires. The correlation between both measures was 0.98, yet departure from linearity was obvious. We, therefore, adjusted the relation by fitting local (i.e., for each SNP) linear regressions based on 3,000 left- and 3,000 right-sided neighboring SNP (corresponding to frequency ranges of ~7.5%). BeadStudio measured B-allele frequencies were converted to adjust using the SNP-specific β0 (intercept) and β1 (slope) estimators (Supplementary Figure 1, available online at http://www.journalofdairyscience.org/). Mean square errors averaged 0.005, corresponding to a residual standard deviations of 0.07 (i.e., comparable to the values used in the simulations). Statistical analyses were conducted in R software (R Foundation for Statistical Computing, Vienna, Austria).

Results

Simulated data.

Figure 1 is showing representative examples confronting actual and estimated SCC. Predictions were very accurate with coefficient of determination > 0.98 (i.e. r^2 = proportion of the variance of true SCC accounted for by predictions) with 3K or more SNPs for 25 cows, 50K or more SNPs for 100 cows and 700K SNPs for 500 cows. Predictors appeared unbiased under all tested conditions. Thus, our simulations indicated that bulk genotyping of tank milk could be effective for SCC monitoring of individual cows, including identification of cows with increased SCC indicative of subclinical mastitis.

Supplementary Figure 1A shows representative examples of frequency distribution of the statistical significance (*-log(p)*) of the $\boldsymbol{pc_i}$'s for cows that did contribute milk to the tank, evaluated with a standard *t*-test. For ~80% of the cows, p-values were ≤ 0.0001 when using arrays interrogating ≥ 3K SNPs for herds with 25 cows, ≥ 50K SNPs for herds with 100 cows and (500 cows) As expected p-values were strongly and inversely correlated with the proportion of SCC contributed by a given cow to the milk tank (data not shown). By comparison, the frequency distribution of corresponding p-values for cows that did not contribute milk to tank (added one at the time in the model), largely followed the uniform distribution expected under the null hypothesis, despite a slight excess of low p-values (two-fold excess of p-values < 0.01, including three-fold excess of p-values < 0.001)(Supplementary Figure 1B).

Actual data.

We selected 8,696 SNP (1) that were interrogated by the 3 SNP panels used and (2) for which genotype information was complete for all 40 analyzed cows (21 that did contribute milk to the tank and 19 that did not). Supplementary Figure 3 (available online at http://www.journalofdairyscience.org/) shows the distribution of minor allele frequencies in the actual data set for these 8,696 SNP. The distribution was fairly uniform, hence comparable to the simulated SNP sets.Figure 2 shows the relation between measured SCC and SSC estimated from the adjusted B-allele proportions (f_j) in the pooled milk sample. The correlation was highly significant ($P < 0.0001$; R2 = 0.60), yet lower than in equivalent simulated data sets. We then evaluated our ability to distinguish cows contributing milk to the tank from those that were not. The P-values associated with $\boldsymbol{pc_i}$ were $\leq 10^{-4}$ for all cows contributing milk, except the one with lowest SCC (10,000/mL; P = 0.006). However, when adding cows that did not contribute milk to the tank one-by-one in the linear model, the P-value of the corresponding regression coefficients ($\boldsymbol{pc_i}$) was ≤0.0025 (i.e., a Bonferroni-corrected 5% threshold) for approximately one-third of the cows (Supplementary Figure 4, available online at http://www.journalofdairyscience.org/).

Discussion

In this work, we present a novel approach for determining SCC for individual cows by genotyping a sample of milk from the milk tank, i.e. without having to perform a separate measurement for each animal. The method presupposes that individual SNP genotypes are available for each cow. This will increasingly correspond to reality as (i) genotyping costs continue to drop, and (ii) applications of genomic selection extend to cows.

The method proved effective using both simulated and real data. Not unexpectedly, for comparable scenarios in terms of number of cows and SNPs, the coefficient of determination (r^2) was considerably better with the simulated than with the real data. Several factors could contribute to this discrepancy, including (i) inaccuracies in counting somatic cells in real samples, (ii) underestimation of the inaccuracies in estimating SNP allele frequencies in the simulation, (iii) absence of linkage disequilibrium between simulated SNPs leading to an overestimation of the informativeness of the simulated SNP sets, (iv) relatedness between real but not simulated cows. For all these reasons, application of the proposed method in the field would certainly benefit from maximizing the number of utilized SNPs to compensate for propagation of inaccuracies from various sources.

The SNP genotypes of the individual cows could either be real genotypes, or - assuming that the SNP panel used on the tank milk is not identical to that used on the cows – imputed genotypes. Methods for genotype imputation are well established and particularly effective in livestock as close relatives are often genotyped, allowing exploitation of Mendelian and within-family linkage information in addition to population-wide linkage disequilibrium information. Alternatively, it might be possible to impute allelic frequencies for the tank milk at missing SNP positions from the allelic frequencies measured at flanking markers and from the known linkage disequilibrium structure in the herd of interest. The impact of genotype (frequency) imputation on the precision of the proposed method will have to be evaluated.

In Belgium, SCC is typically evaluated ~10 times per year at a cost of ~25 Euros/cow/year. In a farm with 100 cows, this amounts to ~2,500 Euros/year. At the present price, this corresponds to > 25 analyses with a 50K SNP array, and ~ 10 analyses with a 700K SNP array. Thus the argument can be made that - even today - the proposed methodology is price-competitive.

Analyzing DNA extracted from tank milk will become increasingly useful and cost-effective as other applications are becoming available. Monitoring the milk's microbiome, including detection of putative pathogens, by means of high-throughput sequencing of – for instance - 16S rRNA amplicons is certainly one application that is virtually immediately practical.

We explored the possibility to determine whether specific cows did or not contribute milk to the tank. This might for instance be useful to monitor the respect of exclusion of milk from treated cows to the tank. Related applications have previously been evaluated in the context of human genetics (Homer *et al.* 2008). In this study, we used the statistical significance of the cow-specific $\boldsymbol{pc_i}$ regression coefficients (measured using a standard t-test) as an indicator of the presence or absence of milk of a specific cow in the pool. While specificity and sensitivity appeared satisfactory with the simulated data, this was not the case with real data. The reasons underlying this discrepancy could be multiple and are presently being examined.

In summary, we herein describe the principles and demonstrate the feasibility of a novel approach to determine SCC of individual cows that has the potential to be a valuable addition to the arsenal of methods to control mastitis including subclinical.

Acknowledgments

We are grateful to Erik Mullaart and CRV for providing us with the milk samples and SNP genotypes used to demonstrate the feasibility of the method on real data. We are grateful to Frédéric Rollin, Christian Hanzen and Leonard Théron for fruitful discussions. This work was funded by the Unit of Animal Genomics.

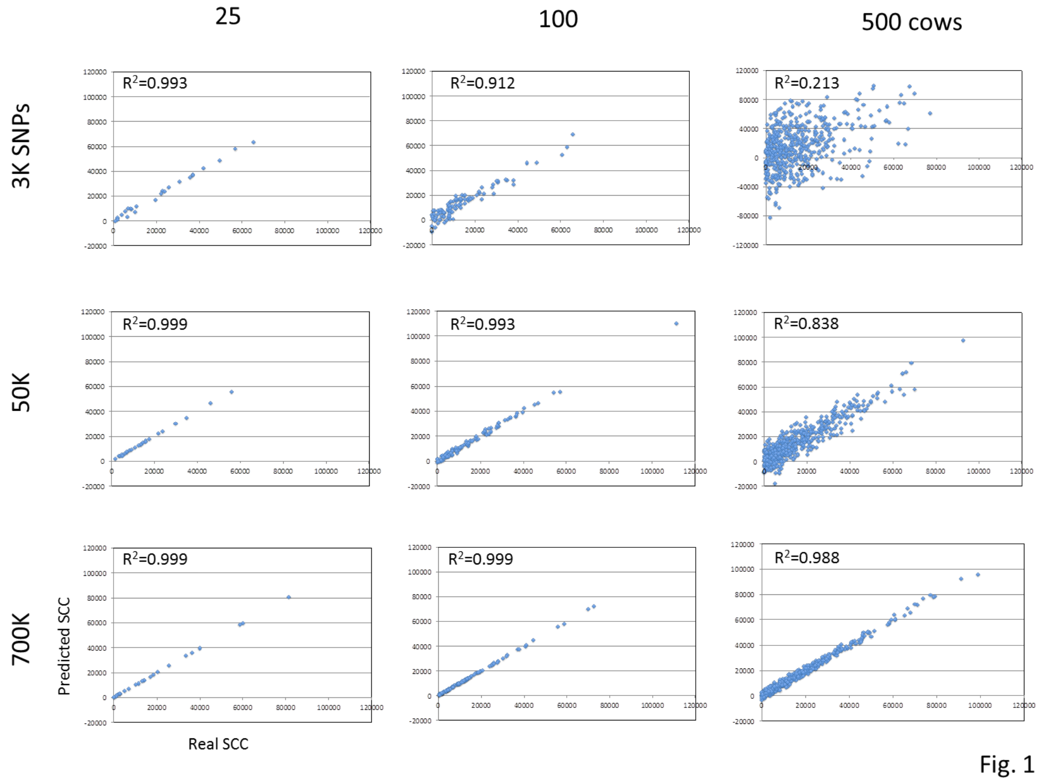


Fig. 1

Figure 1: Simulated data. Representative examples of the relationship between actual SCC and predictions based on SNP genotyping of a sample from the tank milk. Evaluated scenarios consider 25, 100 and 500 cows genotyped for 3,000, 50,000 or 700,000 SNPs. R^2 corresponds to the coefficient of determination.

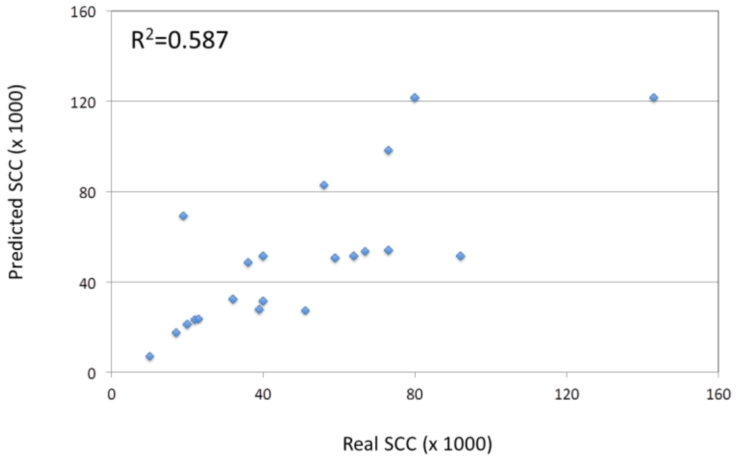
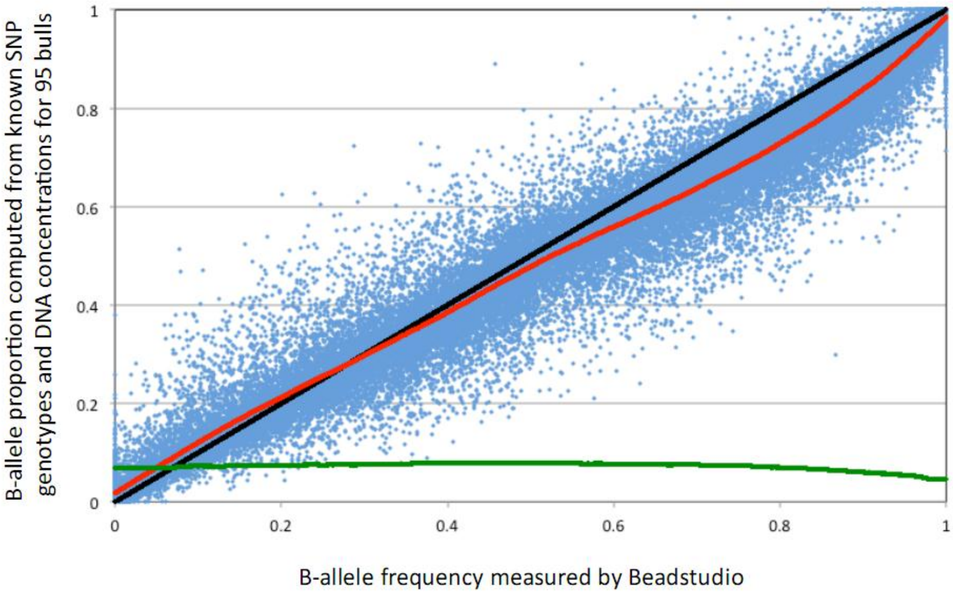
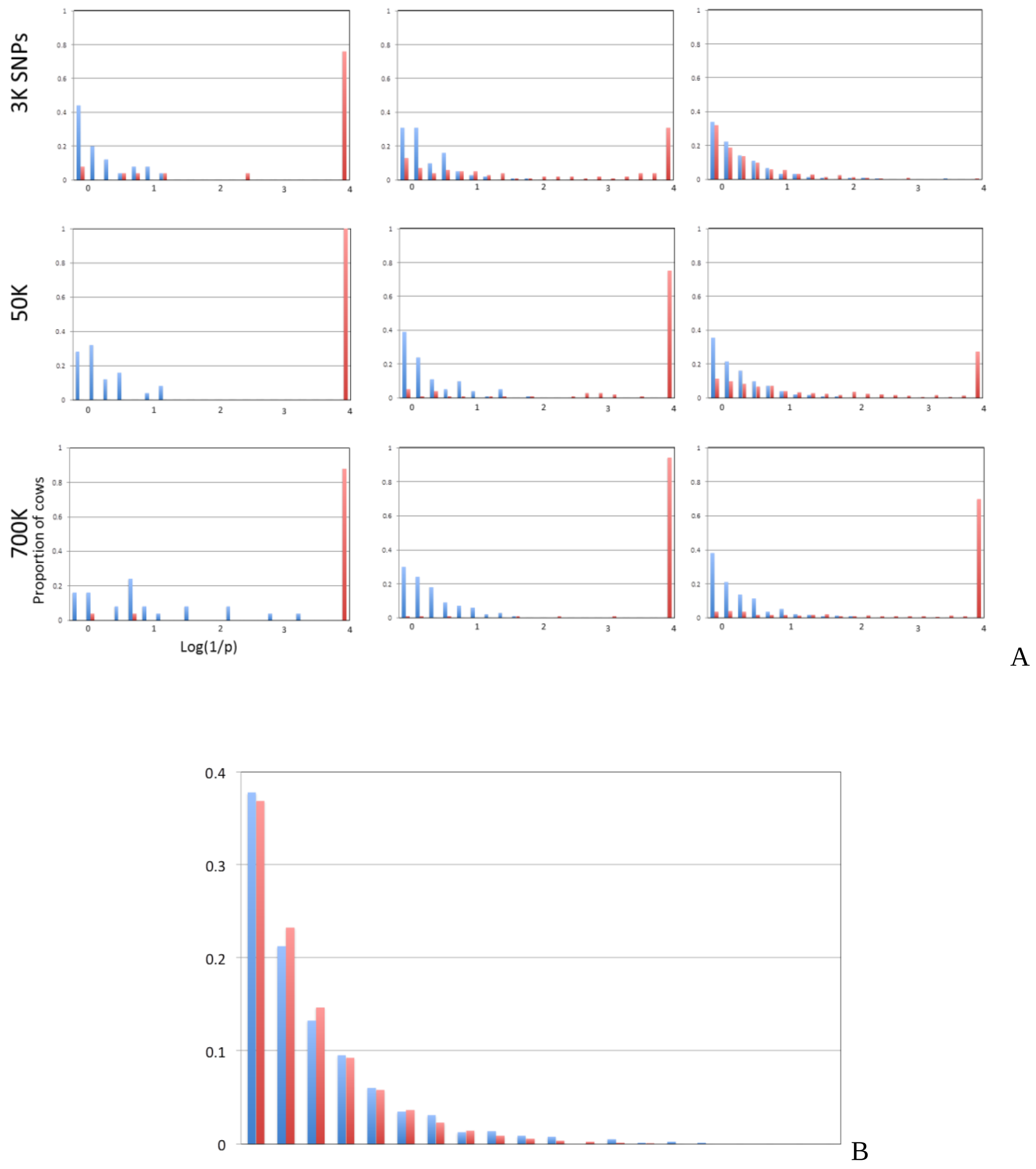


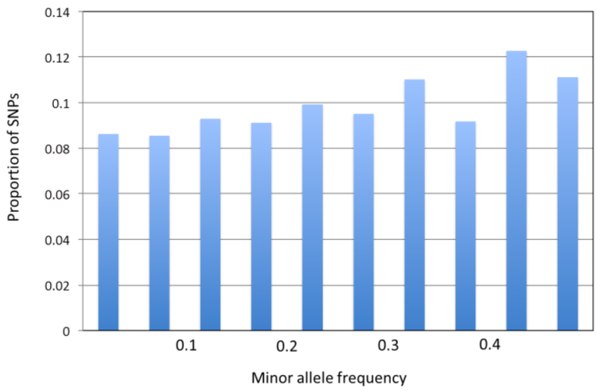
Figure 2: Real data (A) Relationship between measured SCC and SCC estimated by SNP genotyping milk from a tank including milk from 20 cows. **(B)** (Results with imputed genotypes)



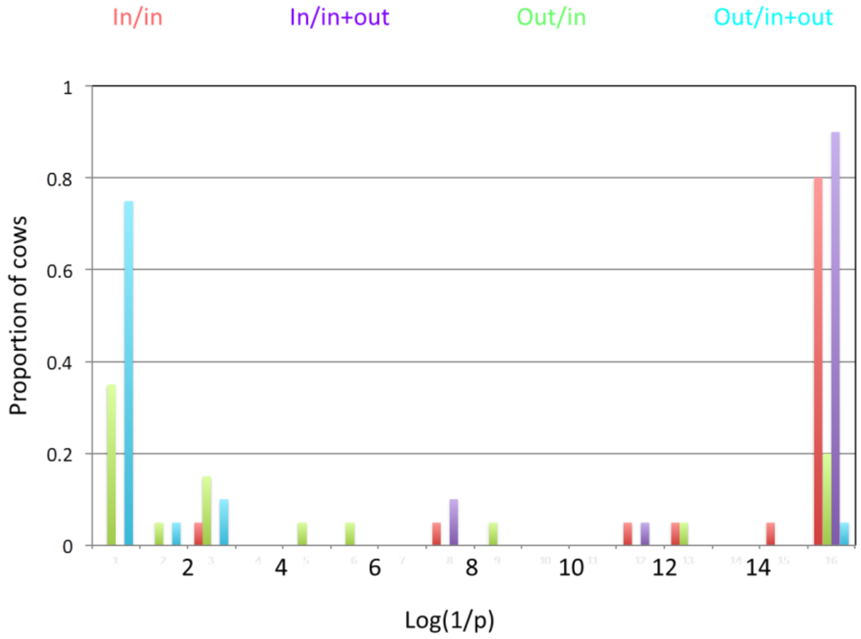
Supplementary Figure 1: Relationship between B-allele frequency (estimated from Illumina BovineSNP50 Beadchip using Beadstudio) and B-allele proportion (computed from the genotype and DNA concentration) in 95 bulls contributing to DNA pool, for 45,248 SNPs with complete genotype data. The black line corresponds to a hypothetical "perfect" regression with 0 intercept and slope of one. The red line corresponds to a curve fitted by performing local linear regressions using information from 6,000 "flanking" SNPs (i.e. having comparable B-allele frequency). The green line shows the corresponding residual standard errors.



Suppplementary Figure 2: Simulated data. (A) Frequency distribution of the statistical significance (t-test) of the ***pc$_i$*** regression coefficients for (i) simulated cows that did not contribute milk to the tank added one-by-one in a linear model including all the cows that did contribute milk to the tank (blue bars), (ii) simulated cows that did contribute milk to the tank and use of a linear model where all cows included in the model contributed milk to the tank (red bars). **(B)** Expected (null hypothesis, red) and observed (blue) distribution of log(1/p) values for simulated cows that did not contribute milk to the tank. In general expected and observed distributions match closely, with however a slight excess of observed low p-values.



Supplementary Figure 3: Real data. Distribution of minor allele frequency amongst 40 analyzed cows for 8,696 SNPs utilized to estimate SCC of individual cows from the SNP allele frequencies in the tank milk.



Supplementary Figure 4: Real data. Distribution of the statistical significance (log(1/p) values estimated using a standard t-test) of the regression coefficient corresponding to (i) the 20 cows that did contribute milk to the tank (Red=In/in: linear model including the 20 cows that did contribute milk to the tank; mauve=IN/in+out: linear model including all 40 cows), and (ii) the 20 cows that did not contribute milk to the tank (Green=Out/in: linear model including the 20 cows that did contribute milk to the tank and one that did not; turquoise=Out/in+out: linear model including all 40 cows).

Conclusions and Perspectives

In this thesis, we describe the development of a novel method for association mapping of variants/genes that influence binary traits segregating in outbred populations, as well as of a software package (GLASCOW) to implement it (Zhang *et al.* 2012).

One of the main features of the proposed approach is that it exploits haplotype information rather than performing single-SNP analysis as most available methods do. Haplotype information is extracted using a previously described Hidden Markov Model (Druet & Georges 2010) that simultaneously phases the genotypes and assigns the ensuing haplotypes to a predetermined number of ancestral haplotypes corresponding to the hidden states of the model. The method does not require the use of "windows" with often arbitrarily defined boundaries. It allows for seamless integration of within-family linkage and across-family linkage disequilibrium information.

The use of haplotype information is expected to endow the method with superior performances - at least - in some circumstances. Superiority of a haplotype-based method is expected when none of the SNPs considered individually is in high LD with the causative variant(s), but when some of the haplotypes are. Other scenarios where haplotype information may be more powerful than single SNP analysis is in the case of epistatic interactions between closely linked variants, i.e. when the biological effects requires a combination of genetic variants. A potential handicap of the haplotype-based methods is related to multiple testing: we typically estimate the effect of 10-20 ancestral haplotypes at each marker position (instead of two alleles). The balance between gains and losses will probably depend on the LD features of the considered locus and the complexity of the underlying biology.

The proposed method may be viewed as a "non-parametric" method in the sense that one tests for deviations from random expectations rather than testing a specific genetic model. As a consequence the method is applicable to a broad range of scenarios and can deal with recessivity and dominance, allelic homo- and heterogeneity, phenocopies, locus heterogeneity, oligogenic and polygenic inheritance. The complexity of the determinism will of course affect the corresponding detection power.

Another important feature of the approach is that it corrects for population stratification using a random "individual animal" or polygenic model. The covariance between the animal effects can either be predicted from pedigree or from genotype data. In the former case, we typically use the available haplotype (rather than single SNP) information to compute genome-wide IBD probabilities between pairs of individuals, which should – intuitively – be more accurate. Stratification issues, including both polygenic and environmental confounding, are commonplace in animal populations. It is essential to correct for these when performing association mapping, and the animal model is viewed by many as the most effective way to achieve this goal.

The efficacy of the proposed method was demonstrated on traits with a broad range of genetic determinisms. It was first applied to simple autosomal recessive genetic defects with allelic homogeneity. This led to the localization of the mutations underlying prolonged gestation and arthrogryposis in the Belgian Blue Cattle Breed (BBCB) (Zhang *et al.* 2012). It was then applied retrospectively to color-sidedness, which is inherited as an autosomal dominant trait BBCB (Durkin *et al.* 2012). It was subsequently applied successfully to stunted growth in BBCB, leading to the detection of an autosomal recessive allele that would explain 40% of reported cases. The method could thus overcome the occurrence of as of yet unexplained phenocopies in this specific case (Sartelet *et al.* 2012). It was also applied to scan the genome for risk loci for recurrent laryngeal neuropathy, a multifactorial disease of the horse (Dupuis *et al.* 2011). We used a related haplotype-based method designed for the analysis of quantitative traits (Druet and Georges, 2010), to fine-map loci influencing hematological traits segregating in a porcine line-cross (Zhang *et al.* 2013). The latter are also complex multifactorial traits.

A non-anticipated use of the proposed method is for the genomic localization of Copy Number Variants (CNV). CNVs are typically assumed to correspond to tandem or at least closely linked copies of a given genomic segment. While this indeed appears to apply to the majority of CNV, a non-negligible proportion of CNV may correspond to non-syntenic or at least not closely linked copies of the given genomic segment. This was unambiguously demonstrated for the Cs_{29} allele causing color-sidedness in BBCB and several other breeds, which was shown to correspond to the translocation of a segment of bovine chromosome 6 to chromosome 29 (Durkin *et al.* 2012). This finding spurred us to search for other examples of “non-syntenic CNV”. We proposed to do this by considering CNV genotype as a binary trait (copy number 2 = “controls” versus copy number >2 = “cases”) and to scan the genome for regions associated with this newly defined trait. In the case of a classical “syntenic CNV” the only genomic region expected to show an association would correspond to the genomic coordinates to which the CNV sequences map in the reference genome. In the case of “non-syntenic CNVs” a significant association would be detected somewhere else in the genome. By doing so in humans, cattle and horse we have detected tens such instances of “trans-association”, identifying CNVs of which some might result from the same mechanism causing translocation via circular intermediates uncovered for color-sidedness (Durkin *et al.* 2012; Dupuis *et al.* 2012).

The last part of this work deals with specific applications of marker information in animal breeding. Despite considerable drops in price witnessed over the last five years, the costs of SNP genotyping remain an obstacle for the systematic application of genomic selection on all or nearly all animals in livestock populations. Methods are therefore being developed to extract a maximum of information from a minimum number of genotyped SNPs. One way to reduce the overall genotyping costs would be to only genotype a selected fraction of the population at high density (or even sequence them), to genotype the remainder of the population with low-density panels, yet to exploit information about relatedness combined with principles of linkage analysis to impute missing SNP genotypes upon the entire population. We have tested the feasibility of this approach in a scenario that would mimic the actual situation in dairy cattle in terms of population and available SNP panels.

We finally propose and evaluate the feasibility of a new application of genome-wide SNP information in dairy cattle breeding. Mastitis remains the most costly health issue in dairy cattle breeding. Rapid detection of cows with subclinical mastitis may contribute to better control of the disease. We have proposed to achieve this by monitoring individual somatic cell counts of all cows in the farm by analyzing one sample of milk from the farm's milk tank, hence containing a mixture of milk from all cows on the farm. The deconvolution of the signal obtained on the milk mixture in order to obtain information for individual cows is achieved by realizing that the combination of SNP allele frequencies (measured from the relative intensities of the fluorescence signals corresponding to the two alleles at each SNP) observed in the milk tank across all interrogated SNPs, can only be obtained by mixing the milk (or rather the somatic cells) of individual cows in computable proportions (provided that all cows have been individually genotyped for the same SNPs). We demonstrate the feasibility of this approach "in principle" by simulation, and perform a pilot experiment that indicates that implementation in the real world will likely be achievable. It seems reasonable to speculate that the analysis of nucleic acids in the milk tank will become routine in the future particularly for the detection, of specific pathogens. The proposed assay may be combined with microbiological analyses to maximize the extraction of information for the farmer's benefit.

The mapping methods developed and applied in this thesis still fit within the classical "positional cloning paradigm", in which: (i) the loci of interest are mapped – one-by-one - by association analysis using genetic markers that correspond to a relatively small subset of all the genetic variants that segregate in the population, and (ii) the identified loci are subject – one-by-one - to a follow-up "fine-mapping" and "functional analysis" to hopefully identify causative variants and genes.

Rapid advances in next generation sequencing may considerably change this paradigm. It is likely that in the near future, the full sequence information will become available for all individuals considered in a "mapping" experiment, whether through actual resequencing or imputation. In this context, the whole issue of the use of haplotype information will become mostly irrelevant as the causative variants will be part of the data. At present, the functional studies are largely disconnected from the actual mapping step. This may also change in the near future as functional information will increasingly be gathered in Encode-like projects, and become accessible in public databases. Thus it will become possible to readily integrate functional information at the mapping stage. Examples of such advances can already be guessed from the intents to use functional information to alter prior probabilities in Bayesian approaches towards genomic selection, or from the prediction of the disruptive nature of variants when performing burden tests.

It seems reasonable to also speculate that the "one locus-by-one locus" paradigm will evolve in more systematic "simultaneous searches". By that we mean that the effect of variants will be estimated conditional on that of as many as possible if not all other variants. In some way this is already accomplished by the Bayesian approaches towards genomic selection. However, in these approaches distinct variants are assumed to act additively, while there is growing evidence that properly account for epistatic interactions may reveal novel biology.

Whichever methods and approaches end-up dominating forward genetics, it can be anticipated that tremendous progress will be achieved in the near future in establishing genotype-phenotype maps, including for the complex traits of relevance to medicine and agriculture.

References

Abecasis G.R., Cherny S.S., Cookson W.O. & Cardon L.R. (2001) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101.

Altshuler D., Daly M.J. & Lander E.S. (2008) Genetic mapping in human disease. *Science* **322**, 881-8.

Amin N., van Duijn C.M. & Aulchenko Y.S. (2007) A genomic background based method for association analysis in related individuals. *PloS one* **2**, e1274.

Aranzana M.J., Kim S., Zhao K., Bakker E., Horton M., Jakob K., Lister C., Molitor J., Shindo C. & Tang C. (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* **1**, e60.

Archer R., Lindsay W. & Duncan I. (1989) Equine laryngeal hemiplegia-endoscopic survey of 400 draft horses. *Vet Surg* **18**, 62-3.

Aulchenko Y.S., De Koning D.J. & Haley C. (2007a) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577-85.

Aulchenko Y.S., Ripke S., Isaacs A. & Van Duijn C.M. (2007b) GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294-6.

Azmi P. & Seth A. (2005) RNF11 is a multifunctional modulator of growth factor receptor signalling and transcriptional regulation. *Eur J Cancer* **41**, 2549-60.

Balding D.J. (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**, 781-91.

Ball P. (2013) DNA: Celebrate the unknowns. *Nature* **496**, 419-20.

Bamshad M. & Wooding S.P. (2003) Signatures of natural selection in the human genome. *Nat Rev Genet* **4**, 99-111.

Bansal V., Libiger O., Torkamani A. & Schork N.J. (2010) Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* **11**, 773-85.

Barakzai S. (2009) Heritability of recurrent laryngeal neuropathy. In: *4th World equine airways symposium, Berne, Switzerland*, pp. 40-2. www.ivis.org.

Barisic N., Claeys K.G., Sirotkovic-Skerlev M., Lofgren A., Nelis E., De Jonghe P. & Timmerman V. (2008) Charcot-Marie-Tooth disease: a clinico-genetic confrontation. *Ann Hum Genet* **72**, 416-41.

Barrett J., Fry B., Maller J. & Daly M. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-5.

Beard W. & Hayes H. (1993) Risk factors for laryngeal hemiplegia in the horse. *Prev Vet Med* **17**, 57-63.

Berndt S.I., Gustafsson S., Mägi R., Ganna A., Wheeler E., Feitosa M.F., Justice A.E., Monda K.L., Croteau-Chonka D.C. & Day F.R. (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* **45**, 501-12.

Bhangale T.R., Rieder M.J. & Nickerson D.A. (2008) Estimating coverage and power for genetic association studies using near-complete variation data. *Nat Genet* **40**, 841-3.

Blard G., Zhang Z., Coppieters W. & Georges M. (2012) Identifying cows with subclinical mastitis by bulk single nucleotide polymorphism genotyping of tank milk. *J Dairy Sci* **95**, 4109-13.

Bloom J.S., Ehrenreich I.M., Loo W.T., Lite T.-L.V. & Kruglyak L. (2013) Finding the sources of missing heritability in a yeast cross. *Nature* **494**, 234-7.

Boldman K. (1995) *A manual for use of MTDFREML. A set of programs to obtain estimates of variances and covariances [DRAFT]*. US Department of Agriculture, Agricultural Research Service.

Borneman A.R., Desany B.A., Riches D., Affourtit J.P., Forgan A.H., Pretorius I.S., Egholm M. & Chambers P.J. (2011) Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of Saccharomyces cerevisiae. *PLoS Genet* **7**, e1001287.

Bradbury P.J., Zhang Z., Kroon D.E., Casstevens T.M., Ramdoss Y. & Buckler E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633-5.

Brakenhoff J.E., Holcombe S.J., Hauptman J.G., Smith H.K., Nickels F.A. & Caron J.P. (2006) The prevalence of laryngeal disease in a large population of competition draft horses. *Vet Surg* **35**, 579-83.

Breslow N.E. & Clayton D.G. (1993) Approximate inference in generalized linear mixed models. *J Am Stat Assoc* **88**, 9-25.

Brooks S.A., Lear T.L., Adelson D.L. & Bailey E. (2007) A chromosome inversion near the *KIT* gene and the Tobiano spotting pattern in horses. *Cytogenet Genome Res* **119**, 225-30.

Brown J.A., Hinchcliff K.W., Jackson M.A., Dredge A.F., O'Callaghan R.A., McCaffrey J.R., Slocombe R.F. & Clarke A.F. (2005) Prevalence of pharyngeal and laryngeal abnormalities in Thoroughbreds racing in Australia, and their association with performance. *Equine Vet J* **37**, 397-401.

Browning S.R. (2008) Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum Genet* **124**, 439-50.

Browning S.R. & Browning B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97.

Buckley B.A., Burkhart K.B., Gu S.G., Spracklin G., Kershner A., Fritz H., Kimble J., Fire A. & Kennedy S. (2012) A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. *Nature* **489**, 447-51.

Cahill J.I. & Goulden B.E. (1986a) Equine laryngeal hemiplegia. Part I. A light microscopic study of peripheral nerves. *N Z Vet J* **34**, 161-9.

Cahill J.I. & Goulden B.E. (1986b) Equine laryngeal hemiplegia. Part II. An electron microscopic study of peripheral nerves. *N Z Vet J* **34**, 170-5.

Cahill J.I. & Goulden B.E. (1986c) Equine laryngeal hemiplegia. Part III. A teased fibre study of peripheral nerves. *N Z Vet J* **34**, 181-5.

Cahill J.I. & Goulden B.E. (1987) The pathogenesis of equine laryngeal hemiplegia--a review. *N Z Vet J* **35**, 82-90.

Chang Y.F., Imam J.S. & Wilkinson M.F. (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76**, 51-74.

Chanock S.J., Manolio T., Boehnke M., Boerwinkle E., Hunter D.J., Thomas G., Hirschhorn J.N., Abecasis G., Altshuler D., Bailey-Wilson J.E., Brooks L.D., Cardon L.R., Daly M., Donnelly P., Fraumeni J., Freimer N.B., Gerhard D.S., Gunter C., Guttmacher A.E., Guyer M.S., Harris E.L., Hoh J., Hoover R., Kong C.A., Merikangas K.R., Morton C.C., Palmer L.J., Phimister E.G., Rice J.P., Roberts J., Rotimi C., Tucker M.A., Vogan K.J., Wacholder S., Wijsman E.M., Winn D.M. & Collins F.S. (2007) Replicating genotype-phenotype associations. *Nature* **447**, 655-60.

Charlier C., Agerholm J.S., Coppieters W., Karlskov-Mortensen P., Li W., de Jong G., Fasquelle C., Karim L., Cirera S. & Cambisano N. (2012) A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina. *PloS one* **7**, e43085.

Charlier C., Coppieters W., Rollin F., Desmecht D., Agerholm J.S., Cambisano N., Carta E., Dardano S., Dive M. & Fasquelle C. (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nat Genet* **40**, 449-54.

Charlier C., Denys B., Belanche J.I., Coppieters W., Grobet L., Mni M., Womack J., Hanset R. & Georges M. (1996) Microsatellite mapping of the bovine roan locus: a major determinant of White Heifer disease. *Mamm Genome* **7**, 138-42.

Cho I.C., Park H.B., Yoo C.K., Lee G.J., Lim H.T., Lee J.B., Jung E.J., Ko M.S., Lee J.H. & Jeon J.T. (2011) QTL analysis of white blood cell, platelet and red blood cell-related traits in an F2 intercross between Landrace and Korean native pigs. *Anim Genet* **42**, 621-6.

Cockett N.E., Jackson S.P., Shay T.L., Farnir F., Berghmans S., Snowder G.D., Nielsen D.M. & Georges M. (1996) Polar overdominance at the ovine callipyge locus. *Science* **273**, 236-8.

Collins N., Milne E., Hahn C. & Dixon P. (2009) Correlation of the Havemeyer endoscopic laryngeal grading system with histopathological changes in equine Cricoarytenoideus dorsalis muscles. *Ir Vet J* **62**, 334-8.

Conrad D.F., Pinto D., Redon R., Feuk L., Gokcumen O., Zhang Y., Aerts J., Andrews T.D., Barnes C., Campbell P., Fitzgerald T., Hu M., Ihm C.H., Kristiansson K., Macarthur D.G., Macdonald J.R., Onyiah I., Pang A.W., Robson S., Stirrups K., Valsesia A., Walter K., Wei J., Tyler-Smith C., Carter N.P., Lee C., Scherer S.W. & Hurles M.E. (2010) Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-12.

Cordell H.J. (2009) Detecting gene–gene interactions that underlie human diseases. *Nat Rev Genet* **10**, 392-404.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-94.

Cutter A.D. & Payseur B.A. (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet* **14**, 262–74.

D íaz S., Fargione J., Chapin III F.S. & Tilman D. (2006) Biodiversity loss threatens human well-being. *PLoS Biol* **4**, e277.

Dalvit C., De Marchi M. & Cassandro M. (2007) Genetic traceability of livestock products: A review. *Meat Sci* **77**, 437-49.

De Koning D.-J., Bovenhuis H. & van Arendonk J.A. (2002) On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. *Genetics* **161**, 931-8.

de los Campos G., Gianola D. & Allison D.B. (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat Rev Genet* **11**, 880-6.

de Roos A., Schrooten C. & Druet T. (2011) Genomic breeding value estimation using genetic markers, inferred ancestral haplotypes, and the genomic relationship matrix. *J Dairy Sci* **94**, 4708-14.

de Roos A., Schrooten C., Mullaart E., Calus M. & Veerkamp R. (2007) Breeding value estimation for fat percentage using dense markers on *Bos taurus* autosome 14. *J Dairy Sci* **90**, 4821-9.

Devlin B. & Roeder K. (1999) Genomic control for association studies. *Biometrics* **55**, 997-1004.

Devlin B., Roeder K. & Wasserman L. (2001) Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* **60**, 155-66.

Dickson S.P., Wang K., Krantz I., Hakonarson H. & Goldstein D.B. (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol* **8**, e1000294.

Dixon P.M., Hahn C.N. & Barakzai S.Z. (2009) Recurrent laryngeal neuropathy (RLN) research: where are we and to where are we heading? *Equine Vet J* **41**, 324-7.

Dixon P.M., McGorum B.C., Railton D.I., Hawe C., Tremaine W.H., Pickles K. & McCann J. (2001) Laryngeal paralysis: a study of 375 cases in a mixed-breed population of horses. *Equine Vet J* **33**, 452-8.

Dixon P.M., McGorum B.C., Railton D.I., Hawe C., Tremaine W.H., Pickles K. & McCann J. (2002) Clinical and endoscopic evidence of progression in 152 cases of equine recurrent laryngeal neuropathy (RLN). *Equine Vet J* **34**, 29-34.

Dohoo I. & Leslie K. (1991) Evaluation of changes in somatic cell counts as indicators of new intramammary infections. *Preve Vet Med* **10**, 225-37.

Donnelly P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728-31.

Drögemüller C., Engensteiner M., Moser S., Rieder S. & Leeb T. (2009) Genetic mapping of the belt pattern in Brown Swiss cattle to BTA3. *Anim Genet* **40**, 225-9.

Druet T. & Farnir F.P. (2011) Modeling of identity-by-descent processes along a chromosome between haplotypes and their genotyped ancestors. *Genetics* **188**, 409-19.

Druet T., Fritz S., Boussaha M., Ben-Jemaa S., Guillaume F., Derbala D., Zelenika D., Lechner D., Charon C., Boichard D., Gut I.G., Eggen A. & Gautier M. (2008) Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. *Genetics* **178**, 2227-35.

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

Druet T., Schrooten C. & de Roos A. (2010) Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *J Dairy Sci* **93**, 5443-54.

Ducharme N.G., Hackett R.P., Fubini S.L. & Erb H.N. (1991) The reliability of endoscopic examination in assessment of arytenoid cartilage movement in horses. Part II. Influence of side of examination, reexamination, and sedation. *Vet Surg* **20**, 180-4.

Duerr R.H., Taylor K.D., Brant S.R., Rioux J.D., Silverberg M.S., Daly M.J., Steinhart A.H., Abraham C., Regueiro M., Griffiths A., Dassopoulos T., Bitton A., Yang H., Targan S., Datta L.W., Kistner E.O., Schumm L.P., Lee A.T., Gregersen P.K., Barmada M.M., Rotter J.I., Nicolae D.L. & Cho J.H. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461-3.

Duncan I.D., Griffiths I.R. & Madrid R.E. (1978) A light and electron microscopic study of the neuropathy of equine idiopathic laryngeal hemiplegia. *Neuropathol Appl Neurobiol* **4**, 483-501.

Dupuis J. & Siegmund D. (1999) Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151**, 373-86.

Dupuis M.C., Zhang Z., Druet T., Denoix J.M., Charlier C., Lekeux P. & Georges M. (2011) Results of a haplotype-based GWAS for recurrent laryngeal neuropathy in the horse. *Mamm Genome* **22**, 613-20.

Dupuis M.C., Zhang Z., Durkin K., Charlier C., Lekeux P. & Georges M. (2013) Detection of copy number variants in the horse genome and examination of their association with recurrent laryngeal neuropathy. *Anim Genet* **44**, 206-8.

Durkin K., Coppieters W., Drogemuller C., Ahariz N., Cambisano N., Druet T., Fasquelle C., Haile A., Horin P., Huang L., Kamatani Y., Karim L., Lathrop M., Moser S., Oldenbroek K., Rieder S., Sartelet A., Solkner J., Stalhammar H., Zelenika D., Zhang Z., Leeb T., Georges M. & Charlier C. (2012) Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* **482**, 81-4.

Durrant C., Zondervan K.T., Cardon L.R., Hunt S., Deloukas P. & Morris A.P. (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am J Hum Genet* **75**, 35-43.

Edfors-Lilja I., Wattrang E., Magnusson U. & Fossum C. (1994) Genetic variation in parameters reflecting immune competence of swine. *Vet Immunol Immunop* **40**, 1-16.

Eding H. (2001) Marker-based estimates of between and within population kinships for the conservation of genetic diversity. *J Anim Breed Genet* **118**, 141-59.

Escribano L., Ocqueteau M., Almeida J., Orfao A. & San Miguel J.F. (1998) Expression of the *c-kit* (CD117) molecule in normal and malignant hematopoiesis. *Leuk Lymphoma* **30**, 459-66.

Evans D.M., Frazer I.H. & Martin N.G. (1999) Genetic and environmental causes of variation in basal levels of blood cells. *Twin Res* **2**, 250-7.

Ewens W. & Spielman R. (2003) The transmission/disequilibrium test. *Handbook of statistical genetics*.

Fasquelle C., Sartelet A., Li W., Dive M., Tamma N., Michaux C., Druet T., Huijbers I.J., Isacke C.M., Coppieters W., Georges M. & Charlier C. (2009) Balancing selection of a frame-shift mutation in the *MRC2* gene accounts for the outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. *PLoS Genet* **5**, e1000666.

Fesus L., Zsolnai A. & Komlosi I. (2005) Influence of porcine coat colour genotypes on haematological parameters, piglet birth weight and body weight gain until weaning. *J Anim Breed Genet* **122**, 127-30.

Feuk L., Carson A.R. & Scherer S.W. (2006) Structural variation in the human genome. *Nat Rev Genet* **7**, 85-97.

Flister M.J., Tsaih S.-W., O'Meara C.C., Endres B., Hoffman M.J., Geurts A.M., Dwinell M.R., Lazar J., Jacob H.J. & Moreno C. (2013) Identifying multiple causative genes at a single GWAS locus. *Genome Res*, doi: 10.1101/gr.160283.113.

Frazer K.A., Murray S.S., Schork N.J. & Topol E.J. (2009) Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**, 241-51.

Fritz S., Capitan A., Djari A., Rodriguez S.C., Barbat A., Baur A., Grohs C., Weiss B., Boussaha M. & Esquerr éD. (2013) Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in *GART*, *SHBG* and *SLC37A2*. *PloS one* **8**, e65550.

Fujii J., Otsu K., Zorzato F., de Leon S., Khanna V.K., Weiler J.E., O'Brien P.J. & Maclennan D.H. (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science* **253**, 448-51.

Furlotte N.A., Eskin E. & Eyheramendy S. (2012) Genome-wide association mapping with longitudinal data. *Genet Epidemiol* **36**, 463-71.

Ganesh S.K., Zakai N.A., van Rooij F.J., Soranzo N., Smith A.V., Nalls M.A., Chen M.H., Kottgen A., Glazer N.L., Dehghan A., Kuhnel B., Aspelund T., Yang Q., Tanaka T., Jaffe A., Bis J.C., Verwoert G.C., Teumer A., Fox C.S., Guralnik J.M., Ehret G.B., Rice K., Felix J.F., Rendon A., Eiriksdottir G., Levy D., Patel K.V., Boerwinkle E., Rotter J.I., Hofman A., Sambrook J.G., Hernandez D.G., Zheng G., Bandinelli S., Singleton A.B., Coresh J., Lumley T., Uitterlinden A.G., Vangils J.M., Launer L.J., Cupples L.A., Oostra B.A., Zwaginga J.J., Ouwehand W.H., Thein S.L., Meisinger C., Deloukas P., Nauck M., Spector T.D., Gieger C., Gudnason V., van Duijn C.M., Psaty B.M., Ferrucci L., Chakravarti A., Greinacher A., O'Donnell C.J., Witteman J.C., Furth S., Cushman M., Harris T.B. & Lin J.P. (2009) Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat Genet* **41**, 1191-8.

Gao Y., Ganss B.W., Wang H., Kitching R.E. & Seth A. (2005) The RING finger protein *RNF11* is expressed in bone cells during osteogenesis and is regulated by *Ets1*. *Exp Cell Res* **304**, 127-35.

Garner C., Tatu T., Reittie J.E., Littlewood T., Darley J., Cervino S., Farrall M., Kelly P., Spector T.D. & Thein S.L. (2000) Genetic influences on F cells and other hematologic variables: a twin heritability study. *Blood* **95**, 342-6.

Garrett K.S., Pierce S.W., Embertson R.M. & Stromberg A.J. (2010) Endoscopic evaluation of arytenoid function and epiglottic structure in Thoroughbred yearlings and association with racing performance at two to four years of age: 2,954 cases (1998-2001). *J Am Vet Med Assoc* **236**, 669-73.

Ge B., Gurd S., Gaudin T., Dore C., Lepage P., Harmsen E., Hudson T.J. & Pastinen T. (2005) Survey of allelic expression using EST mining. *Genome Res* **15**, 1584-91.

George A.W., Visscher P.M. & Haley C.S. (2000) Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics* **156**, 2081-92.

Georges M. (2007) Mapping, fine mapping, and molecular dissection of quantitative trait Loci in domestic animals. *Annu Rev Genet* **8**, 131-62.

Gerstein M.B., Kundaje A., Hariharan M., Landt S.G., Yan K.-K., Cheng C., Mu X.J., Khurana E., Rozowsky J. & Alexander R. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100.

Gibson G. (2012) Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-45.

Girard S.L., Gauthier J., Noreau A., Xiong L., Zhou S., Jouan L., Dionne-Laporte A., Spiegelman D., Henrion E. & Diallo O. (2011) Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nat Genet* **43**, 860-3.

Goddard M.E. & Hayes B.J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* **10**, 381-91.

Goulden B.E. & Anderson L.J. (1981a) Equine laryngeal hemiplegia part II: Some clinical observations. *N Z Vet J* **29**, 194-8.

Goulden B.E. & Anderson L.J. (1981b) Equine laryngeal hemiplegia, Part I: Physical characteristics of affected animals. *N Z Vet J* **29**, 151-4.

Graux C., Stevens-Kroef M., Lafage M., Dastugue N., Harrison C., Mugneret F., Bahloula K., Struski S., Gregoire M. & Nadal N. (2008) Heterogeneous patterns of amplification of the *NUP214-ABL1* fusion gene in T-cell acute lymphoblastic leukemia. *Leukemia* **23**, 125-33.

Grobet L., Martin L.J., Poncelet D., Pirottin D., Brouwers B., Riquet J., Schoeberlein A., Dunner S., Menissier F., Massabanda J., Fries R., Hanset R. & Georges M. (1997) A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nat Genet* **17**, 71-4.

Habier D., Fernando R.L. & Dekkers J.C. (2009) Genomic selection using low-density marker panels. *Genetics* **182**, 343-53.

Hager R., Cheverud J.M. & Wolf J.B. (2008) Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. *Genetics* **178**, 1755-62.

Hahn C.N., Matiasek K., Dixon P.M., Molony V., Rodenacker K. & Mayhew I.G. (2008) Histological and ultrastructural evidence that recurrent laryngeal neuropathy is a bilateral mononeuropathy limited to recurrent laryngeal nerves. *Equine Vet J* **40**, 666-72.

Hanset R., Michaux C. & Boonen F. (1994) Linear classification in the Belgian Blue Cattle Breed: phenotypic and genetic parameters. Ottawa, Canada, International Committee for Animal Recording (ICAR), seminar, Beef performance recording and genetic evaluation. *Proc of the 29th biennial session of ICAR* **75**, 231-7.

Harville D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc* **72**, 320-38.

Hastings P., Lupski J.R., Rosenberg S.M. & Ira G. (2009) Mechanisms of change in gene copy number. *Nat Rev Genet* **10**, 551-64.

Hauser M.-T., Aufsatz W., Jonak C. & Luschnig C. (2011) Transgenerational epigenetic inheritance in plants. *BBA-Gene Regul Mech* **1809**, 459-68.

Hayes B. & Goddard M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genet Sel Evol* **33**, 209-30.

Hayes B.J. & Goddard M. (2008) Technical note: Prediction of breeding values using marker-derived relationship matrices. *J Anim Sci* **86**, 2089-92.

Hayes B.J., Pryce J., Chamberlain A.J., Bowman P.J. & Goddard M.E. (2010) Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet* **6**, e1001139.

He X., Sanders S.J., Liu L., De Rubeis S., Lim E.T., Sutcliffe J.S., Schellenberg G.D., Gibbs R.A., Daly M.J. & Buxbaum J.D. (2013) Integrated model of *de novo* and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet* **9**, e1003671.

Hellemans J., Mortier G., De Paepe A., Speleman F. & Vandesompele J. (2007) qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome biol* **8**, R19.

Hill W.G., Goddard M.E. & Visscher P.M. (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* **4**, e1000008.

Hogeveen H., Huijps K. & Lam T.J. (2011) Economic aspects of mastitis: new developments. *N Z Vet J* **59**, 16-23.

Homer N., Szelinger S., Redman M., Duggan D., Tembe W., Muehling J., Pearson J.V., Stephan D.A., Nelson S.F. & Craig D.W. (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* **4**, e1000167.

Hovinen M. & Pyorala S. (2011) Invited review: udder health of dairy cows in automatic milking. *J Dairy Sci* **94**, 547-62.

Hurles M. (2012) Older males beget more mutations. *Nat Genet* **44**, 1174.

IBI T., MIYAKE T., HOBO S., OKI H., ISHIDA N. & SASAKI Y. (2003) Estimation of heritability of laryngeal hemiplegia in the Thoroughbred horse by Gibbs sampling. *J Equine Science* **14**, 81-6.

Irobi J., De Jonghe P. & Timmerman V. (2004) Molecular genetics of distal hereditary motor neuropathies. *Hum Mol Genet* **13 Spec No 2**, R195-202.

Jackson I.J., Budd P., Horn J.M., Johnson R., Raymond S. & Steel K. (2006) Genetics and molecular biology of mouse pigmentation. *Pigment Cell Res* **7**, 73-80.

Johansson A., Pielberg G., Andersson L. & Edfors-Lilja I. (2005) Polymorphism at the porcine Dominant white/KIT locus influence coat colour and peripheral blood cell measures. *Anim Genet* **36**, 288-96.

Kamatani Y., Matsuda K., Okada Y., Kubo M., Hosono N., Daigo Y., Nakamura Y. & Kamatani N. (2010) Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet* **42**, 210-5.

Kamatani Y., Wattanapokayakit S., Ochi H., Kawaguchi T., Takahashi A., Hosono N., Kubo M., Tsunoda T., Kamatani N., Kumada H., Puseenam A., Sura T., Daigo Y., Chayama K., Chantratita W., Nakamura Y. & Matsuda K. (2009) A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat Genet* **41**, 591-5.

Kang H.M., Sul J.H., Service S.K., Zaitlen N.A., Kong S.Y., Freimer N.B., Sabatti C. & Eskin E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-54.

Kang H.M., Zaitlen N.A., Wade C.M., Kirby A., Heckerman D., Daly M.J. & Eskin E. (2008) Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-23.

Karim L., Takeda H., Lin L., Druet T., Arias J.A., Baurain D., Cambisano N., Davis S.R., Farnir F. & Grisart B. (2011) Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nat Genet* **43**, 405-13.

Kashi Y., Hallerman E. & Soller M. (1990) Marker-assisted selection of candidate bulls for progeny testing programmes. *Anim Prod* **51**, 63-74.

Kimura M. (1983) Rare variant alleles in the light of the neutral theory. *Mol Biol Evol* **1**, 84-93.

Kong A., Frigge M.L., Masson G., Besenbacher S., Sulem P., Magnusson G., Gudjonsson S.A., Sigurdsson A., Jonasdottir A. & Jonasdottir A. (2012) Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471-5.

Kong A., Steinthorsdottir V., Masson G., Thorleifsson G., Sulem P., Besenbacher S., Jonasdottir A., Sigurdsson A., Kristinsson K.T. & Jonasdottir A. (2009) Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868-74.

Kumar Kadri N., Sahana G., Charlier C., Iso-Touru T., Guldbrandten B., Karim L., Nielsen U.S., Panitz F., Aamand G.P., Schulman N., Georges M., Vilkki J., Lund M.S. & Druet T. (2013) A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. *PLoS Genet*, Submitted for publication.

Laird N.M. & Ware J.H. (1982) Random-effects models for longitudinal data. *Biometrics*, 963-74.

Lander E. & Kruglyak L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* **11**, 241-7.

Lane J.G., Ellis D.R. & Greet T.R. (1987) Observations on the examination of Thoroughbred yearlings for idiopathic laryngeal hemiplegia. *Equine Vet J* **19**, 531-6.

Lango Allen H., Estrada K., Lettre G., Berndt S.I., Weedon M.N., Rivadeneira F., Willer C.J., Jackson A.U., Vedantam S. & Raychaudhuri S. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-8.

Lawson H.A., Cheverud J.M. & Wolf J.B. (2013) Genomic imprinting and parent-of-origin effects on complex traits. *Nat Rev Genet* **14**, 609–17.

Lee E.G., Boone D.L., Chai S., Libby S.L., Chien M., Lodolce J.P. & Ma A. (2000) Failure to regulate TNF-induced NF-kappaB and cell death responses in A20-deficient mice. *Science* **289**, 2350-4.

Lewis C.M., Levinson D.F., Wise L.H., DeLisi L.E., Straub R.E., Hovatta I., Williams N.M., Schwab S.G., Pulver A.E. & Faraone S.V. (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *Am J Hum Genet* **73**, 34-48.

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

Lin S., Chakravarti A. & Cutler D.J. (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet* **36**, 1181-8.

Liu G., Ventura M., Cellamare A., Chen L., Cheng Z., Zhu B., Li C., Song J. & Eichler E. (2009) Analysis of recent segmental duplications in the bovine genome. *BMC genomics* **10**, 571.

Liu G.E., Hou Y., Zhu B., Cardone M.F., Jiang L., Cellamare A., Mitra A., Alexander L.J., Coutinho L.L. & Dell'Aquila M.E. (2010) Analysis of copy number variations among diverse cattle breeds. *Genome Res* **20**, 693-703.

Lohr N.J., Molleston J.P., Strauss K.A., Torres-Martinez W., Sherman E.A., Squires R.H., Rider N.L., Chikwava K.R., Cummings O.W., Morton D.H. & Puffenberger E.G. (2010) Human ITCH E3 ubiquitin ligase deficiency causes syndromic multisystem autoimmune disease. *Am J Hum Genet* **86**, 447-53.

Long A.D., Mullaney S.L., Mackay T.F. & Langley C.H. (1996) Genetic interactions between naturally occuning alleles at quantitative trait loci and mutant alleles at candidate loci affecting bristle number in *Drosophila* melanogaster. *Genetics* **144**, 1497-510.

Luo W., Chen S., Cheng D., Wang L., Li Y., Ma X., Song X., Liu X., Li W., Liang J., Yan H., Zhao K., Wang C. & Zhang L. (2012) Genome-wide association study of porcine hematological parameters in a Large White x Minzhu F2 resource population. *Int J Biol Sci* **8**, 870-81.

Lynch M. & Walsh B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Incorporated.

MacArthur D.G., Balasubramanian S., Frankish A., Huang N., Morris J., Walter K., Jostins L., Habegger L., Pickrell J.K. & Montgomery S.B. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8.

Malher X., Beaudeau F. & Philipot J.M. (2006) Effects of sire and dam genotype for complex vertebral malformation (CVM) on risk of return-to-service in Holstein dairy cows and heifers. *Theriogenology* **65**, 1215-25.

Malosetti M., van der Linden C.G., Vosman B. & van Eeuwijk F.A. (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. *Genetics* **175**, 879-89.

Manolio T.A., Collins F.S., Cox N.J., Goldstein D.B., Hindorff L.A., Hunter D.J., McCarthy M.I., Ramos E.M., Cardon L.R. & Chakravarti A. (2009) Finding the missing heritability of complex diseases. *Nature* **461**, 747-53.

Marchini J. & Howie B. (2008) Comparing algorithms for genotype imputation. *Am J Hum Genet* **83**, 535-9; author reply 9-40.

Marchini J., Howie B., Myers S., McVean G. & Donnelly P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13.

Marklund S., Kijas J., Rodriguez-Martinez H., Ronnstrand L., Funa K., Moller M., Lange D., Edfors-Lilja I. & Andersson L. (1998) Molecular basis for the dominant white phenotype in the domestic pig. *Genome Res* **8**, 826-33.

Marti E. & Ohnesorge B. (2002) Genetic basis of respiratory disorders. *L. P, ed. (Ithaca. www. ivis. org)*.

Matukumalli L.K., Lawley C.T., Schnabel R.D., Taylor J.F., Allan M.F., Heaton M.P., O'Connell J., Moore S.S., Smith T.P. & Sonstegard T.S. (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PloS one* **4**, e5350.

McCarroll S.A., Kuruvilla F.G., Korn J.M., Cawley S., Nemesh J., Wysoker A., Shapero M.H., de Bakker P.I., Maller J.B. & Kirby A. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166-74.

McCullagh P. & Nelder J.A. (1989) Generalized linear models. Monographs on statistics and applied probability 37. *Chapman Hall, London*.

Meisinger C., Prokisch H., Gieger C., Soranzo N., Mehta D., Rosskopf D., Lichtner P., Klopp N., Stephens J. & Watkins N.A. (2009) A genome-wide association study identifies three loci associated with mean platelet volume. *Am J Hum Genet* **84**, 66-71.

Meuwissen T.H. & Goddard M.E. (2001) Prediction of identity by descent probabilities from marker-haplotypes. *Genet Sel Evol* **33**, 605-34.

Meuwissen T.H., Karlsen A., Lien S., Olsaker I. & Goddard M.E. (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161**, 373-9.

Momozawa Y., Mni M., Nakamura K., Coppieters W., Almer S., Amininejad L., Cleynen I., Colombel J.F., de Rijk P., Dewit O., Finkel Y., Gassull M.A., Goossens D., Laukens D., Lemann M., Libioulle C., O'Morain C., Reenaers C., Rutgeerts P., Tysk C., Zelenika D., Lathrop M., Del-Favero J., Hugot J.P., de Vos M., Franchimont D., Vermeire S., Louis E. & Georges M. (2011) Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease. *Nat Genet* **43**, 43-7.

Montgomery S.B. & Dermitzakis E.T. (2011) From expression QTLs to personalized transcriptomics. *Nat Rev Genet* **12**, 277-82.

Morgan D.K. & Whitelaw E. (2008) The case for transgenerational epigenetic inheritance in humans. *Mamm Genome* **19**, 394-7.

Nagahata H. (2004) Bovine leukocyte adhesion deficiency (BLAD): a review. *J Vet Med Sci* **66**, 1475-82.

Nalls M.A., Wilson J.G., Patterson N.J., Tandon A., Zmuda J.M., Huntsman S., Garcia M., Hu D., Li R., Beamer B.A., Patel K.V., Akylbekova E.L., Files J.C., Hardy C.L., Buxbaum S.G., Taylor H.A., Reich D., Harris T.B. & Ziv E. (2008) Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. *Am J Hum Genet* **82**, 81-7.

Nathan C. & Ding A. (2010) Nonresolving inflammation. *Cell* **140**, 871-82.

Neale B.M., Rivas M.A., Voight B.F., Altshuler D., Devlin B., Orho-Melander M., Kathiresan S., Purcell S.M., Roeder K. & Daly M.J. (2011) Testing for an unusual distribution of rare variants. *PLoS Genet* **7**, e1001322.

Nejentsev S., Walker N., Riches D., Egholm M. & Todd J.A. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387-9.

Nielsen R. (2010) Genomics: In search of rare human variants. *Nature* **467**, 1050-1.

Ohnesorge B., Deegen E., Miesner K. & Geldermann H. (1993) Hemiplegia laryngis in horses-an examination on stallions, mares and their offspring. *J Vet Med Series* **40**, 134-.

Okada Y., Hirota T., Kamatani Y., Takahashi A., Ohmiya H., Kumasaka N., Higasa K., Yamaguchi-Kabata Y., Hosono N., Nalls M.A., Chen M.H., van Rooij F.J., Smith A.V., Tanaka T., Couper D.J., Zakai N.A., Ferrucci L., Longo D.L., Hernandez D.G., Witteman J.C., Harris T.B., O'Donnell C.J., Ganesh S.K., Matsuda K., Tsunoda T., Kubo M., Nakamura Y., Tamari M., Yamamoto K. & Kamatani N. (2011) Identification of nine novel loci associated with white blood cell subtypes in a Japanese population. *PLoS Genet* **7**, e1002067.

Olshen A.B., Venkatraman E., Lucito R. & Wigler M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-72.

Olson T. (1999) *Genetics of colour variation*. CAB International.

Perkins J.D., Salz R.O., Schumacher J., Livesey L., Piercy R.J. & Barakzai S.Z. (2009) Variability of resting endoscopic grading for assessment of recurrent laryngeal neuropathy in horses. *Equine Vet J* **41**, 342-6.

Piercy R., Gath C., Powell E., Massey C., Stanley R., Barakzai S. & Perkins J. (2009) Examining the association between resting endoscopic grade of recurrent laryngeal neuropathy and both objective and subjective histopathologic assessment of the laryngeal intrinsic musculature. In: *Fourth world equine airways symposium, Berne, Switzerland*, p. 233.

Platt A., Vilhjálmsson B.J. & Nordborg M. (2010) Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**, 1045-52.

Poncet P.A., Montavon S., Gaillard C., Barrelet F., Straub R. & Gerber H. (1989) A preliminary report on the possible genetic basis of laryngeal hemiplegia. *Equine Vet J* **21**, 137-8.

Porter V. (2002) *Mason's world dictionary of livestock breeds, types, and varieties*. CABI.

Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A. & Reich D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9.

Price A.L., Zaitlen N.A., Reich D. & Patterson N. (2010) New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459-63.

Pritchard J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**, 124-37.

Pritchard J.K., Stephens M. & Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-59.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J. & Sham P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75.

Rassoulzadegan M., Grandjean V., Gounon P., Vincent S., Gillot I. & Cuzin F. (2006) RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature* **441**, 469-74.

Raychaudhuri S., Plenge R.M., Rossin E.J., Ng A.C., Purcell S.M., Sklar P., Scolnick E.M., Xavier R.J., Altshuler D. & Daly M.J. (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* **5**, e1000534.

Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., Shapero M.H., Carson A.R. & Chen W. (2006) Global variation in copy number in the human genome. *Nature* **444**, 444-54.

Reich D.E. & Lander E.S. (2001) On the allelic spectrum of human disease. *Trends Genet* **17**, 502-10.

Ren J., Guo Y.M., Ma J.W. & Huang L.S. (2006) Growth and meat quality QTL in pigs with special reference to a very large White Duroc × Erhualian resource population. *Proc. 8WCGALP. Brazil.*, Poster ID 13-1.

Reneau J.K. (1986) Effective use of dairy herd improvement somatic cell counts in mastitis control. *J Dairy Sci* **69**, 1708-20.

Ripke S., O'Dushlaine C., Chambert K., Moran J.L., Kähler A.K., Akterin S., Bergen S.E., Collins A.L., Crowley J.J. & Fromer M. (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**, 1150-9.

Rivas M.A., Beaudoin M., Gardet A., Stevens C., Sharma Y., Zhang C.K., Boucher G., Ripke S., Ellinghaus D. & Burtt N. (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* **43**, 1066-73.

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

Robinson N. (2004) Consensus statements on equine recurrent laryngeal neuropathy: conclusions of the Havemeyer Workshop. *Equine Vet Edu* **16**, 333-6.

Rossin E.J., Lage K., Raychaudhuri S., Xavier R.J., Tatar D., Benita Y., Cotsapas C. & Daly M.J. (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273.

Sabeti P., Schaffner S., Fry B., Lohmueller J., Varilly P., Shamovsky O., Palma A., Mikkelsen T., Altshuler D. & Lander E. (2006) Positive natural selection in the human lineage. *Science* **312**, 1614-20.

Sakurai M., Zhou J.H., Ohtaki M., Itoh T., Murakami Y. & Yasue H. (1996) Assignment of c-KIT gene to swine chromosome 8p12-p21 by fluorescence in situ hybridization. *Mamm Genome* **7**, 397.

Sandor C. & Georges M. (2008) On the detection of imprinted quantitative trait loci in line crosses: effect of linkage disequilibrium. *Genetics* **180**, 1167-75.

Sanyal A., Lajoie B.R., Jain G. & Dekker J. (2012) The long-range interaction landscape of gene promoters. *Nature* **489**, 109-13.

Sartelet A., Druet T., Michaux C., Fasquelle C., Géron S., Tamma N., Zhang Z., Coppieters W., Georges M. & Charlier C. (2012) A splice site variant in the bovine *RNF11* gene compromises growth and regulation of the inflammatory response. *PLoS Genet* **8**, e1002581.

Schaid D.J., Rowland C.M., Tines D.E., Jacobson R.M. & Poland G.A. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* **70**, 425.

Scheet P. & Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**, 629-44.

Scherf B.D. (2001) World Watch List for domestic animal diversity. *Anim Genet Resources Infor* **97**.

Schukken Y. & Kremer W.D.J. (1996) Monitoring udder health: Objectives, materials and methods. *Heard Health and production Management in Dairy Practice, Wageningen Pres*, 351-60.

Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Månér S., Massa H., Walker M. & Chi M. (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-8.

Segurado R., Detera-Wadleigh S.D., Levinson D.F., Lewis C.M., Gill M., Nurnberger Jr J.I., Craddock N., DePaulo J.R., Baron M. & Gershon E.S. (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: bipolar disorder. *Am J Hum Genet* **73**, 49-62.

Seltman H., Roeder K. & Devlin B. (2003) Evolutionary-based association analysis using haplotype data. *Genet Epidemiol* **25**, 48-58.

Selzer R.R., Richmond T.A., Pofahl N.J., Green R.D., Eis P.S., Nair P., Brothman A.R. & Stallings R.L. (2005) Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes, Chromosomes and Cancer* **44**, 305-19.

Shanks R.D. & Robinson J.L. (1989) Embryonic mortality attributed to inherited deficiency of uridine monophosphate synthase. *J Dairy Sci* **72**, 3035-9.

Shembade N., Harhaj N.S., Parvatiyar K., Copeland N.G., Jenkins N.A., Matesic L.E. & Harhaj E.W. (2008) The E3 ligase Itch negatively regulates inflammatory signaling pathways by controlling the function of the ubiquitin-editing enzyme A20. *Nat Immunol* **9**, 254-62.

Shembade N., Parvatiyar K., Harhaj N.S. & Harhaj E.W. (2009) The ubiquitin-editing enzyme A20 requires *RNF11* to downregulate *NF-kappaB* signalling. *EMBO J* **28**, 513-22.

Shuster D.E., Kehrli Jr M.E., Ackermann M.R. & Gilbert R.O. (1992) Identification and prevalence of a genetic defect that causes leukocyte adhesion deficiency in Holstein cattle. *Proc Natl Acad Sci U S A* **89**, 9225-9.

Sonstegard T.S., Cole J.B., VanRaden P.M., Van Tassell C.P., Null D.J., Schroeder S.G., Bickhart D. & McClure M.C. (2013) Identification of a nonsense mutation in *CWC15* associated with decreased reproductive efficiency in Jersey cattle. *PloS one* **8**, e54872.

Soranzo N., Rendon A., Gieger C., Jones C.I., Watkins N.A., Menzel S., Doring A., Stephens J., Prokisch H., Erber W., Potter S.C., Bray S.L., Burns P., Jolley J., Falchi M., Kuhnel B., Erdmann J., Schunkert H., Samani N.J., Illig T., Garner S.F., Rankin A., Meisinger C., Bradley J.R., Thein S.L., Goodall A.H., Spector T.D., Deloukas P. & Ouwehand W.H. (2009a) A novel variant on chromosome 7q22.3 associated with mean platelet volume, counts, and function. *Blood* **113**, 3831-7.

Soranzo N., Spector T.D., Mangino M., Kuhnel B., Rendon A., Teumer A., Willenborg C., Wright B., Chen L., Li M., Salo P., Voight B.F., Burns P., Laskowski R.A., Xue Y., Menzel S., Altshuler D., Bradley J.R., Bumpstead S., Burnett M.S., Devaney J., Doring A., Elosua R., Epstein S.E., Erber W., Falchi M., Garner S.F., Ghori M.J., Goodall A.H., Gwilliam R., Hakonarson H.H., Hall A.S., Hammond N., Hengstenberg C., Illig T., Konig I.R., Knouff C.W., McPherson R., Melander O., Mooser V., Nauck M., Nieminen M.S., O'Donnell C.J., Peltonen L., Potter S.C., Prokisch H., Rader D.J., Rice C.M., Roberts R., Salomaa V., Sambrook J., Schreiber S., Schunkert H., Schwartz S.M., Serbanovic-Canic J., Sinisalo J., Siscovick D.S., Stark K., Surakka I., Stephens J., Thompson J.R., Volker U., Volzke H., Watkins N.A., Wells G.A., Wichmann H.E., Van Heel D.A., Tyler-Smith C., Thein S.L., Kathiresan S., Perola M., Reilly M.P., Stewart A.F., Erdmann J., Samani N.J., Meisinger C., Greinacher A., Deloukas P., Ouwehand W.H. & Gieger C. (2009b) A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet* **41**, 1182-90.

Steinmetz L.M., Sinha H., Richards D.R., Spiegelman J.I., Oefner P.J., McCusker J.H. & Davis R.W. (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**, 326-30.

Storey J.D. & Tibshirani R. (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5.

Su S.Y., Balding D.J. & Coin L.J. (2008) Disease association tests by inferring ancestral haplotypes using a hidden markov model. *Bioinformatics* **24**, 972-8.

Sweeney C.R., Maxson A.D. & Soma L.R. (1991) Endoscopic findings in the upper respiratory tract of 678 Thoroughbred racehorses. *J Am Vet Med Assoc* **198**, 1037-8.

Taberlet P., Valentini A., Rezaei H., Naderi S., Pompanon F., Negrini R. & AJMONE-MARSAN P. (2008) Are cattle, sheep, and goats endangered species? *Mol Ecol* **17**, 275-84.

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437**, 1299–320.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61.

The International HapMap Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8.

Thomsen B., Horn P., Panitz F., Bendixen E., Petersen A.H., Holm L.E., Nielsen V.H., Agerholm J.S., Arnbjerg J. & Bendixen C. (2006) A missense mutation in the bovine SLC35A3 gene, encoding a UDP-N-acetylglucosamine transporter, causes complex vertebral malformation. *Genome Res* **16**, 97-105.

Threadgill D.W., Hunter K.W. & Williams R.W. (2002) Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm Genome* **13**, 175-8.

Thurman R.E., Rynes E., Humbert R., Vierstra J., Maurano M.T., Haugen E., Sheffield N.C., Stergachis A.B., Wang H. & Vernot B. (2012) The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82.

TULLIS J.L. (1952) Separation and purification of leukocytes and platelets. *Blood* **7**, 891-6.

Tyler J.W., Thurmond M.C. & Lasslo L. (1989) Relationship between test-day measures of somatic cell count and milk production in California dairy cows. *Can J Vet Res* **53**, 182-7.

Tzeng J.-Y. & Zhang D. (2007) Haplotype-based association analysis via variance-components score test. *Am J Hum Genet* **81**, 927-38.

Uda M., Galanello R., Sanna S., Lettre G., Sankaran V.G., Chen W., Usala G., Busonero F., Maschio A., Albai G., Piras M.G., Sestu N., Lai S., Dei M., Mulas A., Crisponi L., Naitza S., Asunis I., Deiana M., Nagaraja R., Perseu L., Satta S., Cipollina M.D., Sollaino C., Moi P., Hirschhorn J.N., Orkin S.H., Abecasis G.R., Schlessinger D. & Cao A. (2008) Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc Natl Acad Sci U S A* **105**, 1620-5.

Vandesompele J., De Preter K., Pattyn F., Poppe B., Van Roy N., De Paepe A. & Speleman F. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biol* **3**, RESEARCH0034.

VanRaden P., Olson K., Null D. & Hutchison J. (2011) Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *J Dairy Sci* **94**, 6153-61.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* **91**, 4414-23.

VanRaden P.M., Van Tassell C.P., Wiggans G.R., Sonstegard T.S., Schnabel R.D., Taylor J.F. & Schenkel F.S. (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci* **92**, 16-24.

Verbeke G. & Molenberghs G. (2003) The use of score tests for inference on variance components. *Biometrics* **59**, 254-62.

Villa-Angulo R., Matukumalli L.K., Gill C.A., Choi J., Van Tassell C.P. & Grefenstette J.J. (2009) High-resolution haplotype block structure in the cattle genome. *BMC Genet* **10**, 19.

Visscher P.M. (2008) Sizing up human height variation. *Nat Genet* **40**, 489-90.

Visscher P.M., Hill W.G. & Wray N.R. (2008) Heritability in the genomics era—concepts and misconceptions. *Nat Rev Genet* **9**, 255-66.

Visscher P.M., Medland S.E., Ferreira M.A., Morley K.I., Zhu G., Cornes B.K., Montgomery G.W. & Martin N.G. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* **2**, e41.

Wang J.Y., Luo Y.R., Fu W.X., Lu X., Zhou J.P., Ding X.D., Liu J.F. & Zhang Q. (2013) Genome-wide association studies for hematological traits in swine. *Anim Genet* **44**, 34-43.

Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S.F., Hakonarson H. & Bucan M. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-74.

Warner L.E., Mancias P., Butler I.J., McDonald C.M., Keppen L., Koob K.G. & Lupski J.R. (1998) Mutations in the early growth response 2 (EGR2) gene are associated with hereditary myelinopathies. *Nat Genet* **18**, 382-4.

Weigel K.A., de los Campos G., Gonzalez-Recio O., Naya H., Wu X.L., Long N., Rosa G.J. & Gianola D. (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J Dairy Sci* **92**, 5248-57.

Weigel K.A., Van Tassell C.P., O'Connell J.R., VanRaden P.M. & Wiggans G.R. (2010) Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. *J Dairy Sci* **93**, 2229-38.

Werner F., Durstewitz G.v., Habermann F., Thaller G., Krämer W., Kollers S., Buitkamp J., Georges M., Brem G. & Mosner J. (2004) Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. *Anim Genet* **35**, 44-9.

Workman C., Jensen L.J., Jarmer H., Berka R., Gautier L., Nielser H.B., Saxild H.-H., Nielsen C., Brunak S. & Knudsen S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome biol* **3**, 1-16.

Yalcin B., Willis-Owen S.A., Fullerton J., Meesaq A., Deacon R.M., Rawlins J.N.P., Copley R.R., Morris A.P., Flint J. & Mott R. (2004) Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice. *Nat Genet* **36**, 1197-202.

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E. & Visscher P.M. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9.

Yang S., Ren J., Yan X., Huang X., Zou Z., Zhang Z., Yang B. & Huang L. (2009) Quantitative trait loci for porcine white blood cells and platelet-related traits in a White Duroc x Erhualian F resource population. *Anim Genet* **40**, 273-8.

Yoshida H., Kunisada T., Grimm T., Nishimura E.K., Nishioka E. & Nishikawa S.I. (2001) Review: melanocyte migration and survival controlled by SCF/c-kit expression. *J Investig Dermatol Symp Proc* **6**, 1-5.

Yu J., Pressoir G., Briggs W.H., Vroh Bi I., Yamasaki M., Doebley J.F., McMullen M.D., Gaut B.S., Nielsen D.M., Holland J.B., Kresovich S. & Buckler E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**, 203-8.

Zhang Z. & Druet T. (2010) Marker imputation with low-density marker panels in Dutch Holstein cattle. *J Dairy Sci* **93**, 5487-94.

Zhang Z., Guillaume F., Sartelet A., Charlier C., Georges M., Farnir F. & Druet T. (2012) Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. *Bioinformatics* **28**, 2467-73.

Zhang Z., Hong Y., Gao J., Xiao S., Ma J., Zhang W., Ren J. & Huang L. (2013) Genome-wide association study reveals constant and specific loci for hematological traits at three time stages in a White Duroc x Erhualian F2 resource population. *PloS one* **8**, e63665.

Zhao K., Aranzana M.J., Kim S., Lister C., Shindo C., Tang C., Toomajian C., Zheng H., Dean C., Marjoram P. & Nordborg M. (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**, e4.

Zou Z., Ren J., Yan X., Huang X., Yang S., Zhang Z., Yang B., Li W. & Huang L. (2008) Quantitative trait loci for porcine baseline erythroid traits at three growth ages in a White Duroc x Erhualian F(2) resource population. *Mamm Genome* **19**, 640-6.

Zuk O., Hechter E., Sunyaev S.R. & Lander E.S. (2012) The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* **109**, 1193-8.

Presses de la Faculté de Médecine vétérinaire de l'Université de Liège

4000 Liège (Belgique)

D/2013/0480/19
ISBN 978-2-87543-040-3

ISBN 978-2-87543-040-3

9 782875 430403