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

Université de Liège

Etudes de caractères Mendéliens et complexes à l'aide d'outils de génotypage et de séquençage à haut-débit chez le bovin

Studying the genetics of Mendelian and complex traits with high-throughput genotyping and sequencing in cattle

Wanbo Li

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

ANNEE ACADEMIQUE 2016-2017

Cattle in Chinese

Acknowledgement

I am grateful to my supervisor ***Prof. Michel Georges*** *for designing the experiments, guiding me towards the best way to work in science, and teaching me the basic technique I needed as a student. I also thank him for always pushing me to ask more questions and answering every question I had. I have learned a lot from his wonderful ideas and suggestions.*

I am grateful to my co-supervisor ***Dr. Carole Charlier*** *for her generous help in developing my experimental techniques and polishing my skills of organizing and reporting results. She spent a wealth of time to helping me to understand experiment design and guided me to the right track. I would also thank her for taking care of the schedule of my doctoral study and helping with all the paper work. Her kindness made my life more easier and happier in Liege.*

I am grateful to ***Dr. Wouter Coppieters*** *for introducing me to programming and data analysis, and for help with organization of the samples and sequencing facility.*

I thank ***Cynthia Sandor*** *for making great contributions to the recombination paper. We had a nice collaboration in the project. And her non-stop work style really encouraged me. I also thank* ***Arnaud Sartelet*** *for happy collaboration on several projects with him, and his hard work in those projects.*

I am grateful to ***Latifa Karim*** *for her help with sequencing and all sorts of experiments and her kindness. I thank all my colleagues from the Unit of Animal Genomes and the GIGA Genomics platform for their help or contribution to my thesis work: Chad Harland, Naveen Kumar Kadri, Keith Durkin, Ming Fang, Tom Druet, Haruko Takeda, Zhiyan Zhang, Xuewen Xu, Huijun Cheng, Nico Tamma, Anne-Sophie Van Laere, Corinne Fasquelle, Sandrina Evrard, Nathalie Faust, Marilou Ramos-Pamplona, Benoit Hennuy, Naima Ahariz, Nadine Cambisano, Myriam Mni, Cécile Lam, Rodrigo Gularte, Emilie Théâtre, Ann-Stephan Gori, Mahmoud Elansary.*

I would like to thank all the people who contributed to the published/submitted papers included in this thesis.

I thank that the China Scholarship Council (CSC) and the Unit of Animal Genomics financially support me during the stay in Liege.

And last but not least, I thank the support from my wife and our both families!

List of Abbreviations

3C	chromosome conformation capture assay
4C	3C-on-chip
5C	carbon copy 3C
AI	artificial insemination
ASD	autism spectrum disorder
AWE	Walloon breeding association
BBC	Belgian blue cattle
BBCB	Belgian blue cattle breed
BS	brachyspina syndrome
BTA	*Bos taurus* chromosome
C	carriers
CAGE	cap-analysis of gene expression
CH	chiasmata
ChIA-PET	chromatin interaction analysis by paired-end tag sequencing
ChIP	chromatin immunoprecipitation
ChIP-seq	chromatin immunoprecipitation followed by sequencing
CI	confidence interval
cM	centimorgan
CMD1	congenital muscular dystonias I
CMD2	congenital muscular dystonias II
CO	cross-overs
CTLDs	c-type lectin-like domains
CTS	crooked tail syndrome
CVM	complex vertebral malformation
ddNTP	dideoxynucleotides
DHS	DNase I hypersensitive sites
DM	deleterious missense
DSV	DNA sequence variant

EL	embryonic lethal
ELV	embryonic lethal variant
ENCODE	encyclopedia of DNA elements
ER	endoplasmic reticulum
FISH	fluorescence in situ hybridization
FS	frame-shift
GEBV	genomic estimated breeding values
GHU	genome-wide hot-window usage
GIL	genome-wide interference levels
GJU	genome-wide jungle usage
GPI	glycosylphosphatidyl inositol
GPI-AP	GPI-anchored protein
GPI-GnT	GPI-GlcNAc transferase complex
GRR	genome-wide recombination rate
GS	genomic selection
GWAS	genome-wide association studies
h^2	heritability
H3K27ac	acylation of histone H3 lysine 27
H3K4me1	histone H3 lysine 4 monomethylation
H3K4me3	histone H3 lysine 4 trimethylation
HF	Holstein-Friesian
HGMD	the human gene mutation database
HHS	hidden haplotype states
HTS	high-throughput sequencing
IBD	identity-by-descent
IF	ichthyosis fetalis
IGV	integrative genomics viewer
iHS	integrated haplotype scores
IMPC	international mouse phenotype consortium
J	Jerseys

KO	knock-out
LD	linkage disequilibrium
LoF	loss-of-function
LRR	locus-specific recombination rate
LRT	likelihood ratio test
M	morgans
MAF	minor allele frequency
MAS	marker-assisted selection
ML	the most likely
MS	missense
MUT	mutant
NC	non-carriers
Ne	effective population size
NGS	next generation sequencing
NMD	nonsense-mediated decay
NMRD	nonsense-mediated RNA decay
NS	non-synonymous
NZ	New-Zealand
NZDC	New Zealand dairy cattle
ORF	open reading frame
P-GV	the proportion of genetic variance
P-PV	the proportion of the phenotypic variance
Prop-Sel	the proportion of sons selected
PRRS	porcine reproductive and respiration syndrome
PSE	pale, soft, and exudative
PSS	porcine stress syndrome
QTL	quantitative trait locus
QTN	quantitative trait nucleotide
REML	restricted maximum likelihood
RFLP	restriction fragment length polymorphism

RL	renal lipofuscinosis
RNA-seq	RNA sequencing
RNAPII	RNA polymerase II
RR	recombination rate
S	synonymous
SG	stop gain
SM	skeletal muscle
SNP	single nucleotide polymorphisms
SS	splice-site
TF	transcriptional factor
TM	transmembrane
TSS	transcription start site
WGS	whole genome sequencing
WT	wild-type
ZF	zinc-finger

Table of Contents

Résumé

Background

Depuis des dizaines d'années, identifier les gènes et mutations qui sous-tendent maladies congénitales et phénotypes économiquement importants a été au cœur des études génétiques chez les animaux domestiques. La première étape, étape de cartographie, à l'aide de marqueurs microsatellites ou de puces de génotypage à basse densité en SNP, était fastidieuse, elle était suivie par de laborieuses et souvent peu productives tentatives d'isoler les gènes et les mutations causales, aussi bien pour les caractères monogéniques que polygéniques. Le développement et la mise à disposition de puces de génotypage à haute densité ont considérablement accéléré ce processus. Plus récemment, les avancées technologiques en séquençage à haut-débit (HTS) dit de seconde génération, incluant le séquençage génome entier, exome et transcriptome, ont rendu la dissection moléculaire des phénotypes d'intérêt incroyablement performante. Au cours de cette thèse, nous décrivons l'utilisation de ces nouvelles technologies pour décrypter les bases génétiques de plusieurs caractères monogéniques et polygéniques, à la fois par des approches positionnelles (forward genetics) mais aussi par une démarche qualifiée d'inverse (reverse genetics).

Résultats

Nous décrivons d'abord l'utilisation de puces de génotypage moyenne-densité pour disséquer le déterminisme moléculaire d'une maladie monogénique (le syndrome de la queue tordue, CTS) mais aussi pour identifier les loci qui influencent un caractère complexe, à savoir la recombinaison méiotique chez les bovins. En ce qui concerne le CTS, nous montrons qu'une délétion de 2pb dans le gène *MRC2* est responsable de ce syndrome en race Blanc-Bleu belge (BBB) et que la fréquence de cette mutation dans la race s'explique par son effet pléiotropique positif sur le développement musculaire. Nous avons, dans ce même gène (*MRC2*), identifié une seconde mutation invalidant sa fonction, ce qui supporte l'hypothèse de

la pléiotropie associée à de la sélection balancée. Pour la recombinaison, nous rapportons ici que des variations génétiques dans les gènes *REC8* et *RNF212* affectent le taux de recombinaison global chez les bovins. Chez l'homme, le taux recombinaison global est aussi associé au gène *RNF212*, par contre des données plus récentes ne semblent pas supporter l'implication du gène *REC8*. Enfin, nous démontrons que plusieurs variants du domaine à 'doigts de zinc' d'un paralogue gonosomal du gène *PRDM9* influencent l'utilisation des 'hotspots' chez le bovin, en accord avec des résultats obtenus chez l'homme et la souris. En outre, la mise-à-profit de la technologie HTS en clonage positionnel nous a permis d'identifier (i) une mutation intronique dans le gène *PIGH* qui induit une non-rétention de l'exon 2 et cause un syndrome létal d'arthrogrypose en BBB, (ii) une délétion de 3,3 kb, enlevant les exons 25 à 27 du gène *FANCI* et occasionnant le syndrome brachyspina en race bovine Holstein-Frisonne.

Enfin, nous avons appliqué une approche de génétique réverse pour identifier des mutations, létales pour l'embryon à l'état homozygote (EL) et qui impactent la fertilité en populations viandeuse (BBB) et laitière (Holstein et Jersey). Nous avons criblé le génome ou l'exome de plus de 500 animaux et établi une liste de variations candidates correspondant à des mutations 'perte de fonction' (LoF : non-sense, décalage de phase de lecture, site d'épissage) ou changement d'acide aminé disruptif. Ce faisant, nous avons démontré que, de manière inattendue, les populations de bovins domestiques (*Bos taurus*) sont génétiquement plus variables que les populations humaines, en ce compris les Africains. Par contre, leur charge en variants délétères est similaire, de l'ordre de ~100 mutations LoF par individu. Ce résultat peut être interprété par l'existence d'une sélection purificatrice plus efficace, suite à l'augmentation récente de la consanguinité chez les bovins. Des dizaines de milliers d'animaux ont été génotypés pour ces candidats EL et, à partir des déplétions en homozygotes observées, nous avons pu estimer que ~15% de ceux-ci sont EL et touchent des gènes essentiels durant le développement. Aucune épistasie synthétique entre variants EL n'a pu être mise en évidence. En évaluant la déviation par rapport aux proportions Mendéliennes

attendues au sein d'accouplements entre parents porteurs, la nature embryonnaire létale de neuf mutations, toutes relativement communes, a pu être clairement prouvée. Toutes touchent des gènes engagés dans des processus de base du fonctionnement de la cellule. Seule une de ces neuf mutations aurait pu être détectée par les approches classiques, basées sur les haplotypes. Enfin, le catalogue de variants de type LoF est actuellement investigué plus avant pour tenter de mettre en évidence des phénotypes associés, non létaux, mais ayant des implications en recherche médicale par exemple. Ainsi, de manière plus anecdotique, une délétion de 10 pb, introduisant un codon stop prématuré dans le gène *MLPH* a été caractérisée, elle se traduit par une nouvelle couleur de robe en race BBB que nous avons baptisée 'cool-gray'.

Conclusions and Perspectives

Suite à l'utilisation à large échelle de certains taureaux d'élite, à cause ou grâce à l'utilisation de l'insémination artificielle, une flambée de pathologies à déterminisme génétique récessif simple a été observée au cours de ces dernières années. Nous avons démontré ici que disséquer les bases génétiques de ces maladies Mendéliennes peut être extrêmement rapide et efficace lorsque les développements technologiques de génotypage et de séquençage haut-débit peuvent être mis à profit. La fertilité, en races bovines, mais aussi dans d'autres espèces domestiques, a significativement décliné au cours de ces dernières décades. La mise en évidence d'un nombre non négligeable de variants EL, qui existent à des fréquences relativement élevées dans les populations bovines, apporte un complément d'explication à ce déclin observé de la fertilité, par rapport à l'hypothèse classique d'une balance énergétique négative chez les vaches à haute production. L'approche de génétique réverse développée ici a au moins deux avantages : (i) elle peut s'appliquer à des populations de taille relativement modeste et (ii) elle a une sensibilité très supérieure aux approches utilisant les haplotypes.

Il est communément admis que la plupart des variations qui sous-tendent les caractères complexes doivent être des régulatrices puisqu'elles ont été principalement localisées dans

des régions non-codantes du génome. Elucider les conséquences fonctionnelles de ce type de variants se révéle ardu. Les projets ENCODE chez l'homme et la souris apportent de bons exemples de comment des tests fonctionnels parallélisés, basés sur des technologies HTS, peuvent accroître notre compréhension du rôle des régions non-codantes du génome. On peut raisonnablement prédire que le développement et la disponibilité de données de type ENCODE, dans le domaine de la recherche animale, auront un impact majeur sur l'élucidation des bases moléculaires des caractères complexes.

Summary

Background

For decades, identifying genes and mutations underlying inherited diseases and economically import traits has been at the core of genetic studies in domestic animals. The initial arduous task of gene mapping with microsatellites or low-density SNP panels was then followed by a often fruitless and tedious attempt to isolate the causative mutations for both monogenic and polygenic traits. The availability of high-density SNP arrays considerably accelerated the gene mapping process. More recently, advances in second-generation high-throughput sequencing (HTS) including whole genome, exome, and RNA sequencing have brought tremendous efficiency into dissecting the genetic underpinnings of phenotypes of interest. In this thesis, we describe the use of these new technologies to decipher the genetic basis of several monogenic and polygenic traits, in both forward and reverse genetic settings.

Results

We first describe the use of medium-density SNP arrays to dissect genetic basis of a monogenic disease (Crooked Tail Syndrome, CTS), as well as of complex phenotypes related to meiotic recombination in cattle. With regards to CTS, we first demonstrate that a common 2bp-deletion in the *MRC2* gene causes the condition in Belgian Blue cattle (BBC), and that the frequency of the condition is due to a pleiotropic effect of the mutation on muscle mass. We discover a second loss-of-function mutation in the same gene *MRC2,* supporting the hypothesis of pleiotropy and balancing selection. With regards to recombination, we report that genetic variants in *REC8* and *RNF212* affect the genome-wide recombination rate in cattle. The *RNF212* gene has also been shown to be associated with genome-wide recombination rate in humans, however *REC8* is not supported as a candidate gene by newer data. We also demonstrated that several variants in the zinc-finger domain of a gonosomal *PRDM9* paralogue influence hotspot usage in cattle, reminiscent of previous findings in

human and mice. The subsequent application of HTS technology in forward genetic analyses, allowed us to identify: i) an intronic mutation in the *PIGH* gene that causes the skipping of exon 2, and is the cause of arthrogryposis in BBC, ii) a 3.3 kb deletion removing exons 25-27 of the *FANCI* gene, that causes brachyspina in Holstein-Friesian cattle.

We finally used a reverse genetic approach to identify embryonic lethal (EL) mutations compromising fertility in beef (BBC) and dairy cattle (Holstein-Friesian and Jersey). We mined whole genome and exome sequence data from more than 500 animals for candidate embryonic lethal variants corresponding to loss-of-function (stop gain, frameshift and splice site variants) and disruptive missense variants. By doing so we demonstrate that against expectations domestic *Bos taurus* cattle are genetically more variable than humans including Africans, however, that their burden of disruptive variants is similar, amounting to ~100 such variants per individual. We interpret this as resulting from more effective purifying selection in recent times as a result of increased inbreeding. We have genotyped thousands of animals for thousands of EL candidates, and estimate – from the observed depletion in homozygotes - that no more ~15% of those are developmentally essential. We do not find evidence for synthetic epistasis between candidate EL mutations. By evaluating the departure from Mendelian expectations in matings between carrier sires and dams, we unambiguously demonstrate the embryonic lethal nature of nine relatively common candidate EL variants. All of them affect genes that are involved in basic cellular processes. We demonstrate that only one of these would have been detected using the haplotype-based approach that were applied thus far. The established list of loss-of-function variants is being mined for putative non-lethal, yet major phenotypic effects of medical relevance. Among these, we uncovered a 10-bp deletion in the *MLPH* gene, which results a premature stop codon, that causes diluted color – 'cool gray' – in BBC.

Conclusions and Perspectives

Due to the intensive use of some elite sires by means of artificial insemination in cattle, several outbursts of genetic defects have been observed in recent years. We herein demonstrate that uncovering the genetic basis of such Mendelian disorders can be very effective when integrating high-throughput genotyping and sequencing technology.

Fertility in cattle and other domestic animals has declined over the last decades. Our discovery of a number of embryonic lethal variants segregating in cattle provide a complementary explanation for the decline in fertility, in addition to the theory of negative energy balance of high producing cows. The reverse genetic approach we developed for screening EL variants has at least two advantages. First, our approach also works in small sized populations. Second, our approach has higher sensitivity than the haplotype-based approach.

It has been suspected for some time that most of the variants underlying complex traits map to noncoding regions. Deciphering the functional consequence of these regulatory variants has proven challenging. The ENCODE projects in human and mice provide good examples of how high-throughput functional assays based on NGS technology may enhance our knowledge of the non-coding regions of the genome. It is reasonable to anticipate that the release of ENCODE-like data in the animal research field will assist in the discovery of the genetic basis of complex traits at an unprecedented pace.

Introduction

High-Throughput Sequencing Facilitates the Discovery of Causative Genes and Mutations Underlying Disorders or Economic Traits in Domestic Animals

After proposing the double helix structure of DNA in 1953, Watson and Crick detailed their view on how DNA replicates based on the proposed structure in a following paper. In this paper, they also suggested that the precise sequential order of nucleotides of DNA might be the genetic code itself. Therefore reading the genetic code of life is to find out the sequence of nucleotides of DNA (Watson and Crick, 1953). To determine the sequence of nucleotides of DNA was not an easy task at that time. It was not until the groundbreaking invention of the DNA sequencing technology based on dideoxynucleotides (ddNTP) as terminating inhibitors of DNA chain elongation by Fred Sanger in 1977 (Sanger sequencing), that our journey to crack the genetic code residing in every living organism really commenced (Sanger et al., 1977).

In the mid 1990s, the appearance of automated platforms capable of parallel Sanger capillary sequencing accelerated the completion of the very first human genome by 2003 (http://www.genome.gov), which enabled us to systematically examine the human genome and functional genes. Furthermore, the availability of the reference genome accelerated the speed of discovering polymorphism, such as SNPs. In domestic animals, reference genomes have been generated for chicken, dog, cat, sheep, cattle, horse and pig (reviewed by Nicholas and Hobbs, 2014) following the reference-generating trends pioneered in human genome project. For decades, mapping genes and mutations underlying diseases and economically import traits has been at the core of genetic studies in domestic animals. Linkage analysis was first adopted to map genes and the use of highly variable microsatellite markers achieved success in various traits in these early days. Later on, the availability of high-density SNP array combined with association studies make the gene mapping process relatively fast and

accurate. Combined with the availability of reference genomes, high-quality gene annotation and high-density SNP markers have simplified the fine-mapping and prioritization of candidate genes affecting phenotypic traits in domestic animals. From 2006, advance in second-generation high-throughput sequencing (HTS) technology facilitated cost-effective whole genome sequencing (WGS), exome sequencing (sequencing all annotated exons through capturing), and RNA sequencing (RNA-seq), greatly simplifying the dissection of the genetic underpinnings of phenotypes of interest. More recently, HTS-based functional assays (e.g., DNase-seq and ChIP-seq) opened new windows to look at the genome in a broader context, elucidating additional genomic features rather than myopically focusing on coding regions.

The majority of congenital defects in humans and domestic animals are caused by the disruption of a single gene/locus in the genome, termed monogenetic or Mendelian traits. However, common diseases in human and many economic traits in animals such as stature, growth, and milk production, are influenced by hundreds or even thousands of genes, and are therefore termed polygenetic or complex traits. Occasionally in a polygenetic model, there are several major genes accounting for large proportions of phenotypic variance, but it is almost impossible to fully explain a complex trait with a limited number of genes. There are several successful examples of finding causative genes and mutations prior to HTS era, but HTS has dramatically changed the landscape of such studies since its birth. In this introduction, I will summarize and discuss how the HTS technology changed the landscape of genetic research in the two broad sections – study of Mendelian and polygenetic traits/disorders.

HTS has brought new strategies into studies of Mendelian traits/disorders

HTS has significantly improved the efficiency of the gene-hunting process. Soon after the rediscovery of Mendel's genetic laws in 1900, Bateson (1902) reported five Mendelian traits

in chicken: Pea-comb, Rose-comb, polydactyly, shank color and white plumage. Since then, a variety of Mendelian traits had been documented in domestic animals. However, it took a long time to unequivocally identify causative mutations for these traits. In the 1980s, with the development of DNA sequencing, recombinant technologies and accumulated biochemical knowledge of proteins and enzymes, geneticists started to screen for mutations in postulated causal genes and test for correlation with a specific disorder in families or population. This genetic strategy is known as the "candidate gene approach". The development of two types of DNA markers - restriction fragment length polymorphisms (RFLP) and microsatellites - spurred the first wave of linkage mapping, combined with positional candidate gene cloning. Medium- and high-density microarrays of single nucleotide polymorphism (SNP) markers appeared in the early 2000s, making genome-wide association studies (GWAS) the preferred approach in gene mapping. More recently, the landscape of discovering causative genes and mutations has been dramatically altered by HTS. GWAS and high-throughput sequencing enable researchers to identify causative mutations for Mendelian traits within several months, or even weeks.

Candidate gene approach

The candidate gene approach largely depends on our knowledge of proteins and enzymes obtained through biochemical analysis in cells or tissues, usually in model organisms. Before the gene mapping approach became affordable for the identification of genes associated with a phenotype of interest and the causative variants, this was the most pragmatic method to study a trait of interest. A typical candidate gene approach comprises resequencing of promising genes (on DNA or cDNA) to discover variants, and testing association between variants and phenotypes. The first successful example of the application of this approach in domestic animals was the identification of a nonsense mutation in thyroglobulin transcripts, the cause of congenital goiter in cattle (Ricketts et al., 1987). Another well-known example is the discovery of the causative mutation underlying malignant hyperthermia in pigs, also

known as porcine stress syndrome (PSS) (Fujii et al., 1991). PSS leads to pale, soft and exudative (PSE) meat post mortem, causes large economic loss and was very common in the pig industry for decades before the discovery of its genetic underpinnings. Initial studies have indicated that the calcium-release channel - ryanodine receptor - could be the responsible gene for PSS (Fill et al., 1991; O'Brien, 1986). Screening for mutations in the ryanodine receptor gene showed that the Arg615Cys mutation underpinned PSS and allowed for the development of DNA test to eliminate the deleterious allele from pig populations worldwide (Fujii et al., 1991).

The candidate gene approach succeeded in the discovery of causal mutations for a few genetic disorders, and benefited the animal industry, especially in the PSS case. However just a handful of causative variants underlying economically important traits have been successfully identified via this approach during the past decades.

Positional cloning accelerated by the availability of HTS

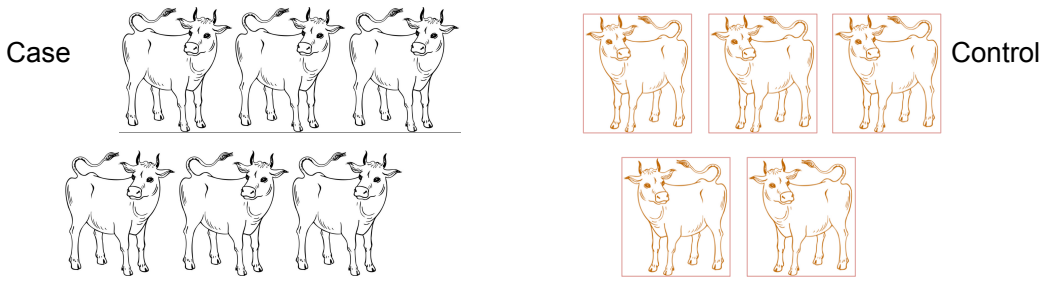
The concept of using linkage information within families to map genes influencing Mendelian or complex traits had been devised in the early 1920s (Sax, 1923; reviewed by Georges, 2012). In essence, linkage mapping examines whether specific genetic maker(s) cosegregate with a disease/phenotype in a pedigree comprising hundreds of individuals. Recombination will disrupt this cosegregation. However, markers evenly spaced every 10 ~ 20 centimorgan (cM) in the genome can capture the majority of linkage information in a study of hundreds of individuals. For instance, two markers 10 cM apart are likely to recombine in only 10 out of 100 offspring. In practice, the linkage mapping approach only started in 1980s, following the development of DNA makers, especially RFLP and microsatellites. A linkage analysis study and follow-up positional cloning approach usually comprise three steps: mapping loci underlying a certain traits/disorders, fine-mapping of the target loci and molecular analysis of plausible causative genes in the fine-mapped region. In the early days, it was very tedious to work through the initial mapping to finally discover the causative mutations, mainly due to

the difficulty of increasing marker density for fine-mapping and the lack of a reference genome. The locus underlying Huntington's disease in humans was first mapped on chromosome 4 in two large families using RFLP markers (Gusella et al., 1983). It then took ten years to identify the causative variant - an expanded trinucleotide repeat in the coding sequence of the *IT15* gene, despite the fact that extensive research efforts had been devoted to the disease (MacDonald et al., 1993). In some cases, when the candidate gene is obvious in the confidence interval, the process of finding the causative mutations can be faster. For instance, double-muscling, a profound phenotype of skeletal-muscle hyperplasia of Belgian Blue cattle (BBC), was mapped to chromosome 2 by linkage mapping in a backcross population (Charlier et al., 1995). After fine-mapping by increasing marker density and using comparative genomic analysis between cattle and human, the *MSTN* gene stood out as a promising candidate as knockout of this gene in mice has an analogous phenotype to that seen in cattle (McPherron et al., 1997). Finally, an 11-bp deletion in the coding region of *MSTN* was identified as causative mutation for the double muscling phenotype in BBC (Grobet et al., 1997). Since then, another 5 mutations within the *MSTN* gene have been found in 9 cattle breeds with the double-muscling phenotype, including 4 LoF and an intronic regulatory variant (Grobet et al., 1998; Bouyer et al., 2014). These results highlight the importance of the *MSTN* gene in muscle development and beef cattle breeding.

In the last decade, the development of medium- or high-density SNP chips in multiple species, including dog, cattle, and pig etc., accelerated the identification of causative mutations as it merged the mapping and fine-mapping into one step, leading to a two-step process – gene mapping and dissection of positional candidates. For mapping genes with medium- or high-density SNP markers, one can use 1) genome-wide association analysis to exploit linkage disequilibrium information, i.e., historical recombination events, usually in case-control studies; or 2) identity-by-descent (IBD) haplotype sharing approach (e.g., homozygosity or autozygosity mapping) in families or pedigrees. Of note, both approaches allow us to map the responsible locus in monogenic traits using tens of individuals or even less. The resulting

regions can cover hundreds Kb to several Mb and enclose tens to hundreds of genes, occasionally one or two genes if encountered a gene-poor region.

In its most basic form, association analysis in case-control studies construct a 2 × 2 contingency table (two alleles by two phenotypic groups) at each site, then perform Pearson's *chi*-squared test across all the markers of the genome, correcting for multiple testing at the end (Figure 1). As an example this efficient association mapping approach has been used to identify the genes underlying white spotting and hair ridge in different dog breeds. Karlsson et al. (2007) mapped the ridgeless allele to a 750-kb region in the Rhodesian ridgeback breed. Targeted sequencing of the association interval revealed a 133-kb duplication as the causative allele for ridgeless (Salmon Hillbertz et al., 2007). In the Karlsson et al. (2007) report, the gene responsible for a white color pattern in white boxers was mapped to a 1-Mb region on canine chromosome 20 containing only one gene – *MITF*. Subsequent sequencing of BAC clones containing the complete *MITF* gene revealed a SINE insertion and length polymorphism upstream of the M promoter of the *MITF* gene as the putatively causative variant.



	A	a	Total
Case	65	35	100
Control	20	80	100
Total	85	115	200

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(65 - 42.5)^2}{42.5} + \frac{(35 - 57.5)^2}{57.5}$$

$$+ \frac{(20 - 42.5)^2}{42.5} + \frac{(80 - 57.5)^2}{57.5}$$

Figure 1. An illustration for 2 x 2 contingence table

To use an IBD sharing approach in the recessive inheritance model, the algorithm implemented in ASSHOM (Charlier et al., 2008) can be used.

$$\sum_{i=1}^{k} -\log (p_i^2)^m$$

In brief, ASSHOM searches for homozygous segments of adjacent *k* markers with only one type homozygote (say 11 or 22) in each marker of *m* cases. p_i is the frequency of the allele in cases, estimated in n controls. Each marker within a segment will receive a value of $-\log(p_i^2)^m$, and a summary score for the segment will be the sum of the values for all the markers. Thus, the longer and rarer homozygous haplotype will receive higher score.

For the dominant inheritance, the ASSDOM algorithm can be applied (Durkin et al., 2012):

$$\sum_{i=1}^{k} -\log (1 - p_i^2)^m$$

ASSDOM scans for segments of adjacent *k* markers without both alternative homozygotes in each marker of *m* cases. P_i is the frequency of the allele missing among *m* cases, estimated in *n* controls. And each segment will receive a score calculated by above-mentioned formula. For both algorithms, the genome-wide statistical significance can be determined by phenotype permutations, which shuffle disease status among all the cases and controls.

This IBD-sharing approach has been adopted to map causative genes and variants for dozens of genetic defects in cattle. Charlier et al. (2008) reported mapping five genetic defects in cattle with 50K SNP array, including congenital muscular dystony type I and II (CMD1 & CMD2) and crooked tail syndrome (CTS) in BBC, renal lipofuscinosis (RL), and ichthyosis fetalis (IF) in other cattle breeds, with a map resolution of 2.12, 3.61, 2.42, 0.87 and 11.78 Mb, respectively. For CMD1, CMD2 and ichthyosis, there were obvious candidate genes in their corresponding critical regions. PCR amplification and sequencing of the coding sequence and exon-intron boundaries of several candidate genes immediately identified three missense

mutation in highly conserved motifs/regions of *ATP2A1*, *SLC6A5*, and *ABCA12* as causative mutations for the three defects, respectively. Further genotyping assays in large cohorts showed that these mutations segregate perfectly between cases and healthy animals, hence suggesting their causality for these diseases. Of note, linkage mapping with 400 microsatellites had failed to map any regions for CMD2, RL and IF disorders previously, clearly illustrating the increased power to detect causative genes of Mendelian traits by using medium or high-density marker panels.

If there is no obvious candidate gene in the critical regions, two straightforward approaches could be applied: 1) fine-mapping with more markers and larger cohorts (e.g., Fasquelle et al. 2009); 2) re-sequencing all annotated genes in the critical regions. The use of larger cohorts allows us to exploit more historical recombination events, existing in the population. It is noteworthy that exploiting linkage disequilibrium across breeds or isolated populations sharing the same phenotypes could increase power to narrow down critical intervals (Charlier et al., 2008; Karlsson et al., 2007; Sutter et al., 2007). This is because divergent breeds separated since a long time might have distinct LD pattern across the genome. For the second approach, it usually requires substantial effort before identifying the causative mutation. For instance, a major locus causing growth-stunting and affecting inflammatory response in the BBC population was mapped to a 3.3 Mb region on bovine chromosome (BTA) 3 encompassing 19 annotated genes without any obvious candidate (Sartelet et al., 2012a). An acceptor splice site mutation in intron 1 of the *RNF11* gene was then identified as the causative mutation after resequencing 14 out of 19 genes.

Exhaustive resequencing of every gene with PCR and Sanger sequencing in a critical region is a tedious process, which could fail if the causative mutations are regulatory (intergenic or intronic). As the cost of HTS decreases, the unbiased method of screening variants in the mapped regions is to sequence the whole region by targeted sequencing, or even whole-genome sequencing of affected and matched control animals. Subsequent tasks will be 1)

discovering DNA sequence variants (DSV, i.e., transitions, transversions) and structural variants (insertions, deletions, inversions, translocations) in the sequenced region; and 2) filtering variants based on whether their genotypes fit the disease status or phenotypic category; and 3) ranking variants according to their perceived relevance regarding the phenotype of interest or disease conditions. Successful examples have been seen in recent years in the application of a "new version" of the two-step strategy – mapping and then sequencing the targeted regions with HTS. This strategy has led to discovery of causative mutations in Mendelian traits at a greatly accelerated pace. In cattle, a missense mutation in the bovine *CLCN7* gene was identified as the cause of osteopetrosis with gingival hamartomas by combining gene mapping and HTS (Sartelet et al., 2014). Moreover, it was shown that a 3.3-kb deletion within bovine *FANCI* gene leads to brachyspina and embryonic death (Charlier et al., 2012) and that serial translocations of chromosome fragments containing the *KIT* gene causes color sidedness (Durkin et al., 2012). In Hampshire pigs, the locus underlying white belt coat color has been mapped to a region near the *KIT* gene in 1999, and it is assumed to be a regulatory element influencing the expression of *KIT* (Giuffra et al., 1999). The corresponding causative mutation - a duplication overlapping regulatory elements of *KIT* - has been identified after 13 years by whole-genome sequencing (Rubin et al., 2012). In chicken, the identification of the causative mutation for the Rose-comb is another interesting example. The Rose-comb is an autosomal dominant and among the earliest recognized Mendelian traits in animals (Bateson 1902). An initial QTL mapping study has only recently mapped the Rose-comb locus to a 7-Mb interval, on chicken chromosome 7, that contained no obvious candidate genes (Dorshorst et al., 2010). A subsequent study showed absence of recombination events within the 7-Mb region, suggesting an inversion. Then a mate-pair library prepared from a pool of Rose-comb homozygous individuals was sequenced and a complex chromosome inversion pattern was characterized. Hence, the genetic basis underlying Rose-comb had been discovered 110 years after it had been described as a Mendelian phenotype (Imsland et al., 2012).

As mentioned above, to examine large DNA structural variants, HTS strategy should be the method of choice. However, there were some examples of detecting duplications/CNV by quantitative Southern blots or quantitative PCR, f.i., the duplication of *KIT* causing dominant white in pigs and CNV within the *SOX5* gene causing the Pea-comb in chickens (Moller et al., 1996; Wright et al., 2009). It should be noted that structural variations could also be deduced from the genotyping data of SNP array. For instance, a suspicious long fragment of homozygosity and inflation of Mendelian inheritance incompatibilities may indicate a large deletion (Kadri et al., 2014) and a large chromosome segment without recombination events may imply an inversion (Imsland et al. 2012). Bioinformatics programs such as PennCNV can infer potential CNVs from raw signal intensity and B allele frequency generated by genotyping platforms (Wang et al., 2007). The preliminary inference of structural variations using SNP array data can provide clues for adopting suitable sequencing technique in the next-round experiment, e.g., constructing pair-end or mate-pair libraries to be used in HTS.

Identifying causative genes in a single step by HTS

As the cost of high-throughput sequences has continuously declined over the past few years, there is almost no need to first map associated regions with genome-wide SNP markers. Instead, one can directly sequence exomes or whole genomes of few individuals in case-control studies or small families, allowing researchers to identify causative mutations in a single step. The general pipeline of identifying the genuine causative mutation amongst the enormous numbers of variants generated by HTS can be addressed in four steps: 1) selecting variants segregated perfectly between cases and controls; 2) filtering variants using publicly available information or databases (e.g. dbSNP); 3) prioritizing variants by annotation with sequence conservation and protein structure information, as implemented in SIFT and Polyphen-2 programs etc.; 4) ideally, only few variants will be left to be validated in a large cohorts and/or to be functionally analyzed in cell lines or laboratory animals (reviewed in details by Koboldt et al., 2013). Although, caution should be taken when filtering against

public database, as the fast growing number of submitted variants may contain disease-related polymorphisms. In humans it is possible to identify causative genes and mutations for monogenic disorders by sequencing the exomes of a few unrelated affected individuals or family members started from 2010 (Bolze et al., 2010; Gilissen et al., 2010; Hoischen et al., 2010; Johnson et al., 2010; Lalonde et al., 2010; Wang et al., 2010). In some cases, sequencing even only one affected animal can lead to finding the putative causative mutation. For example, we recently received a sample form an animal suffering from ichthyosis, for genetic testing. As mentioned before, a missense mutation (H1935R) in exon 39 of the *ABCA12* gene was previously identified as the causal mutation for the disease (Charlier et al., 2008). Our routine genotyping assay revealed that the affected calf did not carry the previously known causative mutation. We then sequenced the exome of this animal and identified a homozygous stop-gain mutation (Q244*) in the *ABCA12* gene, located in a long homozygous region, as the most likely underlying cause of this supposedly recessive disease (unpublished data).

Exome sequencing succeeded in the detection of variants in exons, splice-sites and intronic bases near exon boundary, however it can generally not identify causative mutation located in the majority of noncoding regions. Whole-genome sequencing potentially offers a systematic method to screen all mutations in individuals. As variant discovery becomes nearly trivial, the difficulty becomes how to identify among the candidate variants a plausible functional mechanism that explains how theses variants contribute to the phenotype of interest. Coding variants such as nonsense and frame-shift mutation create premature stop codons, while splice-site disrupting mutations could either lead to skipped exons, or create premature stop codons by using cryptic splice site nearby. Premature stop codons outside the last exons would trigger nonsense-mediated RNA decay (NMRD) (Chang et al., 2007), while missense mutations might disrupt protein functions and/or three-dimensional structure, especially those in highly conserved domains. For synonymous and regulatory variants, it is not easy to demonstrate the functional impact of these variants on gene function. Fortunately, a variety

of functional assays developed in the context of high-throughput sequencing have allowed a systematic approach to validate the functional consequence of regulatory variants. Details outlining these assays are described in the following section dealing with polygenic traits, as the majority of variants underpinning GWAS association peaks in polygenic traits are regulatory.

In conclusion, with the advent of high throughput genotyping and HTS, the number of Mendelian traits in domestic animals with known causative mutations had reached 499 by the end of 2012 (reviewed by Nicholas & Hobbs, 2014). It was estimated that the causative mutation discovery rate is around one per week in domestic animals. The discovery rate still roughly holds true, as 149 new causative mutations had been added to the database of Online Mendelian Inheritance in Animals (http://omia.angis.org.au) during the 40 months (~160 weeks) since the end of 2012 (Table 1).

Table 1. Summary of monogenic traits with known key mutations in main domestic animals

	Total traits/disorders	Mendelian traits/disorders	Mendelian traits/disorders - key mutation known
Dog	666	273	199
Cattle	479	215	117
Cat	324	87	55
Pig	240	60	34
Sheep	237	100	47
Horse	225	48	36
Chicken	210	128	41
Goat	76	15	9
Total	**2457**	**926**	**538**

Data was collected from online database of Mendelian Inheritance in Animals by April, 2016.

A new approach for hunting fertility-related genes and variants

Owing to intensive selection for production traits and the extensive application of artificial insemination in farm animals over the last decades, farm animals have experienced an

increased level of inbreeding and consequently shrinkage in effective population size. The constant emergence of recessive defects and a decline of fertility have been observed in farm animals in recent years. This phenomenon may be partially the result of higher chance of segregating deleterious/lethal alleles, uniting into homozygotes in highly inbreeding populations. The traditional gene mapping approach is not sufficiently powerful to identify fertility-related genes, mainly due to difficulties in phenotype collection and the low heritability of reproduction-related traits.

HTS provides a complementary approach to detect variants affecting fertility related traits. We recently devised a reverse genetic approach to systematically screen for variants related to fertility by using high-throughput sequencing in cattle. This idea grew out of insights gained during our study of the brachyspina syndrome (BS). BS is a rare congenital defect ($< 1/10^5$ birth) in Holstein-Friesian cattle. Affected calves appeared with a significant shortening of the spine, long and thin limbs, growth retardation, usually were delivered as stillbirths (Charlier et al., 2012). By applying IBD mapping and HTS, we found a 3.3-kb deletion removing three coding exons of *FANCI* gene as the causative mutation for the disease. Further examination revealed a ~7.4% carrier frequency of the deletion in Holstein-Friesian dairy cattle. Therefore, many more newborn animals were expected to be affected by BS than those observed (incidence rate at $0.074 * 0.074 / 4 \approx 0.0014$). Using field fertility data, we estimated that at least half of the homozygous mutant embryos/fetuses die during pregnancy. We postulated that other embryonically lethal variants (ELV) might segregate within cattle populations. We thus used a reverse genetic approach to systematically screen the genome for loss-of-function (LoF) and deleterious missense (DM) variants by using HTS. LoF and DM at high allelic frequencies but never encountered at homozygote state in a supposedly healthy population are strong candidates for recessive embryonic lethality. In detail, the approach comprises the following steps: 1) sequencing tens or hundreds of highly used elite sires using exome or whole-genome sequencing; 2) filtering variants based on quality and distribution across breeds; 3) selecting LoF (e.g., nonsense, splice sites and frame-shift) and deleterious missense

variants predicted by SIFT and Polyphen-2 software based on sequence conservation and protein structure information; 4) proving embryonic lethality of the variants by genotyping the candidate variants in large cohort. When a candidate stands out as probable ELV, we then carry out a prospective study by following carrier × carrier mating to test if there is a lack of homozygotes in the offspring. If ELVs could be further intersected with RNAseq datasets (f.i., derived from embryonic or fetus tissues), this would allow us to exclude genes not expressed in embryo and fetus. In addition, a list of embryonic lethal genes in other species such as mice could be used as a reference set.

It is worth mentioning that previous studies have examined recessive fertility traits by using a haplotype-based approach. For example, five haplotypes at high carrier frequencies but with an absence of homozygotes were detected in a large number of cattle populations genotyped with medium-density SNP chips (VanRaden et al., 2011). One concern about this haplotype-based approach is that it needs very large samples. Assuming an embryonically lethal haplotype of 5% carrier frequency in a population, we would expect one homozygote of the haplotype in 1600 ($N = 1/0.05^2 * 4$) animals. Haplotypes with less than 2% of carrier frequency would require >10,000 genotyped animals to expect one homozygote. As a consequence this approach is only practical in a few populations or species, due to large number of genotyped samples needed. Furthermore, the high probability of encountering identity-by-state haplotypes, which could differ at the lethal mutation sites in an inbred population would reduce the sensitivity of this haplotype-based approach.

HTS as the method of choice for traits with high genetic heterogeneity

Genetic heterogeneity is not rare in domestic animals, even in populations with very small effective population size. For example, allelic heterogeneity has been observed in the study of crooked tail syndrome in BBC, despite the fact that population effective size of BBC is ~50. Two causative mutations (c.2904-2905delAG and c.1906T>C) have been found in the *MRC2* gene, causing CTS in the BBC population (Fasquelle et al., 2009; Sartelet et al.,

2012b). Moreover, allelic heterogeneity was also observed in two cattle breeds with double-muscling, with two LoF variants in *MSTN* segregating in each breed (Grobet et al., 1997; Grobet et al., 1998). Single marker association and IBD sharing mapping have low power to detect association signals if allelic heterogeneity exists in a same population.

More importantly, high genetic heterogeneity (f.i. hundreds of loci) in a relatively small sample size will pose great challenge to current genetic mapping methods. Recently HTS has been used to dissect genetic determinisms of disorders with high genetic heterogeneity. Autism spectrum disorders (ASD) in human is one typical example of high locus heterogeneity. Inspired by the discovery of *de novo* CNVs contributing to the risk of autism, researchers further examined the roles of *de novo* SNVs in this disorder using exome sequencing. Independent studies sequenced whole exomes of about 200 trio/quartet simplex families (unaffected parents, one affected proband, and with/without an unaffected sibling), and identified hundreds of *de novo* nonsynonymous mutations in probands with a few recurrently appearing in the same genes (e.g., *SCN2A*, *KATNAL2* and *CHD8*) across families (Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). Assuming there are hundreds of ASD-predisposing genes, the *de novo* mutations contributing to autism in every individual may be due to radically different causative mutations amongst the hundreds of genes, that is, a high degree of locus heterogeneity. The challenge is to verify the causality of the *de novo* mutations statistically. To overcome the obstacle, the authors used simulations to demonstrate that two or more highly deleterious LoF mutations (nonsense, splice-site disruptive and frame-shift) are unlikely to occur in an autism-predisposing gene in cases but not in the controls by chance. It was estimated that 30~60 autism related genes hit by two highly deleterious mutations could be found in 3000 families. So increasing sample size would enhance the detection power of loci with high locus heterogeneity. One worth-mentioning follow-up approach (O'Roak et al., 2012) is resequencing (i.e. targeted sequencing) of candidate autism-predisposing genes in much larger case-control cohorts after

the first round of discovery. In this second round of searching, an excess of deleterious mutations should be found in the initial gene sets in cases rather than controls.

HTS accelerates identification of causative mutations in polygenetic traits

Brief summary of quantitative trait locus (QTL) mapping and GWAS

Polygenetic traits or diseases are complex in their genetic architecture. Research strategies similar to those for monogenic traits have been used to identify genes and mutations: candidate gene approach, QTL mapping with limited genetic markers and GWAS with high-density SNPs. At early stages, QTL mapping was mainly conducted on intercross (F2), backcross populations, or half-sib families. The simplest QTL mapping analysis is quite similar to the linkage mapping for monogenic traits, which sequentially compares trait value difference between groups carrying different alleles at each marker (Soller et al., 1976). Lander and Botstein (1989) developed a mapping strategy based on intervals defined by nearby markers using maximum likelihood method. Haley and Knott (1992) developed a least square method for interval mapping which is much faster than maximum likelihood. Because quantitative traits are controlled by numerous loci, it is suitable to take all of them into account simultaneously when performing QTL mapping. Zeng (1994) proposed a composite interval mapping method, which takes makers outside of the testing interval as a cofactor to absorb multiple QTL effects when performing a genome scan. Noteworthy, Zeng's method is suitable for QTL mapping in line-cross between inbred lines. Kao et al. (1999) suggested a multiple QTL mapping approach that simultaneously fits all QTLs as well as their interacting effect in models. These methods are all based on maximum likelihood estimation. There are also several methods that use Bayesian principles, typically Bayesian

model selection implemented via MCMC, for multiple QTL mapping (Heath, 1997; Sillanpää and Arjas, 1998).

During the last three decades, prodigious efforts have been deployed to map QTLs underlying a variety of traits in domestic animals. By May 2014, 10497, 9180, 4282 and 789 QTLs for hundreds of different traits had been curated into the QTL database of pigs, cattle, chicken and sheep, respectively (http://www.animalgenome.org/cgi-bin/QTLdb/). Despite the great success in QTL mapping efforts, only a handful of quantitative trait nucleotides (QTN) have been convincingly identified. The procedure from QTL to QTN involved the following steps: fine mapping, molecular analysis and functional validation. A typical fine mapping of QTL includes: 1) increasing marker density, 2) using haplotype sharing approaches to narrow down the confidence interval (CI), and 3) exploiting haplotype diversity of diverse breeds if possible. The CI of initial QTL mapping often spans several or more megabases. Increasing marker density in CI was a laborious task before the availability of high-density SNP chips and high-throughput sequencing. Therefore, most previous studies employed fine mapping to narrow down QTL into a region of several hundreds of kilobases and then selected the most functionally relevant candidate genes for further molecular analysis. For example, this approach has been successfully used to identify a protein-altering mutations in *DGAT1* influencing milk fat composition in cattle (Grisart et al., 2002, 2004), a regulatory mutation in *IGF2* affecting muscle growth, fat deposition and heart size in pigs (Van Laere et al., 2003), a missense mutation in *NR6A1* influencing porcine vertebral number (Mikawa et al., 2007), a mutation creating a novel microRNA target site in the myostatin gene influencing meatiness in Texel sheep (Clop et al., 2006), and a coding variant in *PPARD* affecting outer ear size in pigs (Ren et al., 2011). A study on cattle stature is particularly noted in terms of this fine-mapping strategy (Karim et al., 2011). After initial QTL mapping, Karim et al. used high-density SNP genotyping and high-throughput targeting sequencing to highlight 13 putative QTNs from 9,572 variants identified in the critical region. Exploring the haplotype information of 12 diverse cattle breeds enabled the researchers to exclude five out of the 13

variants. A follow-up reporter and gel shift assay and eQTL analysis suggested two variants located in the bidirectional promoter region between the *PLAG1* and *CHCHD7* gene as the most likely causative mutations and *PLAG1* as the most plausible causal gene.

In practice, QTL mapping studies were usually performed in experimental line-crossed population or half-sib families, mainly relying on linkage information. Its power is limited by the number of recent recombination events and sparse markers. The availability of large panels of SNPs could allow researchers to exploit LD information in a population with dense markers, making gene mapping process more faster. In GWAS, one can test association between genotypes and phenotypes with a simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

If a marker with alleles 1 and 2, let X_i = 0, 1, 2 if individual *i* is homozygous (1,1), heterozygous (1,2), and homozygous (2,2), respectively; Y is phenotypic values; Null hypothesis: $\beta_1 = 0$, alternative hypothesis is $\beta_1 \neq 0$. This simple method is only suitable for quantitative traits. For binary traits, the logistic regression must be used.

Taking advantage of increased numbers of sequenced animals, the imputation strategy has been applied to increase marker density in a cost-efficient manner. Imputation uses linkage disequilibrium information between pairs of genetic markers in a 'reference panel of genomes' to compute the genotype probability of markers not genotyped or markers that failed during genotyping procedure. Those reference genomes can be a panel of individuals genotyped with very high-density SNP array, or sequenced with whole-genome sequencing such as the 1000 genomes projects in human and cattle (Daetwyler et al., 2014; The 1000 Genomes Project Consortium, 2012).

Until recently, GWAS has been widely used to map loci affecting polygenic traits such as production, growth, milk yield, and fertility in domestic animals. GWAS in the human population has discovered thousands of genetic loci associated with normal and pathological

traits. Uncovering the functional SNPs from a large set of associated SNPs can be difficult. Typically in fine mapping, a sample size of 1~4 times larger than that in the initial association analysis is needed (Udler et al., 2010). After the fine-mapping, there might still dozens of putative causative variants left in a high LD block. The big challenge then is to find the real functional variant(s) and illustrate their possible functional mechanisms underlying the association peaks. To illustrate this process, our study on a complex trait – recombination – using medium- and high-density SNP array genotypes was introduced in the following section.

Understanding the genetic mechanism controlling recombination

In mammalian sexual reproduction, chromosomes replicate once, then undergo two reduction divisions (meiosis I and II), leading each gamete to obtain a half set of chromosomes of the organism. To achieve this, accurate alignment, pairing and segregation of homologous chromosomes in meiosis are required and highly regulated (Handel and Schimenti, 2010). Recombination (crossover) establishes physical connection and directs precise segregation of homologous chromosomes in meiosis I. Correct segregation of the full chromosome complement demands tight, sex-specific control of the number of crossovers per arm, as well as of their position relative to chromosomal landmarks (centromeres and telomeres) and other CO (in the case of multichiasmatic meioses) (Martinez-Perez and Colaiacovo, 2009). Not surprisingly therefore, genomic mutations that grossly perturb meiotic recombination block gametogenesis, causing complete sterility of affected individuals. In addition, recombination is the force to break haplotypes and hence greatly contributes to evolution. It is hypothesized that by breaking linked favorable and deleterious alleles, recombination can enhance purifying selection ability to efficiently eliminate undesired alleles. Although recent evidence in yeast does not appear to supported this conjecture (Pal et al., 2001).

Recombination occurs at least once in a pair of homologous chromosomes (obligate crossover), and when two or more crossovers occur in paired chromosomes they are widely separated due to the phenomenon of crossover interference. The recombination process is

tightly controlled in meiosis. However, studies have displayed genome-wide recombination rate were highly variable among individuals in several species (Broman and Weber, 2000; Kong et al., 2002; Ma et al., 2015). Studies of Icelandic and other population uncovered some genetic controlling mechanism of recombination in which variants in *RNF212* gene affect genome-wide recombination rate (Chowdhury et al., 2009; Kong et al., 2008). Furthermore, evidence shows that subtle changes in recombination rate might also affect fertility. In Icelandic population, Stefansson et al. showed that a common 17q21.31 inversion affects both recombination and female fertility in the same direction (Stefansson et al., 2005); Kong et al. (2008) demonstrated that recombination rate is positively correlated with fertility in Icelandic women (0.01 child per Morgan).

Recombination is traditionally assumed to occur randomly along chromosomes. However, recombination events at the population scale are not evenly distributed along genome, but usually concentrate in narrow regions and constitute numerous "hotspots". Early sperm typing experiments found dozens of hotspots in human genome (Jeffreys et al., 2001; Jeffreys et al., 1998). Coalescence-based historic recombination analysis on high-density SNP markers revealed >30,000 recombination hotspots in humans (McVean et al., 2004; Myers et al., 2005). More interestingly, a major portion of those hotspots harbors a 13-mer motif CCNCCNTNNCCNC, which strongly drives the nearby hotspot usage (Hinch et al., 2011; Myers et al., 2008). In silico and in vitro analysis has revealed that a purine-rich (PR) domain containing protein PRDM9 contacts the recombination hotspot motif in human and mouse (Baudat et al., 2010; Myers et al., 2010). These findings together made research on recombination a rich and interesting topic.

The recombination mechanism has been well defined in yeast and nematode (Handel and Schimenti, 2010; Martinez-Perez and Colaiacovo, 2009), which share most of fundamental principles with mammals in recombination, but also differ significantly for some characteristics (Lichten and de Massy, 2011). Dissecting the unique mechanism in domestic

animals is essential towards better understanding of possible clinic reproductive problems linked to recombination and how artificial selection has shaped the recombination landscape in farm animals. Farm animals, like cattle, usually have extremely large family size and well-recorded pedigree, which are advantageous for dissecting genetic basis of complex traits. In one of the studies included in this thesis, we seek to characterize the recombination landscape genome-widely in cattle, taking advantage of a large three-generation half-sib dairy cattle pedigrees genotyped with a 50K high density SNP chips for the purpose of genetic selection. Our strategy to identify recombination components is to positionally clone the genes and variants that underlie inherited variation in recombination phenotypes.

Elucidating the mechanisms of regulatory variants by HTS-based functional assays

In humans, about 88% of associated SNPs reside in non-coding or non-transcribed regions, indicating a prominent role for regulatory variants in complex traits and common diseases. Currently, uncovering the functional mechanism caused by regulatory variants is still a big challenge. The most significantly associated SNPs, i.e., lead SNPs, are usually not causative mutations. Typically, a lead SNP is located in a LD block of ~150 kb containing the causal variant(s) driving the biological function, i.e., functional SNP(s) (Georges, 2011). Uncovering the functional SNPs from a large set of associated SNPs can be difficult. As mentioned, there might be still dozens of putative causative variants left in a high LD block after fine-mapping. To illustrate how new technology can aid the finding of real functional variants, in this section, I will discuss how high-throughput functional assays helps to dissect functional mechanisms behind the GWAS signals.

As mentioned before, the major portion of the top GWAS SNPs are likely regulatory variants, located in non-coding functional elements, such as enhancers, promoters and insulators. Proving their functional causality is still a challenge, largely due to the lack of knowledge of regulatory elements in the genome. Recent development of high-throughput functional assays based on HTS enable us to systematically detect regulatory elements in the genome. For

instance, DNase I hypersensitive sites (DHSs), an indication of chromatin accessible regions, can be identified by DNase I-seq (Crawford et al., 2006). Chromatin immunoprecipitation (ChIP) followed by HTS (ChIP-seq) can systematically identify transcription factors (TF) and other key protein binding sites, and regulatory elements related to chromatin modification sites (details in Table 2). Chromatin fragments harboring histone H3 lysine 4 trimethylation (H3K4me3) are indicative of promoters, whereas those marked by H3 lysine 4 monomethylation (H3K4me1) associate with enhancers (Heintzman et al., 2007). However, it was suggested that acetylation of H3K27 (H3K27ac) enriched in active enhancers and H3K4me1 can not distinguish poised or active enhancers (Creyghton et al., 2010; Andersson et al., 2014). The sequencing technology illustrating chromatin 3D conformation allows to identify chromatin interactions via patterns of long-range looping, including interactions between transcription start sites (TSS) and promoters, CTCF-bound sites, enhancers (Sanyal et al., 2012), enhancer-promoter and promoter-promoter interactions (Li et al., 2012). This technology has played an important role in identifying functional variants (see below).

Table 2. Summary of histone modifications and transcriptional factors associated with regulatory elements

Features / TF	Significance	Experimental approach
Open chromatin	sequences harboring regulatory signals	DNase-seq, ATAC-seq
H3K4me1	promoters and enhancers	ChIP-seq
H3K4me2	promoters and enhancers	ChIP-seq
H3K4me3	promoters	ChIP-seq
H3K9me1	active chromatin	ChIP-seq
H3K9me3	repressed chromatin	ChIP-seq
H3K27ac	active enhancers	ChIP-seq
P300	enhancers	ChIP-seq
CTCF	insulator	ChIP-seq
RNA Pol II	active promoters	ChIP-seq

Reviewed in Edwards et al. (2013).

The spatial organization of chromatin is a critical regulatory mechanism of gene transcription (Cook, 1999). Fluorescence in situ hybridization (FISH) was first exploited to illustrate the three-dimensional (3D) organization of chromatin. However, FISH can only examine a few loci simultaneously, and its resolution is limited. Dekker et al. (2002) invented a new method, Chromosome Conformation Capture assay (3C), to detect the spatial conformation of chromatin. Briefly, 3C includes following steps: a) cross-linking spatially contacting regions of DNA and DNA-protein interactions using formaldehyde in cultured cells, b) digestion of cross-linked DNA extracts with a restriction enzyme (usually EcoRI), c) then inducing intramolecular ligation between cross-linked DNA fragments, d) reversing cross-links, and e) amplification of ligation products by quantitative PCR (Figure 1). Chromosome Conformation Capture can only be applied when both interacting fragments are known because it needs sequence information to design primers to amplify the inter-molecularly ligated DNA fragments. Subsequently, advanced versions of 3C have been developed, including 3C-on-chip (4C), also known as circular 3C, which can detect interactions between one genomic site of interest and all its potential interacting positions (Simonis et al., 2006). Carbon copy 3C (5C) has been developed to assay interactions at a high-throughput scale via the use of multiplex PCR (Dostie et al., 2006). The genuine genome-wide high-throughput method of detecting chromatin interactions is the Hi-C assay (Belton et al., 2012), which naturally combines 3C library with high-throughput sequencing technology. However, the resolution of Hi-C is limited due to the vast complexity of chromatin interactions in cells. Very recently, Promoter Capture Hi-C, which captures interactions between all annotated promoters and other genomic elements through oligonucleotides hybridization, has been implemented, largely by increasing the resolution of Hi-C (Schoenfelder et al., 2015). In addition to chromatin interaction analysis, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) can investigate chromatin interactions mediated by specific proteins, for example, RNA polymerase II (RNAPII) and various transcription factors (Fullwood et al., 2009).

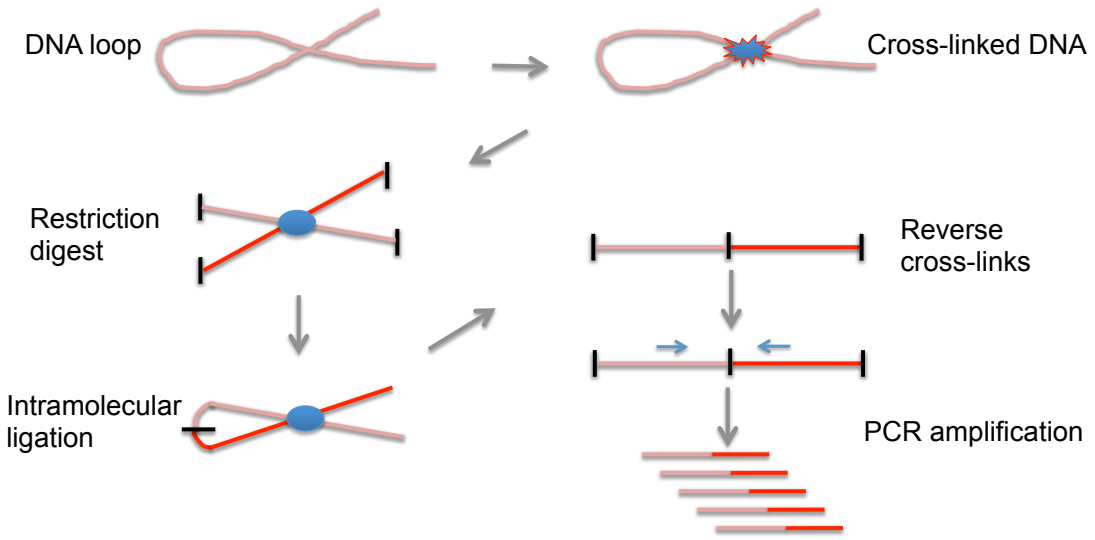


Figure 2. Chromosome Conformation Capture assay (3C)

Application of high-throughput functional assays. High-throughput functional assays have been recently exploited to identify causative mutations underlying GWAS signals. For example, SNP rs6983267 is located 500 kb upstream of the *MYC* oncogene, a protein implicated in many cases of colorectal and prostate cancer (Sur et al., 2012). This SNP resides in an enhancer of *MYC* and affects *MYC* expression via long-range chromatin interactions, as revealed by chromatin conformation capture assays (Ahmadiyeh et al., 2010) and ChIP-seq with transcription factor Tcf7l2 in mice (Sur et al. 2012). Moreover, both epigenomic and chromatin accessibility data showed an enhancer within a 130-kb LD block associated with prostate cancer risk interacts with the *SOX9* gene, explaining the cancer risk at this locus (Zhang et al., 2012). Another example, Zhou et al. (2012) reported that SNPs associated with chronic obstructive pulmonary disease located in an enhancer interacting with *HHIP* promoter, were associated with reduced *HHIP* gene expression. Recently, variants in enhancers of the *IRX3* gene were shown to affect obesity and type 2 diabetes in humans (Smemo et al., 2014). Initially, GWASs have identified SNPs within intron 1 and 2 of the *FTO* gene in a 47-kb LD block that have the strongest association with obesity and type 2 diabetes (Dina et al., 2007; Frayling et al., 2007; Scuteri et al., 2007). Because the *FTO* gene

is known to affect body weight and metabolism in mice, numerous follow-up studies all speculated that the associated SNPs are involved in *FTO* function or regulation of its expression. However, no direct evidence was obtained for these variants affecting *FTO* expression (Gorkin and Ren, 2014). Recently, Smemo et al. (2014) found that enhancers spanning these associated SNPs physically interact with the *IRX3* gene located 500 kb downstream of the first intron of *FTO* by using the 4C technology. Expression QTL analysis in brain tissues indicated the associated SNPs alter the *IRX3* expression but not that of *FTO*. Furthermore, *Irx3*-knockout mice showed a 25-30% reduction in body weight compared to control mice, strongly suggesting that disrupting enhancers of *IRX3* influence body weight and fat deposition. This study highlights that focusing exclusively on genes spatially proximal to specific GWAS peaks as the target genes is not prudent, even if the closest genes are functional relevant to the disease and trait of interest.

The necessity of characterization of functional elements across genomes. High-throughput functional assays have dramatically improved our ability to dissect the molecular mechanisms of regulatory variants in complex traits. Profiling functional elements systematically across the entire genome is essential for understanding how genomes work at transcriptional and regulatory levels. Such a catalogue of elements available in public databases would provide a valuable resource for assigning functional properties to putative regulatory fragments and variants. The human ENCODE project has been conceived to fulfill this goal (http://www.genome.gov/encode/). ENCODE studied 147 different cell types by using various functional elements detection approaches and assigned biochemical functions to a large fraction of the human genome (The ENCODE Project Consortium, 2012). However, the discovery that the high proportion (~80%) of the human genome is biologically functional is still controversial. In summary, by using RNA-seq, cap-analysis of gene expression (CAGE) and paired end tags sequencing (PET), ENCODE observed that processed and primary transcripts cumulatively cover respectively 62.1% and 74.7% of the human genome in 15 cell lines. On average, for each cell line, 22% and 39% of the genome is respectively

covered by processed or primary transcripts, which are mostly noncoding genes. This greatly expanded the catalogue of 'transcribable' sequence and revealed a significant overlap of neighboring genes, and perhaps suggests a need to reexamine the definition of a gene. ENCODE project identified 2.9 million high-confidence DHSs in 125 human cell types by using DNase I-seq, of which most showed cell-specific patterns and 34% were detected in only one cell line (Thurman et al., 2012). Only 5% of the DHSs lie within 2.5 kb of a transcriptional start site (TSS), implying promoter activity; and the remaining 95% lie in distal regions from TSS, pointing to other potential regulatory functions. In addition, 94.4% of the ChIP-seq peaks of ENCODE transcription factors coincide with the identified DHSs, showing highly reproducible results for these technologies. DNase I footprinting assays detected 45 million protein-binding sequences with a size of 6-40 bp in 41 cell lines, significantly overlapping with functional SNPs found in other studies. Furthermore, these footprints lead to the discovery of ~290 more conserved TF binding motifs, almost doubling the previously known number of the TF binding motifs (Neph et al., 2012). ENCODE also systematically applied 5C assay in GM12878, K562 and HeLa-S3 cell lines to map thousands of looping interactions between TSS and distal elements (Sanyal et al., 2012). It shows that elements engaged in long-range interactions are prone to locate at ~120 kb upstream of the TSS. Most importantly, only 7% of the looping interactions are between regulatory elements and its nearest gene, indicating that choosing the nearest genes as target is often incorrect when predicting distal elements-TSS interactions. Additionally, the NIH Roadmap Epigenomics project maps DNA methylation, histone modifications, chromatin accessibility, and small RNA transcripts in stem cells and primary tissues based on next-generation sequencing technologies, provides an epigenetic framework for scientific community (http:// www.roadmapepigenomics.org).

Recently, we have noted the successful applications of functional element databases in elucidating the functional mechanism behind associated SNPs in GWAS (Almeida et al., 2014; Edwards et al., 2013; French et al., 2013; Maurano et al., 2012; Zhang et al., 2012).

For example, in the study of French et al. (2013), after initial association mapping risk loci for breast cancer, fine mapping with stepwise logistic regressions narrowed the associated signal to 5 SNPs representing 3 independent haplotype blocks. Through cis-eQTL analysis, mining public ChIP-seq data against H3K4me1 and H3K4me2, and ERα ChIA-PET data, they found that the 5 most significantly associated SNPs are located in 2 putative regulatory elements (PRE1 and PRE2), which interact with the *CCND1* promoter and terminator. Later, 3C assays validated this finding. Finally, using luciferase reporter assays and electrophoretic mobility shift assay (EMSA) they identified 2 SNPs affecting function of PRE1 as an enhancer and 1 SNP affecting function of PRE2 as a silencer of *CCND1*.

There is a large gap between studies of human and domestic animals regarding the generation of functionally genomic elements profiles. Datasets of Encyclopedia of DNA Elements have been very recently produced for the mouse (Shen et al., 2012). More than 70% of homologous promoters are at high degree of conservation between these two species, but only ~25% of enhancers and CTCF-binding sites are functionally conserved despite the high sequence conservation (Shen et al., 2012). This implies that the research communities of domestic animals probably can not rely on human or mouse ENCODE data by means of comparative genomics, and will need to generate comparable ENCODE-like datasets to understand the consequence of regulatory variants identified in the GWAS in domestic animals.

Of note, not all of the studies mentioned above finally identified causative variants for the diseases through high-throughput functional assays, partially reflecting the difficulty of the fine mapping step for GWASs. But these efforts have narrowed the likely causative variants to just a few and illustrated plausible mechanisms on how highly associated variants mediate expression patterns of target genes. These insights have provided valuable knowledge for clinical practice and pharmaceutical research.

Screening signatures of selection by high-throughput sequencing

Domestic animals exhibit profound phenotypic diversity compared to their wild relatives. The domestication process and artificial selection in domestic animals have dramatically shifted their appearance, behaviors and reproduction patterns. Selection through generations changed the landscape of allele frequency and haplotype structure of selected and adjacent loci, leaving signatures of selection in the genome. More recently, detecting signatures of selection with genome-wide sequence data has been implemented to decipher the genetic basis of polygenic traits in various species. This new approach can be achieved by scanning loci with low heterozygosity, high-frequency extended haplotypes within a population, or highly differentiated genetic signals between populations or breeds.

Selection acts on individuals in three forms: negative, balancing, and positive. Selection imposed by human could increase frequencies of alleles beneficial for production traits, resulting in significantly reduced genetic diversity at target loci, i.e., selective sweeps. This is termed a signature of positive selection. Phenotypes fixed in some breeds by selection usually became the characteristics of the breed such as hyper-muscularity in BBB meat cattle and Belt coat color pattern in Hampshire pigs. Loci under ongoing recent selection still exhibit genetic diversity and the segregation of different phenotypes. Initial attempts to understand the genes underlying fixed breed characteristics prior to genome-wide sequencing era involved QTL mapping approaches, based on the construction of experimental intercrosses between breeds with and without a specific characteristic. This approach has been supplanted by GWAS, which has been employed to find genes involved in ongoing selection within a population. In the context of HTS, ancient positive selection (closed to fixed) can be detected by Fst test, and pooled heterozygosity etc., while recent selection can be identified by the long-range haplotype test (LRT) and integrated haplotype scores (iHS) test (Sabeti et al., 2002, 2007). Compared to QTL and GWAS mapping, searching for signatures of selection in sequencing data is more time-efficient and even cost-efficient, especially for those species that lack a reference genome.

Before WGS data was applied to searching for signatures of selection, medium- and high-density array genotypes were used (Sabeti et al., 2002; Voight et al., 2006). Array-based SNP genotyping data suffers from ascertainment bias of markers and has lower map resolution. Of note, it has reported that none of the results found with SNP arrays has been replicated by later WGS data in the dog (Axelsson et al., 2013). Thus the selective signals detected by using SNP array need to be interpreted with caution.

Detecting selective sweeps is powerful for the identification of domestication-related adaptation. In principle, loci identified as signatures of selection by WGS could be detected either by QTL mapping with experimental cross-lines or GWAS, but there might be one exception – the "domestication syndrome" (Larson et al., 2014). Domestication introduces a variety of changes to the animal which sets it apart from its wild ancestors: reduced fear of humans, increased tameness, morphological and social behavior changes, and altered reproduction patterns, collectively termed "domestication syndrome". Thus, to assess the genes that participated in the domestication events by gene mapping strategy necessitates the breaking down of domestication-related phenotypes into dozens of traits, if not hundreds (e.g., Figure 3). Armed with HTS, one simple strategy for detecting domestication-related

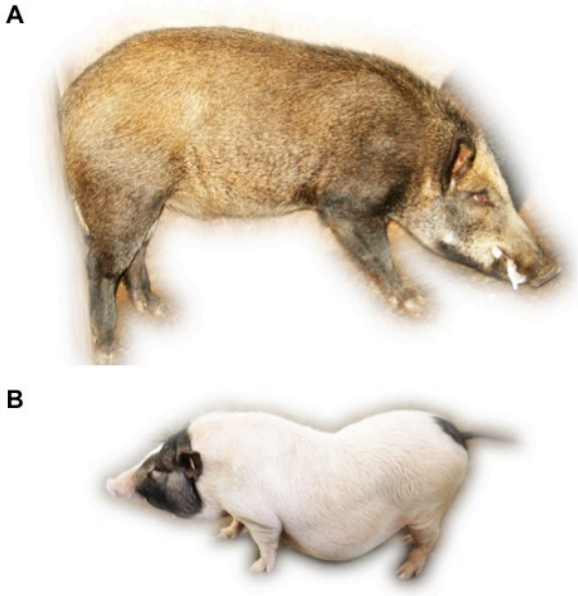


Figure 3. A) A wild boar from China. B) A Bamaxiang miniature pig from south China. The miniature pigs were different in body size, back shape, coat color, growth, tameness, reproduction and possible numerous other traits when compared to wild boars.

adaptation in one species is the comparison of genetic diversity across a pool of sequenced animals from various domestic breeds/populations with their wild relatives. By applying this strategy, Rubin et al. (2010) identified a sweep upstream of semaphoring 3A (*SEMA3A*) that may influence brain development pattern, and another sweep around the *TSHR* gene that may have relaxed the strict regulation of seasonal reproduction in the domestic chicken. A sweep was also observed upstream of V-set and transmembrane-domain-containing protein 2A (*VSTM2A*) of unknown consequence. Rubin et al. (2012) also used the same approach to detect selective sweeps encompassing *NR6A1, PLAG1,* and *LCORL* genes, responsible for body elongation and increased number of vertebrae in European domestic pigs. Another interesting study along these lines has been done in dogs. Axelsson et al. (2013) performed a genome-wide search for selective sweeps, and identified domestication-related genes that were enriched in 'nervous system development' and 'starch metabolic process'. More interestingly, they confirmed that several highly selected genes including *AMY2B* are engaged in the three-step breakdown of starch, indicating that early domestication of dogs was at least partly driven by a shift in diet. Identifying signatures of selection with high-throughput sequencing has been achieved in a number of species and highlighting some very important genes in the process of domestication and adaptation. But there is limitation in this kind of studies, and we will discus below.

Limitation of the implication of scanning signatures of selection. When comparing two different breeds or populations, the phenotypic difference between the comparison-pairs are unlikely to be limited to one or a handful of traits. Instead, breeds often differ in a number of visible morphological and invisible physiological traits. As a result selective sweeps are not informative if no other biological information is available. If selective sweeps overlap with or are adjacent to functional genes, we can infer traits the sweeps act on. However, if sweeps

are located in noncoding regions without adjacent genes, there is no obvious way to interpret their biological significance. These limitation result in studies screening for signatures of selection to usually only discuss sweeps overlapping with a specific gene, QTL or GWAS loci (Qanbari et al., 2014; Rubin et al., 2010, 2012). Some studies typically included genes located within 100 kb up- and downstream of the sweeps, assuming the sweeps might have acted on regulatory elements, which in turn regulates the nearby genes. However, recent work from ENCODE and other projects, show that the view of regulatory elements targeting the nearest genes does not often hold true.

Another challenge for identifying selective sweeps is the statistical difficulty of distinguishing sweeps from sites of genetic drift. To solve this problem, previous studies have usually grouped diverse breeds and populations for analysis assuming the chance of genetic drift happens in an exact same direction in multiple breeds/populations is very low (Axelsson et al., 2013; Rubin et al., 2010, 2012). An alternative approach is to perform sequence simulations to define a p value for a observed sweep (Sabeti et al., 2007).

In summary, we first describe the use of medium-density SNP arrays to dissect genetic basis of a monogenic disease (Crooked Tail Syndrome, CTS), as well as of complex phenotypes related to meiotic recombination in cattle. The subsequent application of NGS technology in forward genetic analyses, allowed us to identify: i) an intronic mutation in the *PIGH* gene that causes the skipping of exon 2, and is the cause of arthrogryposis in BBC, ii) a 3.3 kb deletion removing exons 25-27 of the *FANCI* gene, that causes brachyspina in Holstein-Friesian cattle. We finally used a reverse genetic approach to identify embryonic lethal (EL) mutations compromising fertility in beef (BBC) and dairy cattle (Holstein-Friesian and Jersey), and a loss-of-function non-lethal variant causes diluted color – 'cool gray' – in BBC.

Objectives

The objectives of the present thesis are:

1) Strong selection for meat and milk traits in modern cattle breeds through heavy use of elite sires by means of artificial insemination has significantly reduced the effective population size in cattle. Therefore, it increased the probability of disease-causing or lethal recessive alleles coalescing into homozygotes in the population. We intended to provide accurate genetic tests for outbursts of congenital defects by highly-efficiently mapping of the causative mutations in a classic forward genetic approach. In this thesis, we dealt with three monogenic defects - crooked tail syndrome and arthrogryposis in BBC, and brachyspina syndrome in Holstein dairy cattle.

2) In cattle populations, large amount of animals have been genotyped with medium-density SNP array for the purpose of genomic selection. Combined with a multi-generation pedigree, usually comprising large half-sib families in cattle, it gives us a great opportunity to study a very basic and key mechanism in genetics – recombination.

3) Fertility is a complex trait and difficult to be dissected by traditional forward genetic approach. We here try to dissect genetic basis of fertility decline (caused by embryonic lethality) by using a reverse genetic approach in the next-generation sequencing context.

Part I. Dissecting genetic basis of the crooked tail syndrome and recombination by using SNP array

Balancing selection of a frame-shift mutation in the *MRC2* gene accounts for the outbreak of the crooked tail syndrome in Belgian Blue cattle

Corinne Fasquelle, Arnaud Sartelet, Wanbo Li, Marc Dive, Nico Tamma, Charles Michaux, Tom Druet, Ivo J. Huijbers, Clare M. Isacke, Wouter Coppieters, Michel Georges, Carole Charlier.

PLoS Genetics, 2009, Issue 9, e1000666.

Background

Between 2006 and 2007, hundred of calves born in the Belgian Blue cattle population suffered a defect named "Crooked Tail Syndrome" (CTS). All the case calves were characterized as suffering from growth retardation, abnormal skull, extreme muscular hypertrophy and unusual crooked tails, with the some additional infrequent symptoms among cases. This defect is not lethal, but the result of growth retardation and carcass depreciation causes large economic loss to the breeders.

With the goal of providing an accurate genetic test assay for CTS, we initially sought to map the causative mutation of the defect using SNP array technology. The CTS locus was mapped to a 2.42-Mb interval by using 8 cases and 36 controls that were genotyped with the 50 K SNP chips via identity-by-descent (IBD) mapping (Charlier et al., 2008). The interval was then refined to an 812-kb segment with 5 SNPs genotyped in 135 CTS cases. Sequencing of the coding exons in the seven genes locating in the intervals revealed a 2bp-deletion in the *MRC2* gene conferring the CTS in BBC. The discovery of CTS mutations in the BBC population reveals an interesting example of balancing selection in farm animals, which gives advantages to meat production in heterozygous animals but causes severe healthy problem in homozygotes.

OPEN ACCESS Freely available online

PLoS GENETICS

Balancing Selection of a Frame-Shift Mutation in the *MRC2* Gene Accounts for the Outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle

Corinne Fasquelle[1☯], Arnaud Sartelet[1☯], Wanbo Li[1¤], Marc Dive[1], Nico Tamma[1], Charles Michaux[2], Tom Druet[1], Ivo J. Huijbers[3], Clare M. Isacke[3], Wouter Coppieters[1], Michel Georges[1], Carole Charlier[1]*

1 Unit of Animal Genomics, GIGA-R, Department of Animal Sciences, Faculty of Veterinary Medicine, University of Liège, Liège, Belgium, **2** Unit of Bioinformatics, Department of Animal Sciences, Faculty of Veterinary Medicine, University of Liège, Liège, Belgium, **3** Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London, United Kingdom

Abstract

We herein describe the positional identification of a 2-bp deletion in the open reading frame of the *MRC2* receptor causing the recessive Crooked Tail Syndrome in cattle. The resulting frame-shift reveals a premature stop codon that causes nonsense-mediated decay of the mutant messenger RNA, and the virtual absence of functional Endo180 protein in affected animals. Cases exhibit skeletal anomalies thought to result from impaired extracellular matrix remodeling during ossification, and as of yet unexplained muscular symptoms. We demonstrate that carrier status is very significantly associated with desired characteristics in the general population, including enhanced muscular development, and that the resulting heterozygote advantage caused a selective sweep which explains the unexpectedly high frequency (25%) of carriers in the Belgian Blue Cattle Breed.

Citation: Fasquelle C, Sartelet A, Li W, Dive M, Tamma N, et al. (2009) Balancing Selection of a Frame-Shift Mutation in the *MRC2* Gene Accounts for the Outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. PLoS Genet 5(9): e1000666. doi:10.1371/journal.pgen.1000666

Editor: Gregory S. Barsh, Stanford University School of Medicine, United States of America

Received May 19, 2009; **Accepted** August 27, 2009; **Published** September 25, 2009

Funding: This work was funded by grants from the Walloon Ministry of Agriculture (Rilouke), the Belgian Science Policy Organisation (SSTC Genefunc PAI), the University of Liege, and Breakthrough Breast Cancer Research Centre. CC is Chercheur Qualifié of the Fonds National de la Recherche Scientifique. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: carole.charlier@ulg.ac.be

¤ Current address: Key Laboratory for Animal Biotechnology of Jiangxi Province and the Ministry of Agriculture of China, NanChang, People's Republic of China.

☯ These authors contributed equally to this work.

Introduction

The Belgian Blue Cattle breed (BBCB) is notorious for its exceptional muscular development known as "double-muscling". This extreme phenotype is due in part to an 11-bp loss-of-function deletion in the myostatin gene that has been fixed in the breed (e.g. [1]), as well as to ongoing selection on as of yet unidentified polygenes influencing muscularity. As in other breeds, intense selection has substantially reduced the effective population size. Extensive reliance on artificial insemination (AI), in particular, by allowing popular sires to have thousands of descendants, narrows the genetic basis. The concomitant increase in the rate of inbreeding causes recurrent outbreaks of recessive defects. Inherited defects that have lately afflicted the BBCB include the recently described Congenital Muscular Dystonias (CMD) I and II [2].

As a result of this peculiar demography of domestic animal populations, inherited defects generally involve unique "founder" mutations. Allelic homogeneity greatly facilitates positional identification using identity-by-descent (IBD) mapping, as recently demonstrated using the first generation high density SNP arrays for the bovine [2]. The genes underlying CMD I & II were readily mapped, and the causative mutations in the *ATP2A1* and *SLC6A5* genes identified. The widespread use of the resulting diagnostic tests allowed immediate and effective control of the corresponding pathologies.

We herein report the positional identification of the mutation causing a novel, recently appeared defect referred to as Crooked Tail Syndrome (CTS). The incidence of CTS has risen very suddenly in the BBCB, and 25% of animals now appear to be CTS carriers. We herein provide strong evidence for exacerbated muscular development of carriers of the CTS mutation, conferring "heterozygote advantage" underlying the selective sweep that raised the causative mutation to alarming proportions.

Results

Crooked Tail Syndrome (CTS) exhibits variable expressivity

We recently established a heredo-surveillance platform operating in close collaboration with field veterinarians to rapidly identify emerging genetic defects. As part of these activities, 105 CTS cases were reported to the platform between November 2006 and November 2007. In addition to the striking deviation of the tail (equally likely to be dextro- or levo-rotatory), detailed clinical examination revealed three symptoms shared by all cases: (i) general growth retardation manifesting itself at approximately one month of age, (ii) abnormal skull shape manifested as a shortened

Author Summary

Livestock are being subject to intense artificial selection aimed at ever-increasing, sometimes extreme, production phenotypes. This is well-illustrated by the exceptional muscular hypertrophy characterizing the "double-muscled" Belgian Blue Cattle Breed (BBCB). We herein identify a loss-of-function mutation of the bovine *MRC2* gene that increases muscle mass in heterozygotes, yet causes skeletal and muscular malformations known as Crooked Tail Syndrome (CTS) in homozygotes. As a result of the "heterozygote advantage", the *MRC2 c.2904_2905delAG* mutation has swept through the BBCB population, resulting in as many as 25% carrier animals and causing a sudden burst of CTS cases. These findings highlight one of the risks associated with pushing domestic animals to their physiological limits by intense artificial selection.

broad head, and (iii) extreme muscular hypertrophy including a conspicuous outgrowth of the gluteus medius anchor. Additional symptoms were observed in a substantial proportion but not all cases: (i) spastic paresis of the hind limbs affecting either the quadriceps only (22%), or quadriceps and gastrocnemius (14%), often associated with straight hocks, (ii) short, straight and extended fore limbs (33%), and (iii) pronounced scoliosis with asymmetric development of the muscles of the back (20%). Figure 1 illustrates the corresponding symptomatology. We performed complete necropsy of a few selected cases but detected no additional obvious abnormalities. Moreover, radiological examination of crooked tails and scoliotic spines failed to reveal structural defects of the vertebrae (data not shown).

Although the defect is not lethal by itself, the most severe cases (~25%) were euthanized on welfare grounds. The surviving ~75% nevertheless caused important economic losses to their owners as a result of growth retardation and carcass depreciation.

CTS is caused by a fully penetrant, two base pair deletion in the open reading frame (ORF) of the *MRC2* gene

We previously mapped the CTS locus to bovine chromosome 19, in a 2.4 Mb interval shared homozygous-by-descent by the eight analyzed CTS cases [2]. To refine the map location of the CTS locus we genotyped the 105 reported CTS cases for five SNPs covering the 2.4 Mb interval (Figure 2A). The SNPs were selected on the basis of the low population frequency of the disease-associated allele. Genotyping was achieved by first sequencing 35 pools of three animals, followed by individual sequencing of the pools revealing the presence of the major allele and therefore of one or more recombinants. This approach allowed us to confine the critical region to the 812 Kb rs29010018 - AAFC03034831 interval. It comprises seven annotated genes which were ranked on the basis of their perceived relevance with regard to the CTS condition. Coding exons were sequenced in an affected and a matched healthy control individual.

Figure 1. Clinical spectrum exhibited by CTS cases. Crooked tail, growth retardation, stocky head, extreme muscular hypertrophy, spastic paresis of the hind limbs, straight hock, scoliosis.
doi:10.1371/journal.pgen.1000666.g001

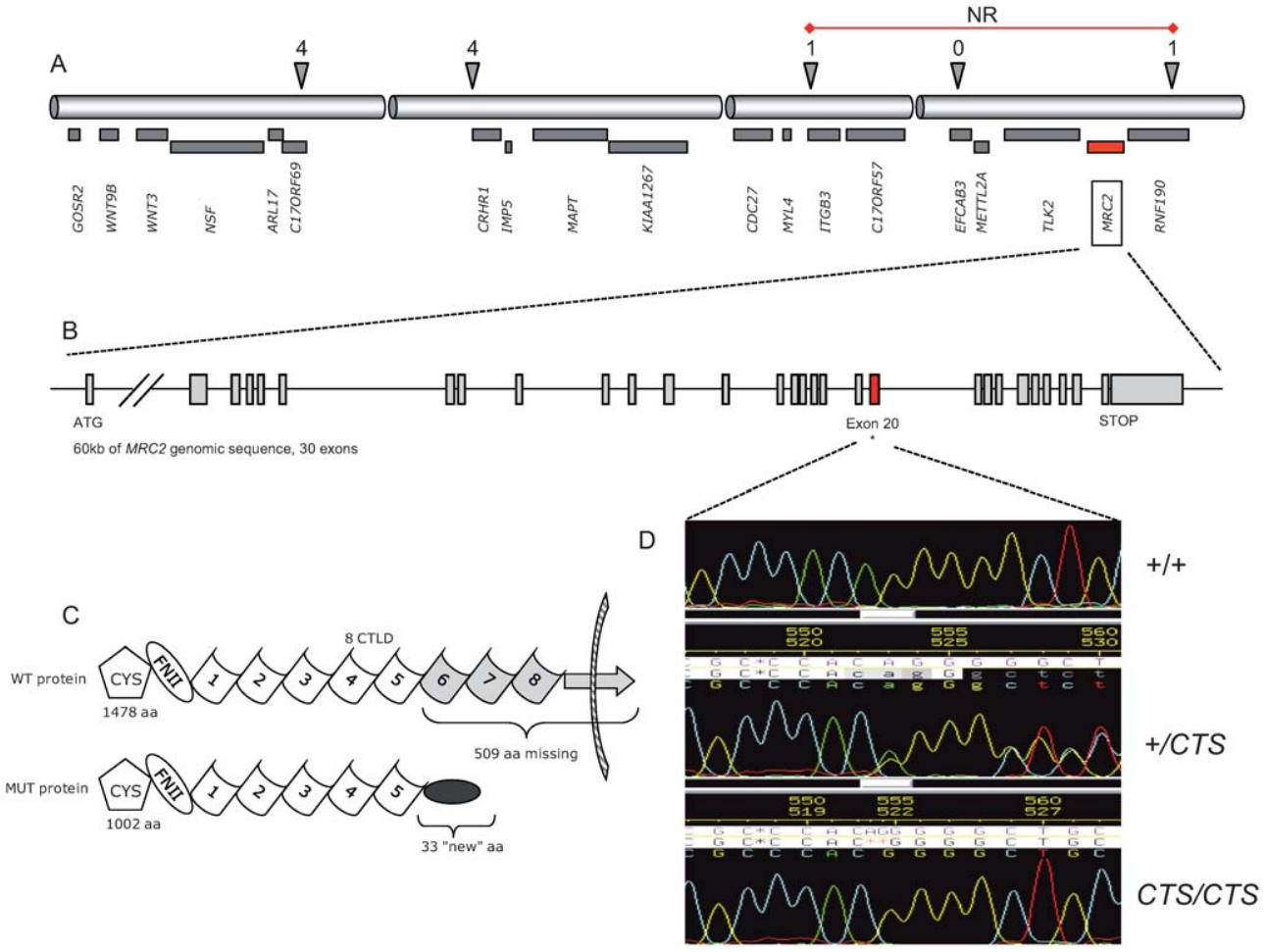


Figure 2. Positional identification of the *c.2904_2905delAG MRC2* mutation causing CTS, and its effect of the Endo180 protein. (A) Gene content of the 2.4 Mb interval in which the CTS mutation was located using identity-by-descent mapping. The triangles correspond to five SNPs used to refine the CTS locus position, with corresponding number of recombinant individuals out of 105 CTS cases. The resulting non-recombinant (NR) interval is marked by the red horizontal line. (B) Structure of the *MRC2* gene within that interval. (C) Domain composition of the wild-type (WT) and mutant (MUT) Endo180 protein. (D) Sequences traced obtained from genomic DNA of a homozygous wild-type, carrier and homozygous CTS animal showing the deletion of the ApG dinucleotide in the mutant.
doi:10.1371/journal.pgen.1000666.g002

During this process, we identified a 2-bp deletion in the ORF of the mannose receptor C type 2 (*MRC2*) gene. The *MRC2* gene encodes the 180 kDa Endocytic Transmembrane Glycoprotein (Endo180), one of the four members of the mannose receptor family [3,4]. Endo180 is a recycling endocytic receptor that is predominantly expressed in mesenchymal cells such as stromal fibroblasts and in the chondrocytes and osteoblasts/osteocytes in the developing bones, and is proposed to play a role in regulating extracellular matrix degradation and remodelling. It has C-type lectin activity, binds collagen and interacts with urokinase-type plasminogen activator receptor (uPAR) in a trimolecular cell surface complex with pro-urokinase plasminogen activator (pro-uPA). The 180 kD Endo180 protein comprises an aminoterminal cysteine-rich domain of unknown function, a fibronectin type II domain which mediates collagen binding, eight C-type lectin-like domains (CTLDs) of which the second mediates Ca^{2+}-dependent lectin activity, a stop-transfer signal anchoring this single-pass transmembrane protein in the membrane, and a carboxyterminal cytoplasmic domain allowing association with adaptor proteins in the clathrin coat. The mutation identified in CTS cases is located in exon 20 and deletes nucleotides 2904 and 2905 of the *MRC2* cDNA (*c.2904_2905delAG*). It is predicted to append a frame-shifted 30-residue peptide to a truncated Endo180 receptor missing the CTLD6-8 domains, the stop-transfer signal and the cytoplasmic domain (Figure 2B,C,D). As a result, the mutated protein should be unable to localize to the plasma membrane and mediate receptor-mediated endocytosis.

We developed a 5′ exonuclease assay for the mutation and genotyped the 105 reported CTS cases. All proved to be homozygous for the *c.2904_2905delAG* mutation. We then genotyped 1,899 healthy Belgian Blue animals. Unexpectedly, 24.7% of animals appeared to be carriers, without a single homozygous mutant ($p<10^{-12}$ assuming Hardy-Weinberg equilibrium). Taken together, these results allowed us to incriminate the *c.2904_2905delAG* mutation as being causal and fully penetrant.

Mutant *MRC2* mRNAs are targeted by the nonsense-mediated decay (NMD) RNA surveillance pathway

C.2904_2905delAG causes a frame-shift resulting in a premature stop codon in the 21st of the 30-exon *MRC2* gene. Mutant mRNAs

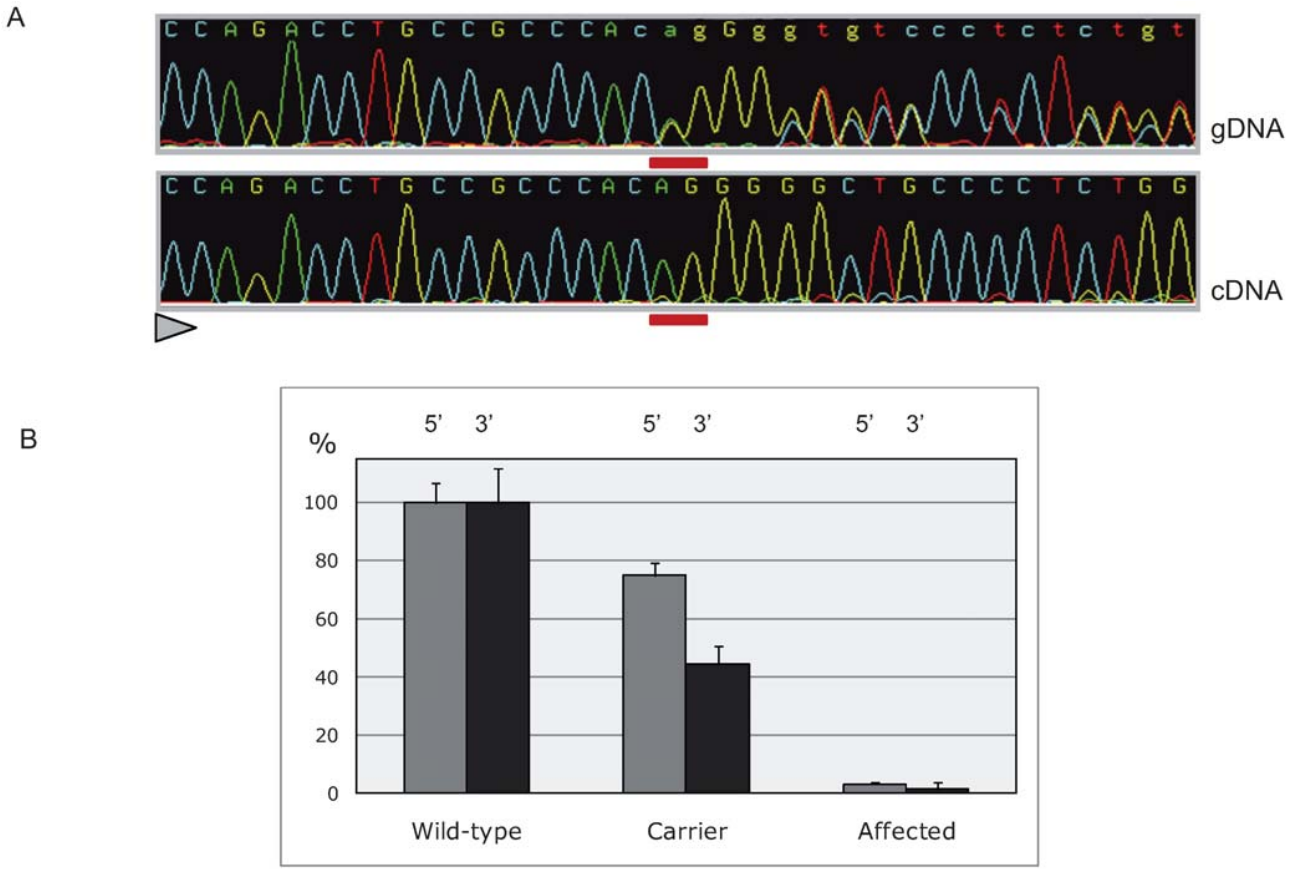


Figure 3. Nonsense-mediated RNA decay of *c.2904_2905delAG* mutant *MRC2* transcripts. (A) Direct sequencing of *MRC2* amplicons spanning the CTS mutation obtained from genomic DNA and pulmonary cDNA of a heterozygous animal, showing the virtually exclusive detection of wild-type allele amongst transcripts (the position of the deleted nucleotides is underlined in red, the sequencing direction is represented by a triangle). (B) Comparing *MRC2* mRNA levels in the lung of +/+, +/CTS and CTS/CTS animals. Data are shown for two amplicons at the 5′ and 3′ ends of the mRNA, respectively. Error bars correspond to standard errors over three replicates per sample.
doi:10.1371/journal.pgen.1000666.g003

are therefore predicted to undergo NMD [5]. To test this, we first compared the levels of wild-type and mutant *MRC2* mRNA in lung and skeletal muscle of a carrier animal by direct sequencing of RT-PCR products encompassing the deletion. As can be seen from Figure 3A, mutant mRNA was barely detectable. We then compared the levels of *MRC2* mRNA in lung tissue of animals of the three genotypes using quantitative RT-PCR performed with primer sets targeting the 5′ and 3′ end of the mRNA respectively. Highly significant reductions in *MRC2* mRNA levels were observed in carriers relative to homozygous wild-type individuals (75%±4% and 45%±6% of control values for the 5′ and 3′ systems respectively), while *MRC2* mRNA levels in cases were less than 5% of homozygous wild-types (Figure 3B). Both the allelic imbalance and qRT-PCR experiments thus supported degradation of the mutant transcripts by NMD.

Amounts of full-length Endo180 protein are halved in tissues of carrier animals

From the ten anti-human Endo180 antibodies tested by Western blotting, only one polyclonal rabbit antibody (CAT2) detected the bovine Endo180 protein with sufficient specificity. The CAT2 antibody is directed against the last 19 amino acids of Endo180 of which the last 18 are perfectly conserved between human and cow [6]. CAT2 was thus predicted to allow recognition of the wild-type but not mutant Endo180. As expected, no wild-type Endo180 was detected in lung tissue of CTS affected animals. In carrier animals, the levels of Endo180 protein were approximately half those observed in homozygous wild-types (Figure 4). Assuming that Endo180 is dosage sensitive,

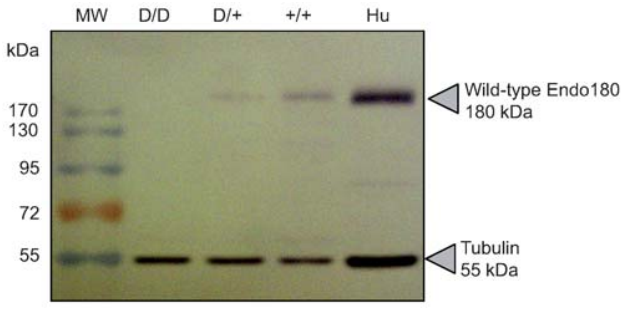


Figure 4. Effect of the CTS mutation on the levels of full-length Endo180. Western blot results from lung of animals of the three genotypes. Hu: human control sample. MW: molecular weight marker. The 55 Kd band corresponds to non-specific binding of the CAT2 antibody to tubulin, used as control for the amount of protein loaded.
doi:10.1371/journal.pgen.1000666.g004

such reduction in the supposedly functional species might affect phenotype.

CTS carrier status increases muscle mass

The unusually high frequency of the CTS mutation in BBCB suggested that it might confer heterozygote advantage in this highly selected population. To test this hypothesis we estimated the effect of carrier genotype on 22 type traits evaluating muscularity, skeletal conformation, size and leg soundness, which are systematically recorded as part of the selection programs implemented in the BBCB. The analysis was conducted on 519 pedigreed bulls, including 148 carrier and 371 homozygous wild-type animals, using a mixed model including fixed effects of *MRC2* genotype, year at scoring, body condition and age at scoring, as well as a random individual animal effect. Variance components and effects were estimated by restricted maximum likelihood (REML) analysis. Highly significant effects were obtained for the four categories of recorded traits (Table 1). CTS carrier animals were smaller, stockier and more heavily muscled. They had a thinner skeleton and more rounded ribs, which are characteristics of beef cattle. *MRC2* genotype accounted for 3.6%, 3.6% and 2.6% for the genetic variance of height, muscularity and general appearance, respectively. These results strongly suggest that CTS carrier frequency is increased by selection programs applied in BBCB.

The *C.2904_2905delAG* mutation is undergoing a selective sweep

To more directly demonstrate the occurrence of a selective sweep, we performed the following analysis. Examination of the available genealogies of the 105 affected individuals indicated that all of them trace back to Précieux, a popular AI sire, via both sire and dam. This suggested that Précieux, born in 1980, was CTS carrier and that its extensive utilization in the mid eighties spread the CTS mutation in BBCB. Genotyping Précieux and three of his sons for the *C.2904_2905delAG* mutation and the 60K Illumina chip, indeed demonstrated that he carried the CTS mutation embedded in the SNP haplotype shared homozygous-by-descent by the examined cases [2]. Thus, the vast majority of *C.2904_2905delAG* mutations encountered in present-day BBCB animals, trace back to Précieux.

We obtained DNA samples from all BBCB sires (174) born between 2003 and 2005, whose semen had been commercialized by one of the ten major Belgian AI studs. Such AI sires are heavily selected for extreme muscularity. Examination of the pedigrees indicated that 160 of the 174 [2003–2005] AI sires were descendents of Précieux. The number of generations separating these sires from Précieux averaged 5.9 (range: 3 to 8). Genotyping the *C.2904_2905delAG* mutation in this cohort identified 45 CTS carriers, all of them amongst the 160 descendents of Précieux.

Assuming that the CTS mutation indeed underwent a selective sweep, 45 carriers out of the 160 Précieux descendants would be significantly higher than expected by chance alone. To verify this assumption we simulated the segregation of a mutation in the true genealogy of the 160 descendants of Précieux and counted the resulting number of carrier bulls. In these simulations, Précieux was systematically assumed to be carrier, while the frequency of the mutation in animals unrelated to Précieux varied from 0 to 0.05. In the absence of selection (i.e. if a carrier animal is equally likely to transmit either the mutation or the wild-type allele to anyone of its descendents), the probability to obtain 45/160

Table 1. Effect of CTS carrier status on type traits in BBCB.

	Trait or syntetic note	Contrast	Std. error	p value	Carrier characteristics
SIZE	Withers height	2.33	0.320	***	Smaller
	Length	0.36	0.181	N.S.	
	Chest width	−0.68	0.319	*	Larger
	Pelvis width	−0.32	0.188	N.S.	
	Pelvis length	0.25	0.163	N.S.	
MUSC.	Shoulder muscling	−0.58	0.283	*	Increased muscularity
	Top muscling	−1.71	0.409	***	Increased muscularity
	Buttock side	−0.35	0.20	N.S.	
	Buttock rear	−0.50	0.251	*	Increased muscularity
	Synthetic note for muscularity	−0.70	0.246	**	Increased muscularity
	General appearance	−1.73	0.124	***	Better
SKELETAL CONFORM.	Skeleton	−1.18	0.386	**	Thinner
	Rib shape	−1.67	0.456	***	Ronder
	Fore legs stance	0.54	0.157	***	More toed-in
	Rear legs stance	−0.43	0.200	*	More toed-out
OTHER	Skin	−0.34	0.442	N.S.	
	Tail set	−0.24	0.529	N.S.	
	Shoulder bone	0.10	0.108	N.S.	
	Rump	1.25	0.425	**	More horizontal
	Top line	−0.36	0.137	**	More convex
	Hocks stance	0.84	0.367	*	Straighter

* p<5%, ** p<1%, *** p<1%.

doi:10.1371/journal.pgen.1000666.t001

carriers was 0.0014, 0.0023 and 0.0130 for mutation frequencies (outside the Précieux lineage) of 0.00, 0.01 and 0.05, respectively (Table S1). Thus we can confidently assert that the *C.2904_2905delAG* mutation indeed underwent a selective sweep in the BBCB.

To have some quantitative assessment of the intensity of the selective sweep, we repeated the "gene dropping" simulations while varying the degree of segregation distortion in favour of the mutant allele. Figure 5 shows the proportion of simulations yielding 45/160 carrier bulls as a function of the transmission probability of the CTS mutation from carrier parents to offspring. It can be seen that the outcome of 45/160 carrier bulls is most likely for a transmission rate between 0.62:0.38 (mutation frequency outside Précieux lineage of 0.05) and 0.67:0.33 (mutation frequency outside Précieux lineage $\leq$0.01). The fact that all 105 CST cases traced back to Précieux both on the dam and sire side, indicates that the mutation frequency outside of the Precieux lineage is closer to 1% than to 5%. Thus, a carrier animal is approximately two times more likely to be selected than a non-carrier sib.

Discussion

We herein describe a frame-shift mutation in the *MRC2* gene causing the CTS syndrome in cattle. Clinical manifestations of CTS are dominated by skeletal and muscular anomalies. Skeletal symptoms including growth retardation, abnormally shaped legs and skulls, are perfectly compatible with the known involvement of *MRC2* in regulating extracellular matrix degradation and remodeling and its strong expression in developing bone [7]. The muscular symptoms, including muscular hypertrophy, tail deviation and spastic paresis are more difficult to rationalize, although a role for the related mannose receptor in myoblast motility and muscle growth has been recently reported [8]. We cannot formally exclude the possibility that the muscular manifestations result from distinct sequence variants in linkage disequilibrium with the CTS mutation, although we favor the more parsimonious hypothesis of a single causative mutation.

It is noteworthy that mice homozygous for a targeted deletion of *MRC2* exons 2 to 6 have been generated in two independent laboratories [9,10]. Both laboratories reported that the mice were viable and fertile, although more recently a minor deficiency in long bone growth, bone mineral density and calvarial bone formation has been demonstrated [7]. Cells derived from these animals show a clear defect in collagen uptake and degradation. One reason for the more pronounced clinical manifestations in cattle than in mice may lie in the distinct nature of the murine and CTS *MRC2* mutations. Cells isolated from the genetically modified mice express a mRNA species in which exon 1 (containing the signal sequence) is spliced in frame onto exon 7 (containing CTLD2), and in embryonic fibroblasts a truncated Endo180 protein missing the cysteine-rich, FNII and CTLD1 domains can be expressed [9]. However, little or no truncated protein is found in postnatal tissues from these knockout mice [10,11]. Alternatively it may be that there are distinct degrees of redundancy between members of the mannose receptor family in different species. Also the more striking phenotype in cattle may be due the different genetic background and particularly the fact that the studied animals were all homozygous for a *MSTN* loss-of-function mutation [1]. This hypothesis could be tested by mating

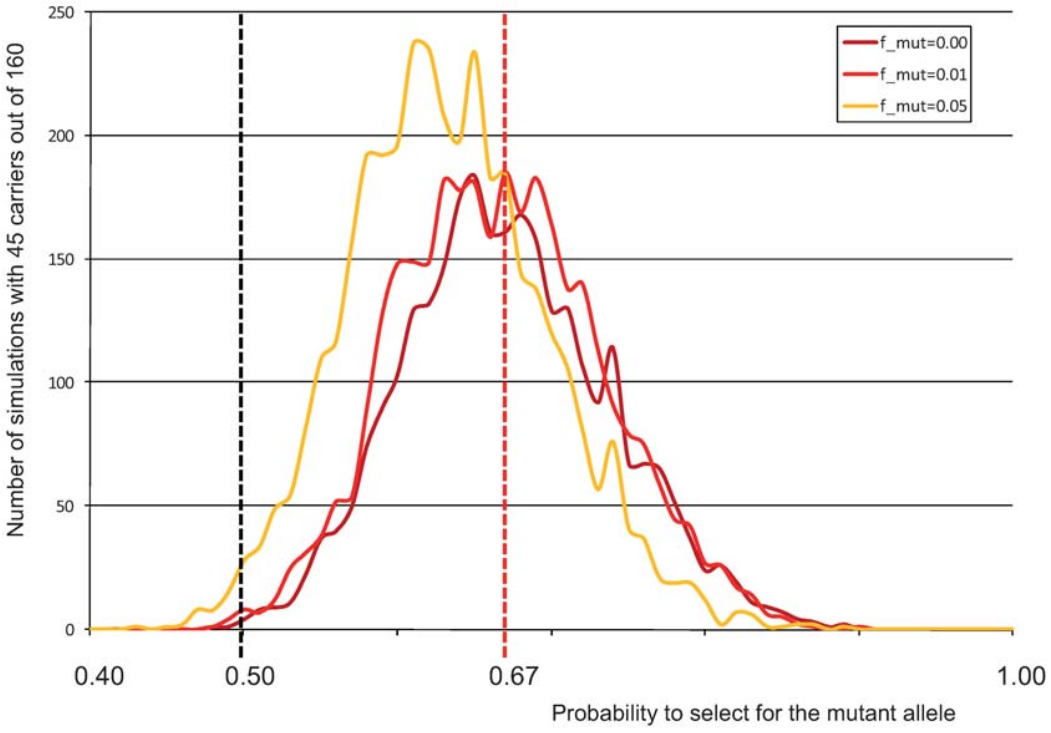


Figure 5. Distribution of the number of simulations (out of 10,000) yielding 45 carriers out of 160 Précieux descendants (Y-axis), as a function of the rate of transmission of the mutation from heterozygous carriers (X-axis). Three curves are given corresponding to frequencies of the mutation outside of the Précieux lineage of 0, 1, and 5%. The dotted red vertical line corresponds to a transmission rate of 67%, maximizing the number of simulations yielding 45 carriers for a mutation frequency (outside of the Précieux lineage) of 1%, considered to be an upper bound in BBCB.
doi:10.1371/journal.pgen.1000666.g005

the available *MRC2* and *MSTN* knock-out mice. Whatever the reason, at this point bovine CTS may be the more informative model to decipher *MRC2* function, and to assist in the identification of as of yet unidentified human pathological conditions resulting from *MRC2* loss-of-function.

We provide very strong evidence of phenotypic manifestations of the CTS mutation in carriers. This is more than likely reflecting dosage sensitivity for Endo180, as NMD causes the mutant protein to be present at near undetectable levels, thus very unlikely to affect cellular function *per se*. Enhanced muscularity of CTS carriers has supposedly contributed greatly to the rapid increase of the CTS mutation in the BBCB. Indeed, we demonstrate that carrier animals have approximately two times more chance to be selected as elite sires than their non-carrier sibs. Note that this is the level of segregation distortion expected for a gene that accounts for ~5% of the genetic variance for a trait with heritability of ~25% and assuming a selection intensity of ~2% (Table S2).

Such selective sweep is reminiscent of the spread of other inherited defects in domestic animals as a result of advantageous traits exhibited by carriers. These include loss-of-function mutations of the porcine ryanodine receptor and equine skeletal muscle sodium channel alpha subunit gene causing, respectively, malignant hyperthermia and hyperkalaemic periodic paralysis in homozygotes, yet increased muscle mass in heterozygotes [12,13], or of a *FGFR3* mutation causing hereditary chondrodysplasia in homozygous sheep and increased size in the carriers [14,15].

A diagnostic test for the CTS mutation has been developed and already applied on more than 4,000 BBCB samples. The resulting information should have an immediate and positive impact on the incidence of CTS, and protect animals and breeders against the pathological condition and ensuing economic losses.

This work is yet another illustration of the value of domestic animal populations in enriching the phenotype-genotype map. It adds to a recent list of positional cloning successes in poultry [16], dog [17,18], horse [19] and bovine [2].

Materials and Methods

Mutation scanning

Coding exons of positional candidate genes were amplified from genomic DNA of a CTS case and a matched control using standard procedures. The primers used for the *MRC2* gene are listed in Table S3. 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, Foster 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 CTS mutation

A Taqman assay was developed to genotype the CTS mutation, using 5′-GCG CAA CAG CAC CAG AGA-3′ and 5′-CTC CCT ACC TTG TTC AGG AAC TG-3′ as PCR primers, and 5′-CTG CCG CCC AC[*** ***] GGG-3′ (CTS) and 5′-CTG CCG CCC AC[**A G**]G-3′ (wild type) as Taqman probes. Reactions were carried out on a ABI7900HT instrument (Applied Biosystems, Foster City, CA) using standard procedures.

Allelic imbalance test of NMD

Total RNA was extracted from lung, heart and skeletal muscle of a two month old heterozygote *c.2904_2905delAG* animal using Trizol (Invitrogen). The RNA was treated with TurboDNase (Ambion). cDNA was synthesized using *SuperscriptTMIII* First Strand Synthesis System for RT-PCR (Invitrogen). A portion of *MRC2* cDNA, encompassing the deletion, was amplified using *MRC2* specific primers (Table S4). The PCR products were directly sequenced as described above.

Real-time quantitative RT-PCR test of NMD

Total RNA from lung and skeletal muscle was obtained from animals of the three genotypes (+/+, +/*CTS* and *CTS*/*CTS*). After *DNase*-treatment (Turbo DNA-free, Ambion), 500 ng total RNA was reverse transcribed in a final volume of 20 µl using the iScript cDNA Synthesis Kit (Bio-Rad). PCR reactions were performed in a final volume of 15 µl containing 2 µl of 2.5-fold diluted cDNA (corresponding to 20 ng of starting total RNA), 7.5 µl of 2× master mix prepared from the qPCR Core Kit for SYBR green I (Eurogentec), 0.45 µl of 1/2000 SYBR green I working solution prepared from the qPCR Core Kit for SYBR green I (Eurogentec), forward and reverse primers (250 nM each) and nuclease free water. PCRs were performed on a an ABI7900HT instrument (Applied Biosystems, Foster City, CA) under the following cycling conditions: 10 min at 95°C followed by 40 cycles at 95°C for 15 sec and 60°C for 1 min. Two primer sets were used to test *MRC2* expression (MRC2_5′QRT_UP/DN and MRC2_3′QRT_UP/DN) and seven genes were included as candidate endogenous controls: (1) Beta Actin *(ACTB)*, (2) Glyceraldehyde-3-phosphate Dehydrogenase (*GAPD*), (3) Hypoxanthine Phosphoribosyltransferase 1 *(HPRT1)*, (4) Ribosomal Protein Large P0 *(RPLP0)*, (5) Ribosomal Protein S18 *(RPS18)*, (6) Succinate Dehydrogenase Complex Subunit A Flavoprotein *(SDHA)*, and (7) Tyr-3- & Trp-5-Monooxygenase Activation Protein Beta *(YWHAB)*. After analyse of the results with geNorm [20], the four following genes were selected as best endogenous controls: *ACTB*, *RPLP0*, *RPS18* and *YWHAB*. The corresponding primer sequences are given in Table S4. All sample/gene combinations were analyzed in triplicate. Relative *MRC2* expression levels, for the 5′ & 3′cDNA parts, in the samples of the three genotypes were computed using the qBase software package (http://medgen.ugent.be/qbase/)(Hellemans et al., 2007).

Western blotting

A series of available antibodies directed against the human Endo180 were tested by Western blotting for cross reactivity with bovine Endo180 on commercial bovine aortic endothelial cells (BAOEC, Cell Applications). A positive control corresponding to a lysate of MRC5 human fibroblast cell line expressing Endo180 was included in each experiment. The tested antibodies were the following: (i) seven mouse monoclonal antibodies (for details see [21–23]), (ii) a rabbit polyclonal antibody (DEX) directed against the full length human protein [21] and (iii) two rabbit polyclonal antibodies (CAT1 and CAT2) against a peptide from the human C-terminal cytoplasmic domain (CATEKNILVSDMEMNEQ-QE) conjugated to KLH [6]. After initial testing, only the CAT2 antibody was retained for further experiments. Flash-frozen skeletal muscle and lung tissues from animals of the three *MRC2* genotypes (see above) were disrupted and homogenized with a tissue lyser system II (Quiagen). Crude protein extracts were obtained and total protein concentrations determined using a colorimetric test (Pierce BCA Protein Assay kit, Thermo Scientific). Fifteen µg were diluted in 15 µl final volume (1× SDS gel-loading buffer) and loaded on a 5% stacking – 10% resolving Tris-glycine SDS-Polyacrylamide gel. Proteins were separated by electrophoresis at 120 V-250 mA during 3 hours, visualized by Coomassie blue staining, and electro-transferred overnight to Hybond P PVDF membranes (GE Healthcare). Membranes were blocked with 5% skim milk in PBS-Tween 20

(PBS-T) followed by incubation with primary CAT2 antibodies (1:200) in a total volume of 3 ml for 1 h30 min. After washing, the specific signal was detected by using Alkaline Phosphatase conjugated secondary rabbit antibodies (Sigma) following the instructions of the manufacturer.

Statistical analysis

Phenotypes corresponded to 22 type traits related to muscularity, skeletal conformation, size and leg soundness that are systematically recorded in the BBCB [24]. These were analyzed using a mixed model including genotype at the *MRC2* locus (2), year at scoring (2), body condition (4) as fixed effects, age at scoring as covariate (quadratic regression), the additive genetic animal effect and the residual effect as random effect [25]. The number of animals in the relationship matrix was 6,356. Variance components were estimated using the DFREML method (Derivative-Free Restricted Maximum Likelihood) [26]. The part of the genetic variance due to *MRC2* genotype was estimated as the difference between the variance due to the animal model with and without *MRC2* genotype in the model. The allele substitution effects (contrast) were calculated as the difference between the genotypic means (+/+ and +/M) obtained from the mixed model equations.

Evidencing a selective sweep

We simulated the segregation of a heterozygous mutation from Précieux to its 160 [2003–2005] sire offspring. Variable parameter values were (i) the transmission rate of the mutation from carriers to their offspring (0.5 to 0.75), (ii) the frequency of the mutation outside of the Précieux lineage. 10,000 simulations were conducted for each set of parameter values. Only non-affected genotypes were sampled from matings between heterozygous parents.

Supporting Information

Table S1 Statistics of number of carriers under the neutral model (no selection)(10,000 simulations).
Found at: doi:10.1371/journal.pgen.1000666.s001 (0.07 MB PDF)

Table S2 The table shows, for varying values of δ, the proportion of the phenotypic (P-PV) and genetic variance (P-GV) explained by the QTN in the general population. Assume a normally disturbed trait with 25% heritability, influenced by a QTN with MAF 1 of 0.25 and with two possible genotypes in the population (+/+ and +/M) as is the case for the CTS mutation. Assume that the average phenotype of the +/+ population is -δ/2 and of the +/M population is +δ/2. Assume also that the residual variance is 1. The table shows, for varying values of δ, the proportion of the phenotypic (P-PV) and genetic variance (P-GV) explained by the QTN in the general population. Assume that one selects future AI sires amongst offspring of popular +/M heterozygous sires. The table shows, for five hypothetical phenotypic threshold values for selection T = 1.00–2.00, the proportion of sons selected (Prop-Sel), and amongst the selected sons, the ratio of carrier (+/M) versus non-carriers (+/+) (C/NC). Dams were assumed to be +/+ for simplicity. It can be seen that the observed ~2:1 segregation ratio observed for CTS implies a selection intensity of the order of 0.02 for a QTN that accounts for ~0.05 of the genetic variance in the general population. The corresponding cells are highlighted in gray.
Found at: doi:10.1371/journal.pgen.1000666.s002 (0.11 MB PDF)

Table S3 Primer pairs for the *MRC2* gene.
Found at: doi:10.1371/journal.pgen.1000666.s003 (0.08 MB PDF)

Table S4 Allelic imbalance and quantitative RTPCR primer pairs for the detection of NMD.
Found at: doi:10.1371/journal.pgen.1000666.s004 (0.08 MB PDF)

Acknowledgments

We are grateful to Dimitri Pirottin and Mallory Draye for their assistance with the quantitative RT-PCR and Western blotting, respectively; to David Robertson for his assistance in characterizing the Endo180 antibodies; and to the Walloon Breeding Association (AWE) and the Belgian Blue Beef Herd-Book (HBBBB) for pedigree information and phenotypic data on the 22 type traits.

Author Contributions

Conceived and designed the experiments: MG CC. Performed the experiments: CF AS WL NT IJH CC. Analyzed the data: CF AS WL CM TD IJH CMI MG CC. Contributed reagents/materials/analysis tools: AS MD IJH CMI WC. Wrote the paper: MG CC. Mutation screening, NMD: CF. Case collection, necropsy, phenotypic description, pedigree analysis: AS. Mutation screening: WL. Case collection: MD. Diagnostic test: NT. Animal model: CM. Selective sweep: TD. Western blotting and immunohistochemistry: IJH, CMI. Custom iSelect illumina panel: WC. Designed experiments, analyzed data, wrote the manuscript, and supervised the project: CC, MG.

References

1. Grobet L, Martin LJ, Poncelet D, Pirottin D, Brouwers B, et al. (1997) A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. Nat Genet 17: 71–74.
2. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, et al. (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet 40: 449–454.
3. East L, Isacke CM (2002) The mannose receptor family. Biochim Biophys Acta 1572: 364–386.
4. Behrendt N (2004) The urokinase receptor (uPAR) and the uPAR-associated protein (uPARAP/Endo180): membrane proteins engaged in matrix turnover during tissue remodeling. Biol Chem 385: 103–136.
5. Chang YF, Imam JS, Wilkinson MF (2007) The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem 76: 51–74.
6. Sturge J, Todd SK, Kogianni G, McCarthy A, Isacke CM (2007) Mannose receptor regulation of macrophage cell migration. J Leukoc Biol 82: 585–593.
7. Wagenaar-Miller RA, Engelholm LH, Gavard J, Yamada SS, Gutkind JS, et al. (2007) Complementary roles of intracellular and pericellular collagen degradation pathways in vivo. Mol Cell Biol 27: 6309–6322.
8. Jansen KM, Pavlath GK (2006) Mannose receptor regulates myoblast motility and muscle growth. J Cell Biol 174: 403–413.
9. East L, McCarthy A, Wienke D, Sturge J, Ashworth A, et al. (2003) A targeted deletion in the endocytic receptor gene Endo180 results in a defect in collagen uptake. EMBO Rep 4: 710–716.
10. Engelholm LH, List K, Netzel-Arnett S, Cukierman E, Mitola DJ, et al. (2003) uPARAP/Endo180 is essential for cellular uptake of collagen and promotes fibroblast collagen adhesion. J Cell Biol 160: 1009–1015.
11. Curino AC, Engelholm LH, Yamada SS, Holmbeck K, Lund LR, et al. (2005) Intracellular collagen degradation mediated by uPARAP/Endo180 is a major pathway of extracellular matrix turnover during malignancy. J Cell Biol 169: 977–985.
12. Fujii J, Otsu K, Zorzato F, de Leon S, Khanna VK, et al. (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. Science 253: 448–451.
13. Rudolph JA, Spier SJ, Byrns G, Rojas CV, Bernoco D, et al. (1992) Periodic paralysis in quarter horses: a sodium channel mutation disseminated by selective breeding. Nat Genet 2: 144–147.
14. Beever JE, Smit MA, Meyers SN, Hadfield TS, Bottema C, et al. (2006) A single-base change in the tyrosine kinase II domain of ovine FGFR3 causes hereditary chondrodysplasia in sheep. Anim Genet 37: 66–71.
15. Smith LB, Dally MR, Sainz RD, Rodrigue KL, Oberbauer AM (2006) Enhanced skeletal growth of sheep heterozygous for an inactivated fibroblast growth factor receptor 3. J Anim Sci 84: 2942–2949.
16. Wright D, Boije H, Meadows JR, Bed'hom B, Gourichon D, et al. (2009) Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. PLoS Genet 5: e1000512. doi:10.1371/journal.pgen.1000512.
17. Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, et al. (2007) A single IGF1 allele is a major determinant of small size in dogs. Science 316: 112–115.

18. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NH, Zody MC, et al. (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet 39: 1321–1328.
19. Rosengren Pielberg G, Golovko A, Sundström E, Curik I, Lennartsson J, et al. (2008) A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. Nat Genet 40: 1004–1009.
20. 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.
21. Isacke CM, van der Geer P, Hunter T, Trowbridge IS (1990) p180, a novel recycling transmembrane glycoprotein with restricted cell type expression. Mol Cell Biol 10: 2606–2618.
22. Sheikh H, Yarwood H, Ashworth A, Isacke CM (2000) Endo180, an endocytic recycling glycoprotein related to the macrophage mannose receptor is expressed on fibroblasts, endothelial cells and macrophages and functions as a lectin receptor. J Cell Sci 113: 1021–1032.
23. Wienke D, MacFadyen JR, Isacke CM (2003) Identification and characterization of the endocytic transmembrane glycoprotein Endo180 as a novel collagen receptor. Mol Biol Cell 14: 3592–3604.
24. 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. In: Milk and beef recording : State of the art, 1994. Proceedings of the 29th biennial session of the International Committee for Animal Recording (ICAR); Ottawa, Canada; July 31–August 5, 1994. EAAP Publication No. 75. pp 231–237.
25. Lynch M, Walsh B (1997) Genetics and analysis of quantitative traits. Sunderland Massachusetts: Sinauer Associates, Inc.
26. Boldman KG, Kriese LA, Van Vleck LD, Kachman SD (1995) A manual for Use of MTDFREML. A set of Programs to obtain Estimates of Variances and Covariances. United States Department of Agriculture: Agricultural Research Service.

SUPPORTING MATERIAL

Supplemental Table 1: Statistics of number of carriers under the neutral model (no selection)(10,000 simulations).

f_{mut}	Min	25%	median	mean	75%	max	n>44
0.00	0	2	7	9.54	15	62	14
0.01	0	4	10	11.76	17	58	23
0.05	0	13	19	20.12	26	63	130
0.10	4	22	29	29.41	36	72	688

Supplemental table 2: Assume a normally disturbed trait with 25% heritability, influenced by a QTN with MAF of 0.25 and with two possible genotypes in the population (+/+ and +/M) as is the case for the CTS mutation. Assume that the average phenotype of the +/+ population is -δ/2 and of the +/M population is +δ/2. Assume also that the residual variance is 1. The following table shows, for varying values of δ, the proportion of the phenotypic (P-PV) and genetic variance (P-GV) explained by the QTN in the general population. Assume that one selects future AI sires amongst offspring of popular +/M heterozygous sires. The following table shows, for five hypothetical phenotypic threshold values for selection (T=1,00-2,00), the proportion of sons selected (Prop-Sel), and amongst the selected sons, the ratio of carrier (+/M) versus non-carriers (+/+) (C/NC). Dams were assumed to be +/+ for simplicity. It can be seen that the observed ~2:1 segregation ratio observed for CTS implies a selection intensity of the order of 0,02 for a QTN that accounts for ~0,05 of the genetic variance in the general population. The corresponding cells are highlighted in gray.

QTN	Population		1/2-sib pedigrees - carrier sire									
			T = 1,00		T = 1,25		T = 1,50		T = 1,75		T = 2,00	
δ	Prop-PV	Prop-GV	Prop-Sel	C/NC	Prop-Sel	C/NC	Prop-Sel	C/NC	Prop-Sel	C/NC	Prop-Sel	C/NC
0,00	0,00	0,00	0,15	1,00	0,10	1,00	0,06	1,00	0,04	1,00	0,02	1,00
0,05	0,00	0,00	0,15	1,08	0,10	1,10	0,06	1,11	0,04	1,12	0,02	1,13
0,10	0,00	0,01	0,15	1,17	0,10	1,20	0,06	1,23	0,04	1,26	0,02	1,29
0,15	0,00	0,02	0,15	1,27	0,10	1,31	0,06	1,36	0,04	1,41	0,02	1,46
0,20	0,01	0,03	0,15	1,38	0,10	1,44	0,06	1,51	0,04	1,58	0,02	1,65
0,25	0,01	0,05	0,15	1,49	0,10	1,58	0,06	1,67	0,04	1,77	0,02	1,87
0,30	0,02	0,07	0,15	1,62	0,10	1,73	0,06	1,85	0,04	1,98	0,02	2,12
0,35	0,02	0,09	0,15	1,75	0,10	1,89	0,06	2,05	0,04	2,22	0,02	2,41
0,40	0,03	0,12	0,16	1,90	0,10	2,07	0,06	2,27	0,04	2,49	0,02	2,73
0,45	0,04	0,15	0,16	2,06	0,10	2,27	0,07	2,51	0,04	2,79	0,02	3,10
0,50	0,04	0,18	0,16	2,23	0,11	2,49	0,07	2,78	0,04	3,12	0,02	3,51
0,55	0,05	0,21	0,16	2,42	0,11	2,72	0,07	3,08	0,04	3,50	0,02	3,98
0,60	0,06	0,25	0,16	2,62	0,11	2,98	0,07	3,41	0,04	3,92	0,02	4,51
0,65	0,07	0,29	0,16	2,84	0,11	3,27	0,07	3,78	0,04	4,39	0,02	5,12
0,70	0,08	0,34	0,17	3,08	0,11	3,58	0,07	4,19	0,04	4,92	0,03	5,81
0,75	0,10	0,38	0,17	3,33	0,11	3,92	0,07	4,64	0,05	5,52	0,03	6,58
0,80	0,11	0,43	0,17	3,61	0,12	4,30	0,08	5,15	0,05	6,19	0,03	7,47
0,85	0,12	0,48	0,17	3,92	0,12	4,71	0,08	5,70	0,05	6,93	0,03	8,47
0,90	0,13	0,53	0,18	4,25	0,12	5,16	0,08	6,32	0,05	7,77	0,03	9,60
0,95	0,14	0,58	0,18	4,60	0,12	5,66	0,08	7,00	0,05	8,71	0,03	10,89
1,00	0,16	0,63	0,18	4,99	0,13	6,20	0,09	7,76	0,05	9,77	0,03	12,35

Supplemental table 3: Primer pairs for the *MRC2* gene

Name	Primer sequence (5'-3')	Gene part	Size (bp)
gUP1	CCGGAGGAAGACGCGAGCCCCT	exon 1 (ATG)	327
gDN1	GGGGGAAAGGAGGAAAAAGTCG		
gUP2	CACAGCCCACTACCAGCGTCAG	exon 2	610
gDN2	CCATGACGATGAAAGAGCTGAC		
gUP3	ACCCTGTGAGAAGCCTTTCCTG	exons 3, 4	907
gDN3	GATGTAGGTCTGCTCGTGGATC		
gUP4	TGTCATGGTGGCAGGTAACGAC	exons 4, 5	576
gDN4	GGGTGGAAATCTGCTGGTCTAG		
gUP5	GGAGGAGGCAAGAGAGCCGAAG	exon 6	366
gDN5	CCTTGTGCTGTGAGGGTGGGTG		
gUP6	AAAGCGTGGTCCCTGTCCCAGC	exon 7	813
gDN6	AACGGTAGCACTCCTTTGGTGGT		
gUP7	GGCTTGGTGGGAAGAGTGGATCT	exons 8, 9	725
gDN7	GGGGAGGAGGGATTCCGAGAGG		
gUP8	TTGGAGGCATCTGCACAGCTAC	exon 10	407
gDN8	TACCACAGGAGGCTGCGGATTC		
gUP9	CCTGTGCTCAAGCCTGCAGAAA	exon 11	375
gDN9	GCCCTGGAGATAGTTGAAGCTCA		
gUP10	GGTCCCCACTTCCCTGAGCAAG	exon 12	375
gDN10	ATGAAGCCCTAGGTCTCGGTCAT		
gUP11	CAACCCCACAGCACATGTCCCT	exon 13	449
gDN11	CTGCTCGGATCATGGCTGGGTC		
gUP12	CCTCGACACCCTGTCCACTGAA	exon 14	398
gDN12	CCTGGCACTAGCAGCAGACACA		
gUP13	CCCGAGTCACAACTGCAAGGATT	exons 15, 16, 17	751
gDN13	GGGTGTGGGATGGACAGGAAGC		
gUP14	CTGCAGCGTGTCTGTCCCTGTT	exon 17, 18	697
gDN14	CGAGTCCCTGCTAGCCATCCAC		
gUP15	GGTCTAACCTGGTGCCCTGTACT	exons 19, 20 (*)	693
gDN15	AGGGGAGAGGGTGGTAGGTTCAG		
gUP16	GCCAGGTCGGGAGGGTATCAGAG	exons 21, 22, 23	1042
gDN16	CTGCAACCCCTGGATGCTCACT		
gUP17	GAGTTGGTCTCTGCCTGCTGTTC	exons 24, 25, 26	995
gDN17	CAGAGTGCAGCACGGGGACTATA		
gUP18	CGTCTCCATGCCATCCTCTATTC	exons 27, 28	743
gDN18	CCCAGGCCTCCCATCCACTGTG		
gUP19	ACTGTATTGTTACTACCACTGTTGTT	exons 29, 30 (STOP)	610
gDN19	AAGGAAACGCCATGCTGCACTC		

Supplemental table 4: Allelic imbalance and quantitative RT-PCR primer pairs for the detection of NMD.

Gene	Forward primer	Reverse primer	Size (bp)
MRC2 del	GACAAGAAGTGCGTGTACATGATG	AGAACTGTGCCTCTGACCACTTC	232
MRC2 5' part	CGAGTCTCTCCAGCCTGCAATG	ACTCAGTGCCTCGCGGTCACAC	168
MRC2 3' part	TCCTGCTCCTGGCTCTGCTGAC	CTGCTGCTCGTTCATTTCCATG	167
ACTB	TCGCGGACAGGATGCAGAAAGA	GCTGATCCACATCTGCTGGAA	149
GAPDH	TGACCCCTTCATTGACCTTCA	GATGGTGATGGCCTTTCCATT	127
HPRT1	TGCTGAGGATTTGGAGAAGG	CAACAGGTCGGCAAAGAACT	154
RPLP0	TGGGCAAGAACACGATGATG	TGAGGTCCTCCTTGGTGAACA	123
RPS18	GCAGAATCCACGCCAATACAA	TCTTCAGGCGCTCCAGGTCTTC	135
SDHA	GCAGAACCTGATGCTTTGTG	CGTAGGAGAGCGTGTGCTT	185
YWHAZ	GCATCCCACAGACTATTTCC	GCAAAGACAATGACAGACCA	120

Genetic variants in *REC8*, *RNF212*, and *PRDM9* influence male recombination in cattle

Cynthia Sandor, Wanbo Li*, Wouter Coppieters, Tom Druet, Carole Charlier, Michel Georges.*

* Contributed equally to this work

PLoS Genetics, 2012, Issue 7, e1002854.

.

Background

In mammalian sexual reproduction, chromosomes replicate once, then undergo twice separation (meiosis I and II), leading each gamete to obtain a half set of chromosomes compared to somatic cell. To achieve this, accurate alignment, pairing and segregation of chromosomes in meiosis are required and highly regulated (Handel and Schimenti, 2010). Recombination (or crossover) establishes physical connection between homologous chromosomes and directs their precise segregation in meiosis I. Crossovers (CO) occur at least once in a pair of homologous chromosomes (obligate crossover), and are well separated when more than one CO occurs in the paired chromosomes (interference). If less recombination happens in meiosis, it may lead to non-disjunction of chromosomes and aneuploidy. Fertilized aneuploid gametes mainly experience early fetal death or mental retardation after birth (Hassold and Hunt, 2001). Accumulated evidence shows that abnormal chromosome disjunction and aneuploidy are prone to occur during female meiosis I. Furthermore, about 5% of clinical observed pregnancy loss is due to aneuploidy in humans and aneuploidy incidence increases with maternal age (Hassold et al., 2007; Hassold and Hunt, 2001; Penrose, 2009). In addition, recombination is the main force that breaks down linkage disequilibrium, and consequently is a cornerstone of genetic research.

Recombination mechanisms have been well defined in yeast and nematode (Handel and Schimenti, 2010; Martinez-Perez and Colaiacovo, 2009), and share most of the fundamental principles with mammals, but also differ significantly for some characteristics (Lichten and de Massy, 2011). Dissecting the unique mechanism of mammalian recombination is essential to better understanding the clinic reproductive problems suffered by mammals.

Farm animals, like cattle, usually have extremely large family size and well-recorded pedigrees, which are advantageous for dissecting the genetic basis of complex traits. We attempted to characterize the recombination landscape genome-widely in cattle, taking advantage of a large three-generation half-sib dairy cattle pedigrees genotyped with a 50K

high density SNP chip, generated for genomic selection. Our strategy to identify recombination components is to positionally clone the genes and variants that underlie inherited variation in recombination phenotypes. To define recombination events, marker phasing was conducted with the Phasebook software package which exploits Mendelian rules and linkage information simultaneously (Druet and Georges, 2010). Crossover events were then identified as phase switches in the gametes transmitted by the GII sires to their GIII sons. We set out to map quantitative trait loci (QTL) influencing genome-wide recombination rate, genome-wide hotspot usage, locus-specific recombination rate, and genome-wide crossover interference.

OPEN ACCESS Freely available online

PLoS GENETICS

Genetic Variants in *REC8*, *RNF212*, and *PRDM9* Influence Male Recombination in Cattle

Cynthia Sandor[☯¤a¤b], Wanbo Li[☯], Wouter Coppieters, Tom Druet, Carole Charlier, Michel Georges*

Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, Liège, Belgium

Abstract

We use >250,000 cross-over events identified in >10,000 bovine sperm cells to perform an extensive characterization of meiotic recombination in male cattle. We map Quantitative Trait Loci (QTL) influencing genome-wide recombination rate, genome-wide hotspot usage, and locus-specific recombination rate. We fine-map three QTL and present strong evidence that genetic variants in *REC8* and *RNF212* influence genome-wide recombination rate, while genetic variants in *PRDM9* influence genome-wide hotspot usage.

Citation: Sandor C, Li W, Coppieters W, Druet T, Charlier C, et al. (2012) Genetic Variants in *REC8*, *RNF212*, and *PRDM9* Influence Male Recombination in Cattle. PLoS Genet 8(7): e1002854. doi:10.1371/journal.pgen.1002854

Editor: Kenneth Paigen, The Jackson Laboratory, United States of America

Received November 30, 2011; **Accepted** June 7, 2012; **Published** July 26, 2012

Funding: CRV (http://www.crv4all.com/) and LIC (http://www.lic.co.nz) provided the SNP genotype and pedigree information. CS benefitted from financial support of the Fonds National de la Recherche Scientifique (FNRS; FRIA fellowship), of the Communauté Française de Belgique (BIOMOD Action de Recherche Concertée), and is supported through a Marie Curie Fellowship (IOF program) from the European Union. CC and TD are respectively Maître de Recherche and Chercheur Qualifié FNRS. This work has been supported by the Belgian Science Policy Organisation (SSTC Genefunc PAI) and by the Walloon Ministry of Agriculture. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: michel.georges@ulg.ac.be

¤a Current address: Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America
¤b Current address: Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, United States of America

☯ These authors contributed equally to this work.

Introduction

Reciprocal recombination between homologues fulfills an essential mechanistic role during meiosis in most organisms [1,2]. It is required for proper bivalent alignment on the metaphase I plate preceding disjunction and segregation at anaphase I. Correct segregation of the full chromosome complement demands tight, sex-specific control of the number of cross-overs (CO) per arm, as well as of their position relative to chromosomal landmarks (centromeres and telomeres) and other CO (in the case of multichiasmatic meioses) [3,4]. Failures in this process underlie aneuploidies affecting as many as 5% of human oocytes [5].

At the population level, recombination affects the rate of creation and loss of haplotypes with *cis*-configured favorable alleles, placing second order selection pressure on modifiers of global and/or local recombination including inversions [3].

Components of the recombination apparatus are well described in yeast and *C. elegans*, but remain largely undefined in most other organisms including mammals [4,6]. One strategy to identify such components is to positionally clone the genes and variants that underlie inherited variation in recombination phenotypes. Genome-wide recombination rate (GRR) is characterized by considerable inter-individual variation, which is in part inherited [7–10]. Genome-wide association studies (GWAS) have identified several loci influencing GRR in human [11–13]. These include the 17q21.31 inversion [11], as well as the *RNF212* gene harboring common variants with antagonistic effects on GRR in males and females [12]. Of note, women's recombination rate correlates positively with reproductive success[9]. In human, ~80% of CO events map to ~10–20% of the genome, encompassing >25,000 recombination hotspots [3,14–16]. Hotspot usage differs considerably between individuals [17] and this was shown to involve variation in *cis*-acting hotspot-triggering sequences [18], as well as in the *trans*-acting *PRDM9* H3K4 trimethyltransferase and hotspot regulator [19–22]. Recombination hotspots and their *PRDM9* regulator undergo accelerated evolution (explained in part by the self-destructive drive of hotspot motifs due to biased gene conversion) [18,21,23,24], and *PRDM9* has been identified as a hybrid sterility gene in the mouse [25]. Genome-wide levels of cross-over interference were also suggested to differ between individuals [26,27], but corresponding genetic variants – if existing - have not been identified thus far.

We herein describe our efforts to take advantage of (i) the large multigenerational half-sib pedigrees typifying dairy cattle population and (ii) the systematization of genome-wide SNP genotyping with ~50 K medium density arrays for "genomic selection" purposes [28], to quantify inter-individual variation in recombination phenotypes as well as to map contributing genetic loci. The bovine haploid genome is estimated at 2.87 Gbp distributed over 29 acrocentric chromosomes and a pair of metacentric sex chromosomes [29]. Total map length was previously estimated at ~31M and shown (contrary to most other mammals) not to differ between sexes [30]. The potential correlation between recombination rate and fertility, as well as the hypothesized effect of domestication on recombination rates [31] adds to the interest of a detailed characterization of recombination phenotypes in livestock.

Author Summary

Homologous recombination is an essential cellular process that determines proper chromosome segregation during meiosis, affects fertility, and influences evolvability. Nevertheless, the components of the recombination apparatus remain incompletely characterized in mammals. One approach to identify such components is to identify the genes that underlie inherited variation in recombination phenotypes. In addition to providing mechanistic insights, this would allow the study of the evolutionary forces that shape the recombination process. In this paper, we take advantage of genotypes for 50,000 genome-wide SNP markers to measure four recombination phenotypes (genome-wide recombination rate, genome-wide hotspot usage, locus-specific recombination rate, genome-wide cross-over interference) for >750 bulls on the basis of >250,000 cross-overs detected in sperm cells transmitted to >10,000 sons. We quantify the heritability and scan the genome for Quantitative Trait Loci (QTL) influencing each one of these recombination phenotypes. We perform a detailed genetic analysis of three such QTL, thereby providing evidence that genetic variants in *REC8* and *RNF212* influence genome-wide recombination rate, while genetic variants in an X-linked *PRDM9* paralogue influence genome-wide hotspot usage.

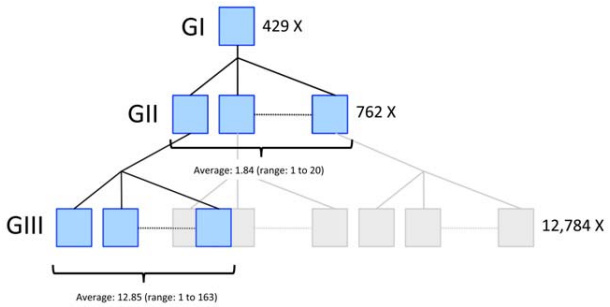


Figure 1. Three-generational pedigrees used to map genetic determinants of variation in male recombination rate in cattle. 10,192 (Dutch population) and 3,783 (NZ population) bulls, genotyped for 50K SNP panels, assorted in 429 three-generational pedigrees of the kind illustrated. All Dutch bulls were from the Holstein-Friesian breed, while in NZ 61% of the bulls were Holstein-Friesian and 39% Jerseys. Each pedigree is composed of one grand-sire with 1.84 GII sons on average (range: 1 to 20). Each GII sire has 12.85 GIII sons on average (range: 1 to 163). We have used the available SNP genotypes to identify 259,752 CO events that occurred in the sperm cells transmitted by the GII sires to their 10,106 GIII sons. QTL affecting variation in recombination rates were mapped by exploiting linkage information (effect of the homologues transmitted by the GI grand-sires to their GII sons) and LD information (effect of haplotypes transmitted by the GI-grand-sires and ungenotyped GI-grand-dams to their GII sons).
doi:10.1371/journal.pgen.1002854.g001

Results

Characterizing recombination in male cattle

The dataset available for analysis comprised 10,192 bulls from the Netherlands (H) and 3,783 bulls from New-Zealand (NZ), that were genotyped for marker panels comprising respectively 50,876 [32] and 51,456 [33] SNPs of which 19,487 in common. The 13,975 bulls assorted in 429 three-generational paternal half-sib pedigrees of the structure shown in Figure 1. All Dutch bulls were from the Holstein-Friesian (HF) breed, while in NZ 61% of the bulls were HF and 39% Jerseys (J). SNP genotypes were phased [34], and CO events identified in the gametes transmitted by generation II (GII) bulls to their GIII sons. We identified 259,752 CO in 10,106 gametes, corresponding to an average genome size of 25.7 M(organs).

Average number of CO for each of the 29 acrocentric chromosomes was remarkably well predicted ($r^2 = 0.96$) by (i) size in bp ($\beta_1 = 0.07$CO/10 Mb) and (ii) the requirement for at least one chiasma per meiosis ($\beta_0 = 0.48$ CO) (Figure S1A). Also in agreement with the obligate chiasma theory, the frequency distribution of gametes with 0, 1, 2, … CO-events was best explained [35] assuming near absence of nullichiasmatic meioses for all autosomes. Moreover, under a truncated Poisson model forcing the proportion of nullichiasmatic meioses at zero [36], the most likely frequency of meioses with one chiasma was considerably lower than expected, and this was largely due to an excess of meioses with two chiasmata. This supports the preferred occurrence of a second chiasma, particularly for the larger chromosomes (Figure S1B).

Recombination rate (RR) computed in 60-Kb windows averaged 0.00062 (i.e. ~1 cM/1 Mb), but was strongly overdispersed with an excess of "hot" and "cold" windows (defined as windows with RR>2.5 standard deviations from the mean) (Figure S2A–S2C). Note that hot windows as defined here (60 Kb) cannot be compared with recombination hotspots as defined in human and mouse genetics (≤5 Kb) [14,15,37]. On average, 34% of CO events could be assigned to hot windows representing 13% of the genome. Hot and cold windows differed in base pair composition and repeat content (Table S1). Hot windows tended to concentrate in sub-terminal (proximal chromosome end) and terminal regions (distal chromosome end), while cold windows concentrated in the middle of the chromosome arms as well as in terminal regions (proximal chromosome end) coinciding with the centromeres (Figure S2D). Hot and cold windows tended to cluster in what we refer to (following Chowdhury et al. [13]) as recombination "jungles" and "deserts", respectively.

We measured chromosome-specific levels of cross-over or chiasma interference using the shape parameter (ν) of a gamma distribution [26,38]. We used a maximum likelihood approach extracting information from the frequency distribution of (i) the number of CO per gamete, (ii) CO-position (in centimorgan (cM)) for gametes with one CO, (iii) inter-CO distance (in cM) for gametes with two CO, and (iii) inter-CO distance (in cM) for gametes with three CO. Positive interference was evident for all chromosomes, manifesting itself by (i) a paucity of gametes with zero CO, (ii) less uniform than expected distribution of single CO position, and (iii) inflated distance between CO for gametes with multiple CO. The value of ν that maximized the overall likelihood averaged 2.6 (range: 1.5–3.1) across all chromosomes (versus 4.5 in human [26]). It was primarily determined by the inter-CO distance for gametes with two recombination events. Values of ν maximizing the likelihood of the frequency distribution of number of CO events and of CO-position for gametes with one recombination tended to be larger that the value of ν maximizing the likelihood of the inter-CO distance for gametes with two recombinations, while values of ν maximizing the likelihood of the distance between CO for gametes with three recombinations tended to be smaller. Of note, the observed distribution of CO events per gamete and hence of chiasmata per meiosis, was remarkably well accounted for by positive interference. There was no evidence for an effect of chromosome length on ν, whether maximizing the overall likelihood or that of the constituent parameters (Figure S3A–S3B).

Genetic analysis of genome-wide recombination rate (GRR)

Average genome-wide recombination rate (GRR) (corrected for family size - M&M) differed significantly between GII sires ($p<0.0001$; range: 18.7–32.1)(Figure S4A). We took advantage of the fact that 72 of the GII bulls had non-overlapping sets of GIII sons in H and NZ, to estimate the repeatability of GRR as the correlation between these independent measurements, yielding a highly significant Spearman's correlation coefficient of 0.58 ($p<3.7\times10^{-7}$)(Figure S4B). We estimated the heritability (h^2) of GRR at 0.22 in the Dutch HF breed.

We used a Hidden Markov Model-based approach that simultaneously exploits linkage and linkage disequilibrium [34] to scan the genome for QTL influencing GRR. At each SNP position, all chromosomes in the dataset (i.e. 2*n* chromosomes for a data set with *n* animals) were assigned to one of 20 hidden states corresponding to "ancestral haplotype states". The effect of these hidden haplotype states (HHS) on the GRR was then estimated using a mixed model including a polygenic effect to correct for population stratification (M&M). We only used HF animals (from both H and NZ) in these analyses. We identified two genome-wide significant QTL, respectively on BTA10 ($z=5.8$) and BTA19 ($z=4.9$)(Figure 2A).

The lod-2 drop-off confidence interval (CI) of the BTA10 QTL spanned ~1.4 Mb encompassing 47 genes. Three of these are strongly expressed in testis: *TBC1D21*, *TSSK4*, and *REC8*. *REC8* is a particularly appealing positional candidate as it codes for a member of the kleisin family of SMC (structural maintenance of chromosome) proteins, which localizes to the axial elements of chromosomes during meiosis in both oocytes and spermatocytes. The mouse homologue is a key component of the meiotic cohesion complex, which regulates sister chromatid cohesion and recombination between homologous chromosomes [39,40]. We therefore re-sequenced 7.2 Kb encompassing the *REC8* gene (including 1.2 Kb upstream of the start codon and 0.9 Kb downstream of the polyadenylation site; Figure S5A and Table S2) for animals selected to obtain the sequence of three HHS associated with an increase in GRR and one associated with a decrease in GRR (as HHS with divergent effect on GRR should differ at the causative variant positions)(Figure 2B). We identified five SNPs located respectively in the 5′UTR (ss428897146 and ss418642851), intron 5 (ss418642852), exon 10 (ss418642853 = E287K) and intron 12 (ss418642854). Of note, two of these (ss418642852 and ss418642854) segregated perfectly between the high and low GRR haplotypes. We developed 5′exonuclease assays for ss418642851, ss418642852, ss418642853 and ss418642854 (Table S3), and genotyped the GI and GII sires. We performed single point association analysis using a mixed model including the (random) effect of the SNPs as well as a polygenic animal effect to correct for stratification. Ss418642854 yielded a lod score of 9.12, i.e. 3.7 units higher than any other BTA10 SNP, and 3.3 units higher than the highest BTA10 haplotype-based signal. The difference in GRR between alternate homozygotes at the ss418642854 SNP was 1.8 CO/genome (Figure 2B). To provide additional support for the causality of the *REC8* gene, we took advantage of the fact that 121 HF GII sires had also been genotyped with a recently developed high-density Illumina 777K SNP array, including 45 SNPs spanning the QTL CI. When performing single point association analysis using the same mixed model for all SNPs in the CI, the lod score still clearly maximized on top of the *REC8* gene and for SNP ss418642854 (Figure S4D). Taken together, these results support the fact that variation in the *REC8* gene indeed underlies the identified QTL.

The CI of the BTA19 QTL spans ~0.6 Mb encompassing two genes: *KCNJ2* and *KCNJ16*. Neither is knowingly related to recombination, yet both are expressed in testes (data not shown). Preliminary sequence analysis of the *KCNJ2* and *KCNJ16* open reading frames (ORF) of animals carrying haplotypes with significantly different effect on GRR did not reveal obvious variants that might underlie the observed effects (data not shown).

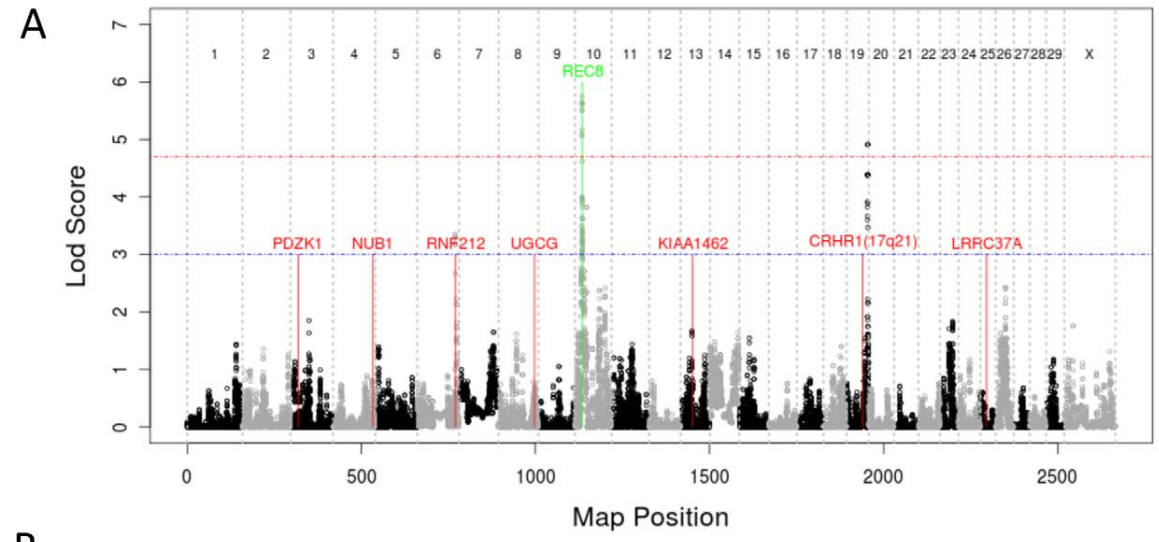
In addition to these two significant QTL, we obtained a suggestive lod score of 3.2 on BTA6 that maximized at the exact position of the *RNF212* gene (Figure 2A). This suggests that variation in *RNF212* affects GRR in cattle as it does in human [12,13]. Homologues of *RNF212* in *C. elegans* (*ZHP3*) and yeast (*ZIP3*) are known to be involved in meiotic recombination [12]. We re-sequenced 10 amplicons encompassing the entire *RNF212* ORF and intron-exon boundaries (Figure S5B and Table S2) in animals selected to obtain the sequence of two haplotypes increasing and two decreasing GRR (Figure 2C). We identified eight SNPs located respectively in the 5′UTR (ss418642855, ss418642856 and ss418642857), intron 1 (ss418642858), exon 3 (ss418642859), intron 4 (ss418642860), intron 9 (ss418642861) and exon 12 (ss469104611 = P259S). Five of these (ss418642855, ss418642858, ss418642860, ss418642861 and P259S) segregated perfectly between the high and low GRR haplotypes. We developed five 5′exonuclease assays (Table S3) and genotyped the GI and GII sires. Ss469104611 (= P259S) yielded a lod score of 18, i.e. 15.3 units higher than any other BTA6 SNP and 14.8 units higher than the highest BTA6 haplotype-based signal. The difference in GRR between alternate homozygotes at the P259S variant was 3.3 CO/genome (Figure 2C). We took advantage of the same 121 GII sires genotyped with the high-density 777K Illumina array, including 27 SNPs in the ~1 Mb CI of the BTA6 QTL. Lod scores clearly maximized on top of the *RNF212* gene, at the position of the ss469104611 variant (Figure S4E). Taken together, these results strongly supported the causality of the *RNF212* gene.

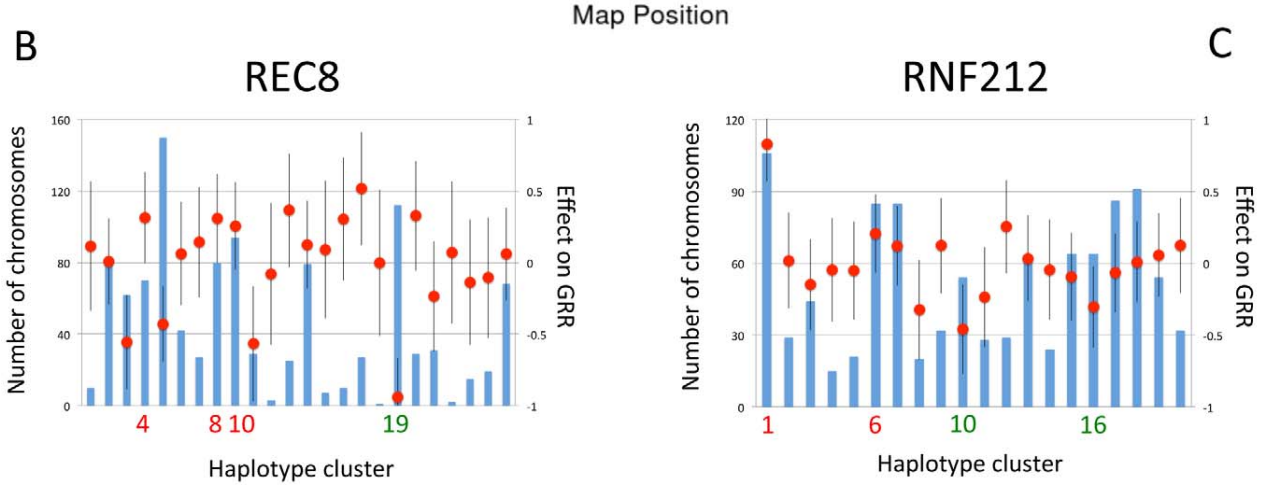
Genetic analysis of genome-wide hot window usage (GHU)

We then computed, for each GII bull, the proportion of CO falling in hot windows (i.e. the genome-wide hot-window usage or GHU). GHU differed significantly between GII sires ($p<0.002$; range: 4%–58%), was repeatable (Spearman's correlation: 0.46; $p<0.0008$) and had a heritability of 0.21 in Dutch HF (Figure S6).

We scanned the genome for QTL affecting GHU in HF, and identified three suggestive QTL, respectively on BTA3 ($z=3.7$), BTA25 ($z=4.1$) and BTAX ($z=2.8$)(Figure 3A). The CI of the BTA3 QTL spans ~2.1 Mb and encompasses three genes (*LOC781798*, *LOC522984* and *OLFM3*) not obviously related to recombination. The CI for the BTA25 QTL (UMD3 31.29–33.62 Mb) contains 25 genes of unknown function.

Most interestingly, the lod score peak on the X chromosome coincided with the position of two adjacent gonosomal *PRDM9* paralogues (hereafter referred to as *PRDM9-XA* and *–XB*). In mice and human, genome-wide hotspot usage has been shown to be genetically controlled, with variation in the *PRDM9* C-terminal tandem array of Cys_2His_2 zinc-finger (ZF) domains having a major effect [19–22]. We therefore designed amplicons allowing specific amplification and sequencing of the complete *PRDM9-XA* and *-XB* ZF arrays. The C-terminal ZF arrays of the *PRDM9-XA* and *PRDM9-XB* reference sequences (UMD3 build) contain respectively eight and 20 ZF domains in tandem (Table S2). Sequence analyses indicate that bovine *PRDM9* ZF arrays are rapidly evolving (as they are in human and rodents but not in dogs [41–43]), and this is predicted to increase allelic heterogeneity. We thus





Haplotype	Effect on GRR	ss428897146 5'UTR	ss418642851 5'UTR	ss418642852 Intron 5	ss418642853 Exon 10 E287K	ss418642854 Intron 12
4	+0.63	C	C	C	A	A
8	+0.62	T	G	C	G	A
10	+0.56	C	C	C	A	A
19	-0.67	T	G	G	G	G

Haplotype	Effect on GRR	ss418642855 5'UTR	ss418642856 5'UTR	ss418642857 5'UTR	ss418642858 Intron 1	ss418642859 Exon 3 A64A	ss418642860 Intron 4	ss418642861 Intron 9	ss469104611 Exon 12 P259S
1	+0.83	G	A/G	C/G	C	G	T	A	T
6	+0.21	G	A/G	C	C	G	T	A	T
10	-0.46	:	G	G	A	G/A	C	G	C
16	-0.30	:	G	G	A	G/A	C	G	C

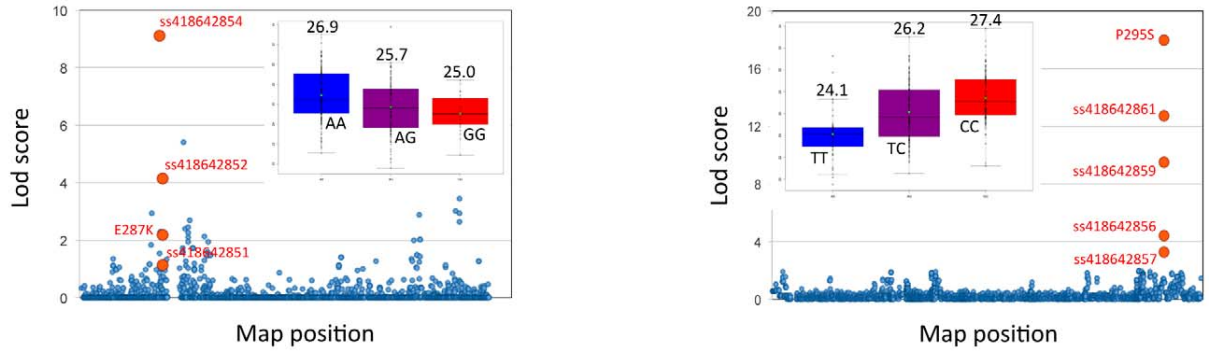


Figure 2. Genome-wide lod score profiles, observed counts and effects on global recombination rate. (A) Genome-wide lod score profiles obtained for GRR in the Holstein-Friesian (H+NZ) sample set. The red and blue horizontal lines mark the genome-wide significant and suggestive thresholds determined by permutation testing. The position of seven loci that have been previously implicated as determinants of variation in GRR [12,13] are shown in red, while the position of *REC8* is shown in green. (B) Observed counts (blue bars), and effects on global recombination rate (GRR) (red circles) ± standard error (black vertical lines) for 25 hidden haplotype clusters at map position of *REC8* gene. Effect on GRR and genotype at five DNA sequence variants in *REC8* for four sequenced hidden haplotype clusters. Combined linkage+LD analysis of BTA10 SNPs on GRR. Sequence variants in *REC8* are shown in red. The inset shows the distribution of GRR for GII sires sorted by ss418642854 genotype, with indication of the average GRR (in M) per genotype. (C) Observed counts (blue bars), and effects on global recombination rate (GRR) (red circles) ± standard error (black vertical line) for 20 hidden haplotype clusters at map position of *RNF212* gene. Effect on GRR and genotype at five variant positions in *RNF212* for four sequenced hidden haplotype clusters. Combined linkage+LD analysis of BTA6 SNPs on GRR. Sequence variants in *RNF212* are shown in red. The inset shows the distribution of GRR for GII sires sorted by ss469104611 (=P259S) genotype, with indication of the average GRR (in M) per genotype.
doi:10.1371/journal.pgen.1002854.g002

decided to determine the sequence of the *PRDM9-XA* and *–XB* ZF arrays for 80 individuals representing all 20 hidden haplotype states. Not a single polymorphism, whether synonymous or not, was observed for the *PRDM9-XA* array. For *PRDM9-XB*, however, we detected (i) a VNTR-like length polymorphism (as we detected a common allele with 22 ZF), and (ii) nine SNPs (Figure 3B). Notably, eight of the nine SNPs were non-synonymous. Two affected residues that are predicted to mediate DNA binding, located respectively in ZF 11 out of 22 (11/20) (ss5 = I23K, position +6) and 16/22 (ss7 = L17T; position -1). Four corresponded to R<->Q amino-acid substitutions at position 13 of ZF 2/22 (ss1), 9/22 (ss2), 11/22 (ss5) and 18/22 (ss8). Two corresponded to A<->Y amino-acid substitutions at position 7 of ZF domains 10/22 (ss3) and 14/22 (ss6). Based on these results, we decided to sequence the *PRDM9-XB* ZF array for all GI and GII sires. The 10 polymorphisms assorted in eight haplotypes observed at least five times, jointly accounting for 98.6% of the sequenced chromosomes (Figure 3B). We tested the effect of *PRDM9-XB* haplotype on GHU using the mixed model described above, and obtained a lod score of 7.3, i.e. 4.5 units higher than in the initial scan, hence strongly supporting the causality of the *PRDM9-XB* paralogue. Analysis of the effects of individual haplotypes indicates that: (i) ss5 has a major effect, the K allele decreasing GHU ~30-fold when compared to the I allele (hap1-hap3 contrast), (ii) ss1 has no effect on GHU (hap1-hap2 contrast), (iii) the VNTR affects GHU as the loss of two ZFs decreases GHU ~6-fold (hap2-hap6 contrast), (iv) ss2, ss3, ss4, ss7, ss8 and ss9 have no effect on GHU (hap6-(hap5,hap7,hap8) contrasts), (v) ss6 affects GHU, the Y allele increasing GHU ~6-fold when compared to the A allele (hap4-hap8 contrast)(Figure 3C). The major effect of the ss5 variant was also apparent from single-point analyses, yielding a lod score of 4.6 (Figure 3D).

Genetic analysis of locus-specific recombination rate (LRR)

Rapid *PRDM9* evolution presupposes accelerated turn-over and hence high polymorphism of recombination hotspots [41]. To test this hypothesis, we scanned the genome for *cis*-acting haplotype effects on LRR (in HF). We tested the effect of hidden haplotype state of the GII sires on the recombination rate in an 800-Kb window centered on the interrogated SNP position (M&M). We obtained one genome-wide significant effect on BTA6 (Figure 4A). The observed signal was primarily driven by two haplotype clusters (HS2 and HS9), increasing recombination ~4 to 5-fold (Figure 4A). The association signal maximized in the middle of a 840-Kb recombination jungle, for which the observed recombination rate exceeded expectation by up to ~8.5 SD (Figure 4B). LRR in the corresponding 800-Kb window was of the order of 8–9% for GII sires heterozygous for either the HS2 or HS9 haplotypes.

Eight additional peaks exceeded the genome-wide suggestive threshold (by definition, expected by chance only once per genome scan), supporting the common occurrence of *cis*-acting haplotype effects on local recombination rate, and presumably reflecting polymorphisms in *cis*-acting recombination-triggering motifs [18].

Genetic analysis of genome-wide interference (GIL)

We finally evaluated inter-individual variation in genome-wide interference levels (GIL). As interference levels were primarily determined by inter-CO distance for gametes with two CO (cfr. above), we used this metric for QTL mapping. Distances between CO were measured both in centimorgan (GIL_{cM}) and base-pairs (GIL_{bp}), and expressed in standardized deviations from the chromosome mean. Both measures proved to significantly differ between GII sires (GIL_{cM}: $p<0.002$; GIL_{bp}: $p<0.001$), to be repeatable (GIL_{cM}: $\rho=0.36$, $p<0.03$; GIL_{bp}: $\rho=0.53$, $p<0.00003$) but to have low heritability (GIL_{cM}: 0.045; GIL_{bp}: 0.052) in Dutch HF (Figure S7). We nevertheless scanned the genome for QTL affecting GIL in HF. We identified no QTL when using GIL_{cM}, yet one genome-wide suggestive QTL ($z=4.1$) on BTA25 when analyzing GIL_{bp}. The CI of the QTL encompassed four genes (*E-NPP7*, *LOC100297064*, *FOX1* and *TMEM114*) not knowingly involved in recombination (Figure 5; Figure S7).

Discussion

We herein estimate the male map length in domestic cattle at 25.7 Morgan based on the analysis of CO events in >10,000 sperm cells. This is ~5 M lower than previous estimates [30], but in better agreement with the relationship between number of chromosome arms and map length observed in other species [3]. Our findings suggest re-evaluation of (i) the presumed equal male and female recombination rate in cattle, and (ii) the inflation of recombination as a result of domestication.

We demonstrate that GRR is repeatable, that it differs between sires, and that ~21% of the observed variation is inherited in the HF breed. We identify two significant and one suggestive QTL influencing GRR. We provide evidence that two strong positional candidate genes, namely *REC8* and *RNF212*, are very likely causative. We reach this conclusion by targeting resequencing efforts to haplotype clusters with significantly different effect on GRR, leading to the identification of SNPs that exhibited highly significant increases in association signal. While variation in *RNF212* has been previously shown to affect GRR in human [12,13], the implication of *REC8* is novel. For *RNF212*, the variant yielding the strongest association is a missense variant resulting in a proline to serine substitution. Despite the fact that the corresponding protein segment is poorly conserved, P259S is a

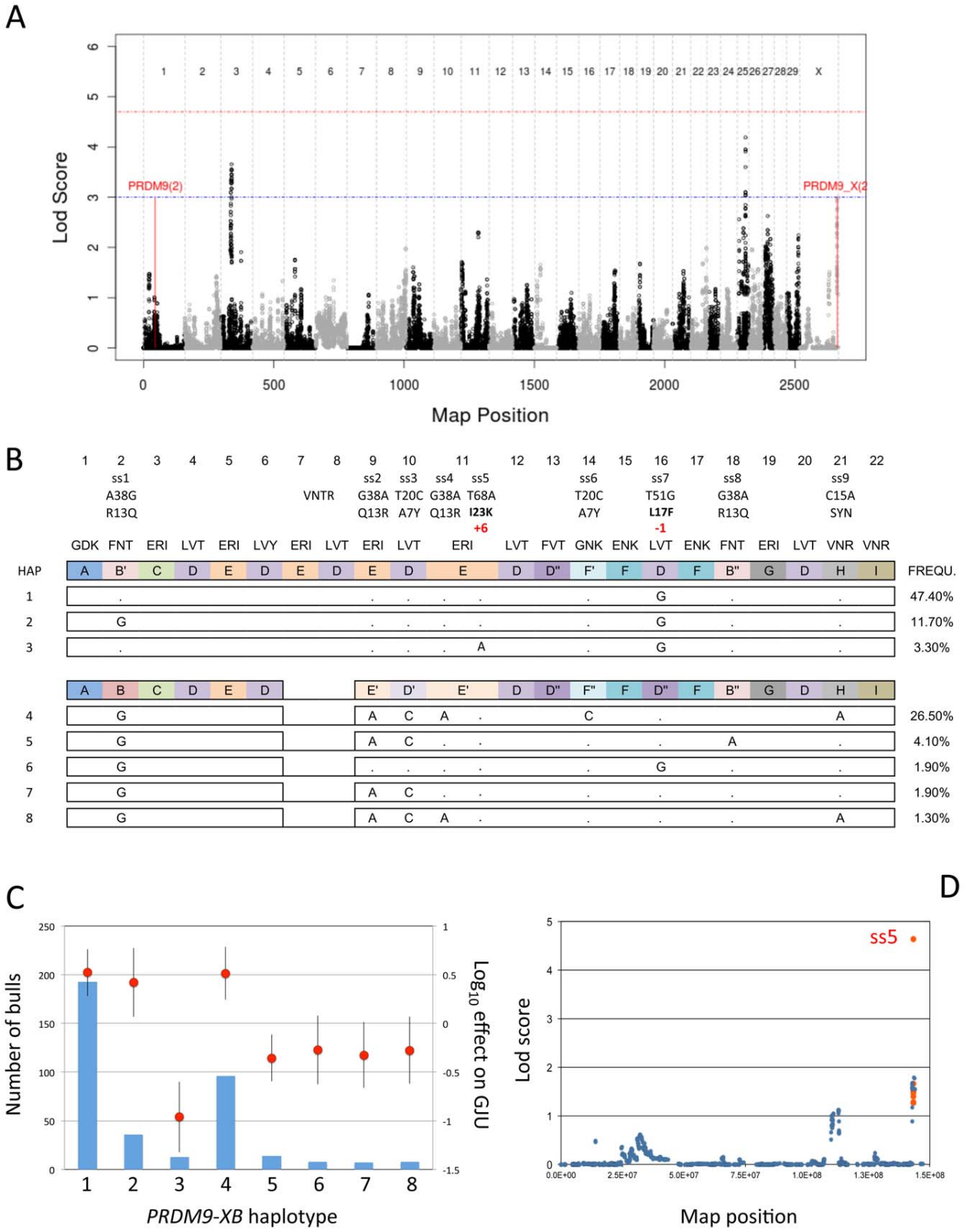


Figure 3. Results of the genome-scan, polymorphisms, observed counts and effects, results of single point analysis. (A) Results of genome-scan for QTL affecting genome-wide jungle usage (GJU) using a method that simultaneously extracts linkage and LD signal [34]. The red and blue horizontal lines mark the genome-wide significant and suggestive thresholds determined by permutation testing. The position of two pairs of PRDM9 parologues, respectively on BTA1 and BTAX, are highlighted. (B) Polymorphisms detected in the ZF array of the *PRDM9-XB* paralogue. Nine

SNP are labeled ss1 to ss9, while the length polymorphism corresponding to the loss of two ZF domains is labeled VNTR. For each SNP, we define the position and the nature of the nucleotide substitution in the corresponding ZF domain (labeled in top row). For non-synonymous substitutions, we also define position and nature of the amino-acid substitution. Ss9 is synonymous and labeled as such (SYN). We represent the eight detected haplotypes sorted by VNTR genotype: haplotypes with 22 ZFs above, haplotypes with 20 ZFs below. For each length class, the top row represents the corresponding ZF domains labeled A, B, C, … I, and colored accordingly. ZFs differing by 'or '' (f.i. D, D' and D'') have an amino-acid similarity of $\geq$92%. The triplet of amino-acids at DNA binding positions -1,3 and 6 [f.i.41] are shown above each ZF. The frequency of the eight haplotypes in the HF population are given in the column on the right. (C) Observed counts (blue bars), and effects on genome-wide jungle usage ($\log_{10}$ of GJU) (red circles) $\pm$ standard error (black vertical line) for the eight *PRDM9-XB* haplotypes defined in (B). (D) Result of single point analysis on GJU for BTAX. The ten *PRDM9-XB* variants are highlighted in red.
doi:10.1371/journal.pgen.1002854.g003

strong candidate causative variant. However, we cannot exclude that the causative variant is regulatory, lying outside of the sequenced *RNF212* segments and in LD with P259S, nor that additional causative *RNF212* variants exist. For *REC8*, the causative variants are most likely regulatory, as coding variants strongly associated with GRR could not be detected despite the sequencing of haplotypes with opposite effects. The most strongly associated SNP (ss418642854) is potentially causal, although the affected sequence is not strongly conserved. Thus, it remains possible that other variants outside the sequenced regions will show equal or even stronger association with GRR. Further sequencing and functional studies are required to achieve complete molecular understanding of these two QTL.

Confirming previous findings in human and mice, we observed an overdispersion of LRR, CO tending to preferentially occur in hot windows (exhibiting sequence features reminiscent of human recombination hotspots), while avoiding cold windows. As expected from human and mouse, hot windows tended to concentrate in sub-terminal regions, while cold windows were enriched at centromeres and in the middle of chromosome arms. The propensity for CO to occur in hot windows (GHU) was shown to be a repeatable and heritable phenotype in HF ($h^2 \approx 21\%$). We identified three genomic loci with suggestive evidence for an effect on GHU. Strikingly, one of these co-localized with two X-linked *PRDM9* paralogues. By resequencing bulls representing all hidden haplotype clusters, we identified nine SNPs and a VNTR-type polymorphism in the *PRDM9-XB* paralogue. Using a haplotype-based approach, we provide strong evidence that an I to K substitution at DNA binding position +6 of ZF 11 decreases GHU ~30-fold, without affecting GRR. Moreover, we provide suggestive evidence that the VNTR-like polymorphism as well as an A to Y amino-acid substitution at position 7 of ZF domain 14 independently modulate GHU ~6-fold. Surprisingly, four of the eight non-synonymous variants correspond to R<->Q substitutions at amino-acid position 13 of four distinct ZF domains. None of these variants appear to affect GHU. While this could indicate that the corresponding position is highly mutagenic, we believe that it is more likely that this finding reflects the spreading of a variant within the ZF array by a process of concerted evolution of tandem repeats [44]. Likewise, ss3 and ss6 both correspond to A<->Y substitutions at amino-acid position 7. Surprisingly, no polymorphisms were observed in the equivalent (although shorter) *PRDM9-XA* array. The reason for this striking difference remains unknown, especially given the fact that both *PRDM9-XA* and *PRDM9-XB* appear to be expressed in bovine testes (data not shown).

In further support of the rapid coevolution of *PRDM9* and recombination hotspots in the bovine, we identify haplotypes with significantly different propensity to engage in recombination at a specific BTA6 jungle. We hypothesize that this results from sequence differences at recombination triggering motifs. This model predicts epistatic interactions between *PRDM9* variation and BTA6 haplotype, and analyses to uncover such effects are ongoing.

We demonstrate that, as expected, all chromosomes are subject to positive interference, multiple CO being more distant that expected by chance alone. By applying a gamma-model to the distance between MLH1 foci, Lian et al. [27] observed that interference might increase with decreasing chromosome size. It was subsequently indicated, however, that the observed trend might be due to inappropriate modeling of finite chromosome size [45]. It has been suggested that crossovers might involve two pathways [f.i. 46]: (i) the pairing pathway not subject to interference, and (ii) the disjunction pathway undergoing interference. As the proportion of pairing over disjunction CO increases with decreasing chromosome size, the two-pathway model predicts a decrease in interference levels with decreasing chromosome size as observed in budding yeast [47]. However, we did not find evidence for an effect of chromosome size on levels of interference, in general agreement with Broman and Weber [26] for human. We devised a novel metric to quantify genome-wide interference, and showed that it is repeatable and differs significantly between individuals, yet modestly heritable. We obtain preliminary evidence for the existence of a QTL influencing this trait on BTA25. The corresponding signal was observed when measuring inter-CO distance in base pairs but not when measured in centimorgan. Further studies will be required to verify the genuine nature of this QTL.

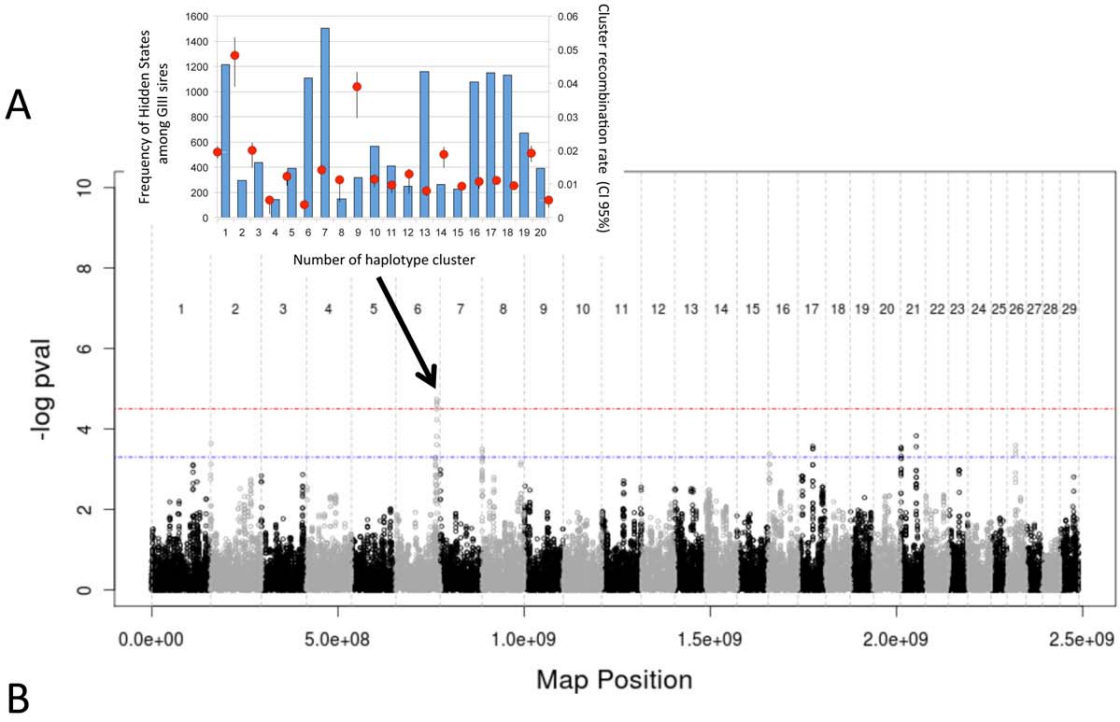
Methods

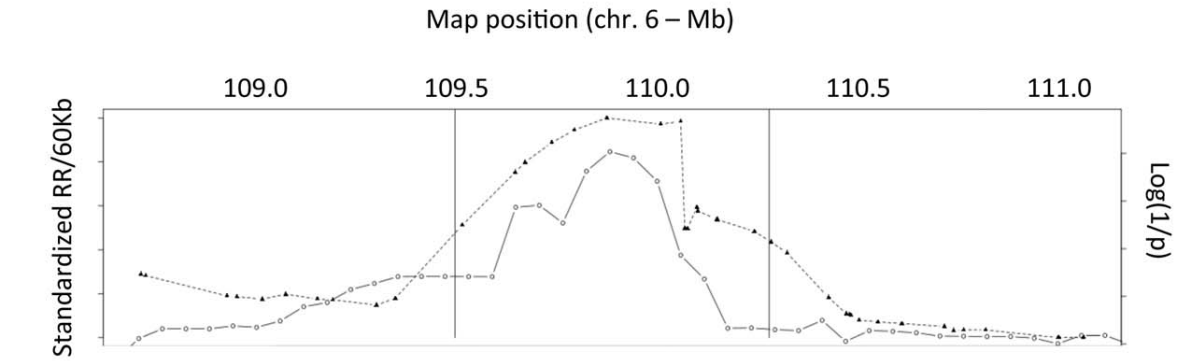
Identifying CO events and data cleanup

Marker phasing was conducted with the Phasebook software package [34]. We exploited Mendelian rules to phase SNP genotypes in sons (GII and GIII), and linkage information to phase SNP genotypes in sires (GI and GII). CO events were then identified as phase switches in the gametes transmitted by the GII sires to their GIII sons. Double-CO occurring in intervals that were separated by less than three informative markers were attributed to genotyping errors and ignored. CO in 2-Mb windows for which the recombination rate of the GII sire was significantly >5% were attributed to GII phasing errors and ignored. The distribution of CO-events was surveyed using a graphical interface to identify as many other artifacts as possible.

Estimating chromosome-specific proportions of meioses with 0, 1, 2, … chiasmata from the proportion of gametes with 0, 1, 2, … crossovers (CO)

Assuming absence of chromatid interference, the proportion of gametes with i CO from meioses with j chiasmata ($p_{i|j}$), follows the binomial distribution:


A
Frequency of Hidden States among GIII sires
Cluster recombination rate (CI 95%)
Number of haplotype cluster
-log pval
Map Position
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
2.5e+09


B
Map position (chr. 6 – Mb)
109.0
109.5
110.0
110.5
111.0
Standardized RR/60Kb
Log(1/p)

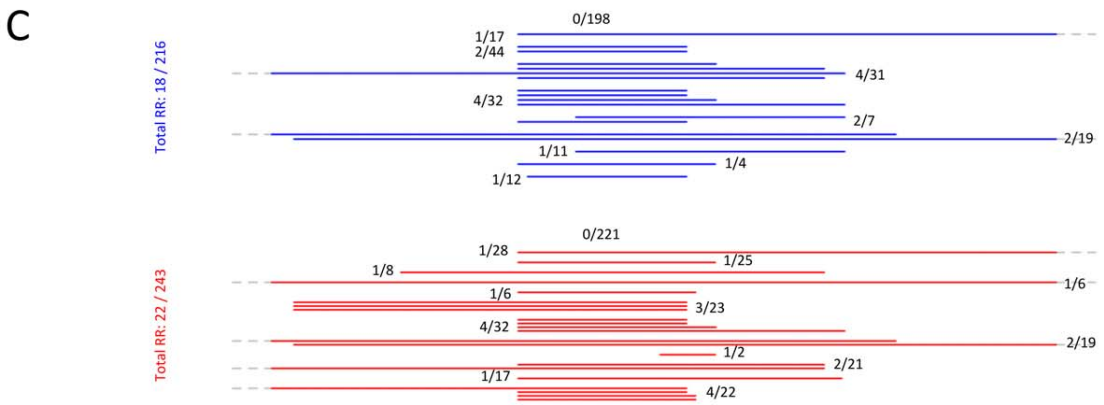

C
Total RR: 18 / 216
0/198
1/17
2/44
4/31
4/32
2/7
2/19
1/11
1/4
1/12
Total RR: 22 / 243
0/221
1/28
1/25
1/8
1/6
1/6
3/23
4/32
2/19
1/2
2/21
1/17
4/22

Figure 4. Log(1/p) values for *cis*-acting haplotype effects on local recombination rate, normalized recombination rate, and marker intervals. (A) Log(1/p) values for *cis*-acting haplotype effects on local recombination rate. The horizontal lines correspond to the significant (red) and suggestive (blue) thresholds. The inset shows the frequency (blue bars, left axis) and effect on recombination rate (red circles, right axis) with 95% CI obtained by bootstrapping (black vertical lines) for the 20 modeled hidden haplotype clusters at the most significant BTA6 position. (B) Normalized recombination rate (triangles, dotted line) and log(1/p) (circles, continuous line) values around the BTA6 QTL. The vertical lines mark the limits of the 800 Kb window in which the *cis*-acting haplotype effect was strongest. (C) Marker intervals to which the recombination events underlying the QTL were mapped. Recombinant individuals are sorted by GII sire (red: heterozygous for HS 2; blue: heterozygous for HS 9) with indication of the number of recombinant/total number of GIII sons for the corresponding GII sire. Numbers were summed for all GII sires without GIII sons recombining in the interval of interest (0/198 and 0/221).

$$p_{i|j} = \binom{i}{j}\left(\frac{1}{2}\right)^j$$

As a consequence, the proportion of gametes with i CO from all meiosis (p_i) equals

$$p_i = \sum_{j=1}^{\infty} p_j p_{i|j}$$

in which p_j correspond to the proportion of meiosis with j chiasmata.

The likelihood of a dataset with n_0 gametes with 0 CO, n_1 gametes with 1 CO, n_2 gametes with 2 CO, etc. equals

$$L = \Pi_{i=1}^{\infty} p_i^{n_i}$$

L is a function of the unknown parameters p_j. We determined the values of p_j that maximized L. Values considered for j were limited to six.

Measuring and normalizing 60-Kb window-specific recombination rates

The recombination rate in a defined 60 Kb window was computed as $\left(\sum_{i=1}^{n} o_i/x_i\right)/T$ where n is the total number of CO events identified on the corresponding chromosome in the analyzed population, x_i is the size (in bp) of the marker interval to which CO i has been mapped, o_i the overlap (in bp) between the 60-Kb window and CO interval i, and T is the total number of analyzed gametes. To normalize window-specific recombination rates for local marker density and informativeness, we simulated (1,000 times) genotypes for the GIII sons by randomly "dropping" CO events on the phased GII chromosomes assuming a uniform distribution of CO events following a Poisson process (with mean corresponding to the real data), randomly sampling one of the two paternal chromosomes, while keeping the original maternal chromosome intact. The entire phasing and CO mapping process was then reinitiated with these *in silico* generated SNP genotypes. The corresponding simulations yielded an average recombination rate with standard deviation for each window. This allowed us to express the actual recombination rate measured for a given window in standardized deviations from the mean (across simulations).

Quantifying crossover interference

Chromosome-specific levels of CO interference were quantified using the shape parameter v of a gamma distribution, following Broman and Weber [26]. We determined – for each chromosome – (i) the frequency distribution of CO events per gamete, (ii) the CO position (in cM) for gametes with one CO, (iii) the inter-CO distance for gametes with two CO, (iv) the inter-CO distance for gametes with three CO. We then compared these distributions with theoretical expectations under various levels of interference, accounting for chromosome size. To that end, we simulated series of "chiasmata" (CH) along four stranded bundles with gamma-distributed intervals. The shape parameter v was varied from 1 (no interference) to 15 with 0.1 increments. The rate parameter was

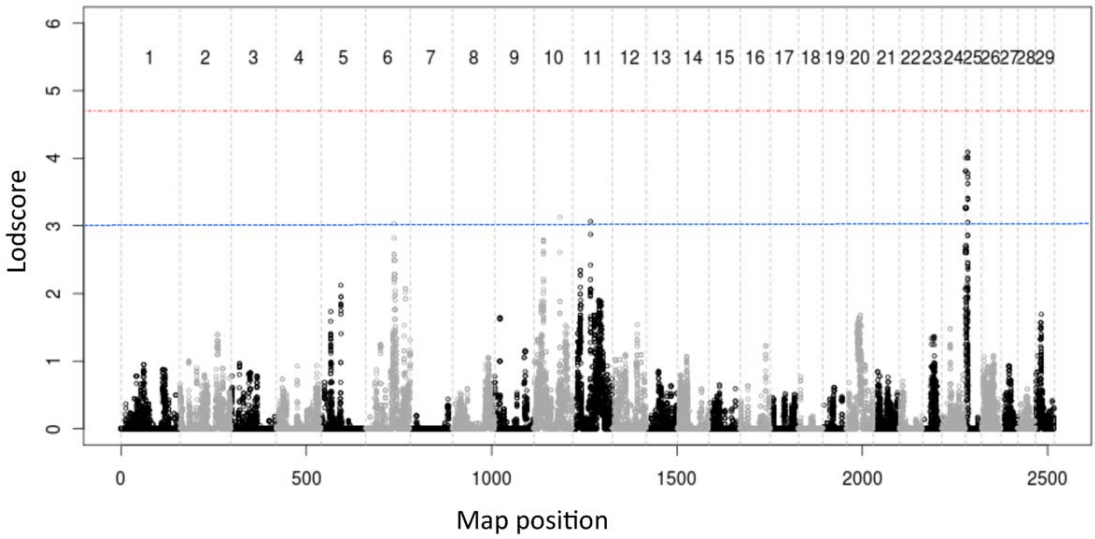


Figure 5. Results of genome-scan for QTL affecting the normalized distance between pairs of CO events measured in base-pairs (GIL_{bp}), using a method that simultaneously extracts linkage and LD signal [34]. The red and blue horizontal lines mark the genome-wide significant and suggestive thresholds determined by permutation testing.
doi:10.1371/journal.pgen.1002854.g005

always said at $2v$. The values of the gamma variables were multiplied by 25, to obtain an average inter-CH distance of 50. The CH-series were then converted to single chromatid CO-series, by retaining CH events with a probability of 0.5. The average inter-CO distance was therefore 100 ("cM"). We then randomly sampled at least 500,000 independent segments of n cM from these chains, where n corresponds to the actual size of the studied bovine chromosome in cM. For these 500,000 "gametes", we computed the frequency distribution of (i) CO-events, (ii) CO-position for gametes with one CO (5 cM bins), (iii) inter-CO distance for gametes with two CO (5 cM bins), and (iv) inter-CO distance for gametes with three CO (5 cM bins). We then evaluated the goodness-of-fit between the real and simulated data by maximum likelihood. The likelihood of the data (L) was assumed to be:

$$L = \Pi_{i=1}^{N} P(Nr_i) \times P(D_i|Nr_i)$$

in which N is the total number of studied gametes, Nr_i is the number of CO-events characterizing gamete i, $P(Nr_i)$ is the probability of having Nr CO-events (which is determined by the value of v), D_i is the CO-position (gametes with one CO) or inter-CO distance(s)(gametes with >1 CO), and $P(D_i|Nr_i)$ is the probability of D_i given Nr_i (which is determined by the value of v). $P(D_i|Nr_i)$ was computed for gametes with 1, 2 and 3 CO and set at 1 for the other gametes (there is no additional information to be extracted from gametes with 0 CO; gametes with >3 CO are rare and their information likely to be less reliable). For simplicity, the probability of the two inter-CO distances for gametes with two CO were considered independent.

Accordingly, this likelihood equation can be reformulated as:

$$L = L_{Nr} \times L_{D|1} \times L_{D|2} \times L_{D|3}$$

in which L_{Nr} is the likelihood of the observed frequency distribution of number of CO per gamete, and equals $L_{Nr} = \Pi_{i=1}^{\infty} f_i^{Ni}$, where f_i is the expected frequency (given v) of gametes with i CO (given v) and N_i is the observed number of gametes with i CO; $L_{D|1}$ is the likelihood of the observed frequency distribution of CO-positions for gametes with one CO, and equals $L_{D|1} = \Pi_{i=1}^{l/5} f_i^{N1i}$, where f_i is the expected frequency (given v) of single-CO gametes with CO-position i (5 cM bin), $N1_i$ is the observed number of single-CO gametes with CO-position in bin i, and l is the length (in cM) of the considered chromosome; $L_{D|2}$ is the likelihood of the observed frequency distribution of inter-CO distance for gametes with two CO, and equals $L_{D|2} = \Pi_{i=1}^{l/5} f_i^{N2i}$, where f_i is the expected frequency (given v) of inter-CO distance i (5 cM bin) for double-CO gametes, $N2_i$ is the observed number of inter-CO distance in bin i for double-CO gametes, and l is the length (in cM) of the considered chromosome; $L_{D|3}$ is the likelihood of the observed frequency distribution of inter-CO distance for gametes with three CO, and equals $L_{D|3} = \Pi_{i=1}^{l/5} f_i^{N3i}$, where f_i is the expected frequency (given v) of inter-CO distance i (5 cM bin) for triple-CO gametes, $N3_i$ is the observed number of inter-CO distance in bin i for triple-CO gametes, and l is the length (in cM) of the considered chromosome.

The values of f_i needed to compute the corresponding likelihoods were obtained from the simulations performed under varying values of v.

The inter-CO distance for gametes with two CO events appeared to be the most influential parameter in determining v (Figure S3). We therefore focused on this measure to perform genetic analysis and QTL mapping (see also hereafter) of crossover interference. Inter-CO distances for gametes with two CO were measured in centimorgan ("GIL_{cM}") or in base-pairs ("GIL_{bp}"), and were normalized by subtracting the mean inter-CO distance for that chromosomes and multiplying by the standard deviation.

Correcting GRR for family size

We noted that estimates of GRR decreased with increasing family size (Figure S4C) and attributed this to errors in determining the sire's phase. To correct GRR for this factor we used 10 paternal half-sib families with >100 G III sons. From these families we randomly sampled (1,000 times) from 1 to 10 sons with corresponding SNP genotypes. Phasing of the GI, GII and GIII bulls was conducted with Phasebook on these purposely limited data-sets, including determination of CO events in the paternal gametes transmitted to GIII sons. For each of the 10 families we then compared average GRR estimated with 1, 2, … 10 sons (over the 1,000 simulations) with GRR estimated with all sons (>100), yielding a set of $\overline{\Delta_{ij}}$ values where i corresponds to the number of used sons (1 to 10) for family j. These values were averaged across families to generate $\overline{\Delta_i}$, i.e. an overall effect on GRR of family size i, used to correct the actual GRR estimates obtained from families with <10 half-sibs.

Estimating h^2

Narrow sense heritabilities (h^2) of recombination phenotypes (measured in the GIII sons) were estimated using two mixed models [48]. The first modeled average phenotypes of GII sires, and included an overall mean, a random individual animal effect (with variance-covariance structure proportionate to twice the coefficient of kinship between corresponding GII sires), and a random error proportionate to the inverse of the number GIII sons per GII sire. The second modeled the individual phenotypes of the gametes transmitted to GIII sons. It included an overall mean, a random individual animal effect (with variance-covariance structure proportionate to twice the coefficient of kinship between corresponding GII sires), a random permanent GII sire effect, and a random error. Variance components were estimated by restricted maximum likelihood (REML) analysis [49].

QTL mapping

QTL were mapped using a previously described mixed model approach that simultaneously exploits linkage and LD information [34]. At each SNP position, homologues in the data set were assigned to one of 20 hidden states corresponding to "ancestral haplotype clusters". The utilized mixed models was the same as the first one used to estimate h^2 (i.e. modeling average phenotypes of the GII sires and adjusting the random error such that it would be proportionate to the inverse on the number of GIII observations per GII sire), with addition of a random "ancestral haplotype cluster" effect. The covariance between the effects of the 20 possible "ancestral haplotype clusters" was assumed to be zero. Significance thresholds were empirically determined by phenotype permutation [50], following standard guidelines [51]. Phenotypic values were permuted amongst half-sibs, a genome-scan conducted, and the highest (across the genome) value of the likelihood ratio test (LRT) stored. QTL were considered significant if the corresponding LRT exceeded the 95% percentile of the LRT-values obtained by permutation (i.e. if it exceeded the value of the LRT expected to occur by chance alone once every twenty genome scans). QTL were considered suggestive if the corresponding LRT exceeded the 63% percentile of the LRT-values obtained by permutation. To see the latter, a LRT that is not exceeded in $100-63=37\%$ of genome scans is exceeded on

average once per genome scan as $0.37 = e^{-1}$ (assuming that such events are Poisson distributed). A linkage signal is defined as being suggestive if it is obtained by chance alone on average once per genome scan.

Scanning the genome for cis-acting haplotype effects on local recombination rate

To identify cis-acting haplotype effects on local recombination rate, we defined 800 Kb windows centered around the interrogated marker position. At that marker position, we selected the GII sires that were heterozygous for "ancestral haplotype clusters" [34] and tested the additive effect of "ancestral haplotype cluster" of the GII sires on the recombination phenotype of their GIII sons by ANOVA. The recombination phenotype of GIII sons was defined as the probability that a paternal CO event would have occurred in the interrogated window measured as the degree of overlap between CO encompassing marker intervals and interrogated window.

Re-sequencing positional candidate genes and genotyping of candidate QTN

We designed primer pairs to amplify and sequence either the entire gene (*REC8*), the ORF (*RNF212*, *KCNJ2*, *KCNJ16*), or the ZF array (*PRDM9-XA* and *PRDM9-XB*) (Table S2). Animals to re-sequence were selected based on the ancestral haplotype clusters they carried at the most likely position of the corresponding QTL. Amplifications, purification of the amplicons and direct sequencing of the amplicons were carried out using standard procedures. Genotyping of candidate QTN was conducted using 5′ exonuclease (Taqman) assays for *REC8* and *RNF212* (Table S3), or by amplicon sequencing for *PRDM9-XB*.

Supporting Information

Figure S1 (A) Linear relationship between chromosome length in Mb (from UMD3.0 build) and average number of CO-events for the 29 bovine autosomes. The least square regression is characterized by a Y-intercept $\beta_0 = 0.48$ and a slope $\beta_1 = 0.07$CO/10 Mb. The slope of the regression is intermediate between the slopes characterizing male and female recombination in human [35]. (B) Proportion of meioses with zero (black), one (gray), two (blue) and three (red) chiasmata for the 29 bovine autosomes. Plain lines: proportions maximizing the likelihood of the data (assuming no chromatid interference). Dotted lines: expected proportions assuming a truncated Poisson distribution of number of chiasmata (proportion of meioses with zero chiasmata forced at zero) [36]. The data are best explained assuming near absence of nullichiasmatic meioses for autosomes 1 to 16, and frequencies <5% for the smaller chromosomes. For the largest chromosomes, the most likely (ML) frequency of meioses with at least two chiasmata is considerably higher than expected under a truncated Poisson model, supporting the preferred occurrence of a second chiasma for larger chromosomes.
(PPTX)

Figure S2 (A) Representative example of the variation in male recombination in 60-Kb windows across a bovine autosome (BTA14). The plain black like (upper halve) corresponds to recombination rate estimated in the Dutch population, while the dotted black line (lower halve) corresponds to the recombination rate estimated in the NZ population. The red and blue horizontal lines correspond to "hot" and cold" windows, respectively, i.e. segments in which the observed recombination rate deviates by more than 2.5 standard deviations from the local recombination rate expected under a model of uniform distribution of CO events. (B) Variation in male recombination in 60 Kb windows across the bovine genome. The plain black like (upper halve) corresponds to recombination rate estimated in the Dutch population, while the dotted black line (lower halve) corresponds to the recombination rate estimated in the NZ population. The correlation between window-specific recombination rate in the Dutch and NZ population was high ($r^2 = 0.80$; $p<0.0001$), despite the use of distinct SNP panels. The red and blue horizontal lines correspond to positions of "hot" and "cold" windows, respectively, i.e. segments in which the observed recombination rate deviates by more than 2.5 standard deviations from the local recombination rate expected under a model of uniform distribution of CO events. (C) Bar graphs: Frequency distribution of local (60-Kb window) recombination rate normalized for local marker density and informativeness as described in M&M. Curve: Standard normal distribution. Red and Blue vertical lines mark the thresholds defining "hot" (mean+2.5 SD) and "cold" (mean – 2.5 SD) windows, respectively. (D) Location of "hot" (red) and "cold" (blue) windows, relative to normalized chromosome length. All 29 acrocentric autosomes were aligned with their centromere towards the left of the graphs. Hot windows tend to concentrate in sub-terminal (proximal chromosome end) and terminal regions (distal chromosome end), while cold windows concentrate in the middle of the chromosome arms as well as in terminal regions (proximal chromosome end) coinciding with the centromeres.
(PPTX)

Figure S3 (A) For each of the 29 bovine autosomes (BTA1-29), *column I:* frequency distribution of gametes with 0, 1, 2, … CO-events expected in the absence of cross-over interference (blue), expected given the value of ν maximizing the likelihood of the overall data (light red), expected given the value of given the value of ν maximizing the likelihood of the frequency distribution of CO-events (dark red), as observed (green). The gray bars correspond to the frequency distribution of meioses with 0, 1, 2, … chiasmata expected given the value of ν maximizing the likelihood of the frequency distribution of CO-events. The number following the BTA number corresponds to the ν-value maximizing the overall likelihood. The inset illustrates the profile of the $\log_{10}$ of the overall likelihood for varying values of ν. The number in brackets correspond to the ν-value maximizing the likelihood of the observed frequency distribution of CO number. *Column II:* Frequency distribution (5 cM bins) of position of single CO-events for gametes with one CO (green bars). The curves correspond to the distributions expected in the absence of interference (blue), assuming the ν-value maximizing the overall likelihood (light red), and assuming the ν-value maximizing the likelihood of the frequency distribution of single CO-positions (dark red). The numbers between brackets correspond the ν-value maximizing the likelihood of the frequency distribution of single CO-positions, and the number of observed gametes (out of a total of 7,277 used in this analysis) with one CO. *Column III:* Frequency distribution of the distance (5 cM bins) between CO events for gametes with two CO (green bars). The curves correspond to the distributions expected in the absence of interference (blue), assuming the ν-value maximizing the overall likelihood (light red), and (if different from the previous ones) assuming the ν-value maximizing the likelihood of the frequency distribution of inter-CO distance for gametes with two CO (dark red). The numbers between brackets correspond the ν-value maximizing the likelihood of the frequency distribution of inter-CO distance, and the number of observed gametes (out of a total of 7,277 used in this analysis) with two CO. *Column IV:* Frequency distribution of the distance (5 cM bins) between CO events for gametes with three CO (green bars). The curves correspond to the distributions expected in the absence of

interference (blue), assuming the ν-value maximizing the overall likelihood (light red), and (if different from the previous ones) assuming the *ν*-value maximizing the likelihood of the frequency distribution of inter-CO distance for gametes with three CO (dark red). The numbers between brackets correspond the *ν*-value maximizing the likelihood of the frequency distribution of inter-CO distance, and the number of observed gametes (out of a total of 7,277 used in this analysis) with three CO. (B) Chromosome-specific levels of chiasma interference measured using the shape parameter nu (*ν*) of a gamma distribution (cfr. M&M). Dark blue (All): *ν*-value maximizing the likelihood of all data. Red (NrCO): *ν*-value maximizing the likelihood of the frequency distribution of CO-events per gametes. Green (SCO): *ν*-value maximizing the likelihood of the frequency distribution of CO-position (in cM) for gametes with one CO. Purple (DCO): *ν*-value maximizing the likelihood of the frequency distribution of inter-CO distance (in cM) for gametes with two CO. Light blue (TCO): *ν*-value maximizing the likelihood of the frequency distribution of inter-CO distance (in cM) for gametes with three CO. Chromosomes are ordered (left to right) from 1 to 29. The numbers under the X-axis correspond to the size of the corresponding chromosome in cM.
(PDF)

Figure S4 (A) Black dots correspond to the total number of CO events identified in the paternal genome of 10,192 GIII sons sorted by GII sire. The red dots mark the average GRR for each GII sire. GRR did not differ significantly between Holstein-Friesian and Jersey bulls. (B) Correlation between the GRR estimated for 72 GII sires separately from the number of CO events transmitted to non-overlapping sets of GIII sons from H and NZ, respectively. Spearman's rank correlation was 0.58 ($p<3.7\times10^{-7}$). (C) Total number of CO events (GRR) in the genome transmitted by GII sires to their GIII sons. GIII sons are sorted according to the number of half-brothers in the data set. The increase of GRR with decreasing family size is clearly visible. (D) Lod scores obtained for GRR using 121 HF GII sires, and (i) 45 SNPs from the Illumina bovine high-density 777K SNP array mapping to the confidence interval of the BTA10 QTL (blue dots) and (ii) *REC8* SNPs (red dots). The highest lod score was obtained for *REC8* variant ss418642854. (E) Lod scores obtained for GRR using 121 HF GII sires, and (i) 27 SNPs from the Illumina bovine high-density 777K SNP array mapping to the confidence interval of the BTA6 QTL (blue dots) and (ii) *RNF212* SNPs (red dots). The highest lod score was obtained for *RNF212* variant ss469104611 (= P259S).
(PPTX)

Figure S5 Position of the amplicons used to scan the *REC8* (A), and *RNF212* genes (B) (cfr. Table S2). The corresponding *RNF212* gene model has been submitted to Genbank.
(PPTX)

Figure S6 (A) Black dots: Average overlap (0 to 1) between marker intervals (<800-Kb) with assigned CO events and "hot" 60-K windows for GIII-sons sorted by GII-sire. Red dots: Average overlap for all CO events transmitted by corresponding GII-sire. (B) Correlation between average hot-window usage estimated for the 72 shared GII-sires respectively from gametes transmitted to Dutch versus New-Zealand GIII sons.
(PPTX)

Figure S7 (A) GIII sons inherit chromosomes with 0, 1, 2, 3, … CO from their GII sires. In this analysis, we only use "di-CO" chromosomes (i.e. with 2 CO). We measure the distance between CO-pairs in centimorgan (GIL-cM) or in base-pairs (GIL-bp) prior to normalization (i.e. expressed in standard deviations from the chromosome mean). Thus, the distance between the CO-pair of the di-CO chr. 1 inherited by son *x* from sire *y*, may be "so many" standard deviations above or below the average distance between CO-pairs on di-CO chr. 1's (across all GIII sons receiving a di-CO chr. 1 from their sire). The black dots correspond to the average of the normalized distances between CO-pairs for all di-CO chromosomes inherited by a given GIII son. GIII sons are sorted by GII sire, i.e. they are on the same vertical black line. The red dots correspond to the average of the normalized inter-CO distances across all di-CO chromosomes transmitted by a given GII sire to all its GIII sons. (B) Correlation between average normalized distance between CO events for all homologues with two recombination events transmitted by 72 shared GII-sire to their Dutch GIII-sons (X-axis), and their NZ GIII-sons (Y-axis). Inter-CO distance was measured either in centimorgan (GIL-cM) or in base pairs (GIL-bp). (C) Results of genome-scan for QTL affecting the normalized distance between pairs of CO events measured in centimorgan (GIL_{cM}), using a method that simultaneously extracts linkage and LD signal[34]. The red and blue horizontal lines mark the genome-wide significant and suggestive thresholds determined by permutation testing.
(PPTX)

Table S1 (A) Following Kong et al. [8], we tested the effect of base pair composition and gene content on LRR by multiple regression. As in human, local recombination rate was positively correlated with CpG content, yet negatively correlated with GC, polyA/polyT and gene content (after adjustment for CpG content). CpG content accounted for ~19% of the variance, while the four parameters explained ~28% jointly.(B) We tested whether "hot" and "cold" status correlated with window content in specific interspersed repeats. For each 60-Kb hot (respectively cold) window, we sampled a "regular" window matched for CpG, GC, polyA/polyT and gene content, and compared total counts of 58 types of interspersed repeats. The statistical significance of the count difference was evaluated by permutation with Bonferroni correction for the realization of 58 independent tests. As can be seen from the table (i) some repeat types were enriched in hot and depleted in cold windows, including SINE/BovA (ratio Jungle/control: 1.12; ratio desert/control: 0.84), LTR/ERVL-MaLR (1.04;0.74), RC/Helitron (2.23;0.74) and DNA (1.30;0.75), (ii) some repeats were depleted in hot and enriched in cold windows, including LTR/ERVK (0.87;1.45), LINE/RTE-BovB (0.93;1.34) and LINE/L1 (0.98;1.06), (iii) SINE/RTE-BovB were enriched in hot and cold windows (1.12;1.06), (iv) some repeat types were depleted in hot and cold windows including SINE/tRNA-Glu (0.91;0.91), LINE/L2 (0.95;0.74), LINE/CR1 (0.92;0.64), DNA:-hAT-Charlie (0.95;0.77) and DNA/MER1_type (0.95;0.67), (v) rRNA were depleted in hot windows, and (vi) SINE/MIR and satellite/centr were depleted in cold windows.
(XLS)

Table S2 Primers used for amplification and resesequencing of candidate genes *REC8*, *RNF212*, *KCNJ2*, *KCNJ16* and gonosomal *PRDM9-XA and -XB.*
(XLS)

Table S3 Primer and probes used for genotyping candidate QTN using 5′exonculease (Taqsman) assays.
(XLS)

Acknowledgments

We are grateful to CRV (http://www.crv4all.com/) and LIC (http://www.lic.co.nz) for providing the SNP genotype and pedigree information, and to Latifa Karim and Cécile Lam from the GIGA Genomics platform

for their help with the sequencing. We thank Soumya Raychauduri for critically reviewing this manuscript.

Author Contributions

Conceived and designed the experiments: CS CC MG. Performed the experiments: CS WL. Analyzed the data: CS CC MG. Contributed reagents/materials/analysis tools: TD WC. Wrote the paper: CS MG.

References

1. Roeder GS (1997) Meiotic chromosomes: it takes two to tango. Gene Dev 11: 2600–2621.
2. Page SL, Hawley RS (2003) Chromosome choreography: the meiotic ballet. Science 301: 785–791.
3. Coop G, Przeworski M (2007) An evolutionary view of human recombination. Nat Rev Genet 8: 23–34.
4. Martinez-Perez E, Colaiácovo MP (2009) Distribution of meiotic recombination events: talking to your neighbors. Curr Opin Genet Dev 19: 105–120.
5. Hassold T, Hunt P (2001) To err (meiotically) is human: the genesis of human aneuploidy. Nat Rev Genet 2:280–291.
6. Handel MA, Schimenti JC (2010) Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. Nat Rev Genet 11: 124–136.
7. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am. J. Hum. Genet. 63: 861–869.
8. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA et al. (2002) A high-resolution recombination map of the human genome. Nat Genet 31: 241–247.
9. Kong A, Barnard J, Gudbjartsson DF, Thorleifsson G, Jonsdottir G et al. (2004) Recombination rate and reproductive success in humans. Nat Genet 36: 1203–1206.
10. Lenzi ML, Smith J, Snowden T, Kim M, Fishel R et al. (2005). Extreme heterogeneity in the molecular events leading to the establishment of chiasmata during meiosis in human oocytes. Am J Hum Genet 76:112–127
11. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G et al. (2005) A common inversion under selection in Europeans. Nat Genet 37: 129–137.
12. Kong A, Thorleifsson G, Stefansson H, Masson G, Helgason A et al. (2008) Sequence variants in the RNF212 gene associate with genome-wide recombination rate. Science 319: 1398–1401.
13. Chowdhury R, Bois PRJ, Feingold E, Sherman SL, Cheung VG (2009) Genetic analysis of variation in human meiotic recombination. PLoS Genet 5: e1000648. doi:10.1371/journal.pgen.1000648
14. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321–324.
15. Paigen K, Petkov P (2010) Mammalian recombination hot spots: properties, control and evolution. Nat Rev Genet 11: 221–233.
16. The 1,000 genomes project consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
17. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M (2008). High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science 319:1395–1398.
18. Jeffreys AJ, Neumann R. (2009) The rise and fall of a human recombination hot spot. Nat Genet 41: 625–629.
19. Parvanov ED, Petkov PM, Paigen K. (2010) Prdm9 controls activation of mammalian recombination hotspots. Science 327: 835.
20. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C et al. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836–840.
21. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C et al. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327: 876–879.
22. Berg IL, Neumann R, Lam KWG, Sarbajna S, Odenthal-Hesse L et al. (2010) PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet 42: 859–863.
23. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N et al. (2005) Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet 37: 429–434.
24. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ et al. (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308: 107–111.
25. Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J (2009) A mouse speciation gene encodes a meiotic histone H3 methyltransferase. Science 323: 373–375.
26. Broman KW, Weber JL (2000) Characterization of human crossover interference. Am J Hum Genet 66: 1911–1926.
27. Lian J, Yin Y, Oliver-Bonet M, Liehr T, Ko E et al. (2008) Variation in crossover interference levels on individual chromosomes from human males. Hum Mol Genet 17: 2583–2594.
28. Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
29. Bovine Genome Sequencing and Analysis Consortium (2009) The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science 324: 522–528.
30. Ihara N, Takasuga A, Mizoshita K, Takeda H, Sugimoto M et al. (2004) A comprehensive genetic map of the cattle genome based on 3802 microsatellites. Genome Res 14: 1987–1998.
31. Ross-Ibarra J (2004) The evolution of recombination under domestication: a test of two hypotheses. Am Nat 163: 105–112.
32. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS et al. (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet 40: 449–454.
33. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF et al. (2009) Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4: e5350. doi:10.1371/journal.pone.0005350
34. Druet T, Georges M. (2010) A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and QTL fine mapping. Genetics 184:789–798.
35. Fledel-Alon A, Wilson DJ, Broman K, Wen X, Ober C et al. (2009) Broad-scale recombination patterns underlying proper disjunction in humans. PLoS Genet 5: e1000658. doi:10.1371/journal.pgen.1000658
36. Sturt E, Smith CA. (1976). The relationship between chromatid interference and the mapping function. Cytogenet Cell Genet 17: 212–220.
37. Paigen K, Szatkiewicz JP, Sawyer K, Leahy N, Parvanov ED, Ng SHS, Graber JH, Broman KW, Petkow PM (2008) The recombinational anatomy of a mouse chromosome. PLoS Genet 4: e1000119. doi:10.1371/journal.pgen.1000119
38. McPeek MS, Speed TP (1995) Modeling interference in genetic recombination. Genetics 139: 1031–1044.
39. Bannister LA, Reinholdt LG, Munroe RJ, Schimenti JC (2004) Positional cloning and characterization of mouse mei8, a disrupted allelle of the meiotic cohesin Rec8. Genesis 40: 184–194.
40. Xu H, Beasley MD, Warren WD, Der Horst GTJ van, McKay MJ (2005) Absence of mouse REC8 cohesin promotes synapsis of sister chromatids in meiosis. Dev Cell 8: 949–961.
41. Ponting CP (2011) What are the genomic drivers of the rapid evolution of *PRDM9*? Trends in Genetics 27:165–171.
42. Axelsson E, Webster MT, Ratnakumar A, Consortium L, Ponting CP, Lindblad-Toh K. (2011) Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. Genome Res Oct 17 [Epub ahead of print].
43. Munoz-Fuentes V, Di Rienzo A, Vila C. (2011) PRDM9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes. PLoS ONE 6: e25498. doi:10.1371/journal.pone.0025498
44. Liao D (1999) Concerted evolution: molecular mechanism and biological implications. Am J Hum Genet 64:24–30.
45. Housworth EA, Stahl FW (2009) Is there variation in crossover interference levels among chromosomes from human males? Genetics 183: 403–405.
46. Housworth EA, Stahl FW (2003). Crossover interference in humans. Am J Hum Genet. 73:188–197.
47. Kaback DB, Barber D, Mahon J, Lamb J, You J (1999) Chromosome-size dependent control of meiotic reciprocal recombination in S. cerevisiae: the role of crossover interference. Genetics 152: 1475–1486.
48. Lynsch M, Walsh B (1998) Genetics and Analysis of Quantitative Traits. S. Associates, ed. (Sunderland).
49. Johnson DL and Thompson R. (1995) Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information. J Dairy Sci 78: 449–456.
50. Churchill GA, Doerge RW. (1996) Permutation tests for multiple loci affecting a quantitative character. Genetics 142: 285–294.
51. Lander E, Kruglyak L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet. 11:241–247.

SUPPLEMENTAL MATERIAL

Genetic variants in *REC8*, *RNF212* and *PRDM9* influence male recombination in cattle.

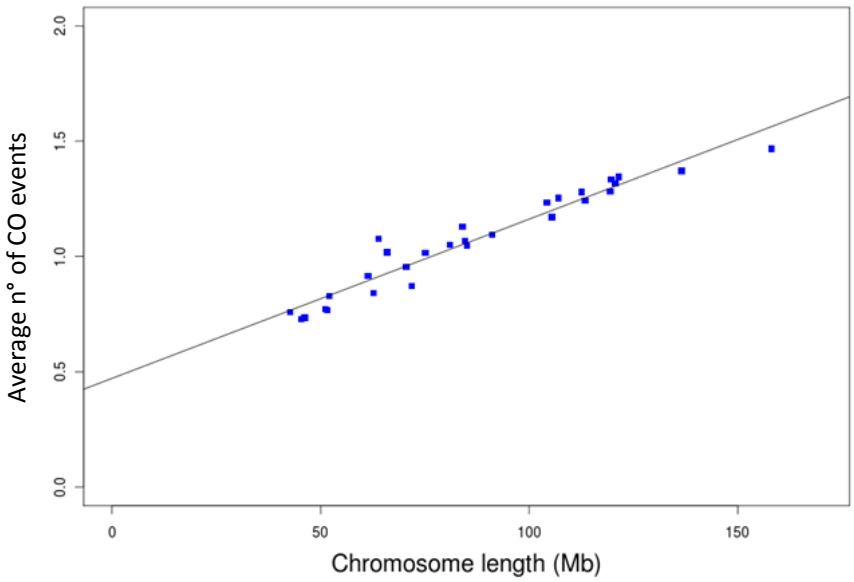
Cynthia Sandor#, Wanbo Li, Wouter Coppieters,

Tom Druet, Carole Charlier & Michel Georges.

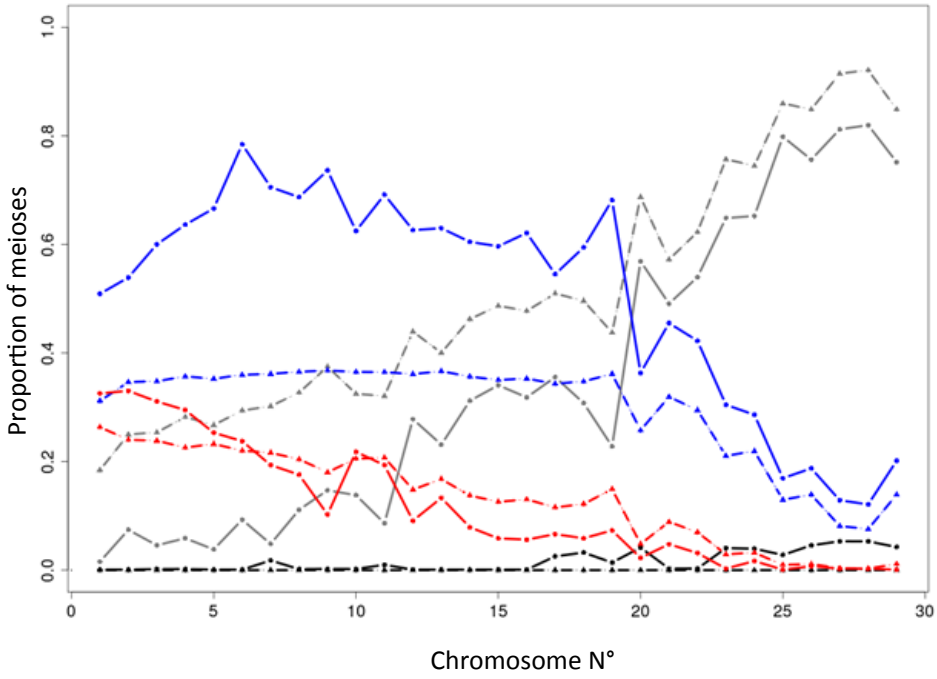
Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège (B34), 1 Avenue de l'Hôpital, 4000-Liège, Belgium

#Present address and affiliation: Divisions of Genetics & Rheumatology, Department of Medicine, Brigham and Women's Hospital Harvard Medical School, 77 Avenue Louis Pasteur New Research Building, Boston, Massachusetts, United States of America. Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, United States of America.

Supplementary Figure 1:

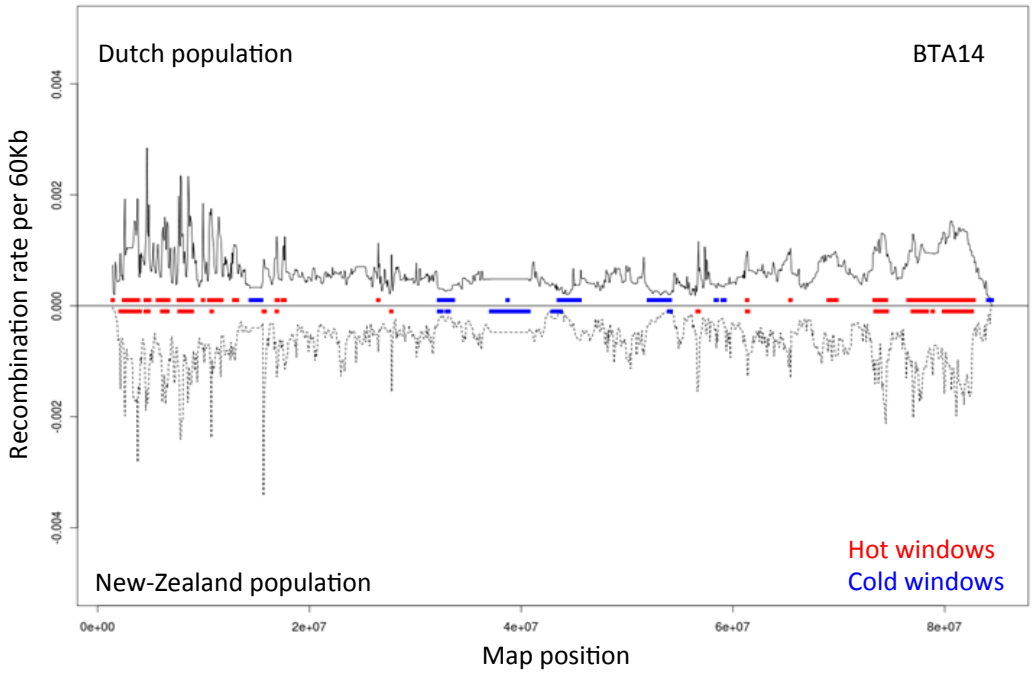


(A) Linear relationship between chromosome length in Mb (from UMD3.0 build) and average number of CO-events for the 29 bovine autosomes. The least square regression is characterized by a Y-intercept β_0 = 0.48 and a slope β_1 =0.07CO/10Mb. The slope of the regression (β_1=0.07CO/10Mb) is intermediate between the slopes characterizing male and female recombination in human[35].

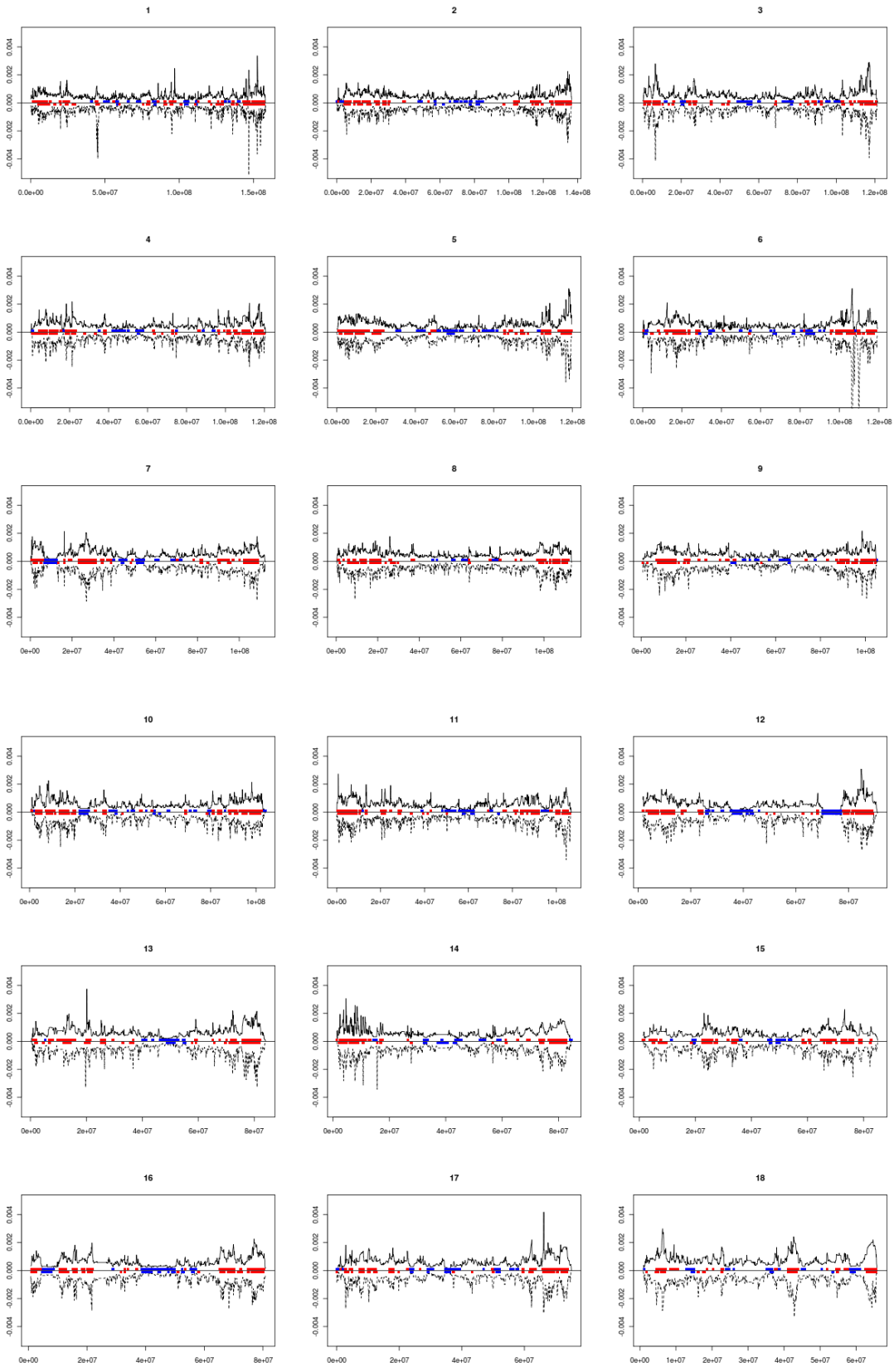


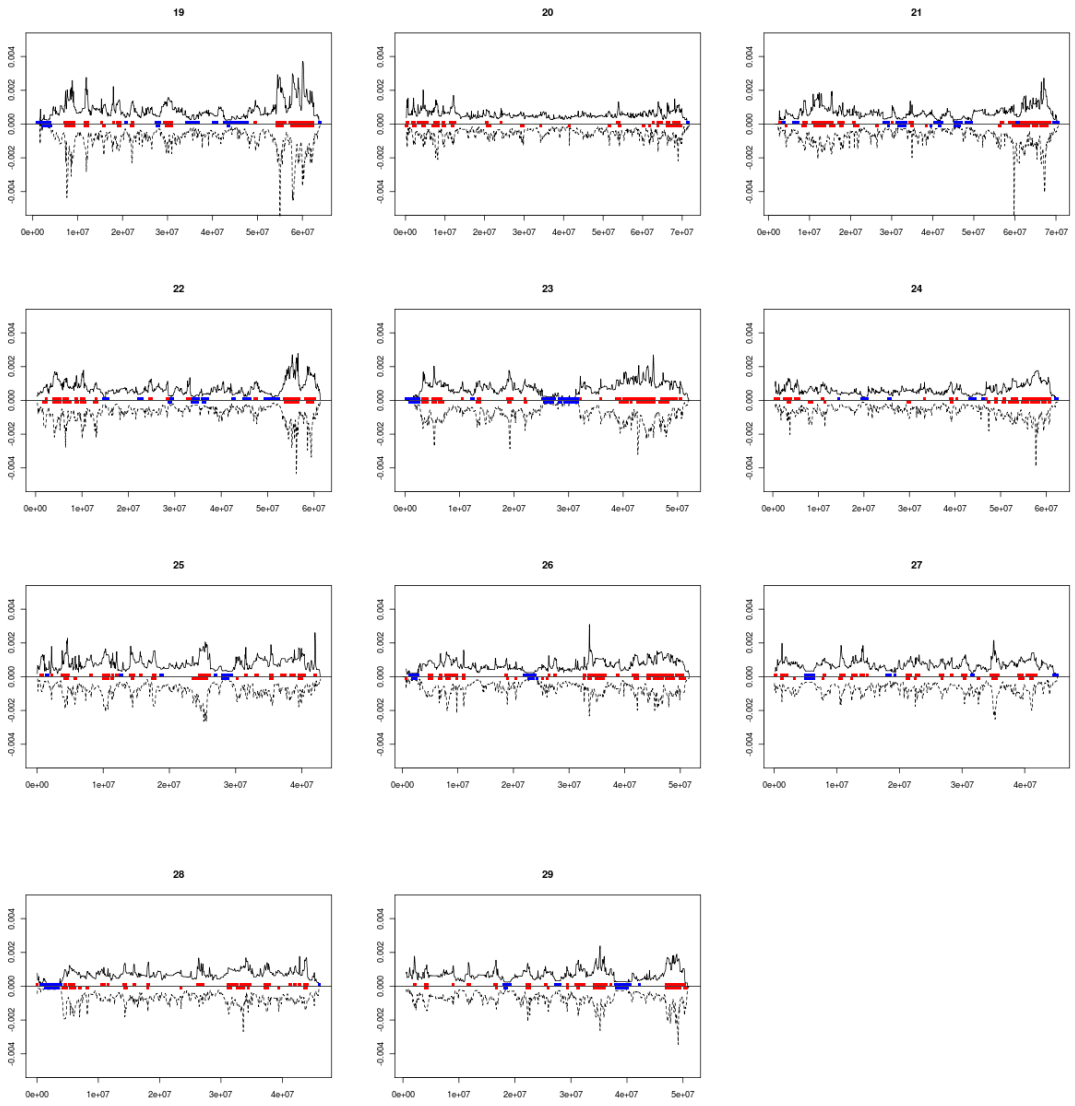
(B) Proportion of meioses with zero (black), one (gray), two (blue) and three (red) chiasmata for the 29 bovine autosomes. Plain lines: proportions maximizing the likelihood of the data (assuming no chromatid interference). Dotted lines: expected proportions assuming a truncated Poisson distribution of number of chiasmata (proportion of meioses with zero chiasmata forced at zero)[36]. The data are best explained assuming near absence of nullichiasmatic meioses for autosomes 1 to 16, and frequencies < 5% for the smaller chromosomes. For the largest chromosomes, the most likely (ML) frequency of meioses with at least two chiasmata is considerably higher than expected under a truncated Poisson model, supporting the preferred occurrence of a second chiasma for larger chromosomes.

Supplementary Figure 2:

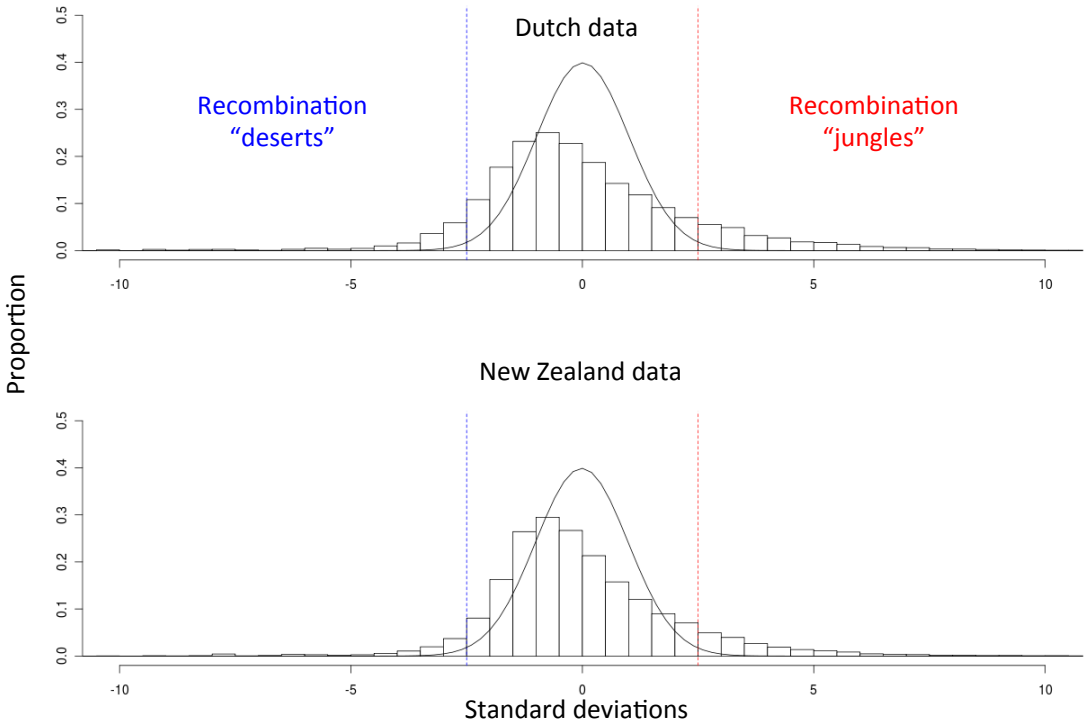


(A) Representative example of the variation in male recombination in 60-Kb windows across a bovine autosome (BTA14). The plain black like (upper halve) corresponds to recombination rate estimated in the Dutch population, while the dotted black line (lower halve) corresponds to the recombination rate estimated in the NZ population. The red and blue horizontal lines correspond to "hot" and cold" windows, respectively, i.e. segments in which the observed recombination rate deviates by more than 2.5 standard deviations from the local recombination rate expected under a model of uniform distribution of CO events.

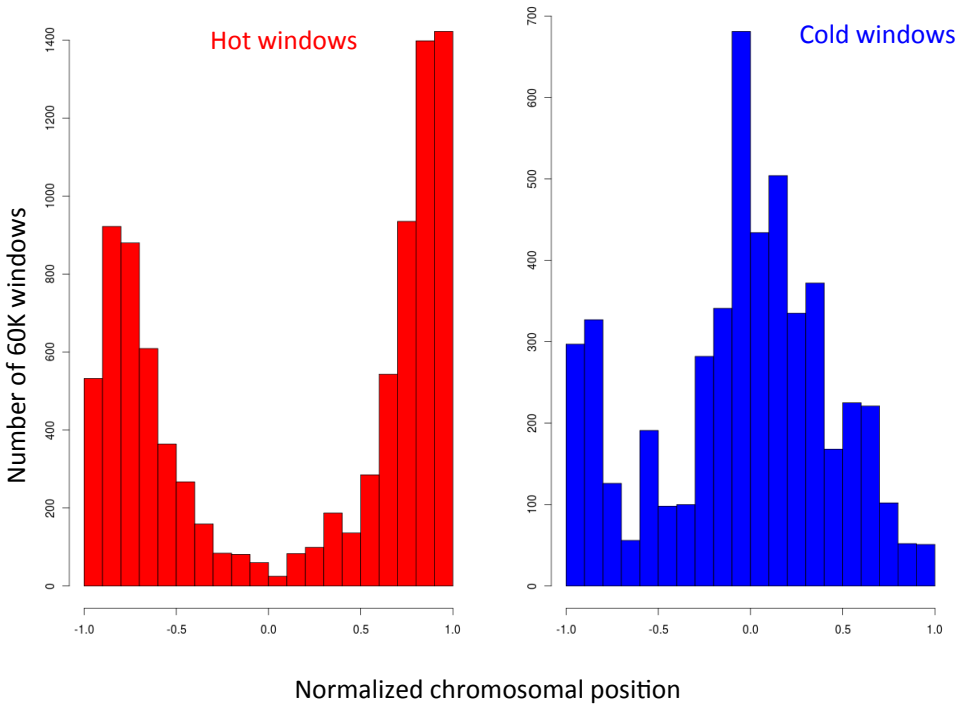

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18



(B) Variation in male recombination in 60Kb windows across the bovine genome. The plain black like (upper halve) corresponds to recombination rate estimated in the Dutch population, while the dotted black line (lower halve) corresponds to the recombination rate estimated in the NZ population. The correlation between window-specific recombination rate in the Dutch and NZ population was high (r^2 =0.80; p<0.0001), despite the use of distinct SNP panels. The red and blue horizontal lines correspond to positions of "hot" and "cold" windows, respectively, i.e. segments in which the observed recombination rate deviates by more than 2.5 standard deviations from the local recombination rate expected under a model of uniform distribution of CO events.

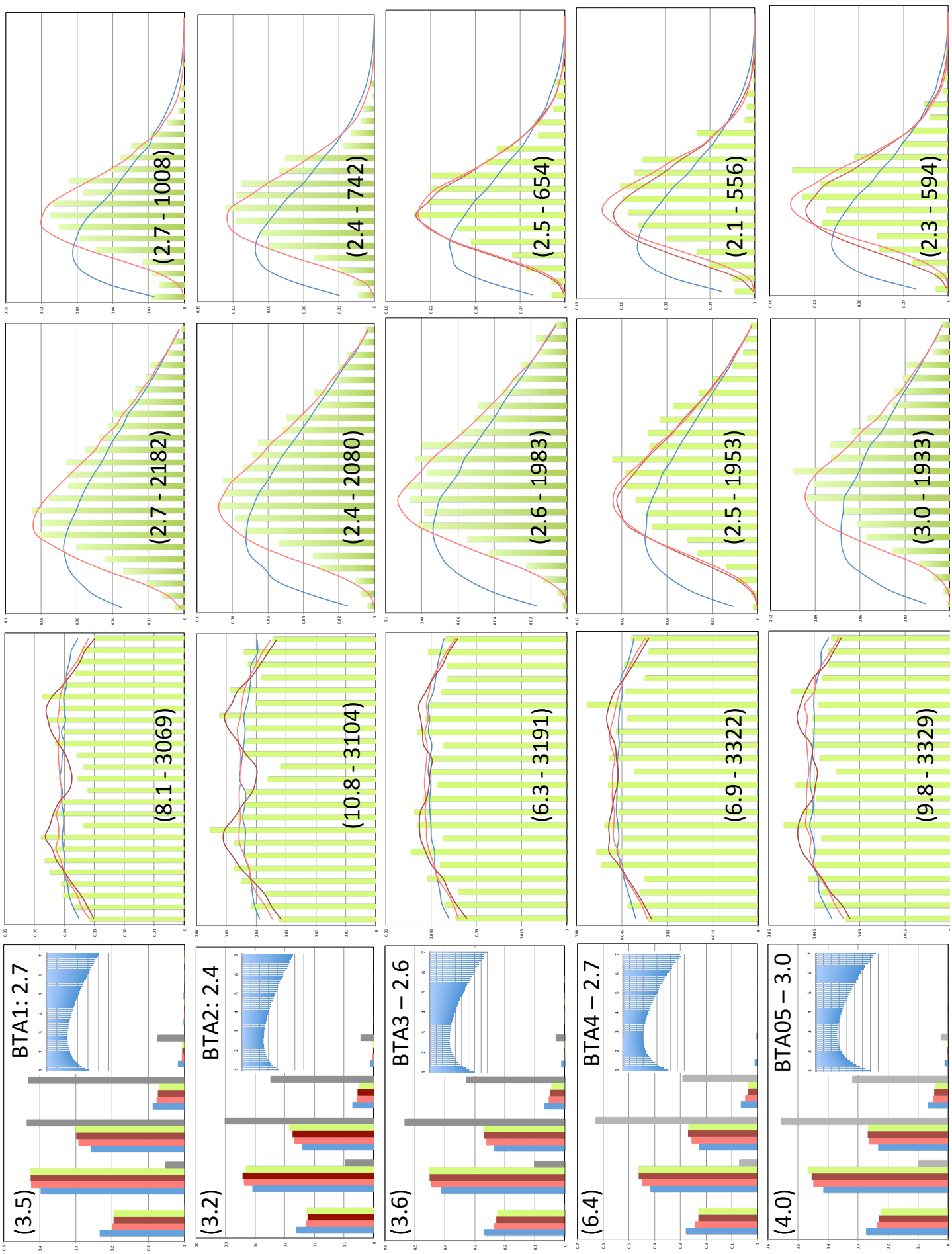


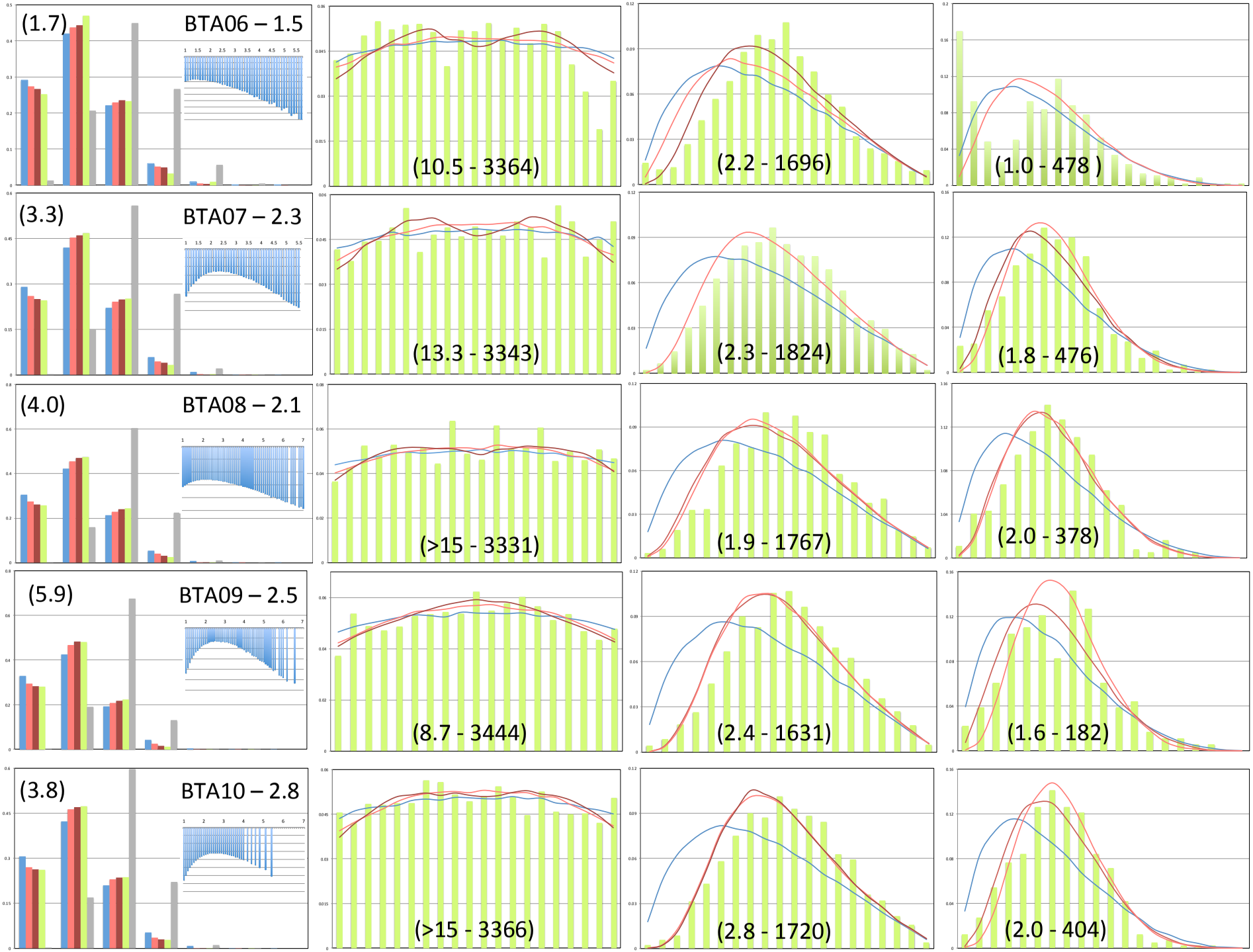
(C) Bar graphs: Frequency distribution of local (60-Kb window) recombination rate normalized for local marker density and informativeness as described in M&M. Curve: Standard normal distribution. Red and Blue vertical lines mark the thresholds defining "hot" (mean + 2.5 SD) and "cold" (mean – 2.5 SD) windows, respectively.

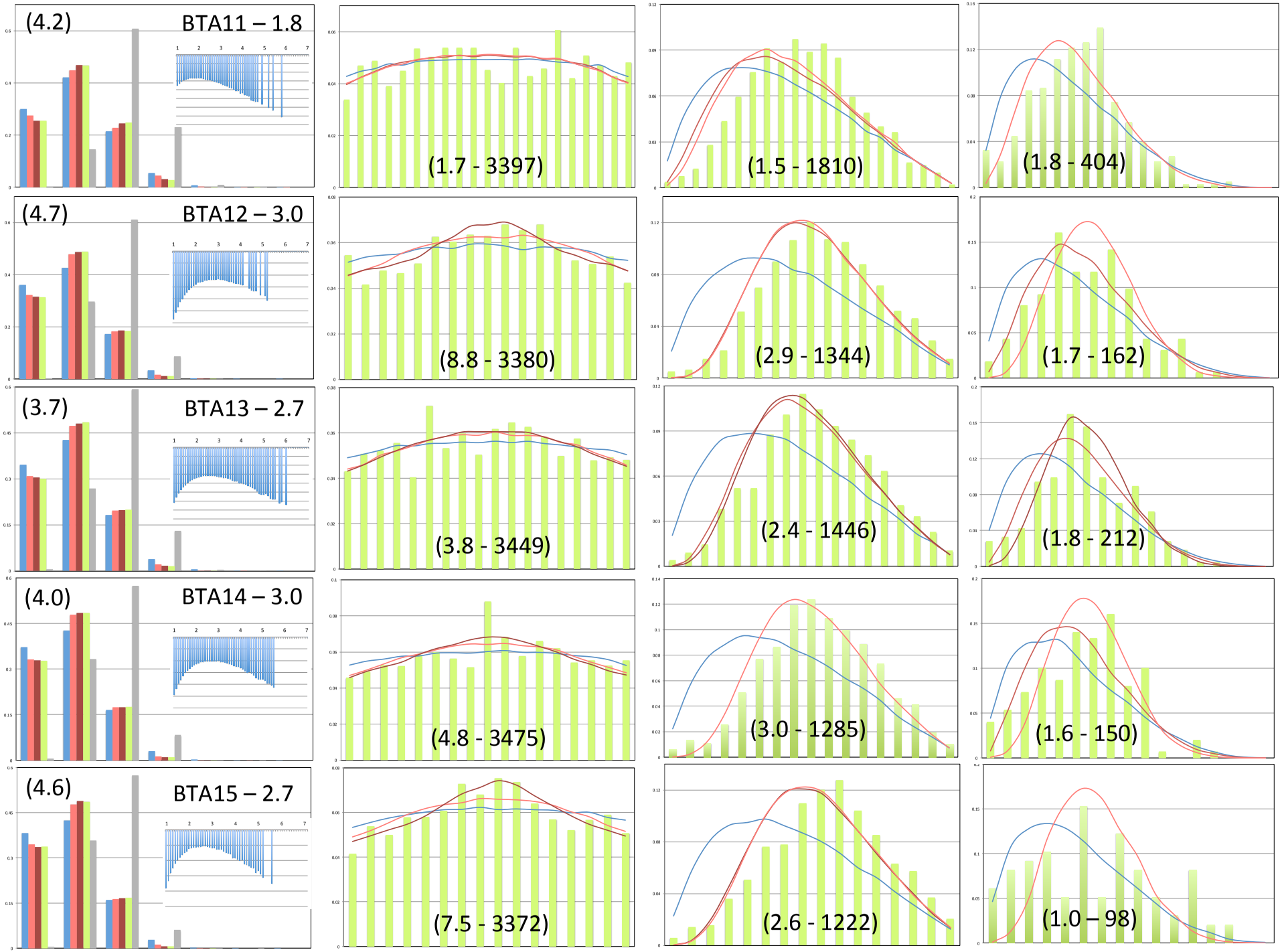


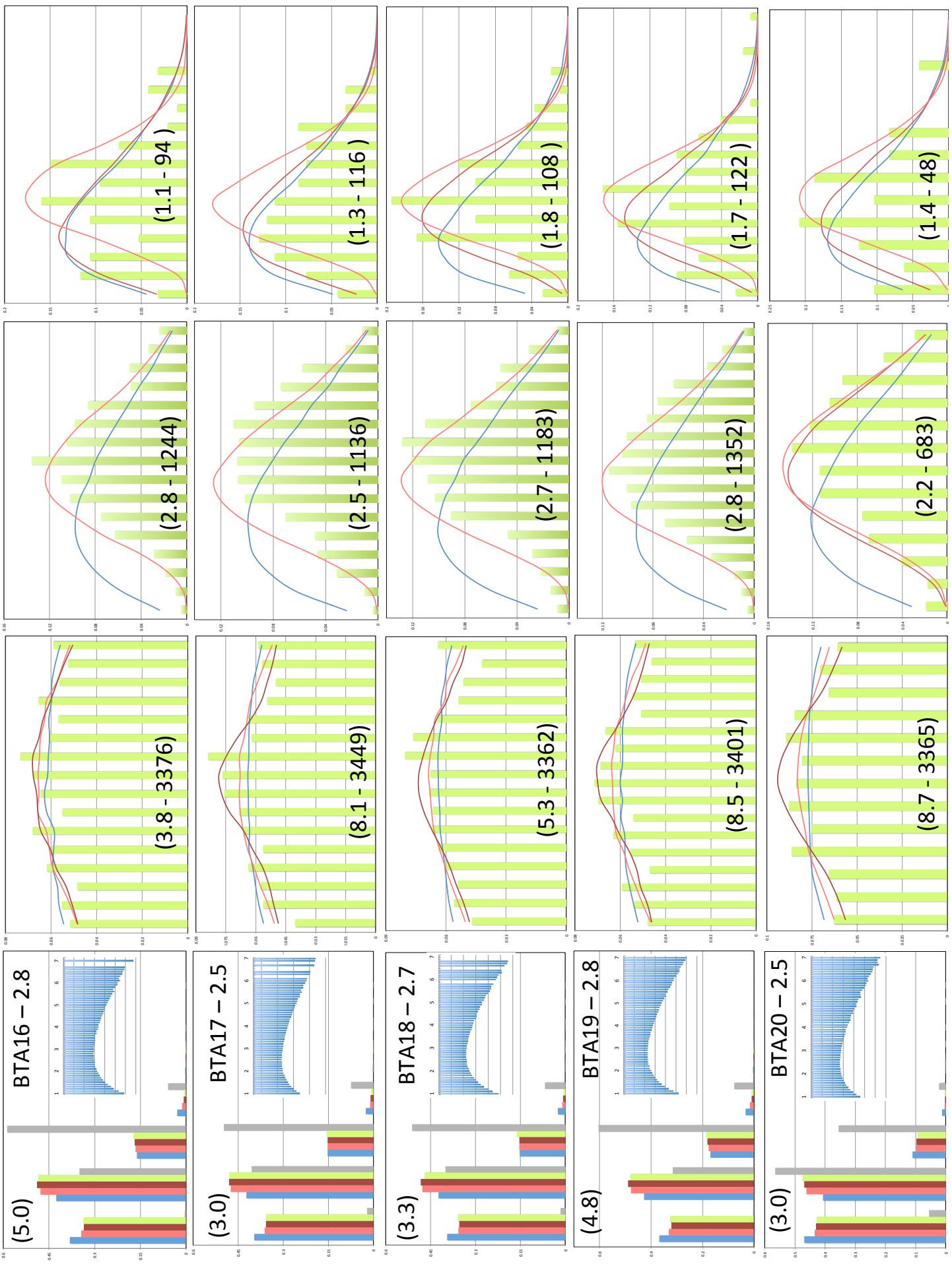
(D) Location of "hot" (red) and "cold" (blue) windows, relative to normalized chromosome length. All 29 acrocentric autosomes were aligned with their centromere towards the left of the graphs. Hot windows tend to concentrate in sub-terminal (proximal chromosome end) and terminal regions (distal chromosome end), while cold windows concentrate in the middle of the chromosome arms as well as in terminal regions (proximal chromosome end) coinciding with the centromeres.

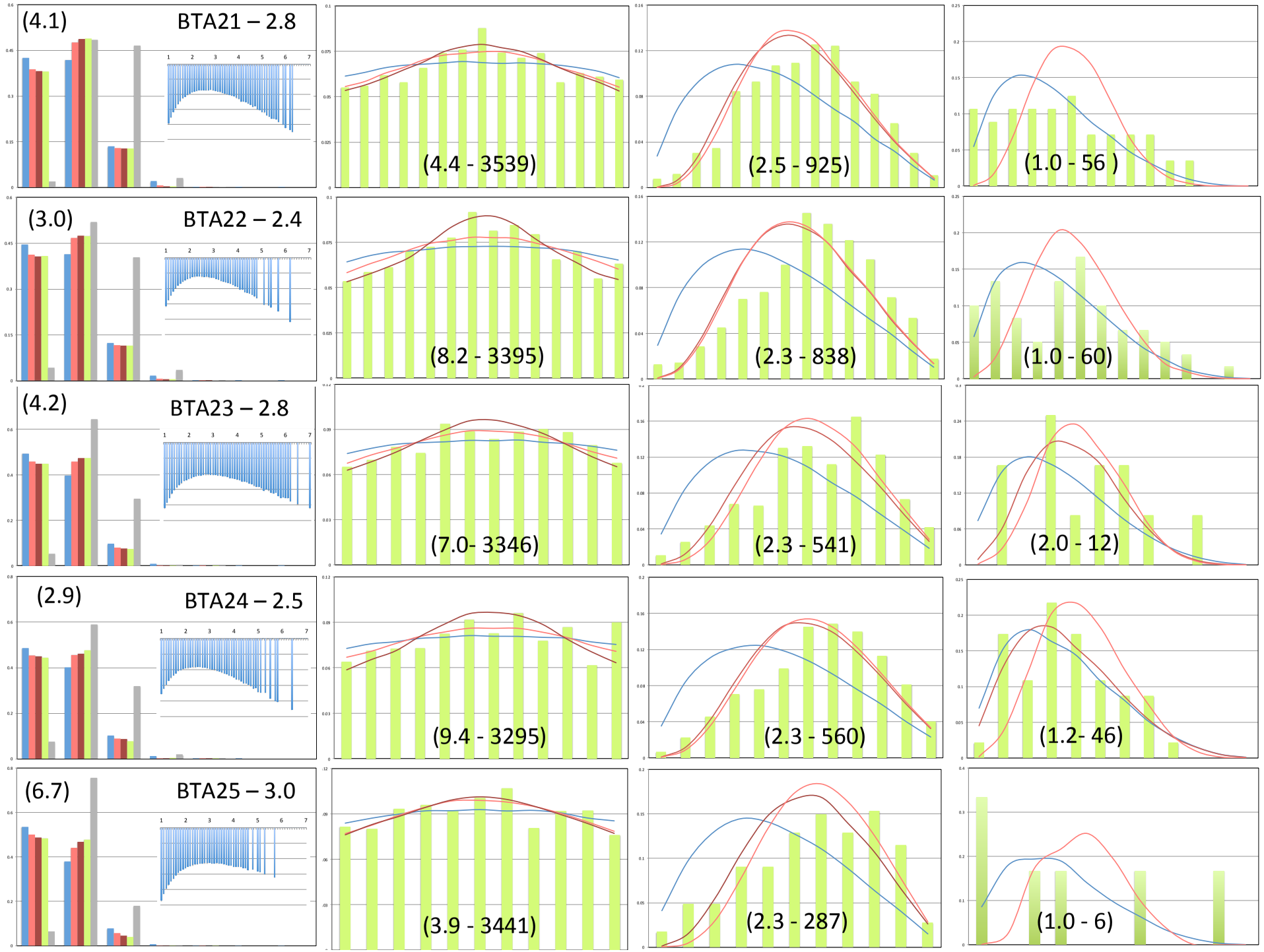
Supplementary Figure 3:

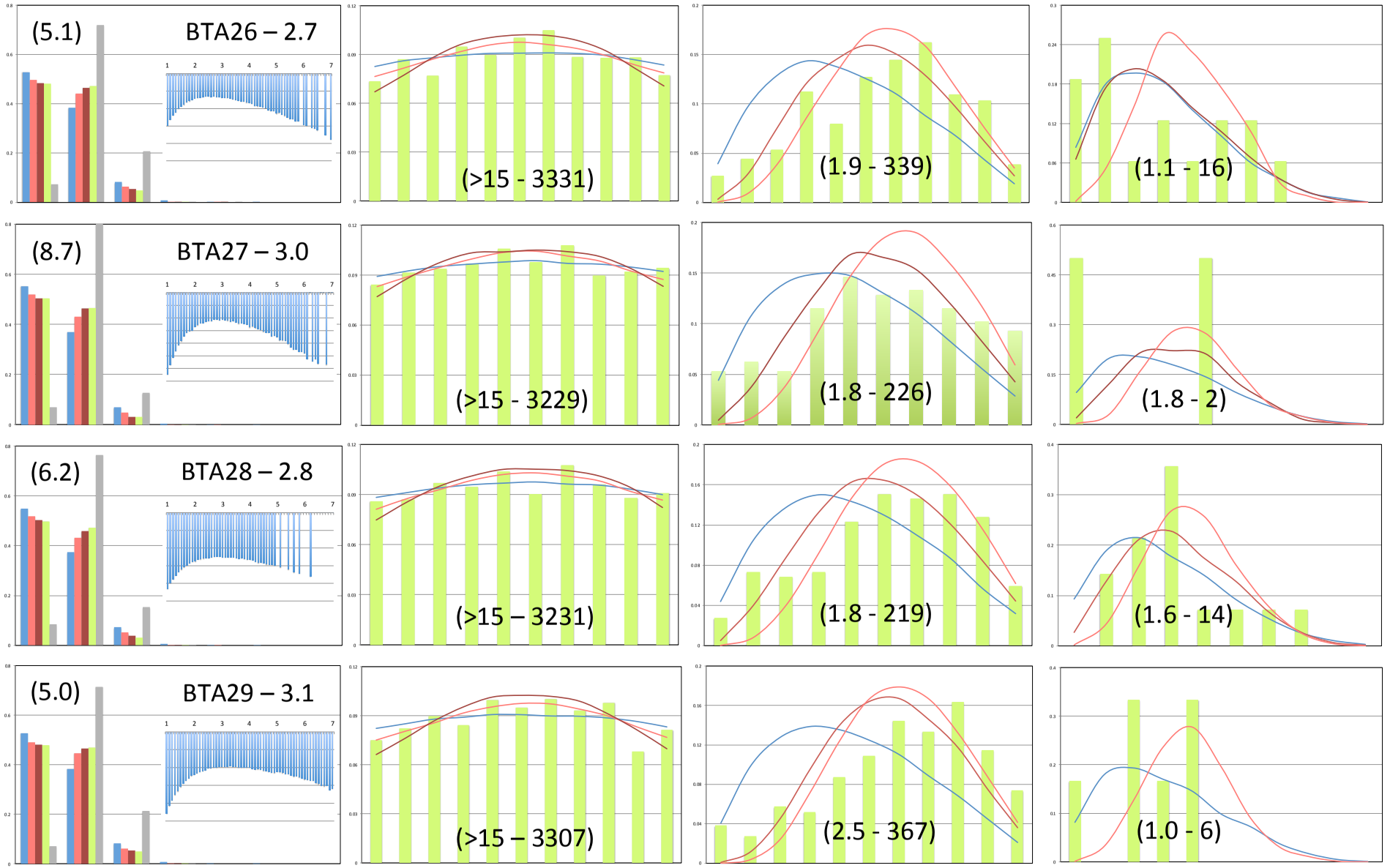



(1.7)
BTA06 – 1.5
(10.5 - 3364)
(2.2 - 1696)
(1.0 - 478)
(3.3)
BTA07 – 2.3
(13.3 - 3343)
(2.3 - 1824)
(1.8 - 476)
(4.0)
BTA08 – 2.1
(>15 - 3331)
(1.9 - 1767)
(2.0 - 378)
(5.9)
BTA09 – 2.5
(8.7 - 3444)
(2.4 - 1631)
(1.6 - 182)
(3.8)
BTA10 – 2.8
(>15 - 3366)
(2.8 - 1720)
(2.0 - 404)


(4.2)
BTA11 – 1.8
(1.7 - 3397)
(1.5 - 1810)
(1.8 - 404)
(4.7)
BTA12 – 3.0
(8.8 - 3380)
(2.9 - 1344)
(1.7 - 162)
(3.7)
BTA13 – 2.7
(3.8 - 3449)
(2.4 - 1446)
(1.8 - 212)
(4.0)
BTA14 – 3.0
(4.8 - 3475)
(3.0 - 1285)
(1.6 - 150)
(4.6)
BTA15 – 2.7
(7.5 - 3372)
(2.6 - 1222)
(1.0 – 98)

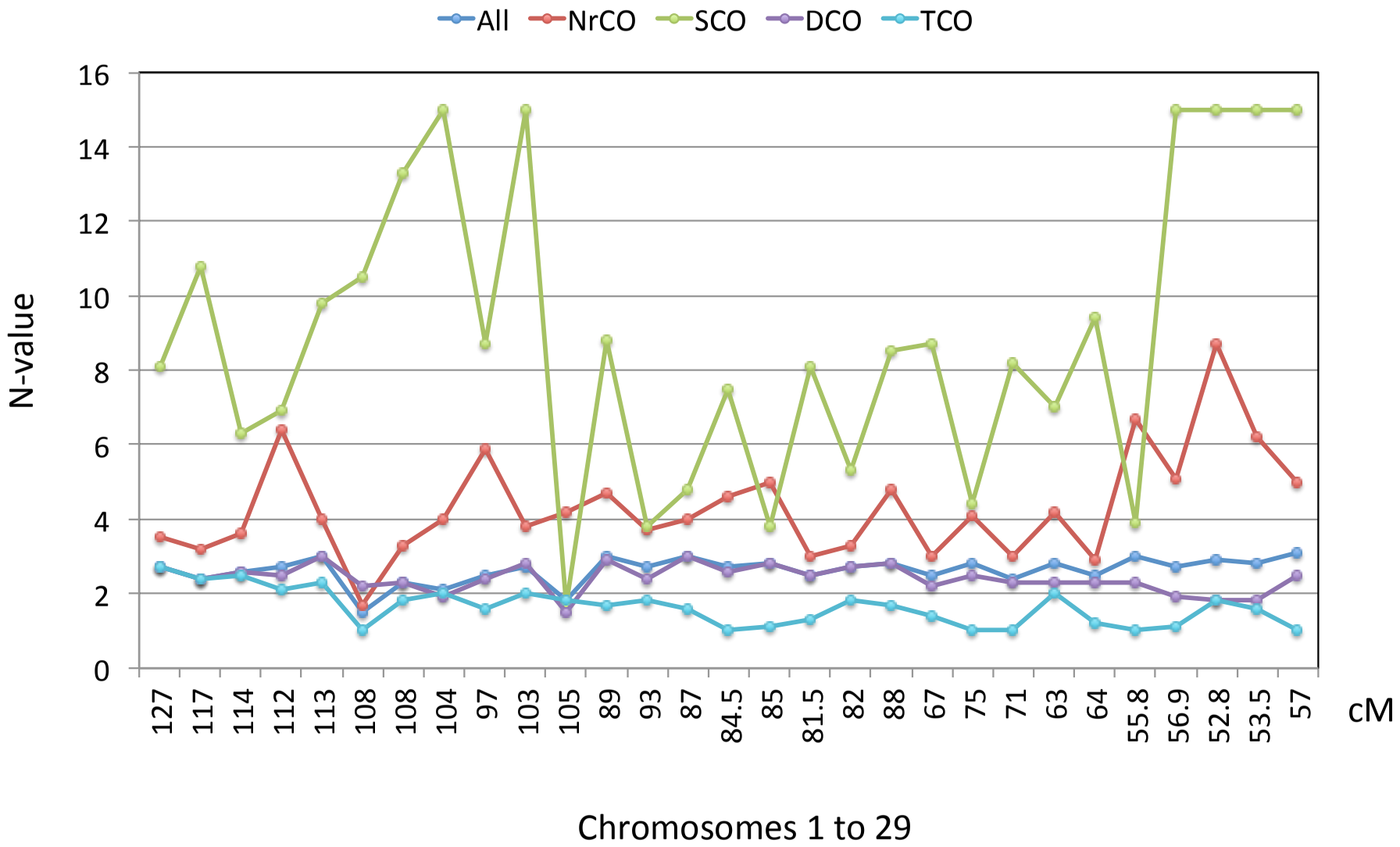

BTA16 – 2.8
(5.0)
(3.8 - 3376)
(2.8 - 1244)
(1.1 - 94)
BTA17 – 2.5
(3.0)
(8.1 - 3449)
(2.5 - 1136)
(1.3 - 116)
BTA18 – 2.7
(3.3)
(5.3 - 3362)
(2.7 - 1183)
(1.8 - 108)
BTA19 – 2.8
(4.8)
(8.5 - 3401)
(2.8 - 1352)
(1.7 - 122)
BTA20 – 2.5
(3.0)
(8.7 - 3365)
(2.2 - 683)
(1.4 - 48)


(4.1)
BTA21 – 2.8
(4.4 - 3539)
(2.5 - 925)
(1.0 - 56)
(3.0)
BTA22 – 2.4
(8.2 - 3395)
(2.3 - 838)
(1.0 - 60)
(4.2)
BTA23 – 2.8
(7.0- 3346)
(2.3 - 541)
(2.0 - 12)
(2.9)
BTA24 – 2.5
(9.4 - 3295)
(2.3 - 560)
(1.2- 46)
(6.7)
BTA25 – 3.0
(3.9 - 3441)
(2.3 - 287)
(1.0 - 6)


(5.1)
BTA26 – 2.7
(>15 - 3331)
(1.9 - 339)
(1.1 - 16)
(8.7)
BTA27 – 3.0
(>15 - 3229)
(1.8 - 226)
(1.8 - 2)
(6.2)
BTA28 – 2.8
(>15 – 3231)
(1.8 - 219)
(1.6 - 14)
(5.0)
BTA29 – 3.1
(>15 – 3307)
(2.5 - 367)
(1.0 - 6)

A

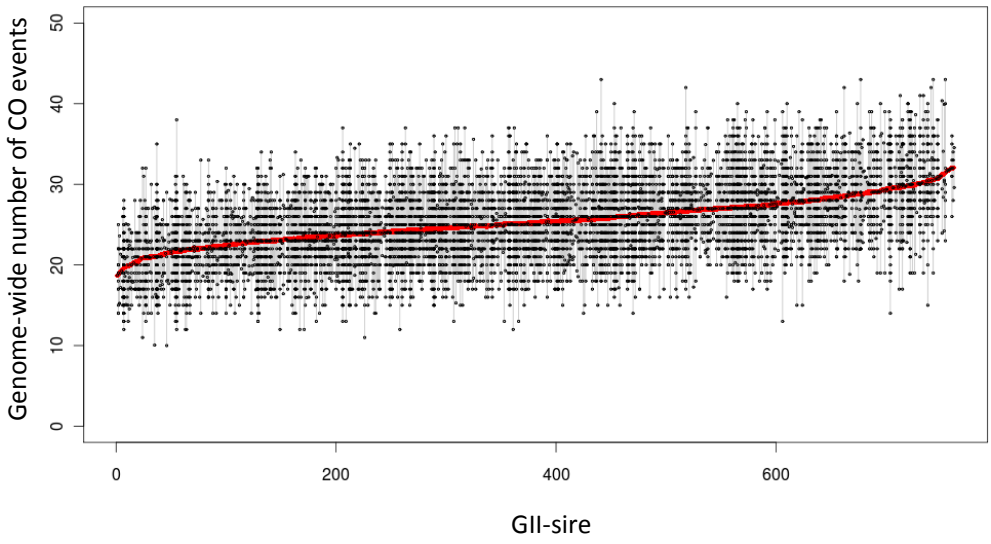
B


All
NrCO
SCO
DCO
TCO
N-value
16
14
12
10
8
6
4
2
0
127
117
114
112
113
108
108
104
97
103
105
89
93
87
84.5
85
81.5
82
88
67
75
71
63
64
55.8
56.9
52.8
53.5
57
cM
Chromosomes 1 to 29

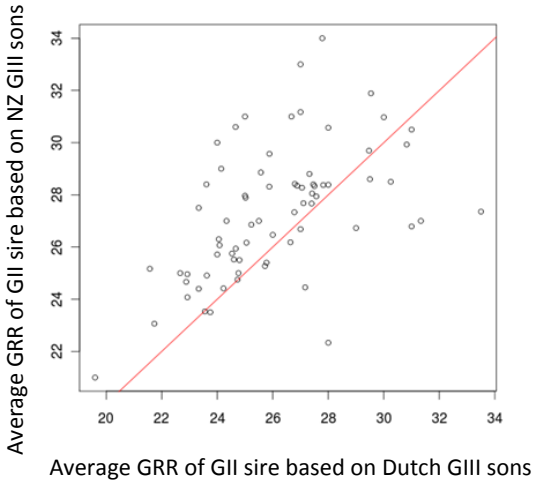
(A) For each of the 29 bovine autosomes (BTA1-29), *column I:* frequency distribution of gametes with 0, 1, 2, … CO-events expected in the absence of cross-over interference (blue), expected given the value of ν maximizing the likelihood of the overall data (light red), expected given the value of given the value of ν maximizing the likelihood of the frequency distribution of CO-events (dark red), as observed (green). The gray bars correspond to the frequency distribution of meiosis with 0, 1, 2, … chiasmata expected given the value of , maximizing the likelihood of the frequency distribution of CO-events. The number following the BTA number corresponds to the , ν maximizing the overall likelihood. The inset illustrates the profile of the $\log_{10}$ of the overall likelihood for varying values of ,. The number in brackets correspond to the ,ν-value maximizing the likelihood of the observed frequency distribution of CO number. *Column II:* Frequency distribution (5 cM bins) of position of single CO-events for gametes with one CO (green bars). The curves correspond to the distributions expected in the absence of interference (blue), assuming the ,ν-value maximizing the overall likelihood (light red), and assuming the ,ν-value maximizing the likelihood of the frequency distribution of single CO-positions (dark red). The numbers between brackets correspond the ν-value maximizing the likelihood of the frequency distribution of single CO-positions, and the number of observed gametes (out of a total of 7,277 used in this analysis) with one CO. *Column III:* Frequency distribution of the distance (5 cM bins) between CO events for gametes with two CO (green bars). The curves correspond to the distributions expected in the absence of interference (blue), assuming the ,ν-value maximizing the overall likelihood (light red), and (if different from the previous ones) assuming the ,ν-value maximizing the likelihood of the frequency distribution of inter-CO distance for gametes with two CO (dark red). The numbers between brackets correspond the ,ν-value maximizing the likelihood of the frequency distribution of inter-CO distance, and the number of observed gametes (out of a total of 7,277 used in this analysis) with two CO. *Column IV:* Frequency distribution of the distance (5 cM bins) between CO events for gametes with three CO (green bars). The curves correspond to the distributions expected in the absence of interference (blue), assuming the ν-value maximizing the overall likelihood (light red), and (if different from the previous ones)

assuming the ν-value maximizing the likelihood of the frequency distribution of inter-CO distance for gametes with three CO (dark red). The numbers between brackets correspond the ν-value maximizing the likelihood of the frequency distribution of inter-CO distance, and the number of observed gametes (out of a total of 7,277 used in this analysis) with three CO. **(B)** Chromosome-specific levels of chiasma interference measured using the shape parameter nu (ν) of a gamma distribution (cfr. M&M). Dark blue (All): ν-value maximizing the likelihood of all data. Red (NrCO): ν-value maximizing the likelihood of the frequency distribution of CO-events per gametes. Green (SCO): ν-value maximizing the likelihood of the frequency distribution of CO-position (in cM) for gametes with one CO. Purple (DCO): ν-value maximizing the likelihood of the frequency distribution of inter-CO distance (in cM) for gametes with two CO. Light blue (TCO): ν-value maximizing the likelihood of the frequency distribution of inter-CO distance (in cM) for gametes with three CO. Chromosomes are ordered (left to right) from 1 to 29. The numbers under the X-axis correspond to the size of the corresponding chromosome in cM.

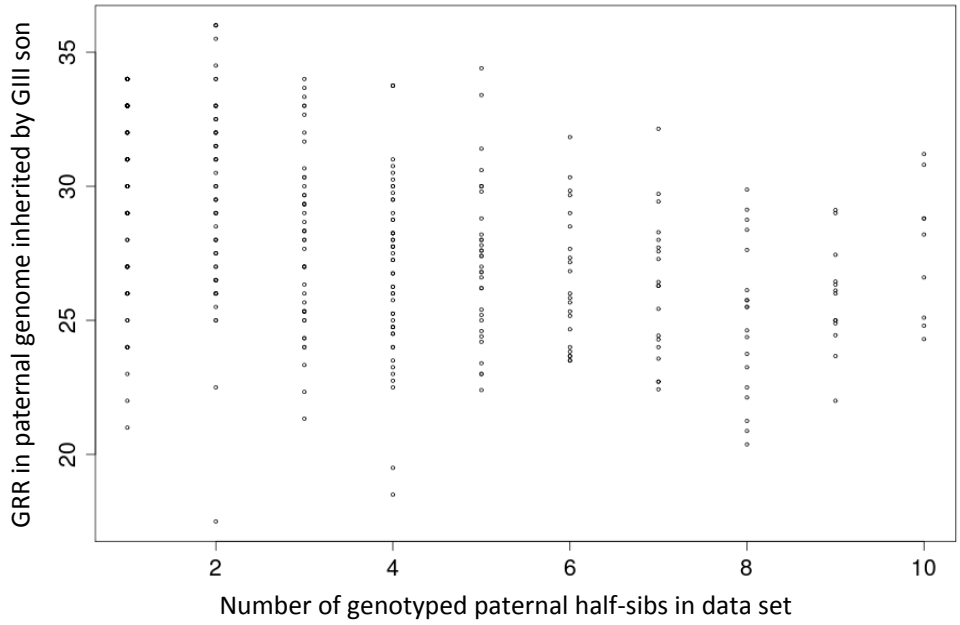
Supplementary Figure 4:



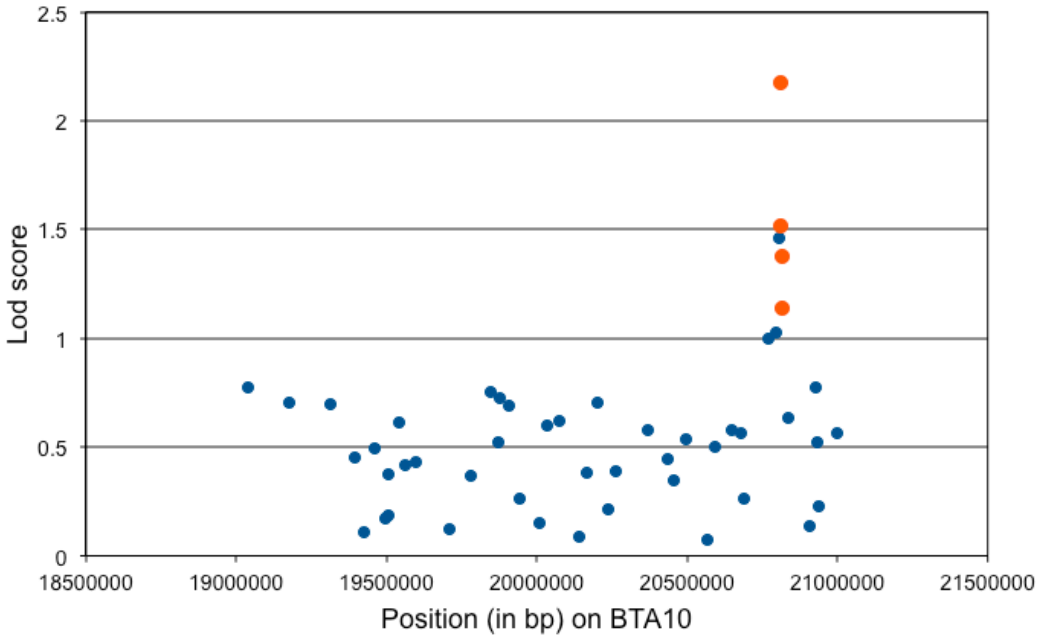
(A) Black dots correspond to the total number of CO events identified in the paternal genome of 10,192 GIII sons sorted by GII sire. The red dots mark the average GRR for each GII sire. GRR did not differ significantly between Holstein-Friesian and Jersey bulls.



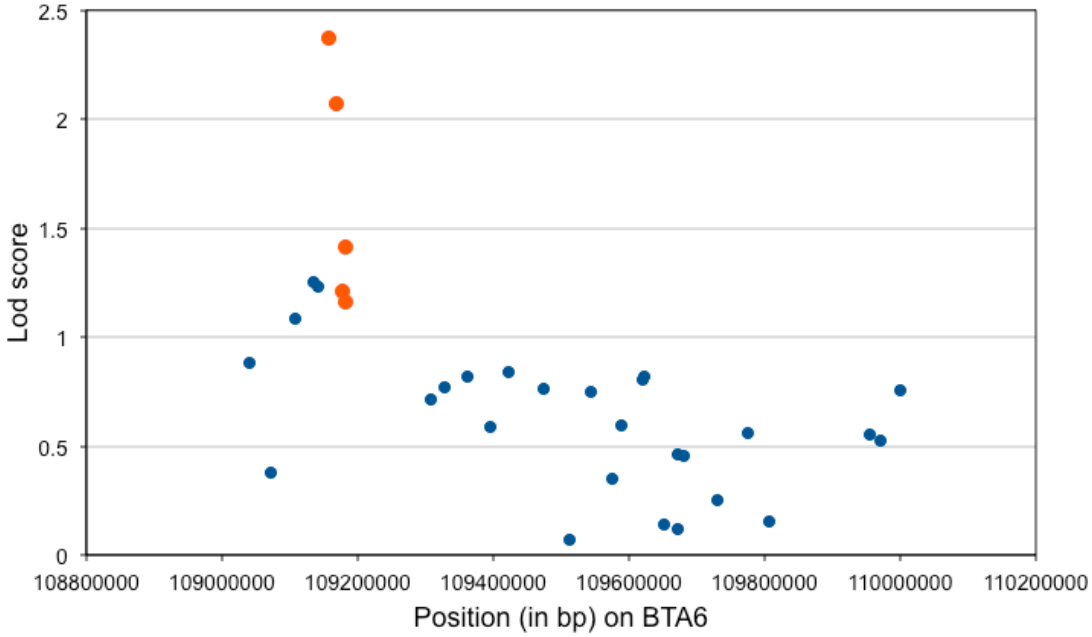
(B) Correlation between the GRR estimated for 72 GII sires separately from the number of CO events transmitted to non-overlapping sets of GIII sons from H and NZ, respectively. Spearman's rank correlation was 0.58 ($p < 3.7 \times 10^{-7}$).



(C) Total number of CO events (GRR) in the genome transmitted by GII sires to their GIII sons. GIII sons are sorted according to the number of half-brothers in the data set. The increase of GRR with decreasing family size is clearly visible.

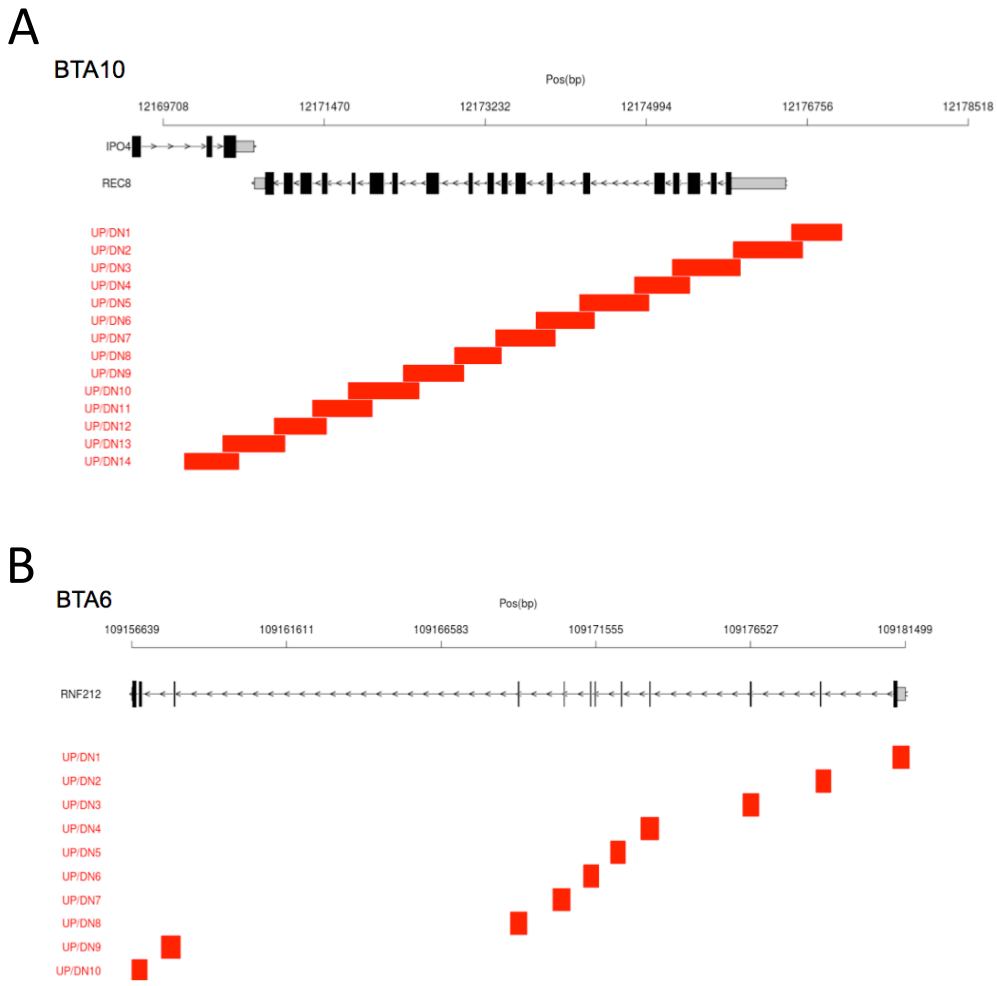


(D) Lod scores obtained for GRR using 121 HF GII sires, and (i) 45 SNPs from the Illumina bovine high-density 777K SNP array mapping to the confidence interval of the BTA10 QTL (blue dots) and (ii) *REC8* SNPs (red dots). The highest lod score was obtained for *REC8* variant ss418642854.



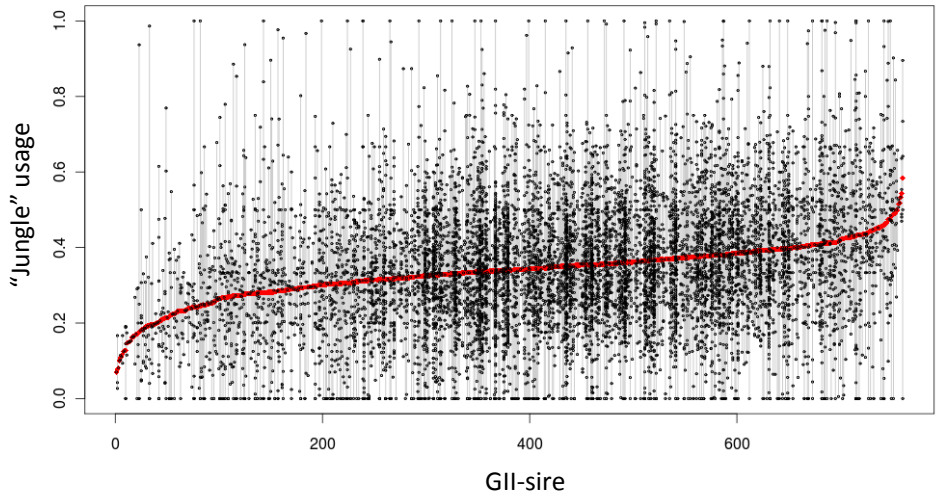
(E) Lod scores obtained for GRR using 121 HF GII sires, and (i) 27 SNPs from the Illumina bovine high-density 777K SNP array mapping to the confidence interval of the BTA6 QTL (blue dots) and (ii) *RNF212* SNPs (red dots). The highest lod score was obtained for *RNF212* variant ss469104611 (=P259S) .

Supplementary Figure 5:

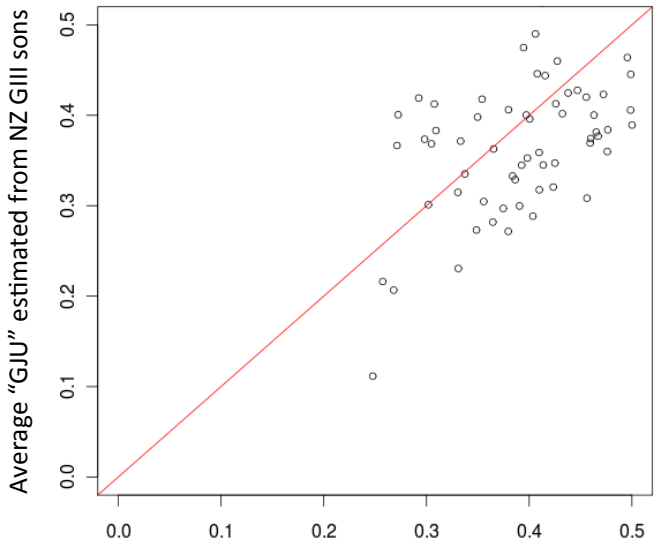


Position of the amplicons used to scan the *REC8* **(A)**, and *RNF212* genes **(B)** (cfr. Suppl. Table 2). The corresponding *RNF212* gene model has been submitted to Genbank.

Supplementary Figure 6:

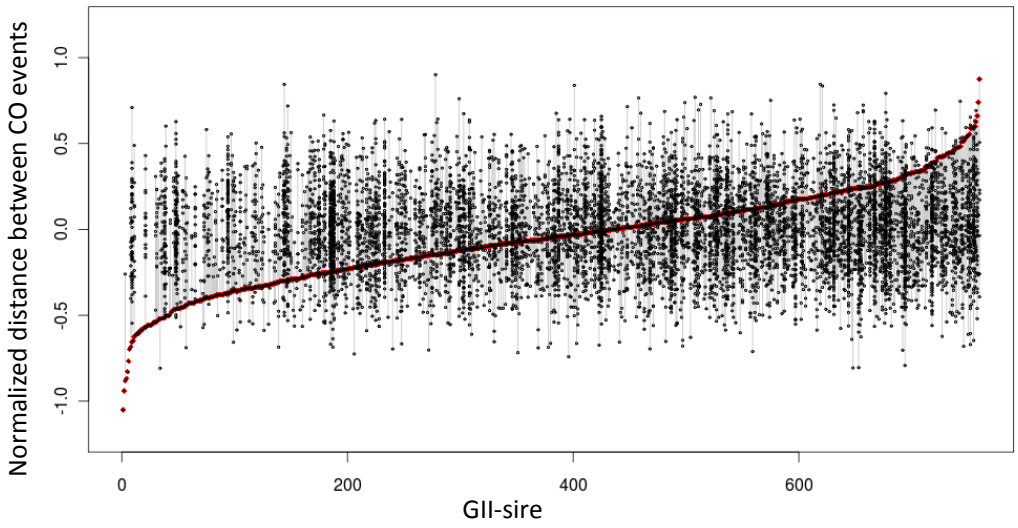


(A) Black dots: Average overlap (0 to 1) between marker intervals (< 800-Kb) with assigned CO events and "hot" 60-K windows for GIII-sons sorted by GII-sire. Red dots: Average overlap for all CO events transmitted by corresponding GII-sire.

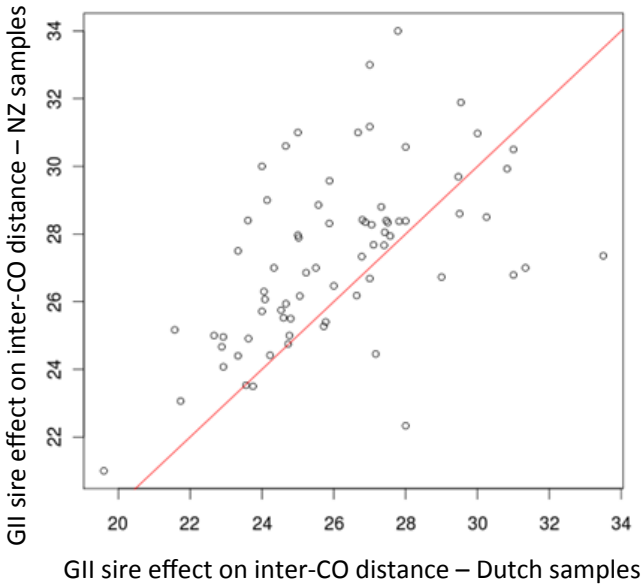


(B) Correlation between average hot-window usage estimated for the 72 shared GII-sires respectively from gametes transmitted to Dutch versus New-Zealand GIII sons.

Supplementary Figure 7:



(A) GIII sons inherit chromosomes with 0, 1, 2, 3, ... CO from their GII sires. In this analysis, we only use "di-CO" chromosomes (i.e. with 2 CO). We measure the distance between CO-pairs in normalized, chromosomes-specific units. Thus, the distance between the CO-pair of the di-CO chr. 1 inherited by son *x* from sire *y*, may be "so many" standard deviations above or below the average distance between CO-pairs on di-CO chr. 1's (across all GIII sons receiving a di-CO chr. 1 from their sire). The black dots correspond to the average of the normalized distances between CO-pairs for all di-CO chromosomes inherited by a given GIII son. GIII sons are sorted by GII sire, i.e. they are on the same vertical black line. The red dots correspond to the average of the normalized inter-CO distances across all di-CO chromosomes transmitted by a given GII sire to all its GIII sons.



(B) Correlation between average normalized distance between CO events for all homologues with two recombination events transmitted by 72 shared GII-sire to (i) their Dutch GIII-sons, and (ii) their NZ GIII-sons.

Supplementary Table 1:

	Simple Regression				Multiple Regression		
Predictor Var.	Reg. Coef.	Std. Error	R^2	Pr(>|t|)	Reg. Coef.	Std. Error	Pr(>|t|)
%G:C/60Kb	0.00157	0.00003	0.17140	<2e-16	-0.00043	0.00011	0.00006
%CpG/60Kb	0.01361	0.00026	0.23870	<2e-16	0.00941	0.00056	< 2e-16
N° of Poly(A):Poly(T)/60Kb	-0.00256	0.00006	0.16158	<2e-16	-0.00085	0.00014	1.22e-09
N° of genes/60Kb	-0.00316	0.00011	0.01963	0.16600	-0.00004	0.00409	< 2e-16
					R^2	0.28	

(A) Following Kong et al.[8], we tested the effect of base pair composition and gene content on LRR by multiple regression. As in human, local recombination rate was positively correlated with CpG content, yet negatively correlated with GC, polyA/polyT and gene content (after adjustment for CpG content). CpG content accounted for ~19% of the variance, while the four parameters explained ~28% jointly.

JUNGLES			
Name	**Ratio**	**Pval**	**Pval***
SINE/BovA	1.12	0.00000	0.00000
SINE/RTE-BovB	1.09	0.00000	0.00000
SINE/tRNA-Glu	0.91	0.00000	0.00000
LTR/ERVK	0.87	0.00000	0.00000
LINE/RTE-BovB	0.93	0.00000	0.00000
LINE/L2	0.95	0.00000	0.00000
LINE/L1	0.98	0.00003	0.00175
rRNA	0.58	0.00005	0.00272
LINE/CR1	0.92	0.00026	0.01481
DNA/hAT-Tip100?	0.51	0.00029	0.01685
DNA/hAT-Charlie	0.95	0.00029	0.01689
LTR/ERVL-MaLR	1.04	0.00051	0.02901
DNA/MER2_type	1.11	0.00171	0.09449
DNA/AcHobo	1.25	0.00446	0.22836
RC/Helitron?	2.23	0.01355	0.54686
LTR?	0.29	0.01842	0.65988
LTR/ERV	0.54	0.02014	0.69268
LTR/ERVL?	0.72	0.03668	0.88552
DNA	1.30	0.03878	0.89912
DNA/MER1_type	0.95	0.04335	0.92347
LTR/Gypsy?	0.85	0.04941	0.94709
DNA/PiggyBac	1.70	0.05676	0.96626
DNA/hAT?	0.76	0.05685	0.96646
LTR/ERV1	0.97	0.05707	0.96691
RC/Helitron	1.32	0.05926	0.97108
DNA/hAT-Blackjack	1.09	0.06750	0.98264
LINE/Dong-R4	0.64	0.07095	0.98600
DNA/hAT	1.09	0.07303	0.98770
LTR	0.79	0.07962	0.99187
tRNA	0.85	0.07963	0.99188
Unknown	1.14	0.08963	0.99569

DESERTS			
Name	**Ratio**	**Pval**	**Pval***
SINE/MIR	0.69	0.00000	0.00000
LINE/RTE-BovB	1.34	0.00000	0.00000
SINE/BovA	0.84	0.00000	0.00000
LINE/L2	0.74	0.00000	0.00000
LTR/ERVK	1.45	0.00000	0.00000
LTR/ERV1	1.41	0.00000	0.00000
LTR/ERVL-MaLR	0.74	0.00000	0.00000
Satellite/centr	7.47	0.00000	0.00000
DNA/hAT-Charlie	0.77	0.00000	0.00000
tRNA	3.88	0.00000	0.00000
LINE/CR1	0.64	0.00000	0.00000
DNA/MER1_type	0.67	0.00000	0.00000
LINE/L1	1.06	0.00000	0.00000
SINE/tRNA-Glu	0.91	0.00000	0.00001
SINE/RTE-BovB	1.06	0.00000	0.00012
RC/Helitron	0.28	0.00001	0.00064
LTR/Gypsy?	0.52	0.00002	0.00136
DNA/TcMar-Tigger	0.81	0.00003	0.00158
LTR/Gypsy	0.60	0.00005	0.00255
DNA	0.39	0.00014	0.00774
LINE/RTE	0.75	0.00027	0.01487
DNA/TcMar?	0.43	0.00035	0.01923
DNA/TcMar-Mariner	0.61	0.00053	0.02906
LTR/ERVL	0.91	0.00079	0.04305
SINE?	0.08	0.00134	0.07237
DNA/MER2_type	0.83	0.00138	0.07438
SINE	0.49	0.00166	0.08874
SINE/tRNA	0.31	0.00202	0.10716
DNA/Tip100	0.75	0.00279	0.14486
Unknown	0.65	0.00301	0.15536
DNA?	0.44	0.00555	0.26760

DNA/PiggyBac?	1.77	0.09558	0.99705	DNA/Tc2	0.63	0.00579	0.27749
snRNA	0.84	0.10272	0.99814	DNA/AcHobo	0.64	0.00855	0.38188
DNA/TcMar	0.63	0.11608	0.99922	DNA/hAT-Blackjack	0.80	0.00873	0.38787
DNA/MuDR	2.00	0.12663	0.99961	DNA/TcMar-Tc2	0.69	0.02697	0.78372
DNA/TcMar?	0.86	0.14655	0.99990	DNA/Charlie	0.31	0.02905	0.80811
Satellite/centr	0.91	0.16233	0.99997	DNA/hAT?	0.45	0.04109	0.90458
DNA?	1.19	0.18148	0.99999	DNA/Tigger	0.65	0.04401	0.91958
DNA/Tip100	0.94	0.21765	1.00000	DNA/PiggyBac	0.48	0.04819	0.93709
DNA/TcMar-Tc2	0.90	0.22732	1.00000	snRNA	1.40	0.05108	0.94692
LINE/RTE	1.05	0.24914	1.00000	DNA/hAT	0.82	0.05263	0.95158
SINE	1.13	0.26638	1.00000	LINE/Dong-R4	0.42	0.08956	0.99477
LTR/ERVL	1.01	0.36859	1.00000	RC/Helitron?	0.33	0.15730	0.99993
SINE?	1.27	0.39614	1.00000	LTR/ERV	0.50	0.15730	0.99993
DNA/TcMar-Tigger	0.98	0.41560	1.00000	SINE/Deu	0.79	0.17319	0.99998
DNA/TcMar-Mariner	0.95	0.47814	1.00000	DNA/hAT-Tip100?	0.65	0.25684	1.00000
LINE/L1?	0.60	0.47950	1.00000	LTR	0.76	0.29351	1.00000
srpRNA	0.60	0.47950	1.00000	DNA/TcMar	2.00	0.31731	1.00000
DNA/Tigger	1.10	0.53465	1.00000	DNA/Mariner	1.24	0.32897	1.00000
LTR/Gypsy	0.96	0.55289	1.00000	DNA/MuDR	0.60	0.47950	1.00000
DNA/hAT-Tip100	0.98	0.55630	1.00000	LTR/ERVL?	0.82	0.48384	1.00000
RNA	1.14	0.66168	1.00000	DNA/PiggyBac?	0.50	0.56370	1.00000
DNA/Charlie	1.08	0.69206	1.00000	RNA	0.75	0.59298	1.00000
SINE/Deu	0.96	0.70420	1.00000	rRNA	1.10	0.65869	1.00000
DNA/Mariner	0.96	0.80837	1.00000	LTR?	0.80	0.73888	1.00000
SINE/MIR	1.00	0.86180	1.00000	DNA/hAT-Tip100	1.02	0.80948	1.00000
SINE/tRNA	0.98	0.93038	1.00000				
DNA/Tc2	1.00	1.00000	1.00000				

(B) We tested whether "hot" and "cold" status correlated with window content in specific interspersed repeats. For each 60-Kb hot (respectively cold) window, we sampled a "regular" window matched for CpG, GC, polyA/polyT and gene content, and compared total counts of 58 types of

interspersed repeats. The statistical significance of the count difference was evaluated by permutation with Bonferroni correction for the realization of 58 independent tests. As can be seen from the table (i) some repeat types were enriched in hot and depleted in cold windows, including SINE/BovA (ratio Jungle/control: 1.12; ratio desert/control: 0.84), LTR/ERVL-MaLR (1.04;0.74), RC/Helitron (2.23;0.74) and DNA (1.30;0.75), (ii) some repeats were depleted in hot and enriched in cold windows, including LTR/ERVK (0.87;1.45), LINE/RTE-BovB (0.93;1.34) and LINE/L1 (0.98;1.06), (iii) SINE/RTE-BovB were enriched in hot and cold windows (1.12;1.06), (iv) some repeat types were depleted in hot and cold windows including SINE/tRNA-Glu (0.91;0.91), LINE/L2 (0.95;0.74), LINE/CR1 (0.92;0.64), DNA:hAT-Charlie (0.95;0.77) and DNA/MER1_type (0.95;0.67), (v) rRNA were depleted in hot windows, and (vi) SINE/MIR and satellite/centr were depleted in cold windows.

Supplementary Table 2: Primers used for amplification and resesequencing of candidate genes *REC8*, *RNF212, KCNJ2, KCNJ16* and gonosomal *PRDM9-XA and -XB.*

REC8	Primer sequence (5'->3')	Product size(bp)	Exon covered
UP1_REC8	TGCAGGAGGCATGAGGAAAG	565	-
DN1_REC8	GTGAGGTGTCTGGACCAATGG		
UP2_REC8	CTGCTAGTTGAAGCACCCTTG	772	-
DN2_REC8	ACCCTCGTCTACACCACTTCC		
UP3_REC8	GTCCACATCCTTCTCTCCAAC	756	exon 1, 2, 3, 4
DN3_REC8	GGATGCACTCACAGTTCAGTC		
UP4_REC8	GGTCTACTCTCAACAATGCCAG	620	exon 3, 4, 5
DN4_REC8	GGAATGGTGTGAAGACAAAGC		
UP5_REC8	AGGGGCATGCTATTTCTGTG	772	exon 6
DN5_REC8	GCGTTCCTCTGTCTCTCACC		
UP6_REC8	CAGGGCACTGAAGCTGAATC	651	exon 6, 7
DN6_REC8	GGGTAGAGAAATGGAGGCAAG		
UP7_REC8	AACAGAGGCTGTGTGACTTCC	667	exon 7, 8, 9
DN7_REC8	AATGGGCAAGAGGATGAAAAC		
UP8_REC8	CTTGTAGGGTAAGGGCAGGAG	525	exon 10, 11
DN8_REC8	CATCTGGGAAGTTTAGGCTGG		
UP9_REC8	TCACCCATTTTCAGTTTCCTC	675	exon 12
DN9_REC8	CACCTGTAATGCCTCGTTCC		
UP10_REC8	GGGAATCTTGAGAACAGTCC	788	exon 13, 14, 15

DN10_REC8	AGAGTGGTGGTTTCCGTTG		
UP11_REC8	GAGGAAAGGAGAAAGGCTGAG	667	exon 14, 15, 16
DN11_REC8	CCTAGAAGAACACACGAGTTGG		
UP12_REC8	CTGGAAGTAGCTGAGGAGGAG	586	exon 16, 17, 18
DN12_REC8	CACTGGTCAACAGAGGAGGAG		
UP13_REC8	CGTGTCTTCTACCTGCTCCTG	694	exon 19
DN13_REC8	CCTCTTCACTCTCCACTCAGAC		
UP14_REC8	CAGGCTTTGTTCTCTCCAGC	608	-
DN14_REC8	CCAGGATGTCAATTATCTGTGG		
RNF212	**Primer sequence (5'->3')**	**Product size(bp)**	**Exon covered**
UP1_RNF212	CACCACCACCACAGAACAAGG	573	exon 1
DN1_RNF212	CCAGGAAGCTCGCCTCAG		
UP2_RNF212	GGGAGCTTTCATTTTGTAGTGC	520	exon 2
DN2_RNF212	ACAGAAAGTCTGCACGATCTCC		
UP3_RNF212	GATTCCACCGATTGCTTTGTAC	548	exon 3
DN3_RNF212	CAGGAGACACAGCGTGAAGAGG		
UP4_RNF212	GTTGTGTCCTGTCCTAAGTCC	612	exon 4
DN4_RNF212	AAGCAAATGGGCACTAAACC		
UP5_RNF212	ATGTCACTTGCATTCTGATTCC	513	exon 5
DN5_RNF212	AACAAGTTGGGTGCTGGTGTG		
UP6_RNF212	GTGTCATGGAGAAATGGGTC	525	exon 6, 7
DN6_RNF212	CAAGTGCAACCTTCCAAGAG		
UP7_RNF212	TGTCTGCTTTATGGGCTTTGC	593	exon 8
DN7_RNF212	ACAGCACAGGGAGAAAACTGC		

UP8_RNF212	CTAATGTCAGCCAAGAACTTCG	568	exon 9
DN8_RNF212	ACTTCCAGAGAGGGACTGAGC		
UP9_RNF212	CAGATGTGGACTGAGAGCTG	645	exon 10
DN9_RNF212	GTAGACAGGCTCTGGGTGAG		
UP10_RNF212	GGGTCACCACAGTCCAGAGT	532	exon 11, 12
DN10_RNF212	GCTGCCTGTAAGGAGGTTCT		
KCNJ2	**Primer sequence (5'->3')**	**Product size(bp)**	**Exon covered**
UP1_KCNJ2	GTTTCTGGCTAGGAACTGTTGC	618	No introns
DN1_KCNJ2	AACACACAGCCGAAGAAGAGC		
UP2_KCNJ2	GAATGGCAAGAGCAAAGTCC	707	No introns
DN2_KCNJ2	TGGGAGACACCAGAAATATGC		
UP3_KCNJ2	AAGCCCAAAAAGAGAAACGAG	767	No introns
DN3_KCNJ2	GTGTTTCAGAGGAAGCACTCC		
UP4_KCNJ2	GCAAAGAGGAAGACGACAGTG	676	No introns
DN4_KCNJ2	GCTAATAGCCAAATACTTGCCAC		
UP5_KCNJ2	AGCTGCTAACTACACCAACACG	713	No introns
DN5_KCNJ2	TCTAAGGGGTCATCATCATCG		
UP6_KCNJ2	GAACTACCCTCTCCTTTTGCAC	681	No introns
DN6_KCNJ2	ACGCTGTTAGGATTGTTGTGG		
UP7_KCNJ2	ATGATAGAGAGCAGGAATGAA	605	No introns
DN7_KCNJ2	CATATCATCCCCTCGCTTG		
UP8_KCNJ2	CTCTTGTCTGCCTTTCGAG	668	No introns
DN8_KCNJ2	AGCATACAGGTACTAGCTCTC		
KCNJ16	**Primer sequence (5'->3')**	**Product size(bp)**	**Exon covered**

UP1_KCNJ16	GCCGTGAAAACCTTAGCAAC	762	exon 1
DN1_KCNJ16	TGTTCCAAAGGGTGGTGAG		
UP2_KCNJ16	CTCGAAAGACGGGAATTGAA	650	exon 2
DN2_KCNJ16	TGATGAAGGTGTTGATGACG		
UP3_KCNJ16	TCTGGCTCATAGCCTTCCAC	771	exon 2
DN3_KCNJ16	CTTTCCAGTCGAGCTGCTTG		
UP4_KCNJ16	TTCATCTACACCGGGGACTC	666	exon 2
DN4_KCNJ16	GCTTCTTGATAGCCTCCCTA		
UP5_KCNJ16	TGTGACACCTGTCTTTGATGG	529	exon 2
DN5_KCNJ16	GTGGCATGTAGGCTCTGTCA		
PRDM9 (chr X)	**Primer sequence (5'->3')**	**Product size(bp)**	**Exon covered**
UP1_PRDM9_XA	GAATAAGAATGGAATGGATAAG	1609	ZF arrays
Seq1_PRDM9_XA	AGCACAAGCCAGAAACTGA		
Seq2_PRDM9_XA	CTGTGGCTAGGATGTGGTT		
DN1_PRDM9_XA	ATTTGCTTATGTGTTTATTTAC		
UP1_PRDM9_XB	TTACGGAGATCATGAGCAAGG	1792 / 1960	ZF arrays
Seq1_PRDM9_XB	TCATGAGCAAGGCTCCAAGGATAG		
Seq2_PRDM9_XB	AAGAAAAATCTCATCACACACA		
DN1_PRDM9_XB	GGTCATGACTCACTCACACTCCA		

Supplementary Table 3: Primer and probes used for genotyping candidate QTN using 5'exonculease (Taqsman) assays.

REC8	Primer sequence (5'->3')	Allele 1(FAM)	Allele 2 (HEX)
ss418642851, Forward	GGT GTA GAC GAG GGT GAA AC	/56-FAM/TGG AG+G +GAC CAT GTT C /3IABkFQ/	/5HEX/TGG AG+G +CAC C+AT G+TT C /3IABkFQ/
ss418642851, Reverse	CCG CCC TTA CCA GAT G		
ss418642852, Forward	TCC CTT TGA TAT CCC TCA GGT AG	/56-FAM/ATGAGCTGA+C+AGT+GCAG/3IABkFQ/	/5HEX/ATGAGCTGA+G+AGT+GCA/3IABkFQ/
ss418642852, Reverse	TGG ACC GGC ACA GAA ATA G		
ss418642853, Forward	GCA GCA GAG CCA CTT CCC AG	/56-FAM/CCC CA+G +AGG AGC TG /3IABkFQ/	/5HEX/CCC +CA+A +AG+G AGC TG /3IABkFQ/
ss418642853, Reverse	AGT AGG TAC TCA CCG GGA AG		
ss418642854, Forward	CAC ATC CAA GGG CAT CAA TTG	/56-FAM/CAT TTT AA+C A+AT +GA+G +CA+G A /3IABkFQ/	/5HEX/TTA A+CA +G+TG +AG+C AGA /3IABkFQ/
ss418642854, Reverse	GCC TCT GGC CTT CCA G		
RNF212	Primer sequence (5'->3')	Allele 1(FAM)	Allele 2 (HEX)
ss418642856, Forward	ACG GAA AGT GGA ATC CTC A	/56-FAM/TGC CTC +G+CG GGC /3IABkFQ/	/5HEX/TGC +CTC +A+CG GGC /3IABkFQ/
ss418642856, Reverse	GGA TCT GAG CCC GGC		
ss418642857, Forward	ACG GAA AGT GGA ATC CTC A	/56-FAM/CAG GCT +GTG GCG CG /3IABkFQ/	/5HEX/CAG GCT +CTG GCG CG /3IABkFQ/
ss418642857, Reverse	GCC ACG CGA GAC CTG		
ss418642859, Forward	GCA GAA GCA GCC CTT TC	/56-FAM/TCC AGG C+G+C TCT TCA /3IABkFQ/	/5HEX/TCC AG+G C+A+C TCT TCA /3IABkFQ/
ss418642859, Reverse	TTG GCG TAC TTC ATG CAC A		
ss418642861, Forward	CTG TCT CCG GGC ACA GAC ATT	/56-FAM/AGT GCC +GTT GCG CC /3IABkFQ/	/5HEX/AGT GC+C +ATT GCG CC /3IABkFQ/
ss418642861, Reverse	AGC CTG GTG CTC ACG CTG AC		
ss469104611, Forward	AGC CCC TCT CAG TAA CCC	/56-FAM/CAG GA+C +CT+G CAG G/3IABkFQ/	/5HEX/CAG G+A+T +CTG +CAG G /3IABkFQ/
ss469104611, Reverse	GCA CGG ACA CCC AGA TTA G		

Part II. Dissecting the genetic basis of the arthrogryposis and the brachyspina syndrome by integrating SNP array and HTS

Genome-wide next-generation DNA and RNA sequencing reveals a mutation that perturbs splicing of the phosphatidylinositol glycan anchor biosynthesis class H gene (*PIGH*) and causes arthrogryposis in Belgian Blue cattle

Arnaud Sartelet, Wanbo Li*, Eric Pailhoux, Christophe Richard, Nico Tamma, Latifa Karim, Corinne Fasquelle, Tom Druet, Wouter Coppieters, Michel Georges, Carole Charlier.*

* Contributed equally to this work

BMC Genomics, 2015, 16:316, DOI 10.1186/s12864-015-1528-y.

Background

Since the 1950s, Belgian Blue cattle have been under intensive selection for meat production traits. The extensive use of artificial insemination of elite sires has increased levels of inbreeding and consequently heavily reduced its effective population size. As a result recessive disease-causing variants from some once-popular sires could spread over the population and led to outburst of several genetic defects in this breed. Up to the present time, dozens of genetic defects have been documented, such as CMD1, CMD2, renal lipofuscinosis (Charlier et al., 2008).

In 2009, twenty-five Belgian-Blue cases of arthrogryposis were reported to the "heredo-surveillance platform". In the process of identifying the causative mutation of this defect, we first mapped the locus to a 2.2 Mb interval on BTA10 using a haplotype-based GWAS approach. We later resequenced the whole genome of four cases and identified 31 candidate mutations. From RNA sequencing data of a heterozygous Belgian Blue fetus, we predicted an intronic variant (c211-10C > G) in the *PIGH* gene affecting its acceptor splice-site, leading to skipping of *PIGH* exon 2 and as a result lacks an essential motif in the PIGH protein. This study highlights the advantage of combining DNA and RNA sequencing to illustrate the consequence of a noncoding variant.

BMC Genomics

Genome-wide next-generation DNA and RNA sequencing reveals a mutation that perturbs splicing of the phosphatidylinositol glycan anchor biosynthesis class H gene (*PIGH*) and causes arthrogryposis in Belgian Blue cattle

Sartelet *et al.*

BioMed Central

Sartelet *et al. BMC Genomics* (2015) 16:316
DOI 10.1186/s12864-015-1528-y

Sartelet et al. BMC Genomics (2015) 16:316
DOI 10.1186/s12864-015-1528-y

BMC Genomics

RESEARCH ARTICLE Open Access

Genome-wide next-generation DNA and RNA sequencing reveals a mutation that perturbs splicing of the phosphatidylinositol glycan anchor biosynthesis class H gene (*PIGH*) and causes arthrogryposis in Belgian Blue cattle

Arnaud Sartelet[1†], Wanbo Li[1†], Eric Pailhoux[2], Christophe Richard[2], Nico Tamma[1], Latifa Karim[1,3], Corinne Fasquelle[1], Tom Druet[1], Wouter Coppieters[1,3], Michel Georges[1] and Carole Charlier[1*]

Abstract

Background: Cattle populations are characterized by regular outburst of genetic defects as a result of the extensive use of elite sires. The causative genes and mutations can nowadays be rapidly identified by means of genome-wide association studies combined with next generation DNA sequencing, provided that the causative mutations are conventional loss-of-function variants. We show in this work how the combined use of next generation DNA and RNA sequencing allows for the rapid identification of otherwise difficult to identify splice-site variants.

Results: We report the use of haplotype-based association mapping to identify a locus on bovine chromosome 10 that underlies autosomal recessive arthrogryposis in Belgian Blue Cattle. We identify 31 candidate mutations by resequencing the genome of four cases and 15 controls at ~10-fold depth. By analyzing RNA-Seq data from a carrier fetus, we observe skipping of the second exon of the *PIGH* gene, which we confirm by RT-PCR to be fully penetrant in tissues from affected calves. We identify - amongst the 31 candidate variants - a C-to-G transversion in the first intron of the *PIGH* gene *(c211-10C > G)* that is predicted to affect its acceptor splice-site. The resulting PIGH protein is likely to be non-functional as it lacks essential domains, and hence to cause arthrogryposis.

Conclusions: This work illustrates how the growing arsenal of genome exploration tools continues to accelerate the identification of an even broader range of disease causing mutations, therefore improving the management and control of genetic defects in livestock.

Keywords: Arthrogryposis syndrome, *PIGH* gene, Splice-site mutation, Glycosylphosphatidyl inositol deficiency, Belgian Blue Cattle breed

Background

The extensive use of elite sires exacerbated by the large-scale exploitation of artificial insemination in cattle breeding causes important reductions in effective population size and the common spread of loss-of-function variants. This in turn is responsible for the periodic outburst of genetic defects that cause considerable economic loss and welfare issues. With the development of genome-wide SNP arrays for all livestock species, it has become possible to rapidly map the underlying locus by means of autozygosity mapping to intervals that typically span 2 to 5 megabases thereby proving the inherited nature and mode of inheritance of the corresponding condition (f.i. [1]). With the advent of targeted or whole-genome next generation sequencing (NGS), it is becoming increasingly facile to identify the causative mutation, needed to develop accurate diagnostic tests, provided that the mutation is a frame-

* Correspondence: carole.charlier@ulg.ac.be
†Equal contributors
[1]GIGA-R & Department of Animal Sciences, Unit of Animal Genomics, Faculty of Veterinary Medicine, University of Liège, Avenue de l'Hôpital 1, 4000 Liège, Belgium
Full list of author information is available at the end of the article

shift, nonsense, canonical splice-site, or severe missense variant. In other cases, the causative mutation may remain elusive for a considerably longer time. We show in this work how the combined use of DNA and RNA NGS data, may accelerate the discovery of an otherwise elusive, novel class of causative mutations.

Results and discussion

Arthrogryposis emerges as a new genetic defect in Belgian Blue Cattle

We recently established an "heredo-surveillance platform" to effectively identify and control inherited defects that recurrently emerge as a result of intensive use of elite sires in Belgian Blue and other cattle breeds (f.i. [1]). Twenty-five Belgian-Blue cases of a new form of arthrogryposis were referred to this platform in 2009 alone. Affected calves were all characterized by arthrogryposis (hooked joints) of the four limbs, severe scoliosis (curved spine), and a stocky head with macroglossy and impaired tooth eruption. A majority of cases suffered from cleft palate (20/25) and upper lip (3/25), omphalocele (abdominal wall defect with umbilical hernia; 19/25) and corneal clouding (21/25) (Figure 1). Several dams developed metritis and peritonitis, caused by hydrops (accumulation of excessive fluid in the allantoïc or amniotic space) of the fetal membranes due to impaired fetal swallowing.

A haplotype-based GWAS maps the culprit locus to a 2.2 Mb interval on bovine chromosome 10

The 25 cases traced back, on sire and dam side, to the artificial insemination (AI) sire *Kalimine du Barsy Fontaine*, suggesting autosomal recessive inheritance. We therefore genotyped 15 cases with the bovine SNP50 beadchip (Illumina, San Diego) to perform a genome-wide association study (GWAS). We used the genotypes from 275 Belgian Blue AI sires, obtained with the bovine HD (700 K) beadchip (Illumina, San Diego), as controls. The analysis was restricted to 34,368 SNPs shared by both arrays, and conducted with GLASCOW as previously described [2]. This yielded a single genome-wide significant signal ($p = 10^{-40}$) on chromosome 10 (Figure 2A). It resulted from autozygosity of the 15 cases for a 2.2 Mb (chr10:78,424,435-80,602,211 bp; *Bos taurus* assembly: BosTau6/UMD3) identical-by-descent haplotype, hence confirming the suspected mode of inheritance (Figure 2B).

Resequencing the whole genome of four cases identifies 31 candidate causative mutations

The corresponding interval is collinear with a 2.4 Mb segment of human chromosome 14 (chr14: 66,474,214-68,918,385 bp) and encompasses 22 annotated genes. As none of these would be obvious candidates, we generated paired-end libraries for four cases and re-sequenced them to average 2.5-fold depth (for a combined total coverage of ~10 fold) on a Illumina GAIIX instrument as described [3]. Equivalent genomic sequences of 15 healthy Belgian Blues that did not carry the incriminated haplotype were used as controls. Sequence reads were mapped on the BosTau6/UMD3 reference genome using BWA [4], and variants called with SAMTools [5]. We detected 2,968 variants with quality score > 100 mapping to the 2.2 Mb interval. As expected, the four cases appeared homozygous for the 21 Beadchip SNPs defining the associated haplotype. Filtering against known bovine dbSNP SNPs and variants observed in at least one of the 15 controls, and demanding homozygosity for the four cases, left us with only 31 candidate variants (Additional file 1). Yet, none of these would alter coding sequences (missense or nonsense), or map within three base pairs from an exon-intron junction.

RNASeq reveals skipping of PIGH exon 2 and pinpoints an intronic mutation as the likely causative variant

The findings described in the previous paragraph suggested that the causative mutation was either regulatory or affecting splicing otherwise. To pursue this hypothesis, we took advantage of available RNASeq data (cfr. M&M) from liver (83 Mb uniquely aligned) and cerebral cortex (105 Mb) of a 60-day post fertilization Belgian Blue fetus shown by SNP genotyping to carry the arthrogryposis risk haplotype. Sequence reads were analyzed using TopHat and Cufflinks [6] and predicted transcripts mapping to the arthrogryposis locus visualized in the IGV browser [7]. We readily noticed skipping of the second exon of approximately half (20/45) of the *PIGH* (phosphatidylinositol glycan anchor biosynthesis class H) transcripts (Figure 3A). To verify whether this observation might be related to the arthrogryposis condition we extracted RNA from available skeletal muscle and kidney of an affected calf and an age-matched control, and performed RT-PCR using primers located in exon 1 and 4 of the *PIGH* gene. While we obtained a unique band of expected 593-bp size for the control, the only band obtained from the case RNA was 377~bp (Additional file 2). Sequencing of the corresponding RT-PCR products revealed the expected *PIGH* exon 1-2-3 sequence for the controls, yet two distinct sequences for the case. The most abundant (~75%) form corresponded to the skipping of exon 2, while the minor (~25%) form was - in addition to missing exon 2 – devoid of the first AAG triplet of exon 3 (Additional file 2). Note the 3′ AG end of this lysine codon susceptible to act as a cryptic acceptor splice-site. This form was also observed in five RNASeq reads, validating this finding (data not shown).

Interestingly, one of the 31 candidate mutations identified by whole genome sequencing is a C-to-G transversion located in the first intron of *PIGH*, 10-bp upstream of the exon 2 junction *"211-10C > G*) (Figure 3B). The

Figure 1 Lethal arthrogryposis syndrome clinical spectrum. **A**. Generalized arthrogryposis. **B**. Brachygnathism and macroglossy. **C**. Impaired tooth eruption. **D**. Omphalocoele. **E**. Corneal clouding. **F**. Hard cleft palate.

corresponding residue is located in the consensus polypyrimidine track defining canonical "GT-AG" type acceptor splice-sites (f.i. [8]), and is conserved in 26 of the 27 sequenced mammals (Additional file 3). We developed a 5′ exonuclease assay to interrogate the *c211-10C > G* variant and genotyped 25 cases, 21 parents and > 10,000 healthy Belgian Blue animals. All cases were homozygous "GG", and all parents "CG" as expected. Six percent of the healthy Belgian Blues were carriers, while none were homozygous "GG" ($p = 0.00263$).

The clinical spectrum of arthrogryposis is compatible with severe glycosylphosphatidyl inositol (GPI) deficiency

The PIGH protein is ubiquitously expressed. It is one of seven highly conserved (from yeast to human) subunits of the complex catalyzing the first step out of eleven in the biosynthesis of the glycophosphatidyl inositol (GPI) anchor, a complex C-terminal posttranslational modification concerning > 150 proteins or ~0.5% of cellular proteins (http://www.uniprot.org/uniprot) (f.i. [9-11]). Within the GPI-GlcNAc transferase complex (GPI-GnT), PIGH is

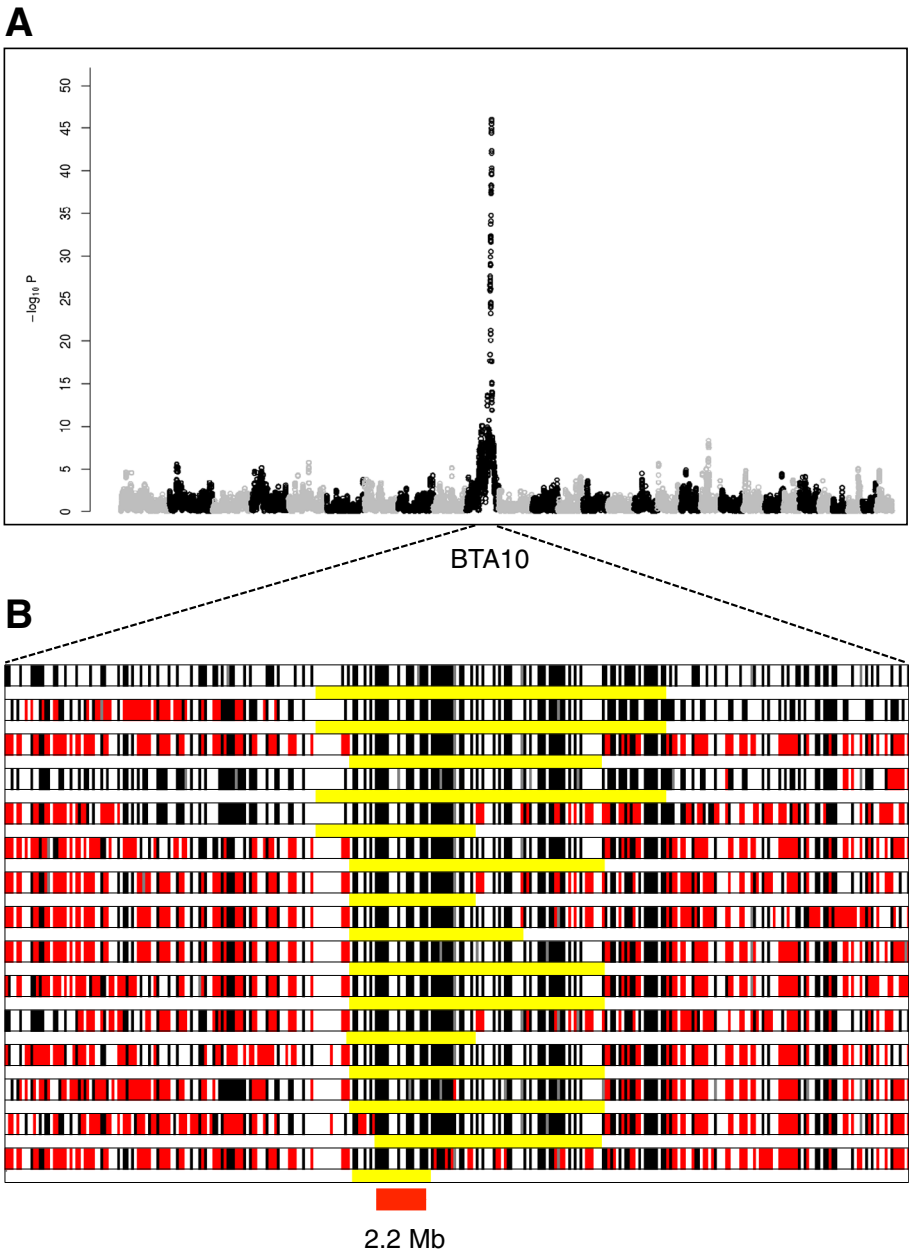


Figure 2 Genetic mapping of the mutation causing the arthrogryposis syndrome in Belgian Blue Cattle. **A**. Manhattan plot for the case–control GWAS study. **B**. Genotypes of 15 cases for a BTA10 segment centered around the most significant GWAS peak, and encompassing 324 SNP (from 70 to 90 Mb). Homozygous genotypes are shown in black or white, heterozygous genotypes in red. The presumed ancestral haplotype encompassing the mutation is underlined in yellow. The 2.2 Mb region of homozygosity shared by all cases is highlighted in red.

anchored in the membrane of the endoplasmic reticulum (ER) by two trans-membrane helices (residues 38–55 and 60–78), connected by a short ER intra-luminal loop (56–59), and bounded by a short N-terminal (1–37) and a long C-terminal cytoplasmic tail (79–188) [12]. Deleting exon 2 generates a protein missing amino-acids 61–130, therefore unlikely to be properly anchored in the ER membrane and to be functional (Additional file 4). In culture, cells with mutations in *PIGH* do not display any GPI-anchored protein (GPI-AP) surface expression (f.i. [13]). In animals, all known GPI-anchor related defects characterized by a complete lack of GPI-anchored proteins are lethal for the embryo. It is thought that this embryonic-lethal phenotype is due to the fact that several cell-to-cell adhesion molecules required for normal development and acting at specific developmental stages are GPI-anchored (reviewed in [11]). The additional file 5 lists the known GPI-AP and their associated knock-out phenotype in mice. Taken

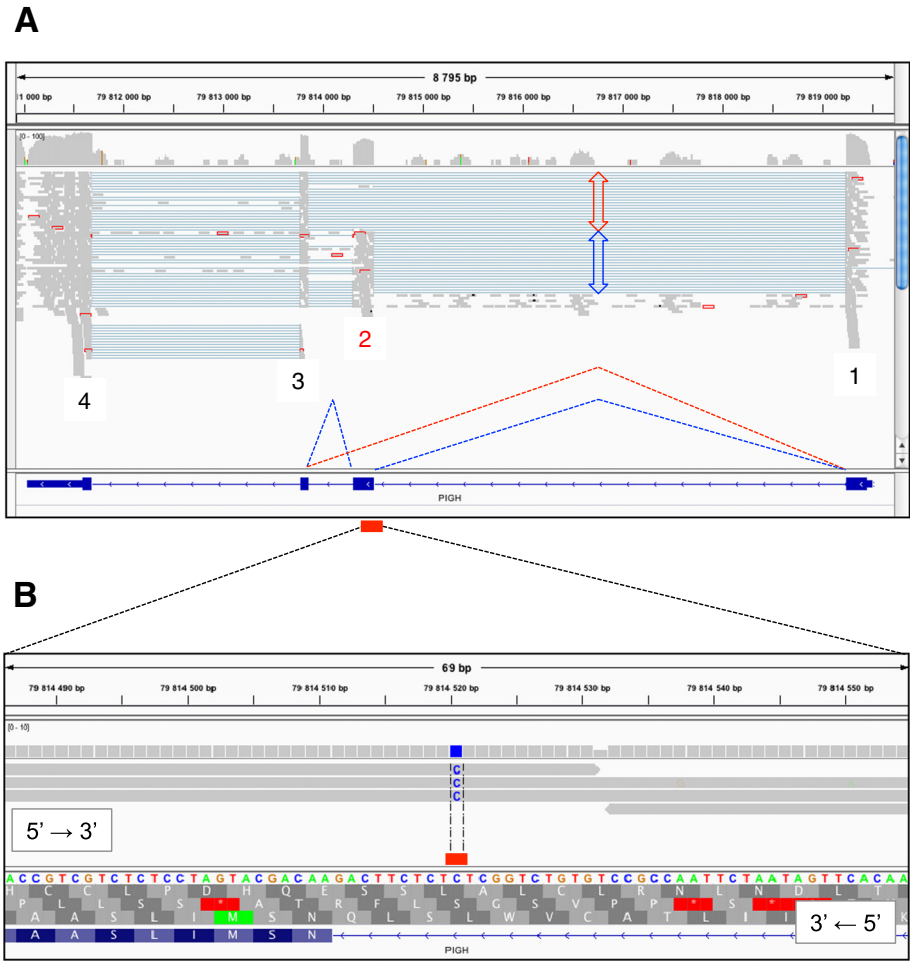


Figure 3 Alternative splicing at the *PIGH* locus and private *c211-10C > G* intronic variation. **A**. Screen capture of an IGV output from liver RNASeq data of a heterozygote mutant embryo aligned on the bovine genomic reference sequence at the *PIGH* locus (top). The four exons appear as stacks of grey reads and splicing is schematically denoted by thin blue lines. Complete skipping of *PIGH* exon 2 in ~ half of the transcripts is noticeable (red versus blue arrow height). *PIGH* intron/exon annotation and the two alternative splicing events are represented by dashed red (skipped exon 2) and blue (incorporated) lines. **B**. Screen capture of an IGV output displaying (i) on the positive strand (top, 5′ to 3′), a private G to C mutation (blue) from genomic DNA sequence reads of four pooled homozygous cases, (ii) on the negative strand (bottom, 3′ to 5′), *PIGH* intron 1/exon 2 annotation showing the private mutation position at −10 nucleotide in the splice acceptor sequence (*c211-10C > G*, red bar).

together, our results strongly suggest that the observed *PIGH c211-10C > G* mutation causes arthrogryposis in Belgian Blue Cattle.

Conclusions

We herein describe the identification of an intronic mutation that disrupts *PIGH* function by causing the skipping of exon 2, thereby causing a severe and lethal form of arthrogryposis in homozygous animals. We mapped the corresponding locus by performing a haplotype-based GWAS, an approach that has proven successful in many instances (f.i. [1]). If the causative mutation is either a stop-gain, frameshift, splice-site variant within 2-bp from an exon-intron boundary, or a severe missense variant, its identification is often relatively straightforward. In our experience, this situation occurs approximately fifty percent of the time, which is in agreement with the data reported by [14]. In the other cases, the identification of the causative mutation may be considerably more arduous. In the present study, it was the concomitant generation of RNASeq data from tissue of a carrier animal that provided the clues for the identification of the causative mutation. The detection of exon skipping in an essential gene that was not observed in non-carrier animals pointed towards one of 31 genetically defined mutations (located in the immediate vicinity of the skipped exon) as being the likely causative mutation. It would have been very difficult to predict the effect of this variant, located at 10-bp from the exon-intron junction, on the splicing reaction without the information on the corresponding transcripts. Our work

therefore illustrates the value of the combined availability of DNA and RNASeq data to rapidly identify a broader panel of disease causing mutations.

Although we consider the supporting evidence to be very strong, we haven't formally proven that skipping of exon 2 of the *PIGH* gene abrogates its function and interferes with GPI-anchor biosynthesis. In human and mice, functional validation of GPI biosynthesis pathway's defects relies on the detection of incomplete or absent cell surface expression of CD59, a well-characterized GPI-anchored protein (f.i. [12]). This quantification is typically performed by FACS on fresh erythrocytes. Unfortunately, homozygous mutant calves are dead at birth, precluding the collection of fresh blood. Producing and collecting homozygous mutant fetuses by mating carriers was beyond the scope of this study.

To the best of our knowledge, this the first report of a naturally occurring loss-of-function mutation in the *PIGH* gene in animals. A genetic test interrogating the *PIGH c211-10C > G* variant has been developed and offered to breeders since 2011. Its widespread use has led to the rapid elimination of this syndrome from the Belgian Blue population.

Methods

Ethics statement

Blood samples were collected from sires, cows and calves, by trained veterinarians following standard procedures and relevant Belgian national guidelines. For bovine fetuses production and collection, experiments reported in this work are in agreement with the ethical guidelines of the French National Institute for Agricultural Research (INRA). Fetuses were produced by artificial insemination of a wild-type Belgian Blue female with semen of a Belgian Blue confirmed-carrier male, 7-days old embryos recovered, transferred to recipient females in an INRA experimental farm (France), then collected at 60 days post-fertilization at the INRA slaughterhouse (France). The protocol (N°: 12/046) was approved by the local ethical committee (COMETHEA) and Eric Pailhoux is the recipient of an official authorization for animal experimentation (N°: B91-649).

SNP array 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 15 cases of the initial genome scan were genotyped using a custom-made 50 K SNP array [1]. The 275 control sires were genotyped with the BovineHD BeadChip (Illumina). SNP genotyping was conducted using standard procedures at the GIGA genomics core facility.

RNAseq data generation and analysis

As part of a companion project, bovine embryos were produced by directed mating between a sire carrier for the arthrogryposis haplotype and a wild-type cow. A panel of tissues - including pituitary, cerebral cortex and liver - from two embryos diagnosed as carrier based on a haplotype test, were collected at 60 days post fertilization (dpf); total RNA was extracted and cDNA librairies were produced using the TrueSeq mRNA kit from Illumina, following manufacturer's instructions. Equal amounts of each indexed library were combined and sequenced 2X100bp on one lane of a HighSeq2000 instrument. Transcriptomes were analyzed using the RNASeq tool kit TopHat and Cufflinks [6]. Mapped RNASeq reads for the carrier embryo were visually evaluated in IGV (Integrative Genome Browser) [7].

Mutation validation at the mRNA level in case and control calves

Total RNA was extracted from kidney and muscle of one homozygous case and one unaffected unrelated individual using Trizol (Invitrogen) following manufacturer's instructions. The obtained total RNA was treated with TurboDNaseI (Ambion) and cDNA was synthesized using Superscript™III First Strand Synthesis System for RT-PCR (Invitrogen). A 593 bp *PIGH* cDNA fragment was PCR amplified using a specific primers respectively located in exon 1 (PIGH_UP : 5′-TCT CTT TGC GCT CGC TCA CC-3′) and exon 4 (PIGH_DN: 5′-GAT CCA CCA CAT CCA TAC TGG-3′). Amplicons were directly sequenced using the Big Dye terminator cycle sequencing kit (Applied Biosystems, Foster City, CA). Electrophoresis of purified sequencing reactions was performed on an ABI PRISM 3730 DNA analyzer (PE Applied Biosystems, Foster City, CA). Sequence traces were aligned and compared to bovine cDNA sequence using the Phred/Phrap/Consed package (http://www.phrap.org/phredphrapconsed.html).

5′ exonuclease diagnostic assay of the PIGH c211-10C > G splice acceptor site mutation

A 5′ exonuclease assay was developed to genotype the *PIGH c211-10C > G* mutation, using 5′- ATG GCA GCA GAG AGG ATC ATG -3′ and 5′- GGA GTT GAC TTA TTA ACC AGC AGA GA -3′ as PCR primers, and 5′-TGT CTG GCT [**C**]TC TCT TC-3′ (wild type *C* allele) and 5′-TCT GGC T[**G**]T CTC TTC -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.

Availability of supporting data

All the variations of the IBD disease haplotype have been submitted to dbSNP (http://www.ncbi.nlm.nih.gov/SNP/).

They will be publicly available with the next dbSNP Build (B145), planned for fall, 2015. Additional files 6 and 7 provide respectively the list of "ss" numbers, genomic positions (*Bos taurus* assembly: BosTau6/UMD3) and the list of corresponding sequence variations.

Additional files

Additional file 1: Chromosomal position and sequence context for the disease risk haplotype private mutations within the IBD interval. Are given, from left to right: the chromosome (Chr.), the variant start position on UMD3.1/bosTau6 assembly (Start), the reference sequence allele (Ref.), the derived allele (Der.), both on the positive strand, the underlying gene (Gene) if any, the regional annotation (Annotation) and the variant sequence context, within a unique sequence (Unique, yes) or in a repetitive region (Unique, no).

Additional file 2: Effect of the *c211-10C > G* mutation at the RNA level. A. Partial schematic representation of the *PIGH* organization (genomic and mRNA) accompanied by wild-type cDNA sequence traces across exon 1/exon 2 and exon 2/exon 3 junctions obtained respectively with a forward primer in exon 1 and a reverse primer in exon 4. B. Corresponding agarose gel showing cDNA amplification products from kidney (K) for a wild-type (WT) animal, from kidney (K) and skeletal muscle (SM) for a mutant (MUT) and for a RT-minus control (RT-); MW: molecular weight marker (SmartLadder, Eurogentec). C. Mutant cDNA sequence traces obtained from the ~ 377 bp amplification product with reverse (top) and forward (bottom) primers respectively; the cryptic "AAG" acceptor site in exon 3 is highlighted in dark orange; height of depicted white and orange bars represents the ratio between the two mutant splice forms, with total skipping of exon 2 in white and additional skipping of a "AAG" triplet in orange.

Additional file 3: *PIGH* first exon acceptor splice-site evolutionary conservation. Alignment and conservation of genomic DNA sequences corresponding to the *PIGH* gene (final part of intron 1 in lower cases and beginning of coding exon 2 in upper cases) is presented for placental mammals, genomic sequences were downloaded from the UCSC genome browser; bovine wild-type (WT) "C" nucleotide (nt) is highlighted in black and corresponding mutant (MUT: *c211-10C > G*) "G" nt in orange; consensus acceptor splice-site sequence, from intronic nt −12 to nt −1, is depicted below the alignment (with Y for C/T; N for A/T/C/G and essential acceptor splice-site (AG) at positions −2 and −1 in black); for every consensus position, nucleotide usage (A/U(T)/C/G in %) is presented; one can notice that at nt position −10, "G" nt is present in only 6% (orange) of the acceptor splice-site sequences (SpliceDB: Burset et al., [8]).

Additional file 4: PIGH protein evolutionary conservation and topology in the endoplasmic reticulum (ER) plasma membrane. A. Protein alignment of the PIGH protein from mammals to yeast; the two transmembrane (TM) domains, as annotated in UniProtKB, are underlined and highlighted in yellow; wild-type (WT) and mutant (MUT) bovine PIGH are presented and the deduced missing amino acid sequence - corresponding to the second TM domain and part of the cytoplasmic C-terminal domain - is boxed in grey. B. Schematic representation of the WT PIGH protein (left), adapted from Watanabe et al. [12] confirmed by the topological prediction obtained with TMHMM2.0 software (screen capture of the TMHMM result, right) [15]; putative topology for the MUT protein is shown accordingly (bottom); TM domains are colored in yellow.

Additional file 5: GPI-anchored proteins reviewed in mice and specific associated knock-out phenotypes (if any). Are given from left to right: accession number in UniProtKB (Entry); gene symbol (Entry name); reviewed or unreviewed status (Status); full protein name (Protein name); alternative gene symbols (Gene names); amino acid number (Length); knock-out if any (Knock-out, KO); phenotype severity in homozygous KO (Phenotype severity); KO phenotype as described in MGI (Mouse Genome Informatics; http://www.informatics.jax.org/marker/).

Additional file 6 List of genomic positions and corresponding "ss" (Submitted SNP) numbers for sequence variations of the IBD disease haplotype.

Additional file 7: List of genomic positions, flanking sequences and corresponding sequence variations of the IBD disease haplotype.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AS performed case and control collection, pedigree analysis, phenotyping, necropsy and genotype/haplotype data analysis; WL produced and analyzed HT-RNASeq data and performed RT-PCR analysis. EP managed animal experiments for recovering fetal biological material, read and approved the manuscript. CR managed embryos recovered and transferred from donor to recipient cows. NT developed the 5′-exonuclease assay and genotyped the Belgian blue population. LK and CF generated whole genome sequences of cases and controls. TD performed the haplotype-based GWAS analysis. WC assisted in analyzing HT sequences. MG and CC conceived and designed the experiments. MG and CC wrote the paper with the help of all co-authors. All authors read and approved the final manuscript.

Acknowledgements
We are grateful to the breeders and practitioners for their collaboration and to the Walloon Breeding Association (AWE) and the Belgian Blue Beef Herd-Book (HBBBB) for pedigree data. We also thank all the members of INRA experimental units, especially Alexandre Neveux, Valérie Gelin, Céleste Le Bourhis, Nicolas Stadler, Johann Kergozien, Gilles Gomot, Jean-Philippe Dubois for cow management (breeding, ultrasound scanning, and euthanasia). TD and CC are respectively Research Associate and Senior Research Associate of the Fonds National de la Recherche Scientifique (Belgium). This work was funded by grants from the Walloon Ministry of Agriculture (Rilouke), the Belgian Science Policy Organization (SSTC Genefunc PAI) and the University of Liège.

Author details
[1]GIGA-R & Department of Animal Sciences, Unit of Animal Genomics, Faculty of Veterinary Medicine, University of Liège, Avenue de l'Hôpital 1, 4000 Liège, Belgium. [2]INRA, UMR 1198, Biologie du Développement et Reproduction, F-78350 Jouy-en-Josas, France. [3]GIGA Genomic Platform, GIGA, University of Liège, Avenue de l'Hôpital 1, 4000 Liège, Belgium.

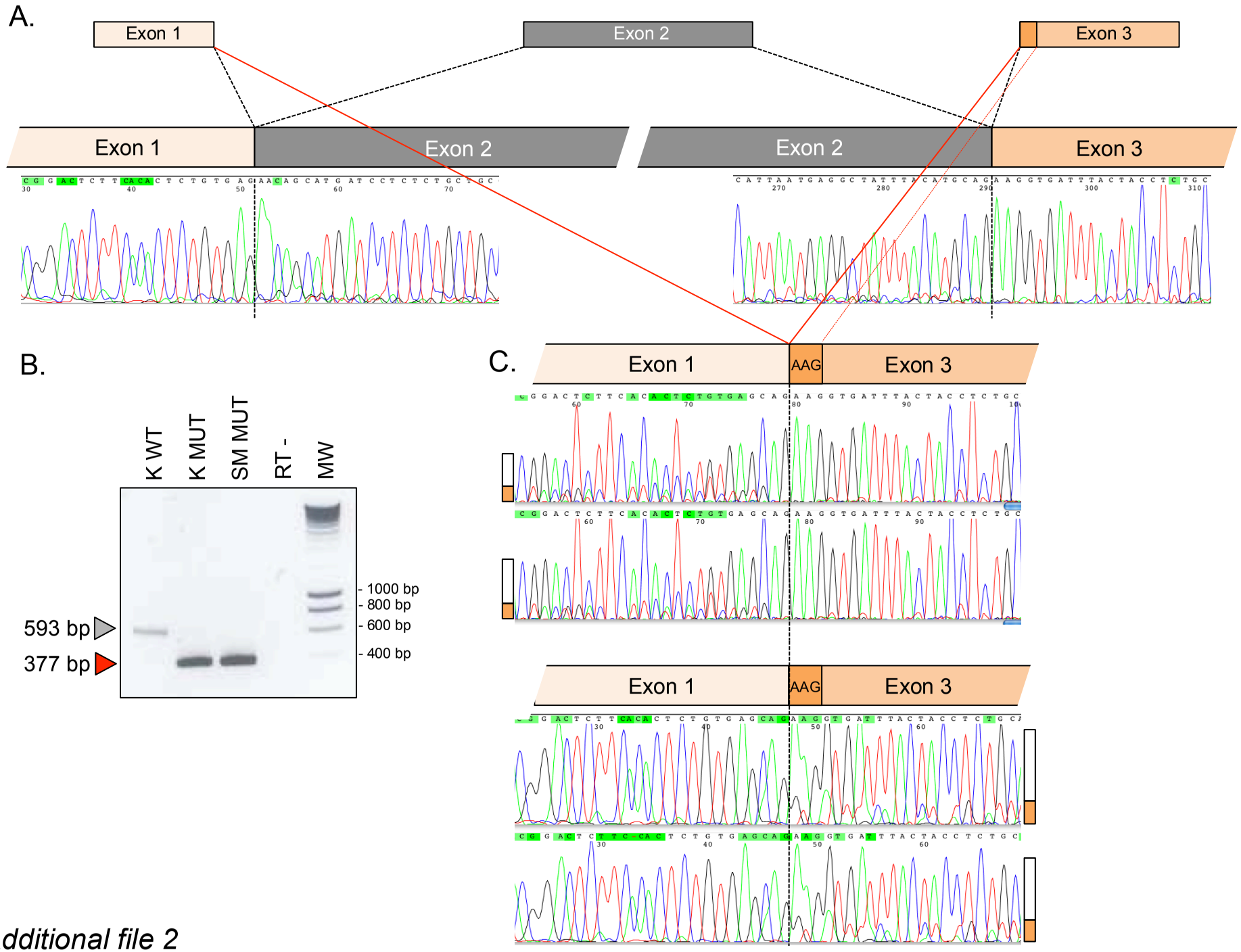
Received: 10 November 2014 Accepted: 13 April 2015
Published online: 18 April 2015

References
1. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, Cambisano N, et al. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet. 2008;40:449–54.
2. Zhang Z, Guillaume F, Sartelet A, Charlier C, Georges M, Farnir F, et al. Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. Bioinformatics. 2012;28:2467–73.
3. Sartelet A, Stauber T, Coppieters W, Ludwig CF, Fasquelle C, Druet T, et al. A missense mutation accelerating the gating of the lysosomal Cl-/H + –exchanger ClC-7/Ostm1 causes osteopetrosis with gingival hamartomas in cattle. Dis Mode Mech. 2014;7:119–28.
4. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
5. Li, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
6. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–78.
7. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
8. Burset M, Seledtsov IA, Solovyev VV. SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res. 2001;29:255–9.
9. Almeida A, Layton M, Karadimitris A. Inherited glycosylphosphatidyl inositol deficiency: a treatable CDG. Biochim Biophys Acta. 2009;1792:874–80.
10. Mayor S, Riezman H. Sorting GPI-anchored proteins. Nat Rev Mol Cell Biol. 2004;5:110–20.

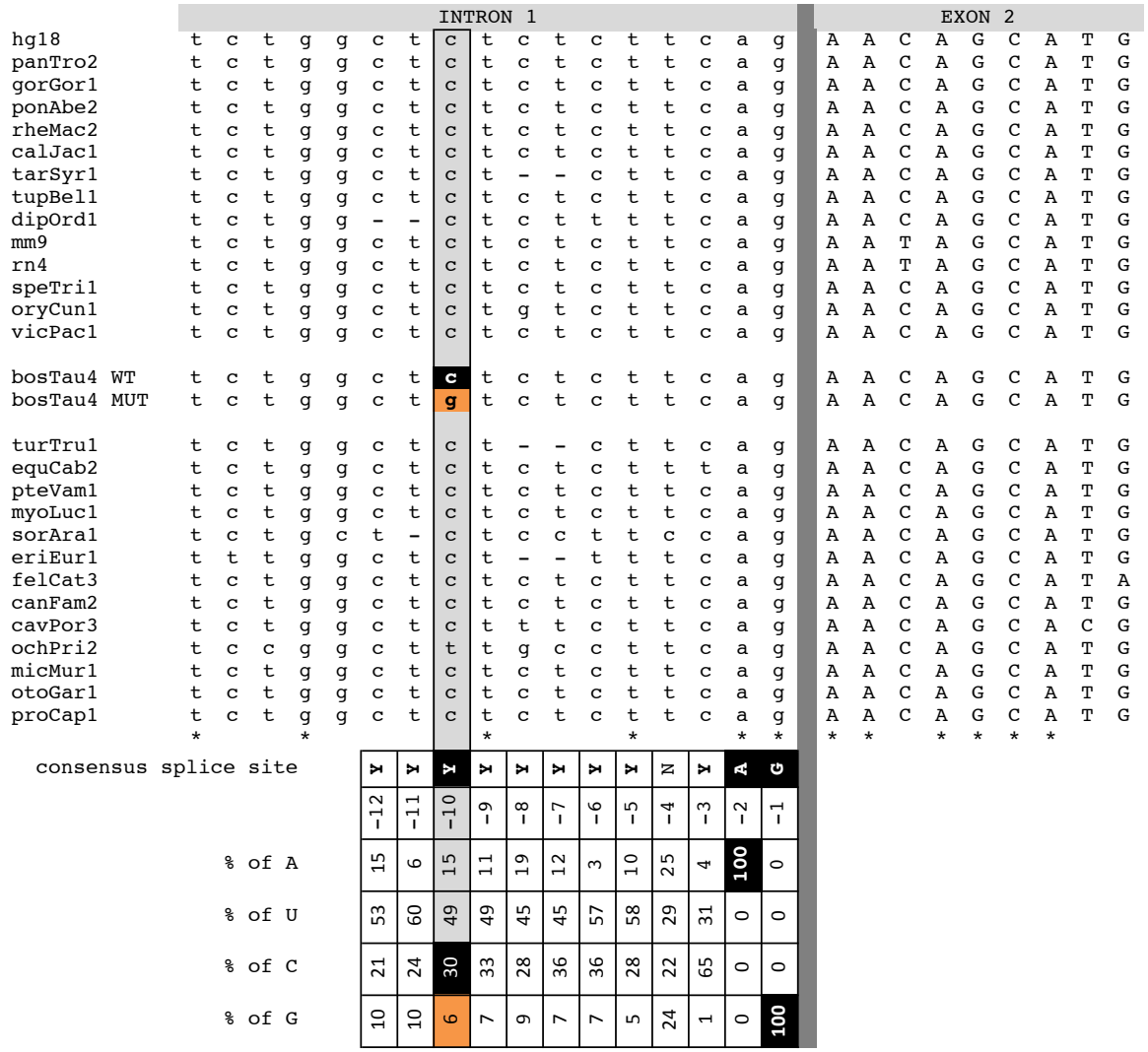
11. Freeze HH. Understanding human glycosylation disorders: biochemistry leads the charge. J Biol Chem. 2013;288:6936–45.
12. Watanabe R, Kinoshita T, Masaki R, Yamamoto A, Takeda J, Inoue N. PIG-A and PIG-H, which participate in glycosylphosphatidylinositol anchor biosynthesis, form a protein complex in the endoplasmic reticulum. J Biol Chem. 1996;271:26868–75.
13. Kamitani T, Chang HM, Rollins C, Waneck GL, Yeh ET. Correction of the class H defect in glycosylphosphatidylinositol anchor biosynthesis in Ltk- cells by a human cDNA clone. J Biol Chem. 1993;268:20733–6.
14. Nicholas FW, Hobbs M. Mutation discovery for Mendelian traits in non-laboratory animals: a review of achievements up to 2012. Anim Genet. 2014;45:157–70.
15. Sonnhammer ELL, von Heijne G, and Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. In J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff, and C. Sensen, editors, Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, pages 175–182, Menlo Park, CA, 1998. AAAI Press. (http://www.cbs.dtu.dk/services/TMHMM/).

Additional file 1: Chromosomal position and sequence context for the disease risk haplotype private mutations within the IBD interval.

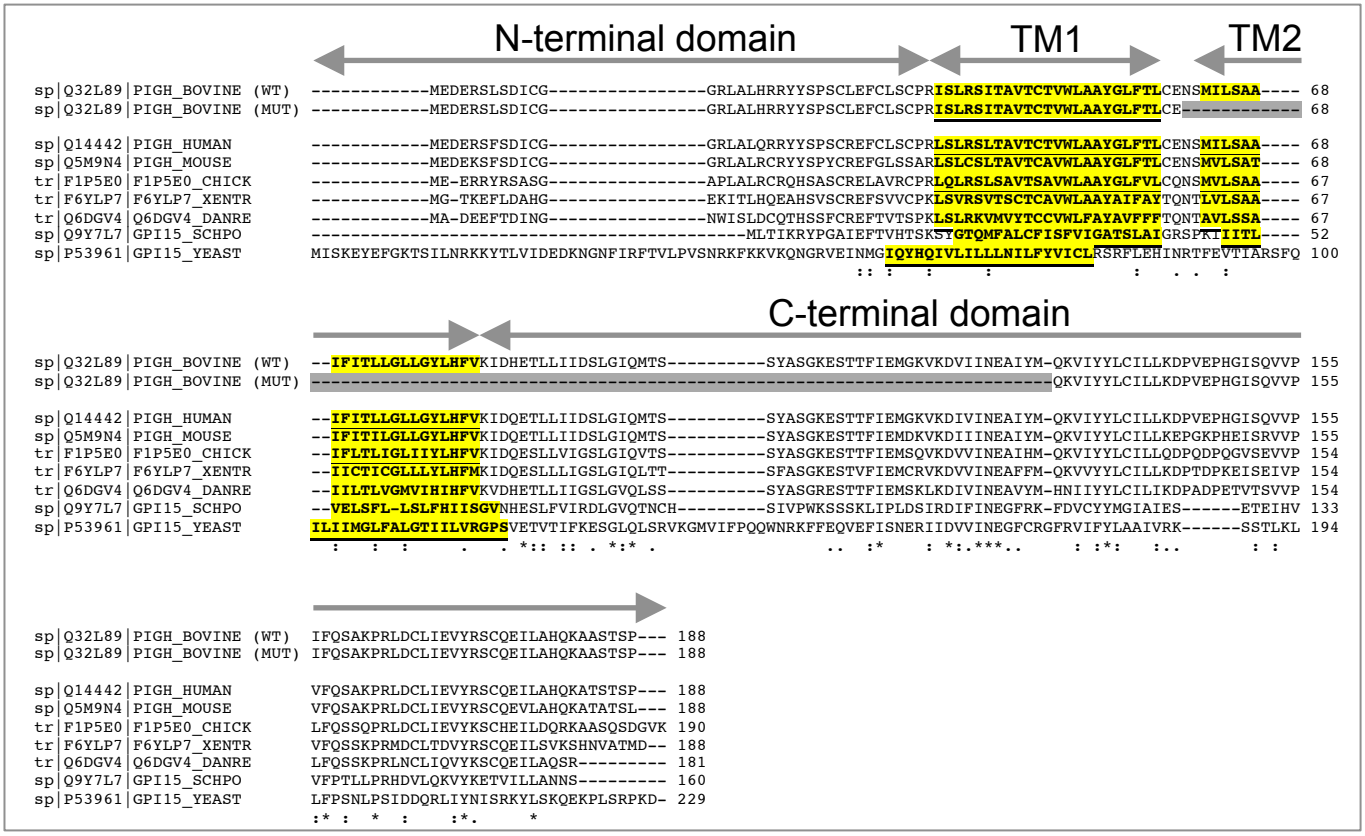
Chr.	Start	Ref.	Der.	Gene	Annotation	Unique
chr10	78427250	G	A	/	intergenic	no
chr10	78454747	A	C	/	intergenic	no
chr10	78674184	T	G	/	intergenic	no
chr10	78714805	C	G	/	intergenic	yes
chr10	78951178	C	T	*GPHN*	intronic	no
chr10	78974848	G	A	*GPHN*	intronic	no
chr10	79151521	T	A	*GPHN*	intronic	yes
chr10	79151943	C	G	*GPHN*	intronic	yes
chr10	79160268	GT	GTT	*GPHN*	intronic	no
chr10	79162630	C	T	*GPHN*	intronic	yes
chr10	79168827	G	C	*GPHN*	intronic	yes
chr10	79170470	C	T	*GPHN*	intronic	yes
chr10	79171445	CTT	C	*GPHN*	intronic	yes
chr10	79172280	A	C	*GPHN*	intronic	no
chr10	79176708	T	A	*GPHN*	intronic	yes
chr10	79181403	GTTACTTA	GTTA	*GPHN*	intronic	yes
chr10	79203039	G	C	*GPHN*	intronic	no
chr10	79234207	G	T	*GPHN*	intronic	no
chr10	79275933	A	G	*GPHN*	intronic	no
chr10	79453112	T	C	*FAM71D*	intronic	yes
chr10	79695337	C	T	*TMEM229B*	intronic	no
chr10	**79814520**	**G**	**C**	***PIGH***	**intronic**	**yes**
chr10	80188354	T	G	*ZFYVE26*	intronic	no
chr10	80194440	T	C	*ZFYVE26*	intronic	no
chr10	80225126	G	A	*ZFYVE26*	intronic	yes
chr10	80231100	C	T	*ZFYVE26*	intronic	yes
chr10	80248464	A	C	/	intergenic	no
chr10	80354103	A	G	*RAD51B*	intronic	no
chr10	80486651	C	T	*RAD51B*	intronic	yes
chr10	80502656	C	T	*RAD51B*	intronic	yes
chr10	80515466	G	A	*RAD51B*	intronic	yes



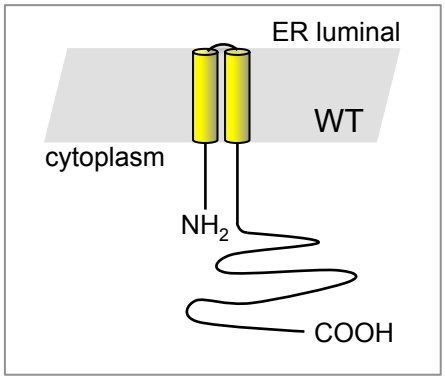
Additional file 2

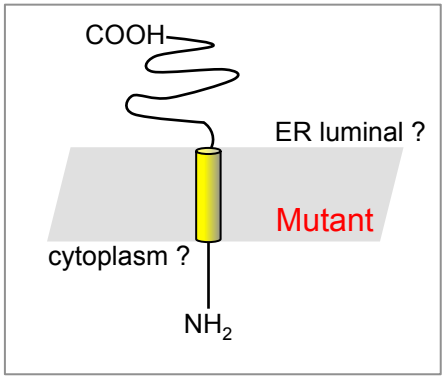

INTRON 1
EXON 2
hg18 t c t g g c t c t c t c t t c a g A A C A G C A T G
panTro2 t c t g g c t c t c t c t t c a g A A C A G C A T G
gorGor1 t c t g g c t c t c t c t t c a g A A C A G C A T G
ponAbe2 t c t g g c t c t c t c t t c a g A A C A G C A T G
rheMac2 t c t g g c t c t c t c t t c a g A A C A G C A T G
calJac1 t c t g g c t c t c t c t t c a g A A C A G C A T G
tarSyr1 t c t g g c t c t - - c t t c a g A A C A G C A T G
tupBel1 t c t g g c t c t c t c t t c a g A A C A G C A T G
dipOrd1 t c t g g - - c t c t t t t c a g A A C A G C A T G
mm9 t c t g g c t c t c t c t t c a g A A T A G C A T G
rn4 t c t g g c t c t c t c t t c a g A A T A G C A T G
speTri1 t c t g g c t c t c t c t t c a g A A C A G C A T G
oryCun1 t c t g g c t c t c t g t t c a g A A C A G C A T G
vicPac1 t c t g g c t c t c t c t t c a g A A C A G C A T G
bosTau4 WT t c t g g c t c t c t c t t c a g A A C A G C A T G
bosTau4 MUT t c t g g c t g t c t c t t c a g A A C A G C A T G
turTru1 t c t g g c t c t c - - c t t c a g A A C A G C A T G
equCab2 t c t g g c t c t c t c t t t a g A A C A G C A T G
pteVam1 t c t g g c t c t c t c t t c a g A A C A G C A T G
myoLuc1 t c t g g c t c t c t c t t c a g A A C A G C A T G
sorAra1 t c t g c t - c t c c t t c c a g A A C A G C A T G
eriEur1 t t t g g c t c t - - t t t t c a g A A C A G C A T G
felCat3 t c t g g c t c t c t c t t c a g A A C A G C A T A
canFam2 t c t g g c t c t c t c t t c a g A A C A G C A T G
cavPor3 t c t g g c t c t t t c t t c a g A A C A G C A C G
ochPri2 t c c g g c t t t g c c t t c a g A A C A G C A T G
micMur1 t c t g g c t c t c t c t t c a g A A C A G C A T G
otoGar1 t c t g g c t c t c t c t t c a g A A C A G C A T G
proCap1 t c t g g c t c t c t c t t c a g A A C A G C A T G
consensus splice site
Y Y Y Y Y Y Y Y Y N Y A G
-12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1
% of A 15 6 15 11 19 12 3 10 25 4 100 0
% of U 53 60 49 49 45 45 57 58 29 31 0 0
% of C 21 24 30 33 28 36 36 28 22 65 0 0
% of G 10 10 6 7 9 7 7 5 24 1 0 100

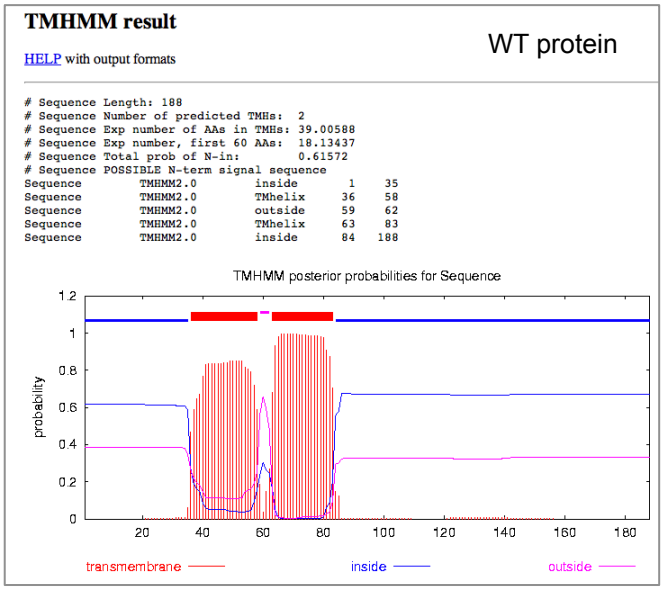
A.


N-terminal domain
TM1
TM2
C-terminal domain
sp|Q32L89|PIGH_BOVINE (WT)
sp|Q32L89|PIGH_BOVINE (MUT)
sp|Q14442|PIGH_HUMAN
sp|Q5M9N4|PIGH_MOUSE
tr|F1P5E0|F1P5E0_CHICK
tr|F6YLP7|F6YLP7_XENTR
tr|Q6DGV4|Q6DGV4_DANRE
sp|Q9Y7L7|GPI15_SCHPO
sp|P53961|GPI15_YEAST

B.


ER luminal
WT
cytoplasm
NH2
COOH


COOH
ER luminal ?
Mutant
cytoplasm ?
NH2


TMHMM result
WT protein
HELP with output formats
Sequence Length: 188
Sequence Number of predicted TMHs: 2
Sequence Exp number of AAs in TMHs: 39.00588
Sequence Exp number, first 60 AAs: 18.13437
Sequence Total prob of N-in: 0.61572
Sequence POSSIBLE N-term signal sequence
Sequence TMHMM2.0 inside 1 35
Sequence TMHMM2.0 TMhelix 36 58
Sequence TMHMM2.0 outside 59 62
Sequence TMHMM2.0 TMhelix 63 83
Sequence TMHMM2.0 inside 84 188
TMHMM posterior probabilities for Sequence
probability
transmembrane
inside
outside

Additional file 5: **GPI-anchored proteins reviewed in mice and specific associated knock-out phenotypes (if any)**
https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-1528-y/MediaObjects/12864_2015_1528_MOESM5_ESM.xls

Additional file 6: **List of genomic positions and corresponding "ss" (Submitted SNP) numbers for sequence variations of the IBD disease haplotype**
https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-1528-y/MediaObjects/12864_2015_1528_MOESM6_ESM.txt

Additional file 7: **List of genomic positions, flanking sequences and corresponding sequence variations of the IBD disease haplotype**
https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-1528-y/MediaObjects/12864_2015_1528_MOESM7_ESM.txt

A deletion in the bovine *FANCI* gene compromises fertility by causing fetal death and brachyspina

Carole Charlier, Jorgen Steen Agerholm, Wouter Coppieters, Peter Karlskov-Mortensen, Wanbo Li, Gerben de Jong, Corinne Fasquelle, Latifa Karim, Susanna Cirera, Nadine Cambisano, Naima Ahariz, Erik Mullaart, Michel Georges, Merete Fredholm.

PLoS ONE, 2012, Issue 8, e43085.

Background

A stillborn calf affected by brachyspina syndrome (BS) was first identified in the Holstein population in 2006 (Agerholm et al., 2006). The affected animals appeared with reduced body weight, shortening of the vertebral column, and abnormal long and thin limbs. In addition, malformation of several inner organs was also found. The incidence of BS in Holstein-Friesian breed is very rare, less than $1/10^5$ birth. And all reported cases could be traced back to a once-popular sire, supporting the suspect of autosomal recessive inheritance.

We intended to uncover the genetic underpinnings of this defect. Combining high-density SNP arrays and next-generation high-throughput sequencing technologies, we were able to identify the causative mutation as a large deletion in the *FANCI* gene. Given the rarity of the defect we were surprised to find that ~7.5% of tested Holstein animals are carriers of this deletion. Using field fertility data we estimated that at least half of the homozygous mutant embryos/fetuses die in utero. The large loss in embryos during gestation is likely to represent an important factor inhibiting female fertility. This interesting finding finally let us to conceive the project of screening for similar variants, causing early embryonic lethality but currently escaping observation in the industry.

OPEN ACCESS Freely available online

PLOS ONE

A Deletion in the Bovine *FANCI* Gene Compromises Fertility by Causing Fetal Death and Brachyspina

Carole Charlier[1]*, Jorgen Steen Agerholm[2], Wouter Coppieters[1,3], Peter Karlskov-Mortensen[4], Wanbo Li[1], Gerben de Jong[5], Corinne Fasquelle[1], Latifa Karim[1,3], Susanna Cirera[4], Nadine Cambisano[1], Naima Ahariz[1], Erik Mullaart[5], Michel Georges[1], Merete Fredholm[4]*

1 Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège (B34), Liège, Belgium, **2** Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark, **3** GIGA-R Genotranscriptomics Core Facility, University of Liège (B34), Liège, Belgium, **4** Division of Genetics and Bioinformatics, Department of Animal and Veterinary Basic Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark, **5** CRV BV, Arnhem, The Netherlands

Abstract

Fertility is one of the most important traits in dairy cattle, and has been steadily declining over the last decades. We herein use state-of-the-art genomic tools, including high-throughput SNP genotyping and next-generation sequencing, to identify a 3.3 Kb deletion in the *FANCI* gene causing the brachyspina syndrome (BS), a rare recessive genetic defect in Holstein dairy cattle. We determine that despite the very low incidence of BS (<1/100,000), carrier frequency is as high as 7.4% in the Holstein breed. We demonstrate that this apparent discrepancy is likely due to the fact that a large proportion of homozygous mutant calves die during pregnancy. We postulate that several other embryonic lethals may segregate in livestock and significantly compromise fertility, and propose a genotype-driven screening strategy to detect the corresponding deleterious mutations.

Citation: Charlier C, Agerholm JS, Coppieters W, Karlskov-Mortensen P, Li W, et al. (2012) A Deletion in the Bovine *FANCI* Gene Compromises Fertility by Causing Fetal Death and Brachyspina. PLoS ONE 7(8): e43085. doi:10.1371/journal.pone.0043085

Editor: Reiner Albert Veitia, Institut Jacques Monod, France

Received April 25, 2012; **Accepted** July 16, 2012; **Published** August 29, 2012

Funding: This work was funded by grants from the Walloon Ministry of Agriculture (Rilouke), and the Belgian Science Policy Organisation (SSTC Genefunc PAI). No additional external funding was received for this study. CC is Senior Research Associate of the Fonds National de la Recherche Scientifique (Belgium). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: GDJ and EM are affiliated with the CRV-BV company. There is a pending patent and, if approved, the number and date of approval will be given prior to publication. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: carole.charlier@ulg.ac.be (CC); mf@life.ku.dk (MF)

Introduction

Fertility is one of the economically most important traits in cattle breeding. It is commonly measured using a series of metric "interval" traits (f.i. calving to first insemination, calving to last insemination (= "days open"), calving to calving, etc.), as well as binary "non-return" (i.e. into oestrus) traits (pregnancy rate at 28, 56, … days after insemination) and stillbirth. Fertility is influenced by three individual animal components: dam (f.i. ability to resume oestrus cycle after calving, oocyte fertilizing capacity, uterine capacity), sire (f.i. semen fertilizing capacity), and offspring (f.i. developmental capacity of the embryo/fetus). The maternal component of fertility (= "female fertility") has received increasing attention as it has been steadily decreasing over the last twenty years (particularly in the most populous Holstein-Friesian dairy cattle population), and has become the primary cause for culling dairy cows. Female fertility is now commonly included in selection indexes (f.i. [1]). Male fertility is obviously an important trait for the artificial insemination (AI) industry, and consequently monitored very closely. The offspring's contribution to fertility is in essence not studied *per se*. Female and male fertility are characterized by low heritabilities (~5–10%), and - with the exception of reciprocal translocations (f.i. [2]) - it has proven difficult to reliably identify QTL, let alone genes, that influence these traits. Significant correlations between genomic and realized breeding values suggest a quasi-infinitesimal architecture involving many genes with individually very small effects (f.i. [3]).

Brachyspina syndrome (BS) is a rare ($<1/10^5$ birth) congenital defect that was recently described in Holstein-Friesian cattle [4–7]). Affected animals are characterized by severely reduced body weight, growth retardation, extensive vertebral malformations causing a significant shortening of the spine (brachyspina) and long and slender limbs. In addition, affected calves exhibit inferior brachygnatism (i.e. uneven alignment of the upper and lower teeth), as well as malformation of the inner organs, in particular the heart, kidneys and gonads. All reported cases trace back on both sire and dam side to *Sweet Haven Tradition*, a once popular sire, suggesting autosomal recessive transmission.

In this work, we use high-density SNP arrays and next generation sequencing (NGS) to identify the mutation causing BS. We make the unexpected observation that carrier frequency is as high as 7.4% in the Holstein-Friesian breed. We provide strong evidence that the discrepancy between disease incidence and carrier frequencies in this population is due – at least in part - to death of affected fetuses during pregnancy. We raise the hypothesis that other, as of yet unrecognized recessive lethal mutations likewise contribute to subfertility in cattle and propose a genotype-driven screen for their detection.

Results

Autozygosity-mapping positions the BS locus in a 2.5 Mb BTA21 interval

Between January 2008 and December 2009, we obtained biological material from six Holstein-Friesian calves diagnosed with BS, originating from Denmark, the Netherlands and Italy. Genomic DNA was extracted using standard procedures and genotyped using a previously described bovine 50 K SNP array [8]. Assuming that BS is inherited as an autosomal recessive defect and genetically homogeneous in Holstein-Friesian, the six cases are predicted to be homozygous for a common haplotype encompassing the causative mutation. We performed autozygosity mapping using the ASSIST program and 15 unaffected Holstein-Friesian bulls as controls, and identified a single, genome-wide significant peak ($p<0.001$) on chromosome 21 (BTA21) (Fig. 1A). The shared haplotype spans 2.46 Mb (bTau4.0: 20,132,767–22,588,403) encompassing 56 annotated genes (Fig. 1B, C).

Targeted and genome-wide resequencing identifies a 3.3 Kb deletion in the FANCI gene

Seven of the 56 genes in the interval are known to cause embryonic lethality when knocked out in the mouse. We amplified the corresponding open reading frames from genomic DNA of cases and controls but did not find any obvious disruptive DNA sequence variant. We then performed targeted sequencing of the entire 2.46 Mb interval. A custom sequence capture array (Roche Nimblegen) was designed based on the bovine bTau4.0 build, and used to enrich the corresponding sequences from total genomic DNA of two affected individuals prior to paired-end sequencing (2×36 bp) on an Illumina GAIIx instrument. Resulting sequence reads were mapped to the bTau4.0 build using Mosaik (http://bioinformatics.bc.edu/marthlab). In the targeted region, the coverage of non-repetitive bases averaged 90.45 (range: 0–336) for the first sample, and 61.28 (range: 0–189) for the second, to be compared with 0.01 (range: 0–24) for the first and 0.01 (range:0–104) for the second sample outside the targeted region. The proportion of targeted non-repetitive bases with coverage <10 was 0.12 for both samples. We used the GigaBayes software (Gabor T. Marth, Boston College, http://bioinformatics.bc.edu/marthlab) to identify polymorphisms and detected 2,368 SNPs and 572 insertion-deletions for a total of 2,940 variants. One thousand thirty two of these corresponded to polymorphisms previously reported in breeds other than Holstein-Friesian (Coppieters, personal communication), and were therefore eliminated as candidate causative mutations. Of the remaining 1,908 variants, only one was coding, causing a serine to glycine substitution in the *LOC516866* gene encoding a myosin light chain kinase-like protein. This variant was not considered to be a credible candidate mutation underlying BS.

We then generated mate-pair libraries from self-ligated 4 to 4.5 Kb fragments of one BS case and three unrelated healthy controls, and generated ~3.4 Gb of sequence on a Illumina GAIIx instrument for each animal. Resulting reads were mapped to the bTau4.0 build using the Burrows-Wheeler Aligner (BWA) [9], and alignments visualized with the Integrative Genomics Viewer (IGV) [10]. The achieved sequence coverage averaged 1.7× per non-repetitive base. Analysis of the reads mapping to the 2.46 Mb interval readily revealed a 3.3 Kb deletion removing exons 25–27 of the 37 composing the *FANCI* (Fanconi anemia complementation-group I) gene (Fig. 1D, Fig. 2A). The deletion was apparent from a cluster of 27 mate-pairs mapping ~8 Kb apart on the bTau4.0 build, and from the complete absence of reads mapping to the deleted segment for the BS case, contrary to the three controls showing normal, uniform coverage in the region (Fig. S1). Retrospective analysis of the sequence reads obtained by targeted capture from affected individuals confirmed the abrupt coverage drop at the exact same location. We designed a primer pair spanning the presumed deletion, allowing productive amplification of a 409 bp product from genomic DNA of affected and carrier animals but not of unrelated unaffected controls from the same or other breeds (Fig. 2A, B). Conversely, primer pairs designed within the deletion did not yield any amplification from DNA of affected individuals compared to unaffected ones (Fig. 2A, C). Sequencing the deletion-specific amplicon defined the breakpoints, confirming a 3,329 bp deletion (Fig. 2B). Analysis of the sequence traces captured from affected individuals identified several reads bridging and confirming the breakpoint. We didn't note any obvious sequence similarity between the breakpoints.

Assuming that the deletion of exons 25 to 27 results in the juxtaposition of exons 24 and 28 in the mRNA, the 3.3 Kb deletion is predicted to cause a frame-shift at amino-acid position 877, substituting the 451 carboxy-terminal amino-acids with a 26-residue long illegitimate peptide (Fig. 3A). Moreover, the ensuing stop codon in exon 28 is expected to cause nonsense mediated RNA decay (NMRD)(Fig. 3B). To examine the effect of the 3.3 Kb deletion on *FANCI* transcripts, we designed primers in exons 24 and 28 and performed RT-PCR experiments using total RNA of leucocytes from carrier and wild-type bulls. In agreement with our prediction, we amplified a smaller, 96 bp fragment from carriers but not from homozygous wild-type animals. Sequencing the 96 bp RT-PCR product indeed confirmed the juxtaposition of exons 24 and 28 in mutated *FANCI* transcripts. Despite its smaller size, the 96 bp fragment was less abundant than the 457 bp fragment corresponding to wild-type *FANCI* transcripts, supporting NMRD (Fig. 3B).

With its homologue FANCD2, the FANCI protein forms the ID complex that localizes to damage-induced chromatin foci. FANCI is essential for DNA interstrand crosslink repair. Like FANCD2, FANCI is mono-ubiquitinated by the ubiquitin ligase FA core complex, and phosphorylated by the ATM/ATR kinase (f.i. [11]). Missense, nonsense and splice-site variants in the *FANCI* gene underlie ~2% of Fanconi anemia (FA) cases in human [11,12]. FA patients exhibit heterogeneous symptoms, including growth retardation, skeletal abnormalities, renal, cardiac, gastrointestinal and reproductive malformations (reminiscent of BS), as well as bone marrow failure, early onset of cancer and mortality at a young age. Thus, *FANCI* qualified as a valid candidate gene.

To further support the causality of the 3.3 Kb deletion in the *FANCI* gene, we developed a genotyping assay that simultaneously interrogates the mutant and wild-type allele (cfr. M&M). As expected, all available BS cases were homozygous for the deletion. The deletion proved to be absent in a sample of 131 healthy animals representing ten breeds other than Holstein-Friesian. We then genotyped a random sample of 3,038 unaffected Dutch Holstein-Friesian animals. Carriers of the deletions accounted for 7.4% of the sample, while no animals were found to be homozygous. Assuming Hardy-Weinberg equilibrium, the absence of homozygous animals in a sample of 3,038 individuals has probability <5%. Taken together, these findings support the causality of the 3.3 Kb *FANCI* deletion.

The 3.3 Kb Fanci deletion compromises fertility in carrier x carrier matings

We readily noticed the discrepancy between the frequency of carriers and the number of reported cases. With a carrier frequency of 7.4%, BS should represent close to 1/730 (≈0.074×0.074×0.25) births. We reasoned that this could indicate

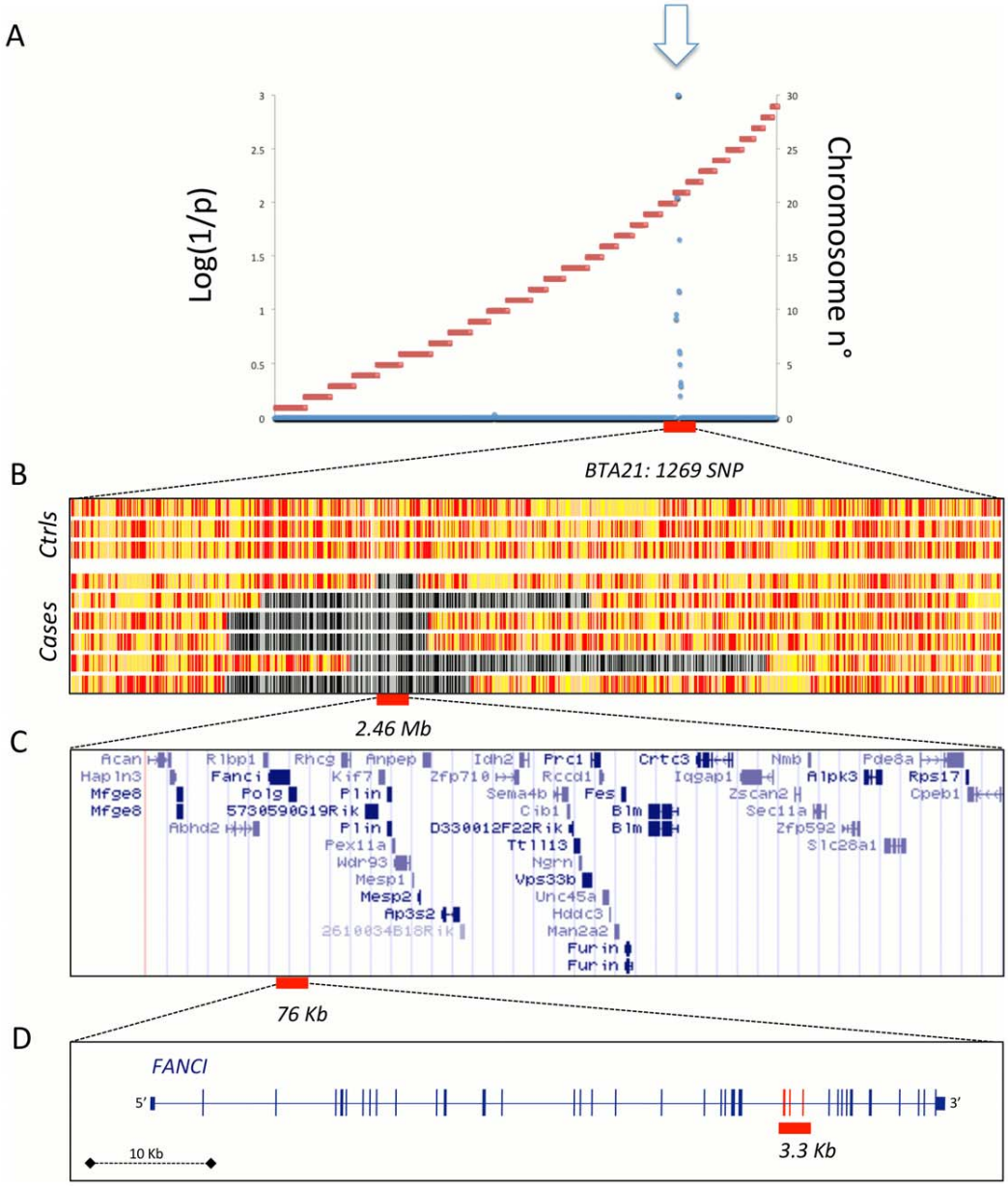


Figure 1. The BS locus maps to a 2.5 Mb interval on BTA21. (**A**) Autozygosity mapping of the BS locus on bovine chromosome 21 using ASSIST (Charlier et al., 2008). Blue dots measure the genome-wide probability that the six BS cases would share the observed segment of autozygosity by chance alone at each SNP position. The horizontal red lines mark the limits between adjacent chromosomes (numbered on the right Y axis). Inset: BS calf. (**B**) Genotype of three unaffected controls and six BS cases at 1,269 BTA21 SNP positions. Homozygous genotypes are shown in yellow or white, heterozygous genotypes in red. Homozygous segments encompassing the BS locus in the six cases are shown in black and white. (**C**) Gene content of the 2.46 Mb segment of autozygosity. (**D**) Map of the bovine *FANCI* gene with indication of the 3.3 Kb BS-causing deletion.
doi:10.1371/journal.pone.0043085.g001

that the majority of homozygous mutant conceptuses die during pregnancy. To test this hypothesis, we compared the pregnancy failure (i.e. return in oestrus) rate at 56, 90, and 270 days post-insemination for matings between (i) a non-carrier dam and non-carrier sire, (ii) a carrier dam and non-carrier sire, (iii) a non-carrier dam and carrier sire, and (iv) a carrier dam and carrier sire. In these carrier/non-carrier status of the sire is known with certainty (from genotyping), while carrier dams are defined as daughters of carrier sires. Thus, dams defined as carrier have ~53.7% probability to carry the BS mutation (50% probability due to the maternal grand-sire, and ~3.7% probability due to the ungenotyped maternal grand-dam), while dams defined as non-carriers still have ~3.7% probability to carry the BS mutation (transmitted by carrier maternal grand-dams).

When compared to non-carrier x non-carrier matings, failure rate at 270 days was increased by ~7% in carrier x carrier matings ($p = 5\ 10^{-259}$)(Fig. 4 and Table S1). This suggests that ~52% of the ~13.4% (≈0.537×0.25) homozygous mutant fetuses expected from such matings die during pregnancy. More than halve of those appear to die before day 56 of gestation, but increased mortality is

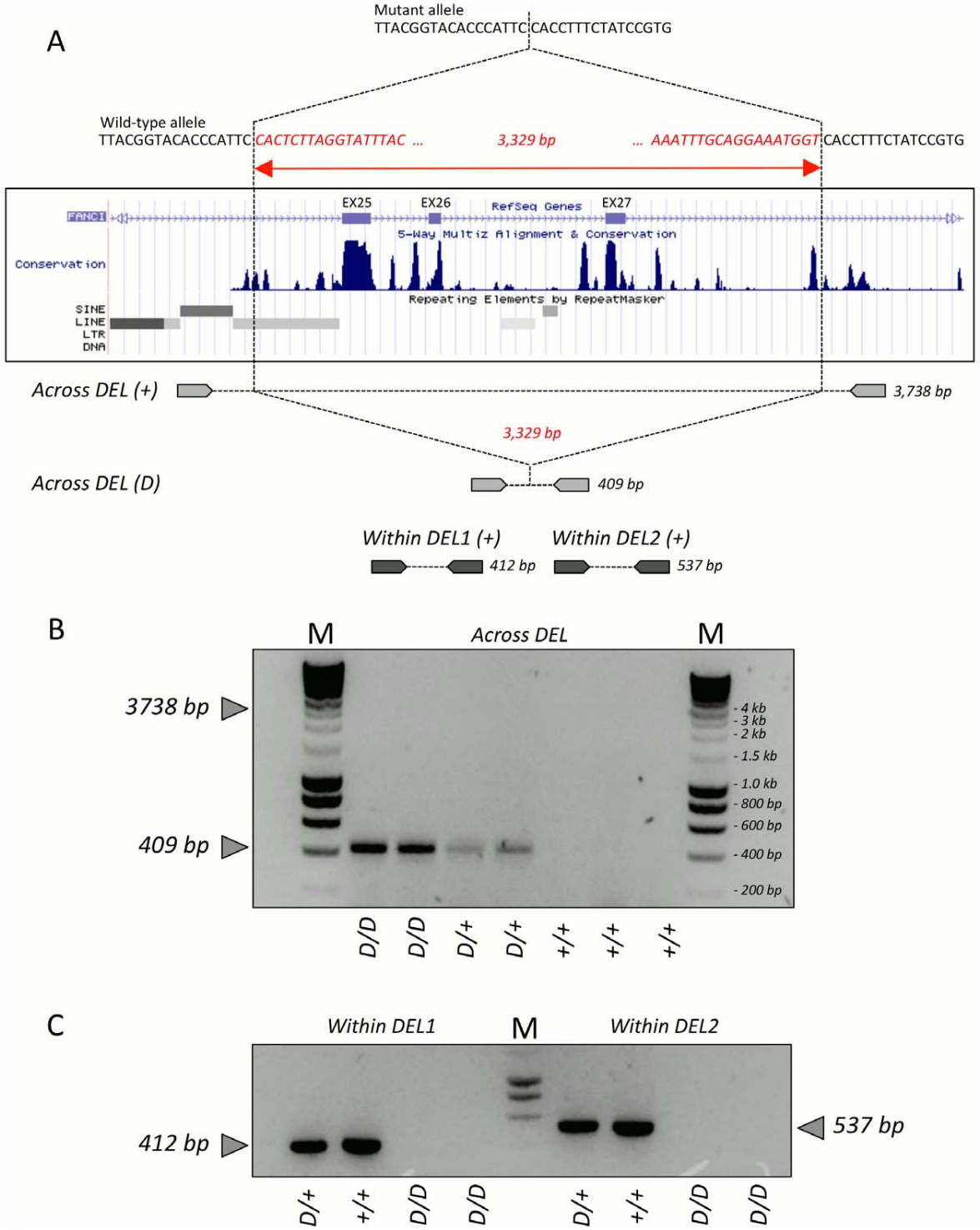


Figure 2. A 3.3 Kb deletion is identified in the *FANCI* gene. (**A**) Sequences and position within the *FANCI* gene of the BS deletion; position and size of amplicons used to validate the deletion. (**B&C**) PCR amplification of amplicons spanning (**B**), and residing within (**C**) the BS deletion in BS cases *(D/D)*, BS carrier *(D/+)* and homozygous wild-type *(+/+)* animals.
doi:10.1371/journal.pone.0043085.g002

still detected between days 56 and 90 as well as between days 90 and 270 of gestation (Fig. 4 and Table S1).

Discussion

In this work, we provide strong evidence that bovine BS, a newly described congenital defect in Holstein dairy cattle, is due to

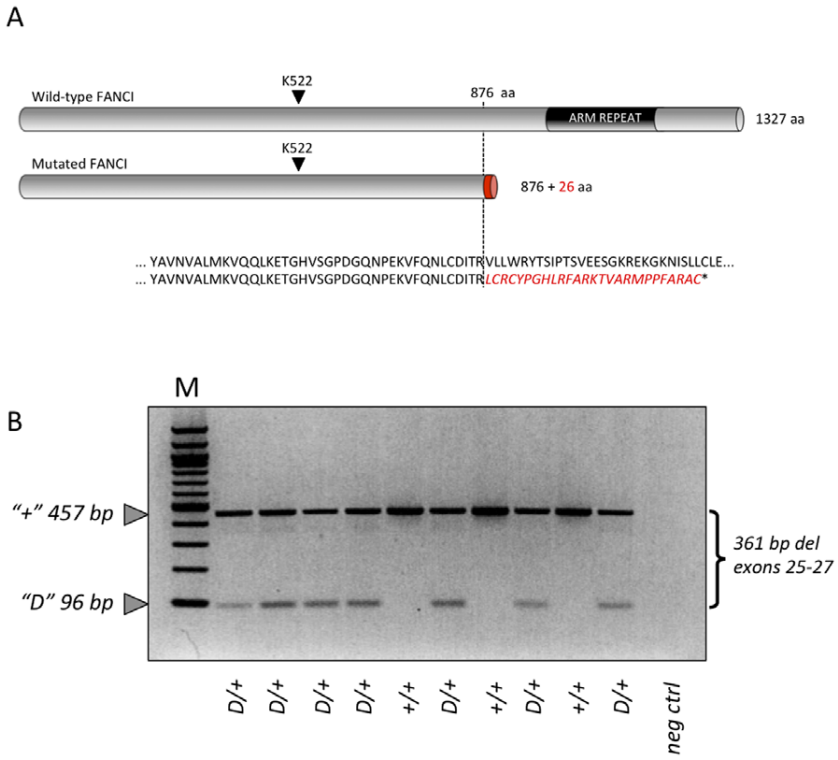


Figure 3. The *FANCI* deletion creates a premature stop codon responsible for mutant mRNA degradation. (**A**) Predicted effect of the BS deletion on the encoded FANCI protein. (**B**) RT-PCR products obtained from leucocyte cDNA from carrier *(D/+)* or wild-type *(+/+)* individuals using primers targeting exons 24 and 28, respectively, showing a less abundant, smaller (96 bp vs 457 bp) amplification product in *D/+* but not *+/+* animals derived from the mutant *FANCI* allele. M: molecular weight marker (100 bp ladder).
doi:10.1371/journal.pone.0043085.g003

a 3.3 Kb deletion encompassing exons 25 to 27 of the bovine *FANCI* gene. The deletion was the only obviously disruptive mutation found by resequencing the entire 2.46 Mb interval shared autozygous by all examined cases. Moreover, several of the lesions of BS, including growth retardation, skeletal defects and malformation of the kidneys, heart and reproductive system, are shared by patients suffering from FA.

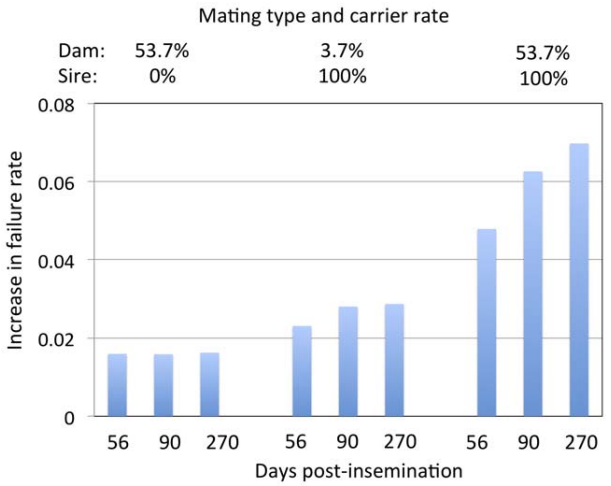


Figure 4. The *FANCI* deletion compromises fertility. Increased pregnancy failure rate detected as a return into oestrus at 56, 90 and 270 days post-insemination in C x NC, NC x C and C x C (over NC x NC matings). All contrasts had p-value$<10^{-4}$.
doi:10.1371/journal.pone.0043085.g004

After initial autozygosity mapping of the BS locus using a medium density 50 K SNP array, we relied on next generation sequencing to identify the causative mutation. We first performed targeted sequencing of two affect individuals using sequence capture arrays, but did not identify the causative mutation despite a sequence depth of ~90 for the first sample, and ~60 for the second. This was due to the fact that we could not be confident that the observed 3.3 Kb deletion in the *FANCI* gene, indeed obvious from the captured sequence reads, was due to the failure to capture the corresponding segment or to a genuine deletion. It demonstrates the need to include a healthy control in the targeted sequencing experiment. Only when analyzing the reads from the genome-wide mate-pair libraries, generated for cases and controls, did the deletion become clear.

A striking feature of BS is the discrepancy between the low incidence of the condition ($\sim 1/10^5$ births), yet the high frequency of carrier individuals in the Holstein-Friesian population (~7.4%). We reconciled these contradictory findings in part by providing strong evidence that at least half of BS homozygous mutant fetuses die before birth. However, this still leaves a large proportion of expected BS cases unaccounted for. Indeed, assuming a mortality of 50% of homozygous conceptuses, and given the 7.4% of carrier animals, one still expects one BS case per 1,500 newborn calves. The reasons for the remaining discrepancy remain unclear, yet we speculate that it might include (i) the fact that matings between carriers occur at a frequency $<0.074\times0.074=0.0055$, as BS farmers control the rate of inbreeding by avoiding matings

between closely related animals (all BS carriers trace back to *Sweet Haven Tradition*), and (ii) the fact that the statistics on failure rates at 56, 90 and 270 days do not account for inseminated cows that were culled without reported calving ("removers"). A significant proportion of these animals is known to be culled because they are not pregnant.

We noticed the significant increase in pregnancy failure rate in crosses between carrier dams and non-carrier sires as well as between non-carrier dams and carrier sires (Fig. 4 and Table S1). While in the latter cross the excess failure could be due to the fact that ~3.7% of non-carrier dams are predicted to be misclassified (having inherited the BS mutation from an expected 7.4% of grand-dams), the excess failure rate in the former is more intriguing. We are presently examining whether this observation reflects a polygenic effect associated with the BS mutation, or whether it involves other biological mechanisms. Preliminary results point towards segregation distortion and effects on global recombination rate of the BS mutation (Charlier, unpublished observations). The relationship between these two observations and increased pregnancy failure rate in crosses involving one carrier parent are currently being examined.

Our findings strongly suggest that the main economic impact of the BS mutation is through its effect on fertility, one of the economically most important traits in cattle breeding. Similar observations were previously reported in the same breed for Complex Vertebral Malformation (CVM) [13–16]. The BS and CVM mutations could only be identified via the genetic analysis of a few surviving homozygous mutant offspring. Genetic defects, for which 100% of homozygous mutant conceptuses would die during early pregnancy, would essentially go unnoticed, as seen for deficiency of uridine monophosphate synthase [17]. Recent data from the 1,000 Genomes Project indicates that naturally occurring null alleles, of which a proportion is bound to be embryonic lethal, may be more common than initially suspected [18], and the same may apply in livestock. Next generation sequencing (NGS) technology offers the opportunity to detect such embryonic lethal mutations using a genotype-driven screening approach. We plan to use emerging capturing reagents for the bovine exome in combination with NGS to screen for coding variants that are predicted to be disruptive in 100–200 individuals from cattle breeds of interest. Subsequent genotyping of ~5,000 animals for a list of candidate mutations with MAF≥0.03 should allow detection of a statistically significant depletion of homozygotes amongst healthy individuals supporting a deleterious effect possibly manifesting itself by its effect on fertility.

Materials and Methods

Ethics statement

Blood samples were collected from sires, cows and calves by trained veterinarians following standard procedures and relevant national guidelines. The samples were collected specifically for this study, with the full agreement of the farmers who owned the animals. According to the Ethics Commission of University of Liège, formal ethical approval is not required under these circumstances.

Autozygosity mapping

DNA extraction and SNP genotyping using a custom-made bovine 50 K SNP array were conducted using standard procedures as previoulsy described [8]. Autozygosity mapping, including permutation testing, was conducted using the previously described ASSIST software [8].

Targeted resequencing

Resequencing of the 2.46 Mb candidate region was performed on two BS affected animals, one Italian and one Danish calf, using a custom made sequence capture array (Roche Nimblgen) followed by sequencing. The capture array was designed based on the bovine bTau4.0 build. Repetitive regions were excluded in the design in accordance with the manufacturer's quality parameters. This resulted in an array with a capacity to capture 84.5% the target region. Twenty-two µg gDNA from each animal was fragmented to a size of 250–500 bp by nebulization. Linkers for sequencing library construction were annealed and capture was performed following the manufacture's protocol. Capture resulted in approximately 500-fold enrichment of targeted DNA and elimination of DNA from outside the target region, as evaluated by qPCR. Paired-end sequencing (2×36 bp) on captured DNA was performed on an Illumina GAIIx. Resulting sequence traces were mapped to the bTau4.0 build using the Mosaik assembler (http://bioinformatics.bc.edu/marthlab) and the GigaBayes software (Gabor T. Marth, Boston College, http://bioinformatics.bc.edu/marthlab) was used to identify DNA Sequence Variation.

Genome wide-resequencing

One affected individual, homozygous for the defined IBD haplotype was selected as well as three unaffected unrelated individuals from other breeds. The Mate Pair Library Prep Kit v2 from Illumina was used to generate a ~400–450 bp paired-end sequencing library from ~4 to 4.5 kb genomic DNA fragments for each animal. Briefly, total genomic DNA was extracted and fragmented by nebulization, 4 to 4.5 kb fragments were end-repaired with biotin labeled dNTPs. After circularization, non-circularized DNA was removed by digestion. Remaining biotinylated circular DNA was fragmented and affinity purified. Purified fragments were end-repaired and ligated to Illumina Paired-End sequencing adapters. Each library was sequenced on one lane of the flow-cell of a Illumina GAIIx with the Paired-End module to generate high-quality reads (2×76 bp). Reads were mapped and analyzed with publicly available software: *Burrows-Wheeler Alignment Tool* (http://bio-bwa.sourceforge.net) and *Samtools* (http://samtools.sourceforge.net). The output files were readily uploaded in the *Integrative Genomics Viewer* (IGV, [10]) and visually scrutinized for structural variation.

Mutation validation and definition of the deletion breakpoint

Three primer pairs were designed, one across the putative breakpoint and two within the putative deleted region. Corresponding primer pairs are listed in Table S2. They were used to amplify products from genomic DNA of homozygous cases, carriers and unaffected unrelated individuals using standard procedures. Amplicons were directly sequenced using the Big Dye terminator cycle sequencing kit (Applied Biosystems Foster City, CA). Electrophoresis of purified sequencing reactions was performed on an ABI PRISM 3730 DNA analyzer (PE Applied Biosystems, Foster City, CA). Sequence traces were aligned and compared to bovine reference using the Phred/Phrap/Consed package (www.genome.washington.edu).

Effect of the 3.3 kb deletion on FANCI transcripts

Whole blood was collected on EDTA from carrier and control bulls, white blood cells were recovered and total RNA was extracted using the RiboPure^{TM}-Blood Kit (Ambion) following manufacturer instrcutions. The RNA was treated with TurboD-

NaseI (Ambion). cDNA was synthesized using Superscript^{TM}III First Strand Synthesis System for RT-PCR (Invitrogen). A portion of *FANCI* cDNA, across exons 25 to 27, was amplified using a pair of *FANCI* specific primers (Table S2). The PCR products were directly sequenced as described above.

Developing a genotyping test for the 3.3 Kb mutation

A 5′ exonuclease assay was developed to genotype the BS deletion, using 5′-TGT TAG CCC AGC AGA GGA-3′ and 5′-ATT CTG AAT CCA CTA GAT GTC-3′ as wild-type PCR primer pair combined with 5′-GCA CAC ACC TAT CTT ACG GTA C-3′ and 5′-GGG AGA AGA ACT GAA CAG ATG G-3′ as mutant PCR primer pair, and 5′HEX-AGT CCC AGT GTG GCT AAG GAG TGA-3′IABkFQ (wild-type) and 5′FAM-CCA TTC CAC/ZEN/CTT TCT ATC CGT GTC CT-3′IABkFQ (mutant) as probes (Integrated DNA Technologies, Leuven, Belgium). Allelic discrimination reactions were carried out on an ABI7900HT instrument (Applied Biosystems, Fosters City, CA) for 40 cycles in 2.5 µl volume with a final concentration of 250 nM for each probe, 500 nM for wild-type primers, 350 nM for mutant primers, Taqman Universal PCR Master Mix 1× (Applied Biosystems, Fosters City, CA) and 10 ng of genomic DNA.

Supporting Information

Figure S1 Targeted and genome-wide resequencing of BS cases and controls. (**A**) Distribution of the genomic distance separating random mate-pairs and mate-pairs flanking the BS deletion. (**B**) IGV screen captures of mate-pair reads mapping to the BTA21 20,536,086–20,541,232 chromosome interval for three unaffected controls (lanes 1–3) and a BS calve (lane 4), as well as paired-end reads obtained from captured DNA of a BS calve (lane 5).
(PDF)

Table S1 Effects on fertility of the *FANCI* deletion. Pregnancy failure rate detected as 100% minus non return into oestrus (NR) at 56, 90 and 270 days post-insemination in the four possible matings. The genotype probabilities of the dams are estimated from the knowledge of the genotype of their sire combined with the known frequency of the BS mutation in the general population.
(PDF)

Table S2 Primer pairs to validate the deletion in the *FANCI* gene.
(PDF)

Acknowledgments

We thank the GIGA genotranscriptomic core facility for their assistance. Prof. A Gentile and Dr. K. Peperkamp are acknowledged for providing materials of Italian and Dutch defective calves, respectively. We are grateful to Dr. A. Sartelet for collecting fresh white blood cell.

Author Contributions

Conceived and designed the experiments: CC MG MF. Performed the experiments: PKM WL CF LK SC NC NA. Analyzed the data: CC PKM WC GDJ MG MF. Contributed reagents/materials/analysis tools: JSA EM. Wrote the paper: CC MG MF.

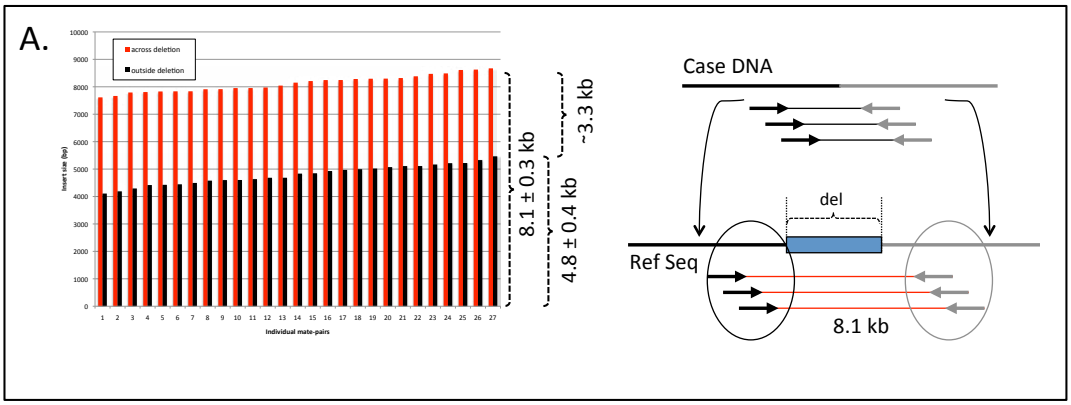
References

1. Berglund B (2008) Genetic improvement of dairy cow reproductive performance. Reprod Dom Anim 43 (Suppl 2): 89–95.
2. Fries R, Popescu P (1999) Cytogenetics and physical chromosome maps. Pages 247–328. in The Genetics of Cattle. Editors: R. Fries A. Ruvinsky. CABI Publishing.
3. Wiggans GR, VanRaden PM, Cooper TA (2011) The genomic evaluation system in the United States: past present and future. J Dairy Sci 94: 3202–3211.
4. Agerholm JS, McEvoy F, Arnbjerg J (2006) Brachyspina syndrome in a Holstein calf. J Vet Diagn Invest 18:418–422.
5. Agerholm JS, Peperkamp K (2007) Familial occurrence of Danish and Dutch cases of the bovine brachyspina syndrome. BMC Vet Res 3:8.
6. Agerholm JS, DeLay J, Hicks B, Fredholm M (2010) First confirmed case of the bovine brachyspina syndrome in Canada. Can Vet J 51:1349–1350.
7. Testoni S, Diana A, Olzi E, Gentile A (2008) Brachyspina syndrome in two Holstein calves. Vet J 177: 144–146.
8. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, et al. (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet 40:449–454.
9. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25:1754–1760.
10. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative Genomics Viewer. Nat Biotechnol 29: 24–26.
11. Smogorzewska A, Matsuoka S, Vinciguerra P, McDonald ER, Hurov KE, et al. (2007) Identification of the FANCI protein, a monoubiquitinated FANCD2 paralog required for DNA repair. Cell 129: 289–301.
12. Dorsman JC, Levitus M, Rockx D, Rooimans MA, Oostra AB, et al. (2007) Identification of the Fanconi anemia complementation group I gene, FANCI. Cell Oncol 29: 211–218.
13. Agerholm JS, Bendixen C, Andersen O, Arnbjerg J (2001) Complex vertebral malformation in Holstein calves. J Vet Diagn Invest 13:283–289.
14. Nielsen US, Aamand GP, Andersen O, Bendixen C, Nielsen VH, et al. (2003) Effects of complex vertebral malformation on fertility traits in Holstein cattle. Livestock Prod Sci 79: 233–238.
15. Berglund B, Persson A, Stålhammar H (2004) Effects of complex vertebral malformation on fertility in Swedish Holstein cattle. Acta Vet Scand 45:161–165.
16. Thomsen B, Horn P, Panitz F, Bendixen E, Petersen AH, et al. (2006) A missense mutation in the bovine SLC35A3 gene, encoding a UDP-N-acetylglucosamine transporter, causes complex vertebral malformation. Genome Res 16:97–105.
17. Shanks RD, Robinson JL (1989) Embryonic mortality attributed to inherited deficiency of uridine monophosphate synthase. J Dairy Sci 72: 3035–3039.
18. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828.

SUPPORTING MATERIAL

Figure S1: Targeted and genome-wide resequencing of BS cases and controls.

(A) Distribution of the genomic distance separating random mate-pairs and mate-pairs flanking the BS deletion. **(B)** IGV screen captures of mate-pair reads mapping to the BTA21 20,536,086 - 20,541,232 chromosome interval for three unaffected controls (lanes 1-3) and a BS calve (lane 4), as well as paired-end reads obtained from captured DNA of a BS calve (lane 5).



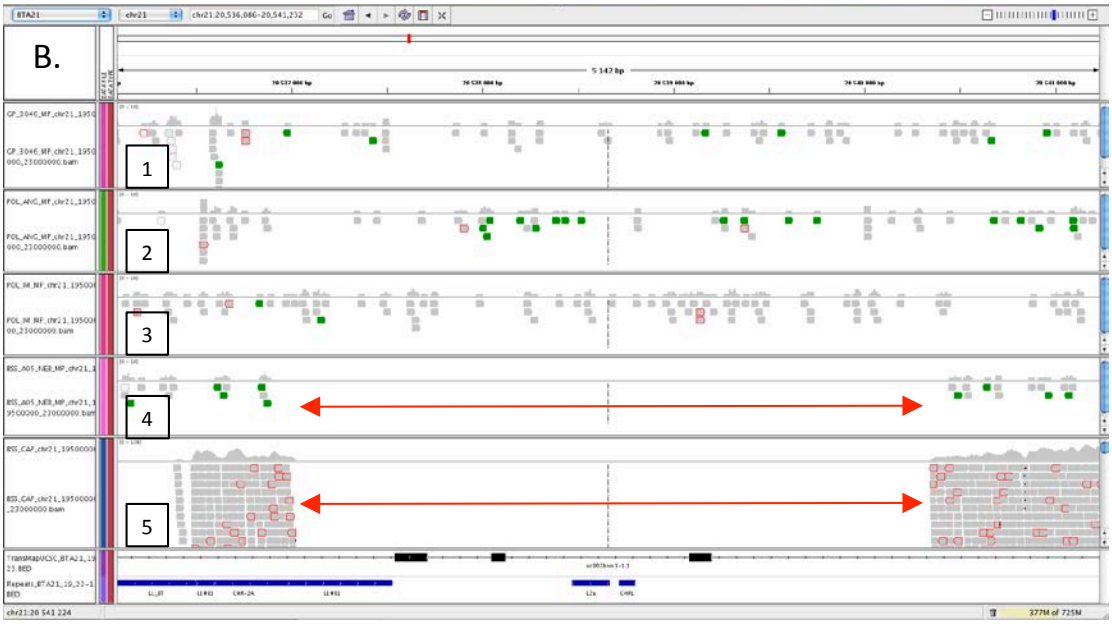


Table S1: Effects on fertility of the *FANCI* deletion.

Pregnancy failure rate detected as 100% minus non return into oestrus (NR) at 56, 90 and 270 days post-insemination in the four possible matings. The genotype probabilities of the dams are estimated from the knowledge of the genotype of their sire combined with the known frequency of the BS mutation in the general population.

Mating type	Phenotype	**WT dam (96.3% +/+)**		**Carrier dam (53.7% *D*/+)**	
WT sire (100% +/+)	NR56	9,391,260	34.69	1,204,592	36.29
	NR90		41.36		42.94
	NR270		45.74		47.37
Carrier sire (100% *D*/+)	NR56	1,025,964	37.00	112,721	39.48
	NR90		44.16		47.61
	NR270		48.61		52.71
		Total	Failure (%)	Total	Failure (%)

Table S2: Primer pairs to validate the deletion in the *FANCI* gene.

Primer name	Primer sequence 5'-3'	Product size *D allele*	Product size *+ allele*	Material
AcrossDEL_UP1	GCTCAAGTAGTTAGTTGCTCCACTG	409 bp	(3738 bp)	gDNA
AcrossDEL_DN1	ATAAATAAATAAAGCAGGATGCTGAAA			
WithinDEL_UP1	TCACAAAAGGGTAGGAGACTACCTG	/	537 bp	gDNA
WithinDEL_DN1	GCTTATTGTTTACCCTTGACAGTGG			
WithinDEL_UP2	ACTGGATTCCATTTAACCACAGATG	/	412 bp	gDNA
WithinDEL_DN2	ATGCATTACCTTTCATTCTCAGAGC			
Exon24_UP	GCCAAAACCCAGAGAAGGTC	96 bp	457 bp	cDNA
Exon28_DN	CAACTGTTTTCCGTGCAAAT			

Part III. A reverse genetic approach to screen embryonic lethal mutations in HTS context

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

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

* Contributed equally to this work

Accepted by Genome Research.

This paper is listed as a "landmark article" by the ONLINE MENDELIAN INHERITANCE IN ANIMALS (OMIA) website, http://omia.angis.org.au/key_articles/landmarks/.

Background

Owing to intensive selection for production traits and the extensive application of artificial insemination in farm animals over the last decades, farm animals have experienced an increased level of inbreeding and consequently shrinkage in effective population size. The repeated emergence of recessive defects and a decline in fertility have been observed in farm animals in recent years. We had uncovered some interesting results when we studied the brachyspina syndrome (BS) in Holstein dairy cattle. BS is a rare congenital defect ($< 1/10^5$ birth) in Holstein-Friesian cattle, characterized by significant shortening of the spine, long and thin limbs, growth retardation and usually delivered as a stillbirth (Charlier et al., 2012). By applying IBD mapping and HTS, we found a 3.3-kb deletion removing three exons of *FANCI* gene as the causative mutation for the disease. Further examination revealed a ~7.4% carrier frequency of the deletion in Holstein-Friesian dairy cattle. Therefore, many more newborn animals were expected to be affected by BS than those observed (incident rate at $0.074 * 0.074 / 4 \approx 0.0014$). Using field fertility data, we estimated that at least half of the homozygous mutant embryos/fetuses die during pregnancy. We postulated that other embryonically lethal variants (ELV) might segregate within cattle population.

High-throughput sequencing provides an approach to detect those ELV in a reverse genetic manner. We set out to sequence hundreds of highly used elite sires using exome and whole-genome sequencing and systematically screen loss-of-function (LoF) and deleterious missense (DM) variants in cattle. We then prove embryonic lethality of some LoF and DM variants by genotyping the candidate variants in large cohort. When a candidate stands out as probable ELV, we went further to carry out a prospective study by following carrier × carrier mating to test if there is a lack of homozygotes in the offspring.

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

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

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

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

[Supplemental material is available for this article.]

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

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

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

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

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

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

Results

Expectations for the number of EL mutations carried per individual

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

Identification of ~94,000 nonsynonymous variants in domestic cattle

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

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

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

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

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

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

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

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

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

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

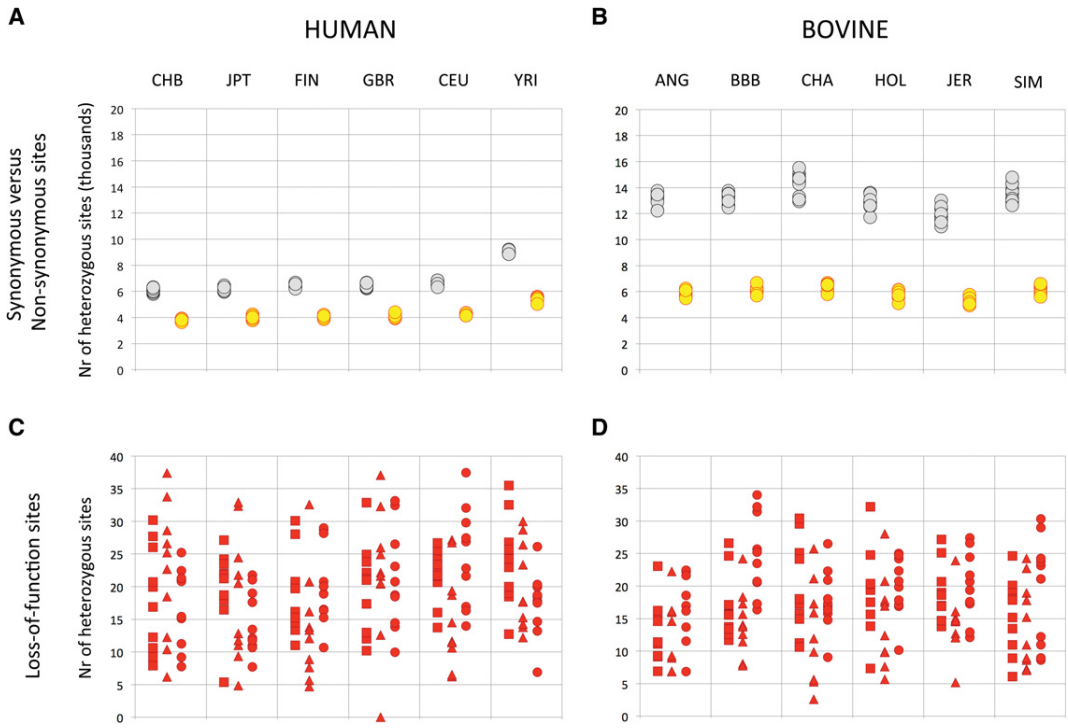


Figure 1. (*A*,*B*) Number of heterozygous synonymous (gray) and nonsynonymous (yellow) sites per individual (*A*: humans; *B*: bovine). (*C*,*D*) Number of heterozygous stop-gain (squares), splice-site (triangles), and frame-shift (circles) sites per individual. CHB: Chinese; JPT: Japanese; FIN: Finns;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

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

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

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

Identifying nonlethal coding variants with phenotypic effects

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

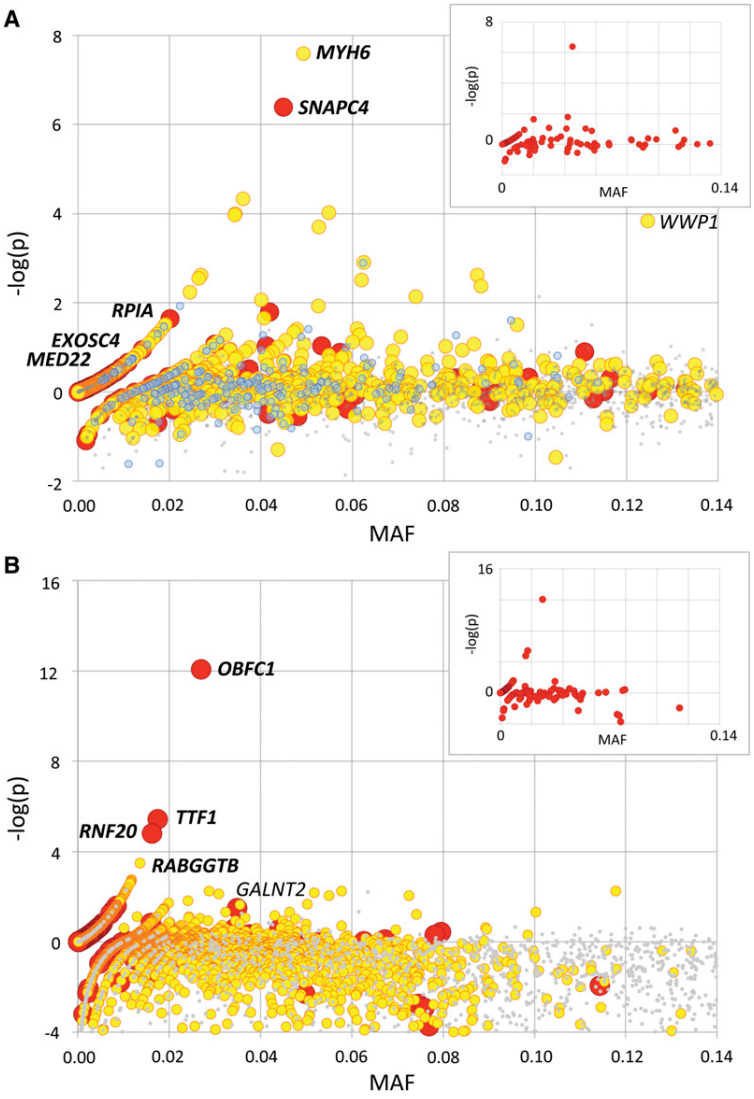


Figure 2. Statistical significance [−log(*p*): *y*-axis] of the depletion (positive values) or excess (negative values) in homozygotes for loss-of-function (red; defined as frame-shift, splice-site, and stop-gain variants), missense (yellow), matched synonymous (blue), and random neutral (small gray)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

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

Lack of evidence for synergistic epistasis

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

Discussion

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

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

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

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

OMIA 002042-9913 Abortion (embryonic lethality), EXOSC4 in Bos taurus (cattle) Gene: EXOSC4.
OMIA 002043-9913 Abortion (embryonic lethality), MED22 in Bos taurus (cattle) Gene: MED22.
OMIA 002039-9913 Abortion (embryonic lethality), MYH6 in Bos taurus (cattle) Gene: MYH6.
OMIA 002035-9913 Abortion (embryonic lethality), OBFC1 in Bos taurus (cattle) Gene: OBFC1.
OMIA 002037-9913 Abortion (embryonic lethality), RABGGTB in Bos taurus (cattle) Gene: RABGGTB.
OMIA 002038-9913 Abortion (embryonic lethality), RNF20 in Bos taurus (cattle) Gene: RNF20.
OMIA 002041-9913 Abortion (embryonic lethality), RPIA in Bos taurus (cattle) Gene: RPIA.
OMIA 002040-9913 Abortion (embryonic lethality), SNAPC4 in Bos taurus (cattle) Gene: SNAPC4.
OMIA 002036-9913 Abortion (embryonic lethality), TTF1 in Bos taurus (cattle) Gene: TTF1.
[a]FS: frame-shift; SG: stop-gain; MS: missense; Del: deletion; SS: splice-site.
[b]BBC: Belgian Blue Cattle Breed; JER: New Zealand Jerseys; HF: New Zealand Holstein-Friesian.
[c]MM/T: *P*-value of MM/total ratio assuming viable MM genotype; ++/+M: *P*-value of ++/+M ratio assuming lethality of MM genotype. (*) $P < 0.05$; (**) $P < 0.01$; (***) $P < 0.001$.
[d]The *MYH6* p.Lys1730del mutation is in high linkage disequilibrium with a p.Thr202Met missense mutation in the *AP1G2* gene for which healthy homozygous individuals have been observed and which can therefore be excluded as being responsible for the observed EL effect.

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

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

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

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

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

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

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

Methods

Simulations

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

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

Next generation sequencing

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

Variant calling

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

Comparing the mutational load of human and bovine exomes

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

Testing for depletion in homozygosity

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

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

and

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

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

Estimating the proportion of ELs among LoF and missense variants

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

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

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

where *b* and *a* are defined as above.

Testing for synergistic epistasis

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

Data access

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

Acknowledgments

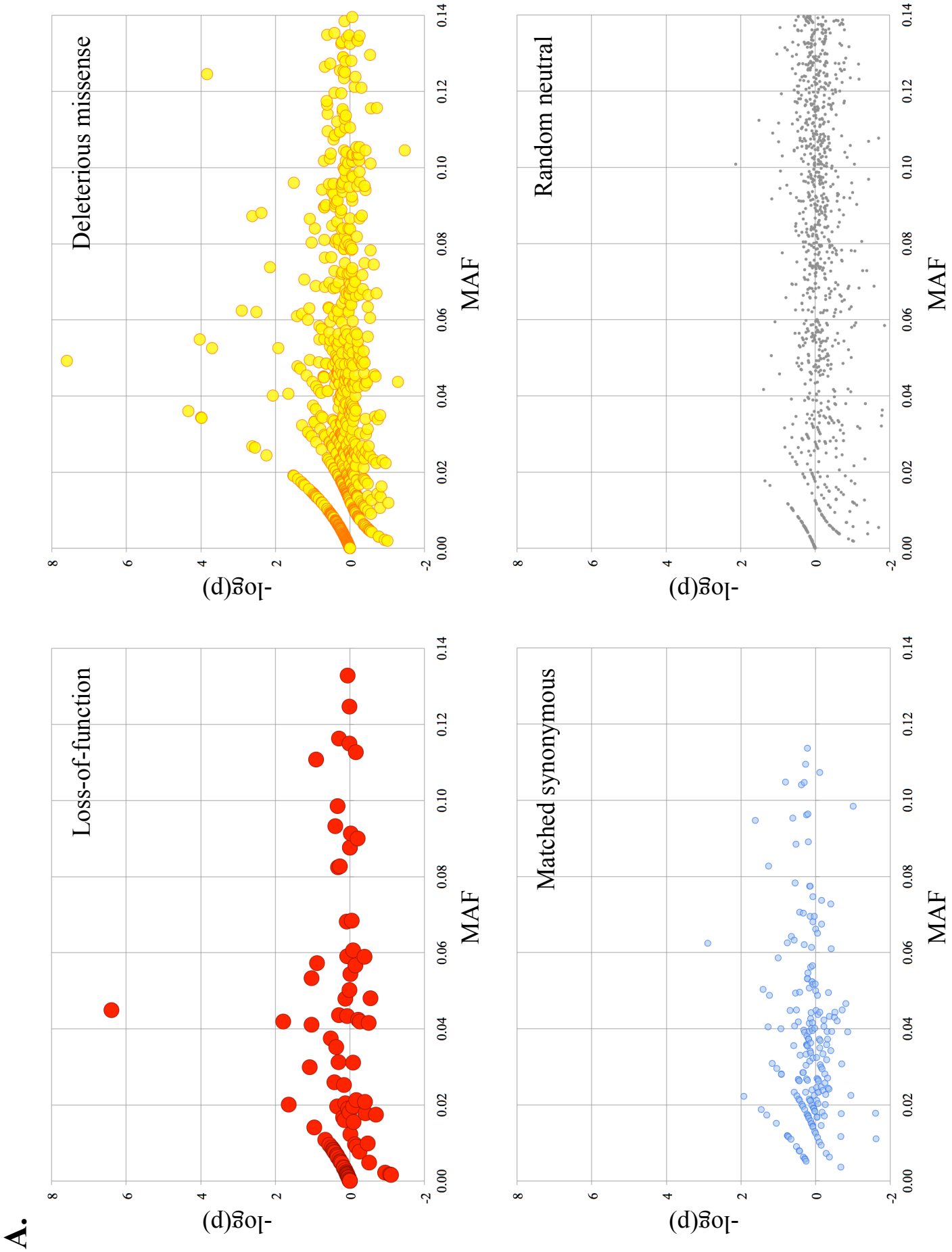
This work was supported by grants from the European Research Council (Damona), the Walloon Region (DGARNE Rilouke), and the EU Framework 7 program (GplusE). We used the supercomputing facilities of the Consortium des Equipements de Calcul Intensitf en Fédération Wallonie Bruxelles (CECI) funded by the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

References

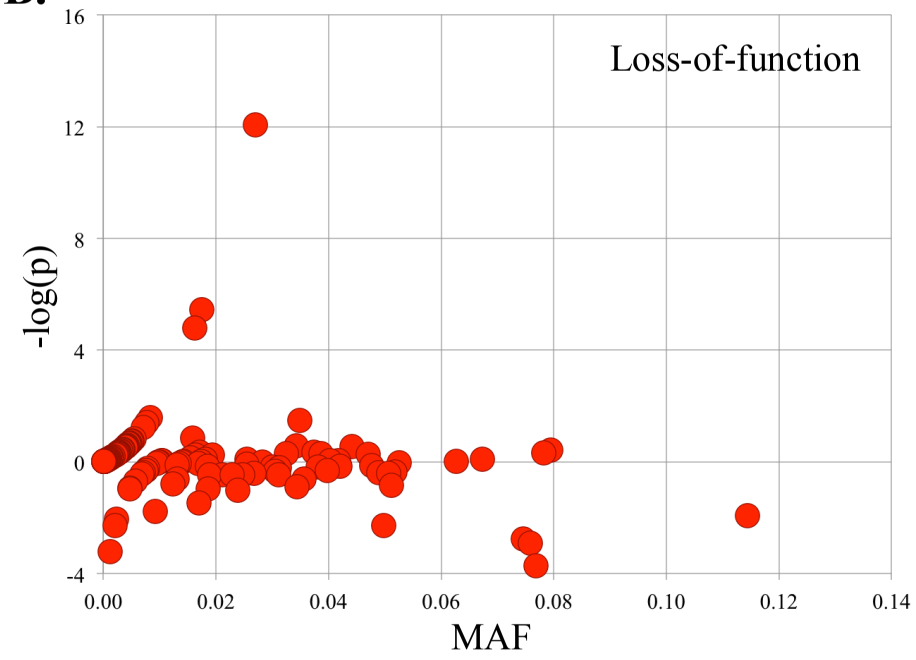
The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.
The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

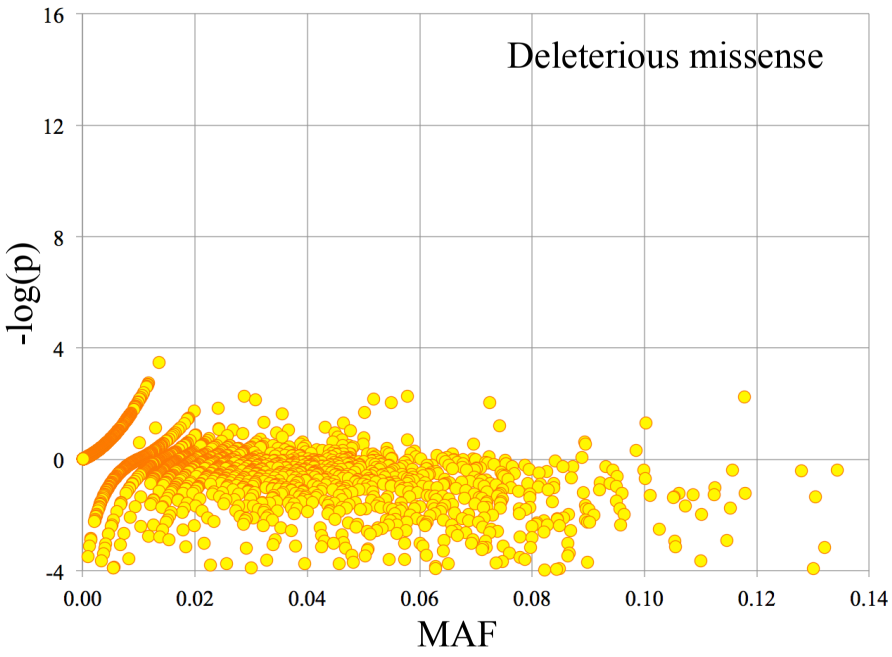
Adzhubei IA, Schmidt S, Peshkin L, Ramensky V, Gerasimova A, Bork P, Kondrashov A, Sunyaev S. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7:** 248–249.
Alkuraya FS. 2015. Human knockout research: new horizons and opportunities. *Trends Genet* **31:** 108–115.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

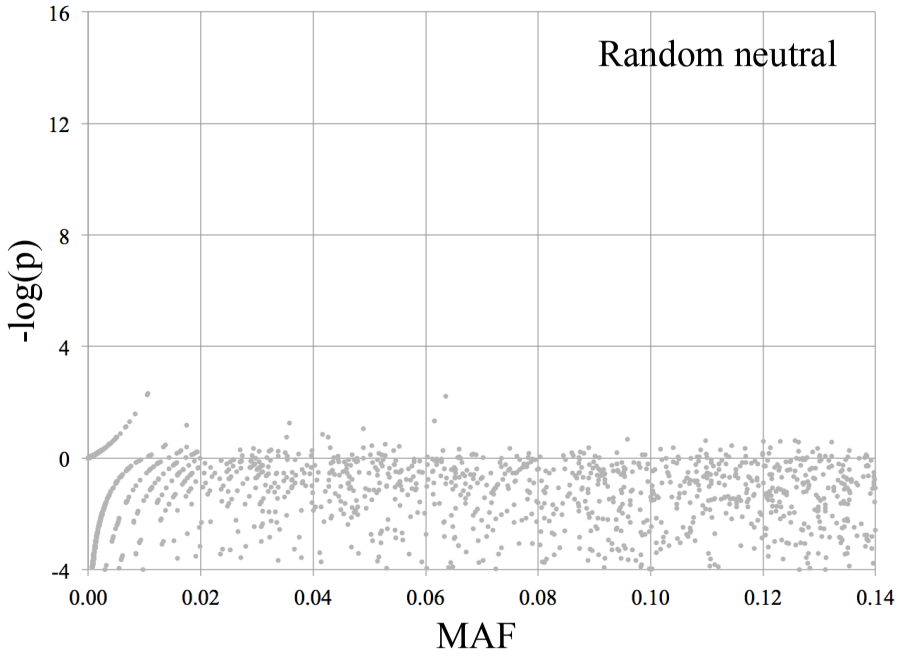
Received March 16, 2016; accepted in revised form August 19, 2016.


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

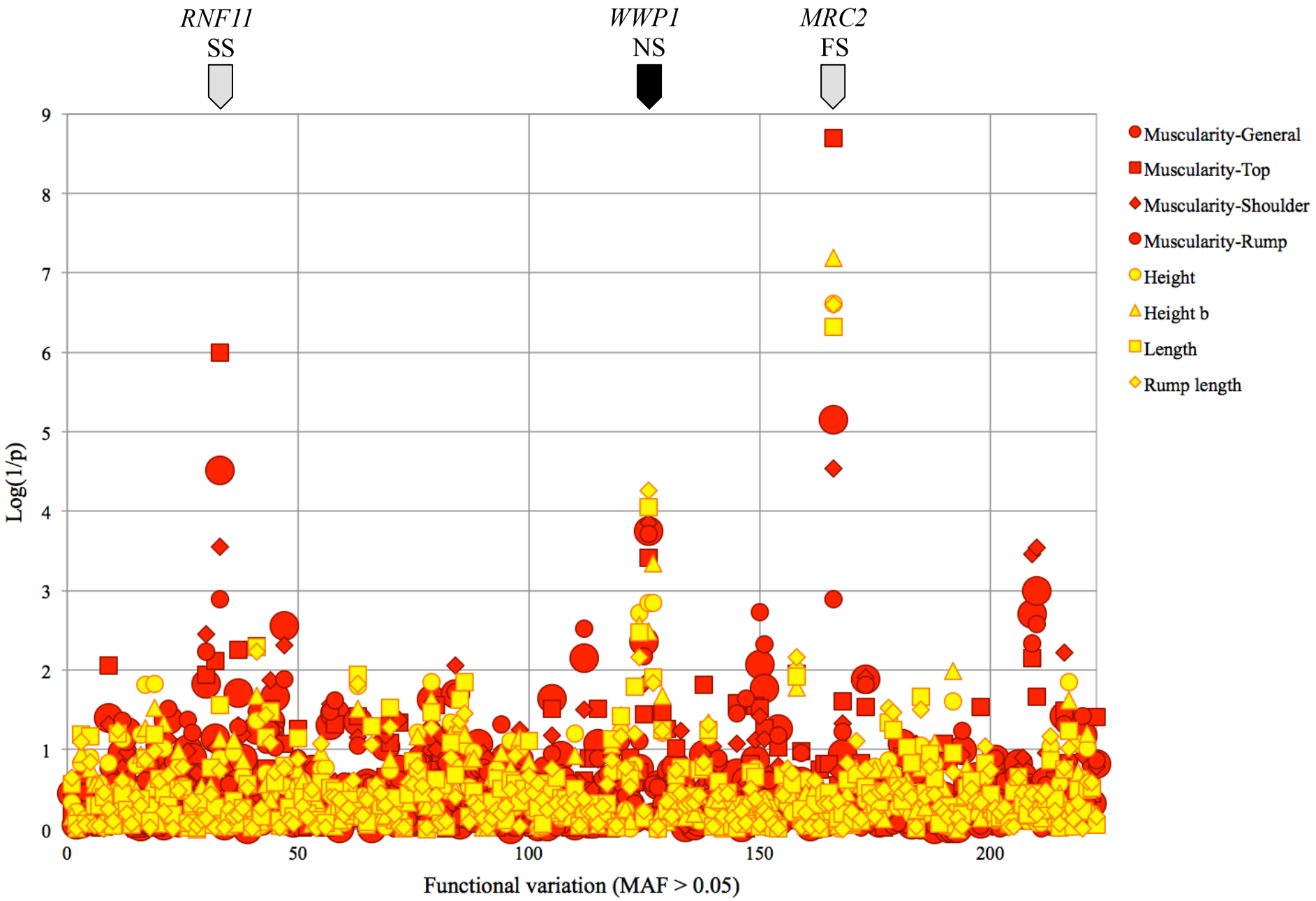
B.


Loss-of-function
16
12
8
4
0
-4
-log(p)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
MAF


Deleterious missense
16
12
8
4
0
-4
-log(p)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
MAF


Random neutral
16
12
8
4
0
-4
-log(p)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
MAF

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


RNF11
SS
WWP1
NS
MRC2
FS
Log(1/p)
Functional variation (MAF > 0.05)
0
1
2
3
4
5
6
7
8
9
0
50
100
150
200
Muscularity-General
Muscularity-Top
Muscularity-Shoulder
Muscularity-Rump
Height
Height b
Length
Rump length

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

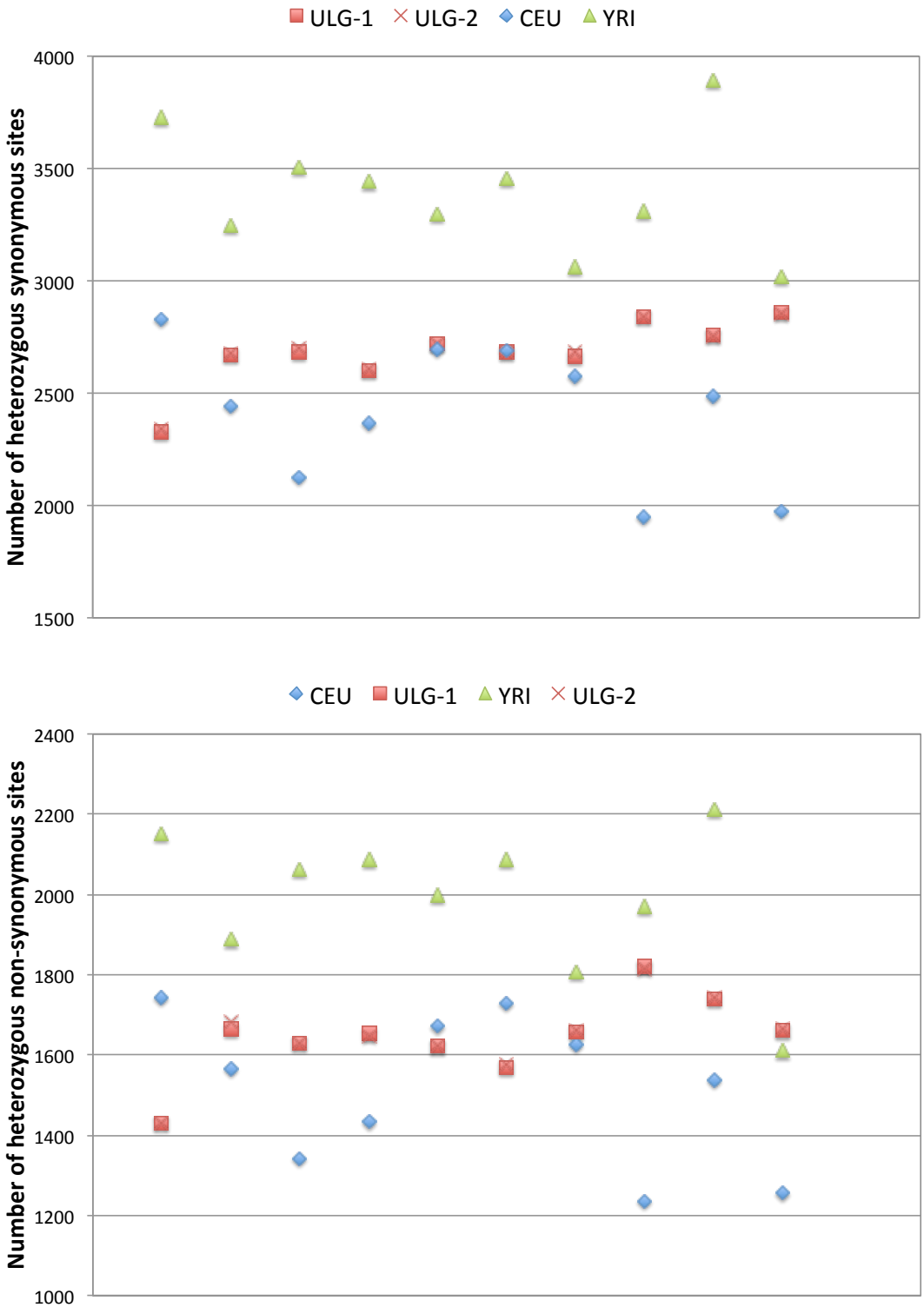
1. Fasquelle et al. Balancing selection of a frame-shift mutation in the MRC2 gene accounts for the outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. *PLoSGenetics* 5: e1000666 (2009).
2. Sartelet et al. A splice site variant in the bovine RNF11 gene compromises growth and regulation of the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

Supplemental Material S1:

Effect of dataset/sequencing center on nucleotide diversity.

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

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



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

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

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

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

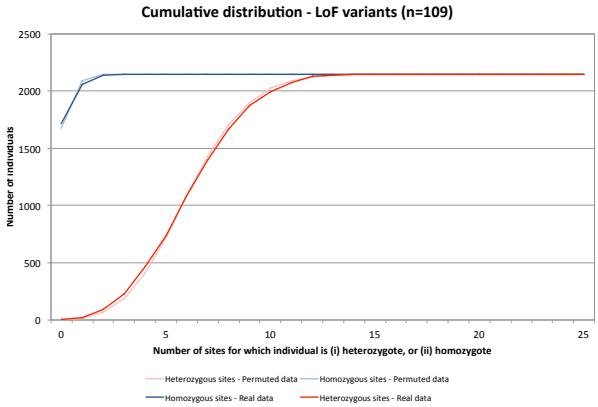
1. Littlejohn, M. D. et al. Expression variants of the lipogenic AGPAT6 gene affect diverse milk composition phenotypes in Bos taurus. *PLoS One* 9, e85757 (2014).
2. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*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

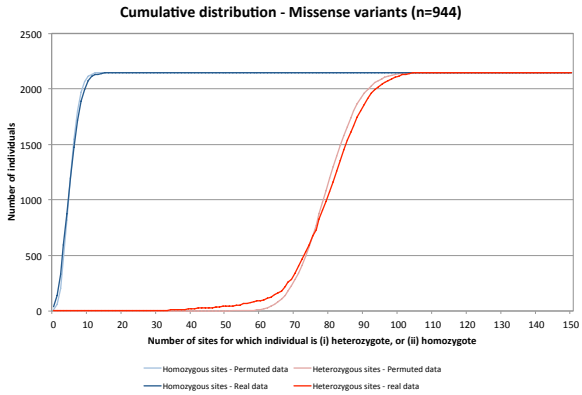
Supplemental Material S3 – Lack of evidence for synergistic epistasis

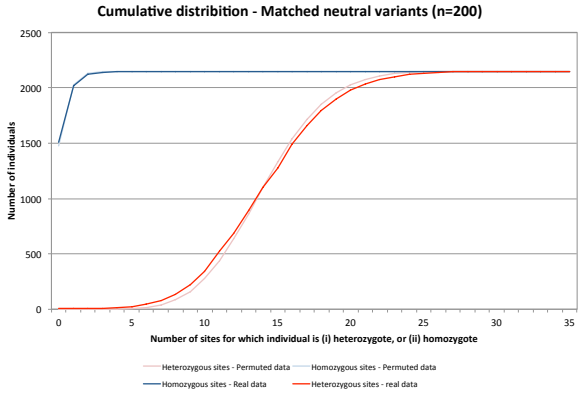
It has been hypothesized that deleterious variants might be purged from the population by synergistic epistasis, i.e. the fact that multiple deleterious variants have a larger cost on fitness than predicted from their multiplicative effect[1]. This hypothesis predicts that (healthy) individuals carrying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

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

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







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

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

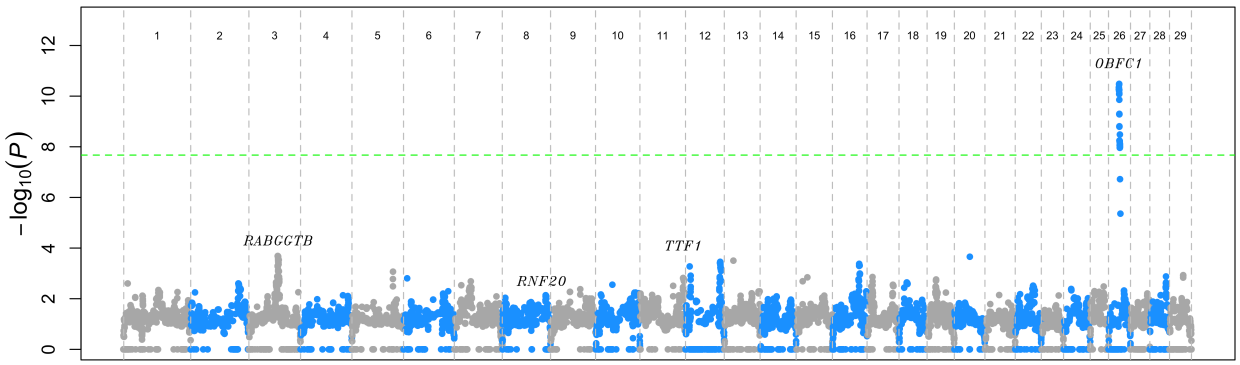
Supplemental Material S4: Haplotype-based genome scan for EL mutations.

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

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

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

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



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

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

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

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

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

Supplemental_Table_S1.xlsx
http://genome.cshlp.org/content/suppl/2016/09/19/gr.207076.11 6.DC1/Supplemental_Table_S1.xlsx

Supplemental_Table_S2.xlsx
http://genome.cshlp.org/content/suppl/2016/09/19/gr.207076.11 6.DC1/Supplemental_Table_S2.xlsx

Reverse genetic screen for loss-of-function mutations uncovers a frameshifting deletion in the melanophilin gene accountable for a distinctive coat color in Belgian Blue cattle

Wanbo Li, Arnaud Sartelet*, Nico Tamma, Wouter Coppieters, Michel Georges and Carole Charlier*

* Contributed equally to this work

Anim Genet. 2016, 47(1):110-113.

Background

In the process of systematical screening embryonic lethal variants in cattle population using aforementioned reverse genetic approach, we found 109 LoF variants after stringent filtering in BBC. We genotyped these LoF with a low-density custom array in 5201 BBC animals. Besides those LoF likely causing embryonic lethality, a 10-bp deletion in the *MLPH* gene has caught our attention. The deletion is predicted to introduce a premature stop codon in the first exon and would likely cause the abnormal transcript to undergo nonsense-mediated decay. More importantly, hypopigmentation phenotypes caused by mutations in *MLPH* have been observed in several species, for example, mouse, cat, dog, mink etc. The *MLPH* gene encodes melanophilin, is the adapter protein links Rab27a-melanosomes to myosin Va. Myosin Va together with Mlph protein transports melanosomes along the actin filament in the dendritic periphery of melanocytes. Myosin Va can transport melanosomes alone. But myosin Va-Mlph complex greatly enhance the transfer of melanosomes to adjacent keratinocyte (Sckolnick et al., 2013). So in theory, knocking out or down of melanophilin in melanocytes could result in few melanin being transferred to keratinocytes, therefore inducing light coat color in all non-white animals. This led us to examine the coat color phenotypes of the homozygotes of the 10-bp deletion in *MLPH*, resulting in the discovery of the mechanism behind the newly characterized 'cool gray' phenotype in BBC.

ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

SHORT COMMUNICATION

doi: 10.1111/age.12383

Reverse genetic screen for loss-of-function mutations uncovers a frameshifting deletion in the *melanophilin* gene accountable for a distinctive coat color in Belgian Blue cattle

Wanbo Li*[1], Arnaud Sartelet[†1], Nico Tamma*, Wouter Coppieters*[‡], Michel Georges* and Carole Charlier*

*GIGA-R and Faculty of Veterinary Medicine, University of Liège, 1, avenue de l'hôpital, 4000 Liège, Belgium. [†]Bovine Clinic, FARAH and Faculty of Veterinary Medicine, University of Liège, 20, boulevard de Colonster, 4000 Liège, Belgium. [‡]GIGA-Genomics platform, University of Liège, 1, avenue de l'hôpital, 4000 Liège, Belgium.

Summary

In the course of a reverse genetic screen in the Belgian Blue cattle breed, we uncovered a 10-bp deletion (c.87_96del) in the first coding exon of the *melanophilin* gene (*MLPH*), which introduces a premature stop codon (p.Glu32Aspfs*1) in the same exon, truncating 94% of the protein. Recessive damaging mutations in the *MLPH* gene are well known to cause skin, hair, coat or plumage color dilution phenotypes in numerous species, including human, mice, dog, cat, mink, rabbit, chicken and quail. Large-scale array genotyping undertaken to identify p.Glu32Aspfs*1 homozygous mutant animals revealed a mutation frequency of 5% in the breed and allowed for the identification of 10 homozygous mutants. As expression of a colored coat requires at least one wild-type allele at the co-dominant Roan locus encoded by the *KIT ligand* gene (*KITLG*), homozygous mutants for p.Ala227Asp corresponding with the missense mutation were excluded. The six remaining colored calves displayed a distinctive dilution phenotype as anticipated. This new coat color was named 'cool gray'. It is the first damaging mutation in the *MLPH* gene described in cattle and extends the already long list of species with diluted color due to recessive mutations in *MLPH* and broadens the color palette of gray in this breed.

Keywords bovine, cool gray, disruptive mutation, *KITLG*, *MLPH*, OMIA 00206-9913, whole-genome/whole-exome sequence

In an attempt to evaluate the fraction of disruptive mutations that could cause embryonic lethality and therefore affect fertility, we embarked on a next-generation-sequencing-based reverse genetic screen in modern cattle populations. As part of this study, we sequenced the whole genome of 50 and the whole exome of 30 sires from the Belgian Blue cattle breed. Sequence data were mined for highly disruptive loss-of-function variants corresponding to frameshift (FS), stop gain (SG) and essential donor/acceptor splice-site (SS) mutations. In this breed, we found 109 loss-of-function variants, including 56 FS, 42 SG and 11 SS. We have demonstrated that only a small fraction of these (~10%) likely causes embryonic lethality in homozygotes, highlighting the importance of molecular redundancy and the high proportion of non-essential genes (Charlier *et al.* 2014).

As a by-product of this study, we are currently searching for distinctive phenotypic features in animals that are apparently normal despite being homozygous for loss-of-function mutations in genes that are evolutionary highly conserved. In this process, we highlighted a 10-bp deletion in the first coding exon of the *melanophilin* gene (*MLPH*) at genomic position 117, 591, 518–117, 591, 527 bp on bovine chromosome 3 (BosTau6/UMD3.1 reference genome assembly) (Fig. 1). It is likely private to the Belgian Blue breed, as it was not listed in dbSNP (http://www.ncbi.nlm.nih.gov/snp/), not found in the 1000 bull genomes project (http://www.1000bullgenomes.com) (run 3: 429 sequenced key ancestors from 15 different breeds) and not found in an additional panel of 10 breeds, including two *bos indicus* breeds. This FS mutation (c.87_96del) is

Address for correspondence

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

[1]These authors contributed equally.

Accepted for publication 23 September 2015

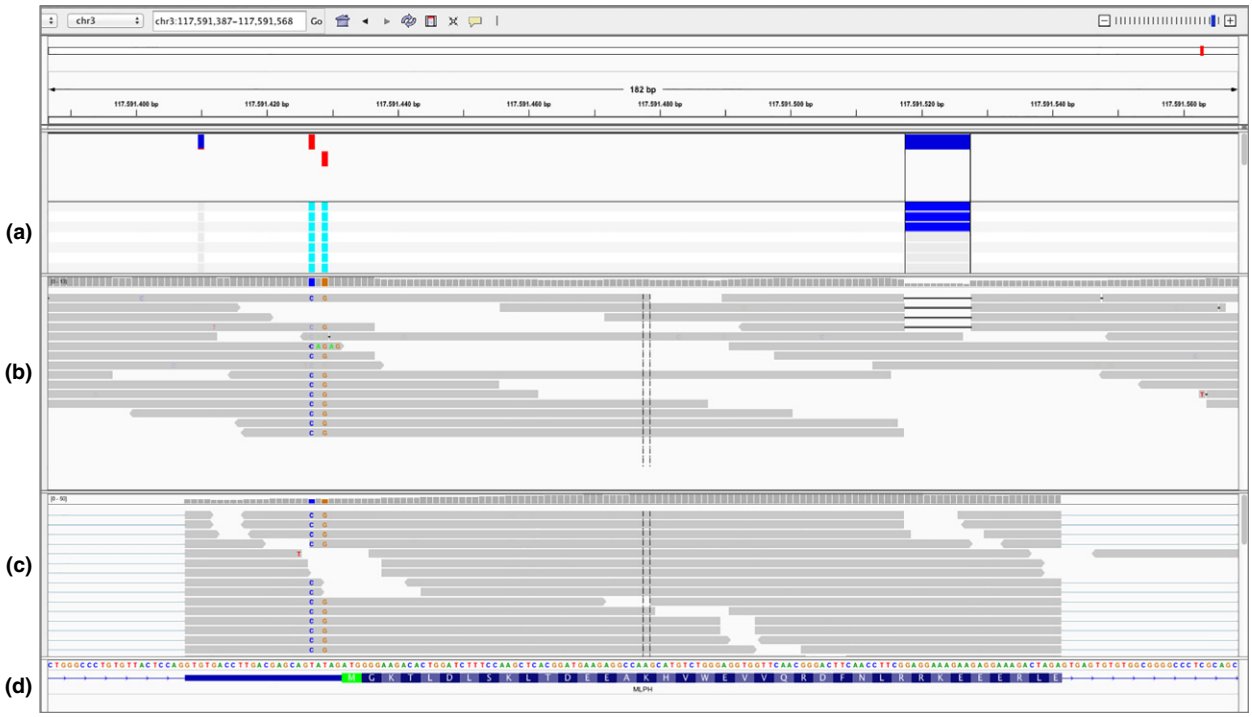


Figure 1 A 10-bp deletion (c.87_96del) in the first coding exon of the *melanophilin* gene. Screen capture of an INTEGRATIVE GENOMICS VIEWER (Robinson *et al.* 2011) output for a 182-bp genomic region encompassing the *MLPH* coding exon 1 with, from top to bottom (a) a 'VCF' file of 50 Belgian Blue sires (Druet *et al.* 2014) displaying three heterozygous animals for the c.87_96del mutation, (b) a 'BAM' file from a whole-genome sequence of a carrier animal, (c) fetal skin cDNA sequence reads from a wild-type animal defining the corresponding exon junctions and (d) a track of Ref-seq gene annotation.

predicted to introduce a premature stop codon in the same exon (p.Glu32Aspfs*1) and a resulting variant protein that is 94% shorter. Furthermore, the nonsense-mediated decay pathway is expected to degrade the variant mRNA.

The *melanophilin* gene encodes a protein expressed primarily in melanocytes, where it plays a central and active role in a tripartite complex (RAB27A-MLPH-MYO5A) (e.g. Strom *et al.*, 2002; Sckolnick *et al.* 2013). This complex is indispensable for active intracellular mature melanosome trafficking. In human, mutations in any of the three subunits of the complex cause Griscelli syndrome, a recessive disorder characterized by variable immune and neurological defects systematically accompanied by skin and hair hypopigmentation. Mutations in *MLPH* cause Griscelli syndrome type 3, which is strictly restricted to pigment dilution without any reported additional phenotypic change (Ménasché *et al.* 2003; Westbroek *et al.* 2012). This dilution is not generated by a defect in melanosome biosynthesis but, rather, caused by an impaired transport of mature melanosomes toward melanocyte dendritic tips where they are transferred to nearby keratinocytes (reviewed by Huizing *et al.* 2008). Similar hypopigmentation phenotypes, explained by recessive mutations in *MLPH*, have been described in numerous species including mouse (Matesic *et al.* 2001), cat (Ishida *et al.* 2006), dog (Philipp *et al.* 2005; Drögemüller *et al.* 2007), mink (Cirera *et al.* 2013), rabbit (Lehner *et al.* 2013; Fontanesi *et al.* 2014), chicken (Vaez *et al.* 2008) and quail (Bed'hom *et al.* 2012)–but not yet cattle. The identified causative mutation(s) and associated dilution phenotype for each species are summarized in Table S1. Therefore, the c.87_96del variant in the bovine *MLPH* gene appeared to be a strong candidate for a yet-to-be-described novel coat color in the Belgian Blue cattle.

The 109 loss-of-function variants detected in Belgian Blue cattle, including the *MPLH:*c.87_96del mutation, were added to the Illumina low-density custom array and used to genotype 5201 Belgian Blue animals as part of a genomic selection program (Charlier *et al.* 2014). The *MLPH:* c.87_96del mutation was shown to segregate at a frequency of 5% in Belgian Blue cattle and to be in Hardy–Weinberg equilibrium ($P = 0.37$; 10 del/del, 501 del/+, 4690 +/+). We identified 10 homozygous mutants. It is worth mentioning that the Belgian 'Blue' cattle breed received its name from the segregation of a co-dominant mutation, known as the Roan mutation, originating from the Shorthorn breed (Jones 1947). Homozygous wild-types, heterozygotes and homozygotes for the Roan mutation are black, roan blue (mixture of black and white hairs) and white, respectively (Charlier *et al.* 1996). These phenotypes are fully explained by a non-synonymous mutation (p.Ala227Asp; inferred annotation after gene model correction) in the *KIT ligand* (*KITLG*) gene (Seitz *et al.* 1999). This missense mutation was also present on our custom

array and was shown to have a frequency of 56% in the breed. Among the 10 *MLPH* homozygous mutants identified, two were black (*KITLG:*Ala227/Ala227), four roan blue (*KITLG*:p.Ala227Asp) and four white *(KITLG:*Asp227/Asp227*)*. As the phenotypic effect of the *MLPH*:c.87_96del mutation would be epistatically masked in white animals, we traced the six non-white animals back to their farms to examine their coat color. As expected, all six displayed a distinctive dilution phenotype, which we called 'cool gray' (Fig. 2, Appendix S1). Non-white animals, heterozygous for the *MLPH*:c.87_96del mutation, were indistinguishable from homozygote wild-type relatives (fully recessive mutation).

Figure 2 *MLPH* knockout dilution phenotypes in Belgian Blue animals and epistasis with the *KITLG* alleles. (a) Two animal homozygotes (del/del) for the c.87_96del mutation at the *MLPH* locus: homozygote wild-type (Ala/Ala, left) and heterozygote (Ala/Asp, right) for the p.Ala227Asp non-synonymous mutation from the *KITLG* locus; the animal on the left displays a black dilute coat color (dark 'cool gray') and the one on the right shows the epistatic 'cool gray' coat color; however, distinguishing a phenotypic difference between these combinations of genotypes remains challenging when animals are not side by side. (b) Two animals heterozygotes (Ala/Asp) for the p.Ala227Asp non-synonymous mutation at the *KITLG* locus: homozygote mutant and wild type (del/del, left; +/+, right) for the c.87_96del mutation at the *MLPH* locus; the animal on the right is under the classical 'roan blue' coat color.

In all the species with loss-of-function mutations in *MLPH*, the reported phenotype appeared to be strictly restricted to skin and hairs. However, it is noteworthy that, in addition to melanocytes, *MLPH* is also highly expressed in mast cells, which are key cellular players mediating allergic and inflammatory reactions. In mice, a recent study by Singh *et al.* (2013) has shown that the Rab27a/Mlph/MyoVa complex seemed to regulate the docking of mast cell granules to the mast cell plasma membrane by modulating its cytoskeleton integrity. It is thus tempting to speculate that loss-of-function mutations in *MLPH* could affect mast cell degranulation and have a pleiotropic effect on an allergic and/or inflammatory reaction's time-course. Moreover, in human, it has been reported that the *MLPH* locus exhibits a strong signal of recent positive selection in non-African populations (Pickrell *et al.* 2009). A link between this sweep in human, underlying alleles and a putative advantageous phenotype–correlated or not with pigmentation–remains to be established.

Up until now, there have been characterized mutations segregating in the Belgian Blue breed at four coat-color loci (*MC1R*, *KITLG*, *KIT* and *MLPH*) (Klungland *et al.* 1995; Charlier *et al.* 1996; Seitz *et al.* 1999; Durkin *et al.* 2012; this study). All known mutations, their modes of inheritance, associated phenotypic effects and respective frequencies within this breed are listed in Table S2. Collectively, they are responsible for the observed variety of colored patterns and subtle blue or gray shades in both Belgian Blue purebred and crossbred animals.

This study stands as proof of the concept that a population-based next-generation-sequencing reverse screen can uncover segregating variations underlying novel phenotypes of biological interest.

Acknowledgements

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

Authors' contributions

WL performed WGS/WES analysis; AS was in charge of cases collection and phenotyping; NT genotyped the samples on the Illumina custom array; WC supervised the

GIGA-genomics platform; and MG and CC designed the study, analyzed the data and wrote the manuscript with the help of all co-authors.

References

Bed'hom B., Vaez M., Coville J.L., Gourichon D., Chastel O., Follett S., Burke T. & Minvielle F. (2012) The lavender plumage colour in Japanese quail is associated with a complex mutation in the region of *MLPH* that is related to differences in growth, feed consumption and body temperature. *BMC Genomics* **13**, 442.

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. *Mammalian Genome* **7**, 138–42.

Charlier C., Li W., Harland C., Littlejohn M., Creagh F., Keehan M., Druet T., Coppieters W., Spelman R. & Georges M. (2014) NGS-based reverse genetic screen reveals loss-of-function variants compromising fertility in cattle. In: *Proceedings of the 10th World Congress of Genetics Applied to Livestock Production*, Vancouver, BC, Canada. Available at http://www.wcgalp.com/.

Cirera S., Markakis M.N., Christensen K. & Anistoroaei R. (2013) New insights into the *melanophilin* (*MLPH*) gene controlling coat color phenotypes in American mink. *Gene* **527**, 48–54.

Drögemüller C., Philipp U., Haase B., Günzel-Apel A.R. & Leeb T. (2007) A noncoding *melanophilin* gene (*MLPH*) SNP at the splice donor of exon 1 represents a candidate causal mutation for coat color dilution in dogs. *Journal of Heredity* **98**, 468–73.

Druet T., Ahariz N., Cambisano N., Tamma N., Michaux C., Coppieters W., Charlier C. & Georges M. (2014) Selection in action: dissecting the molecular underpinnings of the increasing muscle mass of Belgian Blue Cattle. *BMC Genomics*, **15**, 796.

Durkin K., Coppieters W., Drögemüller C. *et al.* (2012) Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* **482**, 81–4.

Fontanesi L., Scotti E., Allain D. & Dall'olio S. (2014) A frameshift mutation in the *melanophilin* gene causes the dilute coat colour in rabbit (*Oryctolagus cuniculus*) breeds. *Animal Genetics* **45**, 248–55.

Huizing M., Helip-Wooley A., Westbroek W., Gunay-Aygun M. & Gahl W.A. (2008) Disorders of lysosome-related organelle biogenesis: clinical and molecular genetics. *Annual Review of Genomics and Human Genetics* **9**, 359–86.

Ishida Y., David V.A., Eizirik E., Schäffer A.A., Neelam B.A., Roelke M.E., Hannah S.S., O'Brien S.J. & Menotti-Raymond M. (2006) A homozygous single-base deletion in *MLPH* causes the dilute coat color phenotype in the domestic cat. *Genomics* **88**, 698–705.

Jones I.C. (1947) The inheritance of red, roan and white coat colour in dairy shorthorn cattle. *Journal of Genetics*. **48**, 155–63.

Klungland H., Våge D.I., Gomez-Raya L., Adalsteinsson S.& Lien S. (1995) The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat colour determination. *Mammalian Genome* **6**, 636–9.

Lehner S., Gähle M., Dierks C., Stelter R., Gerber J., Brehm R. & Distl O. (2013) Two-exon skipping within MLPH is associated with coat color dilution in rabbits. *PLoS One* **8**, e84525.

Matesic L.E., Yip R., Reuss A.E., Swing D.A., O'Sullivan T.N., Fletcher C.F., Copeland N.G. & Jenkins N.A. (2001) Mutations in *Mlph*, encoding a member of the Rab effector family, cause the melanosome transport defects observed in leaden mice. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10238–43.

Ménasché G., Ho C.H., Sanal O., Feldmann J., Tezcan I., Ersoy F., Houdusse A., Fischer A. & de Saint Basile G. (2003) Griscelli syndrome restricted to hypopigmentation results from a melanophilin defect (GS3) or a MYO5A F-exon deletion (GS1). *The Journal of Clinical Investigation* **112**, 450–6.

Philipp U., Hamann H., Mecklenburg L., Nishino S., Mignot E., Günzel-Apel A.R., Schmutz S.M. & Leeb T. (2005) Polymorphisms within the canine *MLPH* gene are associated with dilute coat color in dogs. *BMC Genetics* **6**, 34.

Pickrell J.K., Coop G., Novembre J. *et al.* (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Research* **19**, 826–37.

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

Sckolnick M., Krementsova E.B., Warshaw D.M. & Trybus K.M. (2013) More than just a cargo adapter, melanophilin prolongs and slows processive runs of myosin Va. *Journal of Biological Chemistry* **288**, 29313–22.

Seitz J.J., Schmutz S.M., Thue T.D. & Buchanan F.C. (1999) A missense mutation in the bovine *MGF* gene is associated with the roan phenotype in Belgian Blue and Shorthorn cattle. *Mammalian Genome* **10**, 710–2.

Singh R.K., Mizuno K., Wasmeier C., Wavre-Shapton S.T., Recchi C., Catz S.D., Futter C., Tolmachova T., Hume A.N. & Seabra M.C. (2013) Distinct and opposing roles for Rab27a/Mlph/MyoVa and Rab27b/Munc13-4 in mast cell secretion. *FEBS Journal* **280**, 892–903.

Strom M., Hume A.N., Tarafder A.K., Barkagianni E. & Seabra M.C. (2002) A family of Rab27-binding proteins. Melanophilin links Rab27a and myosin Va function in melanosome transport. *Journal of Biological Chemistry* **277**, 25423–30.

Vaez M., Follett S.A., Bed'hom B., Gourichon D., Tixier-Boichard M. & Burke T. (2008) A single point-mutation within the *melanophilin* gene causes the lavender plumage colour dilution phenotype in the chicken. *BMC Genetics* **9**, 7.

Westbroek W., Klar A., Cullinane A.R. *et al.* (2012) Cellular and clinical report of new Griscelli syndrome type III cases. *Pigment Cell and Melanoma Research* **25**, 47–56.

Supporting information

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

Table S1 Known recessive mutations in the *MLPH* gene and associated effect at the molecular level in various vertebrate species.

Table S2 Known mutations and their associated phenotypic effect in the four coat color genes molecularly characterized in the Belgian Blue breed.

Appendix S1 Photographs of Belgian Blue animals and genotype/phenotype table.

Table S1: Known recessive mutations in the *MLPH* gene and associated effect at the molecular level in various vertebrate species.

Species	Phenotype	Mutation	Mutation effect	References
Mouse	Leaden color	*c.266C>T*	Creation of a new splice donor site (in frame deletion)	Matesic et al., 2001
		Large deletion	Knock-out (KO)	
Human	Griscelli syndrome 3	*c.102C>T; p.Arg35Trp* (recurrent mutation)	Missense mutation responsible for loss of interaction with RAB27A	Ménasche et al. 2003
				Westbroek et al., 2012
Cat	Dilute color	*c.83delT*	Frame-shifting mutation, KO	Ishida et al., 2006
Dog	Dilute color	*c.-22G>A*	Reduced splicing efficiency	Drögemüller et al., 2007
Rabbit	Dilute color	*c.585delG*	Frame-shifting mutation, KO	Fontanesi et al., 2013
		c.111-5C>A	Splice-acceptor mutation, two-exon skipping	Lehner et al., 2013
Mink	Silver blue color	Complex structural mutation	KO	Cirera et al., 2013
Chicken	Lavender color	*c.103C>T ; p.Arg35Trp*	Missense mutation	Vaez et al., 2008
Quail	Lavender color	Complex structural mutation	KO	Bed'hom et al., 2012
Cattle	Cool grey	*c.87_96del; p.Glu32fs*	Frame-shifting mutation, KO	This study

The corresponding OMIA entries are the following:

- **OMIA 000031-9615 Alopecia, colour mutant** in *Canis lupus familiaris* (dog) Gene: MLPH
- **OMIA 000206-9685 Coat colour, dilute** in *Felis catus* (domestic cat) Gene: MLPH
- **OMIA 000206-9986 Coat colour, dilute** in *Oryctolagus cuniculus* (rabbit) Gene: MLPH
- **OMIA 001438-452646 Coat colour, silver** in *Neovison vison* (American mink) Gene: MLPH
- **OMIA 001445-93934 Feather colour, lavender** in *Coturnix japonica* (Japanese quail) Gene: MLPH
- **OMIA 001445-9031 Feather colour, lavender** in *Gallus gallus* (chicken) Gene: MLPH

Table S2: **Known mutations and their associated phenotypic effect in the four coat color genes molecularly characterized in the Belgian Blue breed.**

Gene symbol	Chromosome	Mutation	Allele	Muttaion frequency	Mode of Inheritance	Genotypes	Coat phenotype	References
MC1R	18	*c.296T>C; p.Leu99Pro*	E (Extension)	92.3%	Dominant	*E/E, E/e*	Dominant black	Klungland et al., 1995; Charlier et al., 1996
		c.309delC; p.Ala103fs	e	6.3%	Recessive	*e/e*	Recessive red	
KITLG	5	*c.680C>A; p.Ala227Asp*	R (Roan)	51.4%	Co-dominant	*R/R*	White	Charlier et al., 1996; Seitz et al., 1999
						R/+	Blue	
						+/+	Black	
KIT	6	Duplicative translocation (chr29)	Cs (Color-sided)	< 1%	Semi-dominant	*Cs/Cs*	Color-sided, more white	Durkin et al., 2012
						Cs/+	Color-sided, less white	
MLPH	3	*c.87_96del; p.Glu32fs*	d (dilute)	5.1%	Recessive	*d/d*	Cool grey	This study

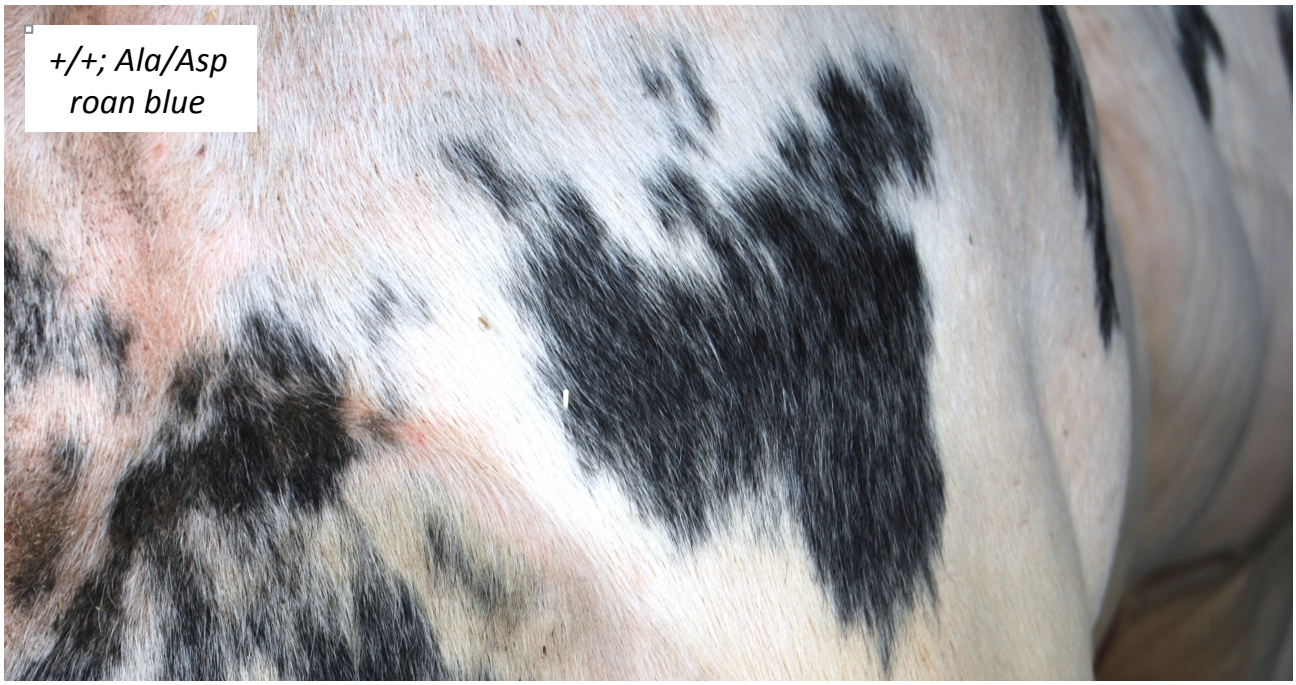
MC1R: Melanocortin 1 receptor gene; the wild-type allele has been observed at very low frequency (~ 1.4%) in this breed.

KITLG: KIT ligand gene or MGF (Mast Cell Growth Factor) or Steel Factor

KIT: V-Kit Hardy-Zuckerman 4 Feline Sarcoma Viral Oncogene Homolog gene

MLPH: Melanophilin gene

Appendix S1 Photographs of Belgian Blue animals and genotype/phenotype table
A collection of photos from Belgian Blue animals under the different combinations of genotypes at *MLPH* and *KITLG* loci. For each animal, corresponding genotypes at *MLPH* (left) and *KITLG* (right) loci are listed and resultant coat-color phenotype is given. A genotype/phenotype table summarizes the results. As some 'cool grey' animals were photographed at different ages, each one is labeled with a unique ID (from 1 to 7, in red numbers).




2
3530
3530
Photo from Chad Harland
del/del; Ala/Asp
cool grey


+/+; Ala/Ala
black
del/del; Ala/Asp
cool grey
2


Photo from Chad Harland
3
3537
3537
del/del; Ala/Asp
cool grey


3
+/+; Ala/Ala
black
del/del; Ala/Asp
cool grey

del/+; Ala/Asp
roan blue

del/del; Ala/Asp
cool grey

+/+; Asp/Asp
white

+/+; Ala/Asp
roan blue

del/del; Ala/Asp
cool grey


4
Photo from Chad Harland
del/del; Ala/Asp
cool grey


+/+; Ala/Asp
roan blue
+/+; Ala/Ala
black
4
del/del; Ala/Asp
cool grey


4

del/del; Ala/Asp
cool grey

+/+; Ala/Asp
roan blue

del/del; Ala/Ala
cool grey (dark)

del/del; Ala/Asp
cool grey

del/del; Ala/Asp
cool grey

+/+; Ala/Asp
roan blue

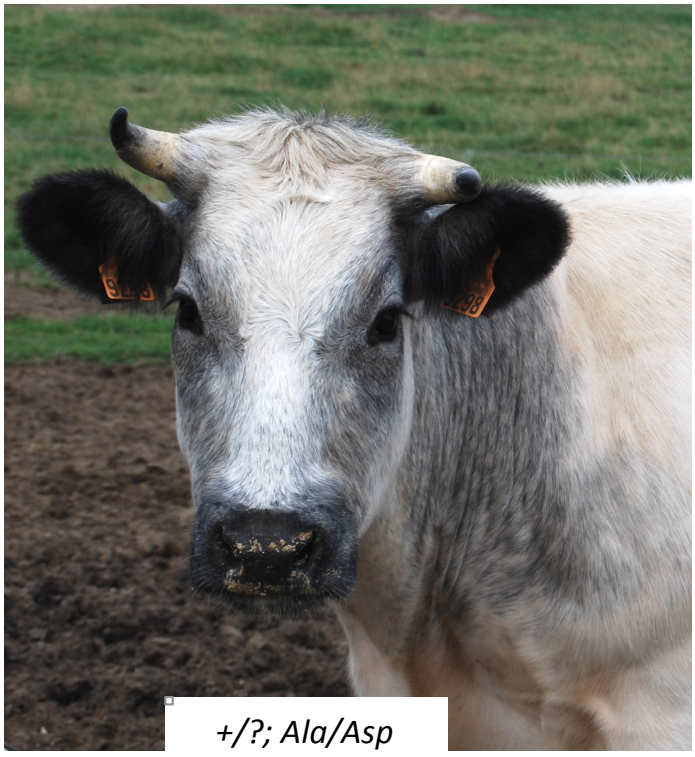
del/del; Ala/Ala
cool grey (dark)

+/+; Ala/Ala
black

del/+; Ala/Asp
roan blue (dark)

+/+; Ala/Asp
roan blue

+/+; Asp/Asp
white



+/?; Ala/Asp
roan blue (light)

+/?; Ala/Asp
roan blue

Genotype/phenotype summary table:

***MLPH* locus genotype**	*+/+ or del/+*	*+/+ or del/+*	*+/+ or del/+*	*del/del*	*del/del*	*del/del*
***KITLG* locus genotype**	*Ala/Ala*	*Ala/Asp*	*Asp/Asp*	*Ala/Ala*	*Ala/Asp*	*Asp/Asp*
Phenotype	black	roan blue (dark to light)	white	cool grey (dark)	cool grey (dark to light)	white
Hairs in colored areas	black hairs	variable ratio of white to black hairs		dilute black hairs	variable ratio of white to dilute black hairs	

Conclusions and perspectives

For decades, identifying causative genes and mutations underpinning phenotypic traits or congenital defects has been at the heart of genetic research. In this thesis, we report our efforts in dissecting the genetic determinants of several Mendelian and complex traits in cattle aided by SNP array, second-generation high-throughput sequencing technologies and advanced statistical methodologies.

The development of SNP arrays has profoundly altered the field of genetic research. It greatly facilitated the hunt for the causative genes in Mendelian traits, while also reduced the complexity of experimental design. It gave birth to the widely used GWAS approach, leading to the discovery of thousands of loci associated with a variety of traits in human and farm animals. The development of second-generation high-throughput sequencing has revolutionized our view of genetic research. It opened the door to the one-step mapping strategy for monogenic traits and give rise to very high-density markers in GWA studies at limited costs by means of imputation.

Armed with the most advanced whole-genome genotyping and sequencing technology, we herein report our efforts carried out in the discovery of a 2bp-deletion in the *MRC2* gene responsible for the crooked tail syndrome and an intronic regulatory mutation in the *PIGH* gene leading to arthrogryposis in BBC breed, as well as a 3.3-kb large deletion in the *FANCI* gene causing brachyspina syndrome in Holstein dairy cattle. Those three mutations underlying genetic defects, together with other causative mutations discovered in the lab, such as in congenital muscular dystony type I and II, renal lipofuscinosis and ichthyosis fetalis, were quickly used to provide genetic tests to breeding farms, avoiding further economic loss to the breeders. Using the causative mutations of monogenic defects, the breeding industry has implemented marker-assisted selection (MAS) very effective to eliminate these diseases from the population. In addition, using a reverse genetic approach implemented in the HTS

context, we verified nine embryonic lethal variants (ELV), segregating at frequencies ranging from 1.2 to 6.6% in the respective populations. The ELV revealed here could also immediately be used in the breeding scheme to avoid at-risky matings, lowering the allele frequency of ELV over subsequent generations.

On the other side, implementation of MAS with associated markers identified by early QTL mapping in complex traits proved largely fruitless as most of the markers displaying association with a trait usually explain a small portion of genetic variance. Meuwissen *et al.* (2001) suggested a genomic selection (GS) approach using genome-wide dense marker sets (e.g., SNP array) to predict genomic estimated breeding values (GEBV) of animals (Meuwissen and Goddard, 2001). GS need a set of reference panel of animals (usually, n>3000) with SNP array genotypes and recorded traits, then effects of every SNP can be estimated. A sire's GEBV is estimated with all SNPs even if those SNPs don't pass stringent significance threshold in association tests. GEBV estimated by GS is much more accurate than those arrived at by MAS as GS could potentially incorporate effects of all QTL for a trait by tagged markers due to linkage disequilibrium (LD). GS also significantly reduces generation time in breeding, especially for those traits either difficult to measure or only record at a late stage of the lifespan (Goddard and Hayes, 2009) (reviewed by Goddard et al., 2009). In the breeding industry, genomic selection has experienced great success in several species, especially in cattle. For instance, the genetic progress in milk yield was estimated to increase about 43% in US Holstein population from 2006 to 2012 compared to the period from 2000 to 2006, mainly due to the application of genomic selection (Hutchison et al., 2014). We have devoted great efforts to identifying causative mutations or QTNs for complex traits besides congenital defects. At this point, we should ask whether it is necessary to continue uncover the causative mutations; will causative mutations still benefit the breeding in complex agronomical traits? Uncovering causative mutations will certainly expand our knowledge of the biological mechanisms underpinning a phenotype, and also will be necessary in the following cases in breeding.

I) Mutations in balancing selection. The discovery of CTS mutations in the BBC population reveals an interesting example of balancing selection in farm animals, the mutation gives an advantage for meat production in the heterozygous animals, however it causes severe health problems in homozygotes. Recently, the evidence of balancing selection on mutations underlying genetic defects is increasing. For example, Kadri et al. (2014) reported that a 660-kb large deletion that increases milk production in carriers, but is embryonic lethal in homozygous embryos in Nordic Red cattle. The c124-2A>G splice variant in the *RNF11* gene, compromises growth and regulation of the inflammatory response in BBC and is also under balancing selection (Sartelet et al., 2012). We found that the frame-shift mutation in the *SNAPC4* gene causing embryonic lethality might be advantageous in heterozygotes through the reverse genetic approach, but further phenotypic data needs to be collected to support this conclusion. These findings expand the list of known mutations under balancing selection, such as mutations underlying malignant hyperthermia in pigs, hyperkalaemic periodic paralysis in the horse, and hereditary chondrodysplasia in sheep (Fujii et al., 1991; Rudolph et al., 1992; Smith et al., 2006). These facts suggest that balancing selection may play more important role in breeding. In this kind of situation, if one takes variants with an advantage for agronomical traits in the GS model but while ignoring the deleterious effects, it will create a flawed breeding scheme.

II) Marker effects need to be re-estimated after generations in GS. GS is implemented base on the principle of LD between markers and QTNs. LD pattern could change over generations in a population, so the marker effects estimated from reference animals need to be renewed with newer panels of animals every few generations. But if most of the causative mutations were known, then re-estimating marker effects is not needed, we can simply type the QTN genotypes to estimate more accurate GEBV. This might be realistic in the situation with few major QTL explaining large amount of genetic variance. Although, as the cost of next-generation HTS continue to decline, the reassessment of marker effects could be potentially surpassed by using sequence-based imputation genotypes or direct sequencing

data in the reference animals because nearly all the QTNs are in the variant sets already. This notion is supported by previous simulation studies (Druet et al., 2014).

III) Bring in favorable alleles from other breeds. GS can only work when there is large genetic variance of a trait in the focused population, i.e., segregating of desire and undesired alleles. But in some cases, traits are fixed or nearly fixed with an undesired allele in the population, leaving no room for the genetic improvement. For example, in pigs, a Chinese local pig breed – Laiwu- has very high intramuscular fat content (IMF, up to 10% of meat mass) in the meat, improving the favor and quality of the pork. If the QTNs conferring the high IMF content were known, the QTN could be brought into other commercial lines with accuracy and efficiency.

IV) Breeding in the era of accurate genomic editing. The genomic engineering technology has developed and updated in an unprecedented speed in last few years. The CRISPR-Cas9 system cleaves and repairs DNA with guide-RNA, allowing us to edit genomes at great specificity and efficiency. It can create knock out/in mutations, introduce insertions, deletions or duplications to the genomes. Now, nearly all model organisms, even human embryos, have been modified with CRISPR-Cas9. It has already been shown that this approach could be applied in animal breeding and have a great impact in the field. For instance, pigs knocked out the *CD163* gene by Cas9 are resistance to porcine reproductive and respiration syndrome (PRRS) (Whitworth et al., 2016). PRRS is the most economically important disease in pigs worldwide and causes enormous economic loss. Using this technology, knocking out the *CD163* gene, while not introducing any vector sequence to the genome could potentially remove marketing barriers of genetic modified animals. If we do not understand the biological pathways or know the QTNs, it is hard to use these state-of-art genome engineering technology in breeding. Certainly, the gene editing technology will be not suitable to use on associated markers which are in high LD with QTNs.

Of note, intensive selection has been applied in domestic animals for decades and there are concerns that if we have reduced the effective population size of domestic animals to a point where the viability of the population is question. However, based on our and others' results, cattle appear to have larger levels of genetic diversity than humans, even including Africans (Daetwyler et al., 2014). This implies that if we control inbreeding properly, we could still have big room to improve productivity in cattle and in other domestic animals.

In this thesis, we studied a fundamental genetic component – recombination. Although it seems to have no direct application in breeding, it could be used to guide artificial selection and might be related to fertility (Kong et al., 2008). We defined the genome-wide recombination rate (GRR) and recombination hotspot-window usages (GHU) in two large cattle population. We reported that genetic variants in *RNF212* and *REC8* affect GRR in cattle. The *RNF212* gene has been demonstrated to be associated with genome-wide recombination rate in human populations (Kong et al., 2008). Recently, the functionality of the *RNF212* gene in recombination has been repeatedly confirmed in very large cattle pedigrees by subsequent work in the lab and by others (Kadri et al., 2016; Ma et al., 2015). However, the intronic variant in the *REC8* gene is no longer supported as the candidate causative variant of the BTA10 QTL by the new data. The fine-mapping with sequence-based imputation genotypes showed variants within a paralogue of the RNF212 gene - RNF212B- might be the causal gene of this QTL (Kadri et al., 2016). This highlights the power of sequence-base imputation in GWAS applications. Indeed, without using sequencing information, most of the candidate genes proposed by Ma et al. (2015) are far away (even up to 1 Mb) from the highest association signals in the corresponding QTL reported by Kadri et al. (2016). Interesting, the study on recombination in the lab revealed that the genetic map in males is larger than in females (23.3 M vs. 21.4 M), consistent with a finding in previous even larger cattle population (Ma et al., 2015). However, female recombination rates in Zebrafish (Singer et al., 2002), human (Chowdhury et al., 2009), mouse (Petkov et al., 2007) and most species examined are higher than that in males. The discrepancy in recombination

between sires and cows may reflect different genetic mechanisms underlying recombination in the different sexes. However, the high genetic correlation of recombination (~0.66) between sires and cows does not support this idea (Kadri et al., 2016). Thus the long male genetic map in cattle need to be further studied.

We also found that variants in a paralogue of *PRDM9* gene in X chromosome dominate the hotspots usage in male cattle as in human and mouse. The results of Ma et al. (2015) did not confirm our findings. Instead, the associated signal for hotspot usage in their study is close to another paralogue of *PRDM9* on BTA1. With the large cattle dataset in the lab, analysis on the hotspot usage is ongoing. Furthermore, they also found, for the first time that *PRDM9* might affect genome-wide recombination rates. However, Ma et al. (2015) could not rule out that the signal is due to other genes in the region. In recent years, the most noted progress in recombination research is that for the *PRDM9* gene, where it has been shown to play a major role in specifying the recombination hotspots across genomes. In silico and in vitro analysis has revealed that the purine-rich (PR) domain containing protein PRDM9 contacts the recombination hotspot motifs in human and mouse (Baudat et al., 2010; Myers et al., 2010). Nevertheless, PRDM9 is not essential for recombination maintenance in mammals, given that the canine counterpart of this gene became extinct during evolution, disrupted by accumulated nonsense mutations (Axelsson et al., 2012). There are still unsolved questions about the detailed mechanism of *PRDM9* and how it recruits double-strand breakage repair machinery. Deciphering the molecular biological mechanism of action of the PRDM9 protein is under intensive research currently (Walker et al., 2015). Since recombination is one of the fundamental determinants of evolution, it will also be interesting to examine the landscape of recombination in other species using the genomes of animals deposited in the public databases.

The ability of annotation of genomic variants in both coding and noncoding regions is important. Our efforts were devoted to uncover causal variants for certain traits or defects.

We identified causal variants for two genetic defects – brachyspina syndrome and arthrogryposis in cattle. According to the experience in our lab, almost 9/12 of the causative mutations of monogenetic diseases were located in the coding regions. However, the underlying mutation of arthrogryposis in the BBC breed is an intronic variant hidden in 31 candidate mutations which segregated perfectly between cases and controls. In the study, RNAseq was performed to rapidly find that the intronic mutation within the *PIGH* gene causes an exon deletion on processed mRNA, strongly suggesting the causality of the mutation. Highlighting that high-throughput sequencing can help decipher the consequence of regulatory variants otherwise difficult to identify. More than 80% of GWAS loci in human populations located in noncoding regions. Deciphering the consequence of these variants is still very challenging. We have discussed at length the need for identifying causative mutations or QTNs in the breeding industry even if we have the powerful tool of genomic selection.

At the present time our knowledge of non-coding regulatory elements is limited in domestic animals, hampering our efforts to identifying causal variants in these regions. Due to these limitations, in our work looking at EL/JL variants our focus has necessarily been limited to the coding parts of bovine genome. It is reasonable to believe that structural variation or key point mutation in noncoding regulatory regions can also cause severe consequence or lethality. The accessibility of the high-throughput functional data, e.g. promoters, enhancers, insulators, etc, will allow us to systematic survey the function of elements in the genome. It will also allow us to better understand the causality of GWAS peaks in noncoding regions. Besides, the biological information from genes and noncoding regulatory elements, could also aid in the QTL mapping or GWAS step (MacLeod et al., 2014). Datasets of Encyclopedia of DNA Elements have been recently produced for human and mouse. More than 70% of homologous promoters are at high degree of conservation between human and mouse, but only ~25% of enhancers and CTCF-binding sites are functionally conserved despite the high sequence conservation of the examined genomic regions (Shen et al., 2012). More recently, an

evolutionary analysis of enhancers and promoters in 20 mammalian species also revealed that enhancers are rarely conserved compared to promoters across mammals (Villar et al., 2015). These results imply that the domestic animal research communities can not rely on human or mouse ENCODE data by means of comparative genomics, and will need to generate comparable ENCODE-like datasets to identify regulatory variants in the GWAS of domestic animals. The establishment of the Functional Analysis of Animal Genome consortium (http://www.animalgenome.org/community/FAANG/) will set the foundation for generating the ENCODE-like datasets for farm animals. It is reasonable to anticipate that the release of the ENCODE-like data in the animal research sphere will speed up pinpointing genetic basis of complex traits at an unprecedented pace.

In conclusion, armed with high-throughput genotyping and sequencing, we can optimistically foresee that the catalogue of causative genes and mutations of Mendelian and complex traits will be expanded largely in the near future in cattle and other species.

Reference

Agerholm, J.S., McEvoy, F., and Arnbjerg, J. (2006). Brachyspina syndrome in a Holstein calf. Journal of veterinary diagnostic investigation: official publication of the American Association of Veterinary Laboratory Diagnosticians, Inc *18*, 418-422.

Ahmadiyeh, N., Pomerantz, M.M., Grisanzio, C., Herman, P., Jia, L., Almendro, V., He, H.H., Brown, M., Liu, X.S., Davis, M., et al. (2010). 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proc. Natl. Acad. Sci. *107*, 9742–9746.

Ai, H., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L., Zhang, F., Zhang, L., Cui, L., He, W., et al. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. Nat Genet *47*, 217-225.

Almeida, R., Ricaño-*Ponce*, I., Kumar, V., Deelen, P., Szperl, A., Trynka, G., Gutierrez-Achury, J., Kanterakis, A., Westra, H.-J., Franke, L., et al. (2014). Fine mapping of the celiac disease-associated LPP locus reveals a potential functional variant. Hum. Mol. Genet. *23*, 2481–2489.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M.T., Perloski, M., Liberg, O., Arnemo, J.M., Hedhammar, Å., and Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature *495*, 360–364.

Axelsson, E., Webster, M.T., Ratnakumar, A., Consortium, L., Ponting, C.P., and Lindblad-Toh, K. (2012). Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. Genome Res *22*, 51-63.

Bateson, W., Saunders, Edith Rebecca (1902). Experimental studies in the physiology of heredity. Rep. Evol. Comm.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science *327*, 836-840.

Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. Methods San Diego Calif *58*, 268–276.

Bolze, A., Byun, M., McDonald, D., Morgan, N.V., Abhyankar, A., Premkumar, L., Puel, A., Bacon, C.M., Rieux-Laucat, F., Pang, K., et al. (2010). Whole-exome-sequencing-based discovery of human FADD deficiency. Am. J. Hum. Genet. *87*, 873–881.

Bouyer, C., Forestier, L., Renand, G., and Oulmouden, A. (2014). Deep intronic mutation and pseudo exon activation as a novel muscular hypertrophy modifier in cattle. PLoS One *9*, e97399.

Broman, K.W., and Weber, J.L. (2000). Characterization of human crossover interference. American journal of human genetics *66*, 1911-1926.

Chang, Y.-F., Imam, J.S., and Wilkinson, M.F. (2007). The Nonsense-Mediated Decay RNA Surveillance Pathway. Annu. Rev. Biochem. *76*, 51–74.

Charlier, C., Coppieters, W., Farnir, F., Grobet, L., Leroy, P.L., Michaux, C., Mni, M., Schwers, A., Vanmanshoven, P., and Hanset, R. (1995). The mh gene causing double-muscling in cattle maps to bovine Chromosome 2. Mamm. Genome Off. J. Int. Mamm. Genome Soc. *6*, 788–792.

Charlier, C., Coppieters, W., Rollin, F., Desmecht, D., Agerholm, J.S., Cambisano, N., Carta, E., Dardano, S., Dive, M., Fasquelle, C., et al. (2008). Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat. Genet. *40*, 449–454.

Charlier, C., Agerholm, J.S., Coppieters, W., Karlskov-Mortensen, P., Li, W., de Jong, G., Fasquelle, C., Karim, L., Cirera, S., Cambisano, N., et al. (2012). A Deletion in the Bovine FANCI Gene Compromises Fertility by Causing Fetal Death and Brachyspina. PLoS ONE *7*, e43085.

Chowdhury, R., Bois, P.R., Feingold, E., Sherman, S.L., and Cheung, V.G. (2009). Genetic analysis of variation in human meiotic recombination. PLoS Genet 5, e1000648.

Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibé, B., Bouix, J., Caiment, F., Elsen, J.-M., Eychenne, F., et al. (2006). A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat. Genet. *38*, 813–818.

Consortium, T.E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Cook, P.R. (1999). The Organization of Replication and Transcription. Science *284*, 1790–1795.

Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., Green, R., Meltzer, P.S., Wolfsberg, T.G., and Collins, F.S. (2006). DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat. Methods *3*, 503–509.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931-21936.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brondum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., et al. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet *46*, 858-865.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. Science *295*, 1306–1311.

Dina, C., Meyre, D., Gallina, S., Durand, E., Körner, A., Jacobson, P., Carlsson, L.M.S., Kiess, W., Vatin, V., Lecoeur, C., et al. (2007). Variation in FTO contributes to childhood obesity and severe adult obesity. Nat. Genet. *39*, 724–726.

Dorshorst, B., Okimoto, R., and Ashwell, C. (2010). Genomic Regions Associated with Dermal Hyperpigmentation, Polydactyly and Other Morphological Traits in the Silkie Chicken. J. Hered. *101*, 339–350.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res *16*, 1299-1309.

Druet, T., and Georges, M. (2010). A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics *184*, 789-798.

Durkin, K., Coppieters, W., Drögemüller, C., Ahariz, N., Cambisano, N., Druet, T., Fasquelle, C., Haile, A., Horin, P., Huang, L., et al. (2012). Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature *482*, 81–84.

Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: Illuminating the Dark Road from Association to Function. Am. J. Hum. Genet. *93*, 779–797.

Fasquelle, C., Sartelet, A., Li, W., Dive, M., Tamma, N., Michaux, C., Druet, T., Huijbers, I.J., Isacke, C.M., Coppieters, W., et al. (2009). Balancing Selection of a Frame-Shift Mutation in the MRC2 Gene Accounts for the Outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. PLoS Genet *5*, e1000666.

Fill, M., Stefani, E., and Nelson, T.E. (1991). Abnormal human sarcoplasmic reticulum Ca2+ release channels in malignant hyperthermic skeletal muscle. Biophys. J. *59*, 1085–1090.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R.B., Elliott, K.S., Lango, H., Rayner, N.W., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science *316*, 889–894.

French, J.D., Ghoussaini, M., Edwards, S.L., Meyer, K.B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M.J., O'Reilly, M., Hillman, K.M., et al. (2013). Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers. Am. J. Hum. Genet. *92*, 489–503.

Fujii, J., Otsu, K., Zorzato, F., de Leon, S., Khanna, V.K., Weiler, J.E., O'Brien, P.J., and MacLennan, D.H. (1991). Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. Science *253*, 448–451.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-α-bound human chromatin interactome. Nature *462*, 58–64.

Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Georges, M. (2011). The long and winding road from correlation to causation. Nat. Genet. *43*, 180–181.

Georges M. (2012). Impact of high-throughput genotyping and sequencing on the identification of genes and variants underlying phenotypic variation in domestic cattle. In Bovine Genomics (Oxford: Wiley-Blackwell), pp. 234–258.

Gilissen, C., Arts, H.H., Hoischen, A., Spruijt, L., Mans, D.A., Arts, P., van Lier, B., Steehouwer, M., van Reeuwijk, J., Kant, S.G., et al. (2010). Exome Sequencing Identifies WDR35 Variants Involved in Sensenbrenner Syndrome. Am. J. Hum. Genet. *87*, 418–423.

Giuffra, E., Evans, G., Törnsten, A., Wales, R., Day, A., Looft, H., Plastow, G., and Andersson, L. (1999). The Belt mutation in pigs is an allele at the Dominant white (I/KIT) locus. Mamm. Genome Off. J. Int. Mamm. Genome Soc. *10*, 1132–1136.

Goddard, M.E., and Hayes, B.J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet *10*, 381-391.

Gorkin, D.U., and Ren, B. (2014). Genetics: Closing the distance on obesity culprits. Nature *507*, 309–310.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., et al. (2002). Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition. Genome Res. *12*, 222–231.

Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J.-J., Kvasz, A., Mni, M., Simon, P., Frère, J.-M., Coppieters, W., et al. (2004). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. U. S. A. *101*, 2398–2403.

Grobet, L., Royo Martin, L.J., Poncelet, D., Pirottin, D., Brouwers, B., Riquet, J., Schoeberlein, A., Dunner, S., Ménissier, F., Massabanda, J., et al. (1997). A deletion in the bovine myostatin gene causes the double–muscled phenotype in cattle. Nat. Genet. *17*, 71–74.

Grobet, L., Poncelet, D., Royo, L.J., Brouwers, B., Pirottin, D., Michaux, C., Menissier, F., Zanotti, M., Dunner, S., and Georges, M. (1998). Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle. Mamm Genome *9*, 210-213.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234–238.

Haley, C.S., and Knott, S.A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity (Edinb) *69*, 315-324.

Handel, M.A., and Schimenti, J.C. (2010). Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. Nat Rev Genet *11*, 124-136.

Hassold, T., Hall, H., and Hunt, P. (2007). The origin of human aneuploidy: where we have been, where we are going. Hum Mol Genet 16 Spec No. *2*, R203-208.

Hassold, T., and Hunt, P. (2001). To err (meiotically) is human: the genesis of human aneuploidy. Nat Rev Genet *2*, 280-291.

Heath, S.C. (1997). Markov Chain Monte Carlo Segregation and Linkage Analysis for Oligogenic Models. Am. J. Hum. Genet. *61*, 748–760.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African Americans. Nature *476*, 170-175.

Hoischen, A., van Bon, B.W.M., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M., de Vries, P., de Reuver, R., Wieskamp, N., Mortier, G., et al. (2010). De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat. Genet. *42*, 483–485.

Hutchison, J.L., Cole, J.B., and Bickhart, D.M. (2014). Short communication: Use of young bulls in the United States. Journal of dairy science *97*, 3213-3220.

Imsland, F., Feng, C., Boije, H., Bed'hom, B., Fillon, V., Dorshorst, B., Rubin, C.-J., Liu, R., Gao, Y., Gu, X., et al. (2012). The Rose-comb Mutation in Chickens Constitutes a Structural Rearrangement Causing Both Altered Comb Morphology and Defective Sperm Motility. PLoS Genet *8*, e1002775.

Jeffreys, A.J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet *29*, 217-222.

Jeffreys, A.J., Murray, J., and Neumann, R. (1998). High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. Molecular cell *2*, 267-273.

Johnson, J.O., Mandrioli, J., Benatar, M., Abramzon, Y., Van Deerlin, V.M., Trojanowski, J.Q., Gibbs, J.R., Brunetti, M., Gronka, S., Wuu, J., et al. (2010). Exome Sequencing Reveals VCP Mutations as a Cause of Familial ALS. Neuron *68*, 857–864.

Kadri, N.K., Sahana, G., Charlier, C., Iso-Touru, T., Guldbrandtsen, B., Karim, L., Nielsen, U.S., Panitz, F., Aamand, G.P., Schulman, N., et al. (2014). A 660-Kb Deletion with Antagonistic Effects on Fertility and Milk Production Segregates at High Frequency in Nordic Red Cattle: Additional Evidence for the Common Occurrence of Balancing Selection in Livestock. PLoS Genet *10*, e1004049.

Kadri, N.K., Harland, C., Faux, P., Cambisano, N., Karim, L., Coppieters, W., Fritz, S., Mullaart, E., Baurain, D., Boichard, D., et al. (2016). Coding and non-coding variants in HFM1, MLH3, MSH4, MSH5, RNF212 and RNF212B affect recombination rate in cattle. Genome Res.

Kao, C.-H., Zeng, Z.-B., and Teasdale, R.D. (1999). Multiple Interval Mapping for Quantitative Trait Loci. Genetics *152*, 1203–1216.

Karim, L., Takeda, H., Lin, L., Druet, T., Arias, J.A.C., Baurain, D., Cambisano, N., Davis, S.R., Farnir, F., Grisart, B., et al. (2011). Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nat. Genet. *43*, 405–413.

Karlsson, E.K., Baranowska, I., Wade, C.M., Salmon Hillbertz, N.H.C., Zody, M.C., Anderson, N., Biagi, T.M., Patterson, N., Pielberg, G.R., Kulbokas, E.J., et al. (2007). Efficient mapping of mendelian traits in dogs through genome-wide association. Nat. Genet. *39*, 1321–1328.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., and Mardis, E.R. (2013). The Next-Generation Sequencing Revolution and Its Impact on Genomics. Cell *155*, 27–38.

Kong, A., Thorleifsson, G., Frigge, M.L., Masson, G., Gudbjartsson, D.F., Villemoes, R., Magnusdottir, E., Olafsdottir, S.B., Thorsteinsdottir, U., and Stefansson, K. (2014). Common and low-frequency variants associated with genome-wide recombination rate. Nat Genet *46*, 11-16.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. Nat Genet *31*, 241-247.

Kong, A., Thorleifsson, G., Stefansson, H., Masson, G., Helgason, A., Gudbjartsson, D.F., Jonsdottir, G.M., Gudjonsson, S.A., Sverrisson, S., Thorlacius, T., et al. (2008). Sequence variants in the RNF212 gene associate with genome-wide recombination rate. Science *319*, 1398-1401.

Lalonde, E., Albrecht, S., Ha, K.C.H., Jacob, K., Bolduc, N., Polychronakos, C., Dechelotte, P., Majewski, J., and Jabado, N. (2010). Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. Hum. Mutat. *31*, 918–923.

Lander, E.S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics *121*, 185–199.

Larson, G., Piperno, D.R., Allaby, R.G., Purugganan, M.D., Andersson, L., Arroyo-Kalin, M., Barton, L., Vigueira, C.C., Denham, T., Dobney, K., et al. (2014). Current perspectives and the future of domestication studies. Proc. Natl. Acad. Sci. 201323964.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. Cell *148*, 84–98.

Lichten, M., and de Massy, B. (2011). The impressionistic landscape of meiotic recombination. Cell 147, 267-270.

Ma, L., O'Connell, J.R., VanRaden, P.M., Shen, B., Padhi, A., Sun, C., Bickhart, D.M., Cole, J.B., Null, D.J., Liu, G.E., et al. (2015). Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. PLoS Genet *11*, e1005387.

MacDonald, M.E., Ambrose, C.M., Duyao, M.P., Myers, R.H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S.A., James, M., Groot, N., et al. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell *72*, 971–983.

MacLeod, I.M., Hayes, B.J., Vander Jagt, C.J., Kemper, K.E., Haile-Mariam, M., Bowman, P.J., Schrooten C., Goddard, M.E. (2014). A Bayesian Analysis to Exploit Imputed Sequence Variants for QTL discovery. In Proceedings, 10th World Congress of Genetics Applied to Livestock Production.

Martinez-Perez, E., and Colaiacovo, M.P. (2009). Distribution of meiotic recombination events: talking to your neighbors. Current opinion in genetics & development *19*, 105-112.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science *337*, 1190–1195.

McPherron, A.C., Lawler, A.M., and Lee, S.-J. (1997). Regulation of skeletal muscle mass in mice by a new TGF-p superfamily member. Nature *387*, 83–90.

McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science *304*, 581-584.

Meuwissen, T.H., and Goddard, M.E. (2001). Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol *33*, 605-634.

Mikawa, S., Morozumi, T., Shimanuki, S.-I., Hayashi, T., Uenishi, H., Domukai, M., Okumura, N., and Awata, T. (2007). Fine mapping of a swine quantitative trait locus for number of vertebrae and analysis of an orphan nuclear receptor, germ cell nuclear factor (NR6A1). Genome Res. *17*, 586–593.

Moller, M.J., Chaudhary, R., Hellmén, E., Höyheim, B., Chowdhary, B., and Andersson, L. (1996). Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding the mast/stem cell growth factor receptor. Mamm. Genome *7*, 822–830.

Myers, S., Bowden, R., Tumian, A., Bontrop, R.E., Freeman, C., MacFie, T.S., McVean, G., and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science *327*, 876-879.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science *310*, 321-324.

Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. Nat Genet *40*, 1124-1129

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242–245.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90.

Nicholas, F.W., and Hobbs, M. (2014). Mutation discovery for Mendelian traits in non-laboratory animals: a review of achievements up to 2012. Anim. Genet. *45*, 157–170.

O'Brien, P.J. (1986). Porcine malignant hyperthermia susceptibility: hypersensitive calcium-release mechanism of skeletal muscle sarcoplasmic reticulum. Can. J. Vet. Res. *50*, 318–328.

Oliver, P.L., Goodstadt, L., Bayes, J.J., Birtle, Z., Roach, K.C., Phadnis, N., Beatson, S.A., Lunter, G., Malik, H.S., and Ponting, C.P. (2009). Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet *5*, e1000753.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.

Pal, C., Papp, B., and Hurst, L.D. (2001). Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. Mol Biol Evol *18*, 2323-2326.

Penrose, L.S. (2009). The relative effects of paternal and maternal age in mongolism. 1933. Journal of genetics *88*, 9-14.

Petkov, P.M., Broman, K.W., Szatkiewicz, J.P., and Paigen, K. (2007). Crossover interference underlies sex differences in recombination rates. Trends Genet *23*, 539-542.

Qanbari, S., Pausch, H., Jansen, S., Somel, M., Strom, T.M., Fries, R., Nielsen, R., and Simianer, H. (2014). Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. PLoS Genet *10*, e1004148.

Ren, J., Duan, Y., Qiao, R., Yao, F., Zhang, Z., Yang, B., Guo, Y., Xiao, S., Wei, R., Ouyang, Z., et al. (2011). A missense mutation in PPARD causes a major QTL effect on ear size in pigs. PLoS Genet *7*, e1002043.

Ricketts, M.H., Simons, M.J., Parma, J., Mercken, L., Dong, Q., and Vassart, G. (1987). A nonsense mutation causes hereditary goitre in the Afrikander cattle and unmasks alternative splicing of thyroglobulin transcripts. Proc. Natl. Acad. Sci. U. S. A. *84*, 3181–3184.

Rubin, C.-J., Zody, M.C., Eriksson, J., Meadows, J.R.S., Sherwood, E., Webster, M.T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., et al. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. Nature *464*, 587–591.

Rubin, C.-J., Megens, H.-J., Barrio, A.M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, Ö., Jern, P., Jørgensen, C.B., et al. (2012). Strong signatures of selection in the domestic pig genome. Proc. Natl. Acad. Sci. 201217149.

Rudolph, J.A., Spier, S.J., Byrns, G., Rojas, C.V., Bernoco, D., and Hoffman, E.P. (1992). Periodic paralysis in quarter horses: a sodium channel mutation disseminated by selective breeding. Nat Genet *2*, 144-147.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.

Salmon Hillbertz, N.H.C., Isaksson, M., Karlsson, E.K., Hellmén, E., Pielberg, G.R., Savolainen, P., Wade, C.M., von Euler, H., Gustafson, U., Hedhammar, Å., et al. (2007). Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. Nat. Genet. *39*, 1318–1320.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.

Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., and Georges, M. (2012). Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. PLoS Genet *8*, e1002854.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. *74*, 5463–5467.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Sartelet, A., Druet, T., Michaux, C., Fasquelle, C., Géron, S., Tamma, N., Zhang, Z., Coppieters, W., Georges, M., and Charlier, C. (2012a). A Splice Site Variant in the Bovine RNF11 Gene Compromises Growth and Regulation of the Inflammatory Response. PLoS Genet *8*, e1002581.

Sartelet, A., Klingbeil, P., Franklin, C.K., Fasquelle, C., Géron, S., Isacke, C.M., Georges, M., and Charlier, C. (2012b). Allelic heterogeneity of Crooked Tail Syndrome: result of balancing selection? Anim. Genet. *43*, 604–607.

Sartelet, A., Stauber, T., Coppieters, W., Ludwig, C.F., Fasquelle, C., Druet, T., Zhang, Z., Ahariz, N., Cambisano, N., Jentsch, T.J., et al. (2014). A missense mutation accelerating the gating of the lysosomal Cl−/H+-exchanger ClC-7/Ostm1 causes osteopetrosis with gingival hamartomas in cattle. Dis. Model. Mech. *7*, 119–128.

Sax, K. (1923). The Association of Size Differences with Seed-Coat Pattern and Pigmentation in PHASEOLUS VULGARIS. Genetics *8*, 552–560.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res.

Sckolnick, M., Krementsova, E.B., Warshaw, D.M., and Trybus, K.M. (2013). More than just a cargo adapter, melanophilin prolongs and slows processive runs of myosin Va. J Biol Chem *288*, 29313-29322.

Scuteri, A., Sanna, S., Chen, W.-M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orrú, M., Usala, G., et al. (2007). Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits. PLoS Genet *3*, e115.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Sillanpää, M.J., and Arjas, E. (1998). Bayesian Mapping of Multiple Quantitative Trait Loci From Incomplete Inbred Line Cross Data. Genetics *148*, 1373–1388.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet *38*, 1348-1354.

Singer, A., Perlman, H., Yan, Y., Walker, C., Corley-Smith, G., Brandhorst, B., and Postlethwait, J. (2002). Sex-specific recombination rates in zebrafish (Danio rerio). Genetics *160*, 649-657.

Smemo, S., Tena, J.J., Kim, K.-H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature *507*, 371–375.

Smith, L.B., Dally, M.R., Sainz, R.D., Rodrigue, K.L., and Oberbauer, A.M. (2006). Enhanced skeletal growth of sheep heterozygous for an inactivated fibroblast growth factor receptor 3. J Anim Sci *84*, 2942-2949.

Soller, M., Brody, T., and Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. *47*, 35–39.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G., et al. (2005). A common inversion under selection in Europeans. Nat Genet *37*, 129-137.

Sur, I.K., Hallikas, O., Vähärautio, A., Yan, J., Turunen, M., Enge, M., Taipale, M., Karhu, A., Aaltonen, L.A., and Taipale, J. (2012). Mice Lacking a Myc Enhancer That Includes Human SNP rs6983267 Are Resistant to Intestinal Tumors. Science *338*, 1360–1363.

Sutter, N.B., Bustamante, C.D., Chase, K., Gray, M.M., Zhao, K., Zhu, L., Padhukasahasram, B., Karlins, E., Davis, S., Jones, P.G., et al. (2007). A Single IGF1 Allele Is a Major Determinant of Small Size in Dogs. Science *316*, 112–115.

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

Udler, M.S., Tyrer, J., and Easton, D.F. (2010). Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. Genet. Epidemiol. 34, 463–468.

Van Laere, A.-S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., Moreau, L., Archibald, A.L., Haley, C.S., Buys, N., Tally, M., et al. (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature *425*, 832–836.

VanRaden, P.M., Olson, K.M., Null, D.J., and Hutchison, J.L. (2011). Harmful recessive effects on fertility detected by absence of homozygous haplotypes. J. Dairy Sci. *94*, 6153–6161.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554-566.

Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A Map of Recent Positive Selection in the Human Genome. PLoS Biol *4*, e72.

Walker, M., Billings, T., Baker, C.L., Powers, N., Tian, H., Saxl, R.L., Choi, K., Hibbs, M.A., Carter, G.W., Handel, M.A., et al. (2015). Affinity-seq detects genome-wide PRDM9 binding sites and reveals the impact of prior chromatin modifications on mammalian recombination hotspot usage. Epigenetics & chromatin *8*, 31.

Wang, J.L., Yang, X., Xia, K., Hu, Z.M., Weng, L., Jin, X., Jiang, H., Zhang, P., Shen, L., Guo, J.F., et al. (2010). TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain *133*, 3510–3518.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and 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*, 1665–1674.

Watson, J.D., and Crick, F.H.C. (1953). Genetical Implications of the Structure of Deoxyribonucleic Acid. Nature *171*, 964–967.

Whitworth, K.M., Rowland, R.R., Ewen, C.L., Trible, B.R., Kerrigan, M.A., Cino-Ozuna, A.G., Samuel, M.S., Lightner, J.E., McLaren, D.G., Mileham, A.J., et al. (2016). Gene-edited pigs are protected from porcine reproductive and respiratory syndrome virus. Nat Biotechnol *34*, 20-22.

Wright, D., Boije, H., Meadows, J.R.S., Bed'hom, B., Gourichon, D., Vieaud, A., Tixier-Boichard, M., Rubin, C.-J., Imsland, F., Hallböök, F., et al. (2009). Copy Number Variation in Intron 1 of SOX5 Causes the Pea-comb Phenotype in Chickens. PLoS Genet *5*, e1000512.

Zeng, Z.B. (1994). Precision mapping of quantitative trait loci. Genetics *136*, 1457–1468.

Zhang, X., Cowper-Sal·lari, R., Bailey, S.D., Moore, J.H., and Lupien, M. (2012). Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome Res. *22*, 1437–1446.

Zhou, X., Baron, R.M., Hardin, M., Cho, M.H., Zielinski, J., Hawrylkiewicz, I., Sliwinski, P., Hersh, C.P., Mancini, J.D., Lu, K., et al. (2012). Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. Hum. Mol. Genet. *21*, 1325–1335.

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

4000 Liège (Belgique)

D/2016/0480/20
ISBN 978-2-87543-092-2