Proceedings, 10th World Congress of genetics Applied to Livestock Production

NGS-based Reverse Genetic Screen Reveals Loss-of-function Variants Compromising Fertility in Cattle.

C. Charlier[1], W. Li[1], C. Harland[1,2], M. Littlejohn[2], F. Creagh[2], M. Keehan[2], T. Druet[1], W. Coppieters[1], R. Spelman[2] & M. Georges[1]
[1]Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, Belgium,
[2]Livestock Improvement Corporation, New Zealand.

ABSTRACT: We herein report the results of a large-scale reverse genetic screen, based on next generation sequencing (NGS) of the exome or whole genome of more than 500 animals, to identify highly deleterious mutations that cause embryonic lethality in domestic cattle. We first demonstrate that - as in human - domestic cattle carry of the order of 100 loss-of-function (LoF) variants per genome. We then present evidence for significant depletion in homozygosity for at least tens of candidate deleterious variants in Belgian Blue Cattle, strongly suggesting that these act as embryonic lethal (EL) or at least juvenile lethal (JL) mutations. We finally formally demonstrate the embryonic/juvenile lethality of a handful of common LoF variants compromising fertility in domestic cattle of Belgium and New Zealand.

Keywords: cattle, fertility, embryonic mortality, loss-of-function mutations, newt generation sequencing.

Introduction

The large-scale use of artificial insemination combined with the estimation of breeding values using the individual animal model has lead to remarkable increases in productivity in livestock, including dairy and beef cattle, over the last 50 years. As an example, average milk yield per lactation has increased from ~6,300 kgs in 1960 to ~11,800 kgs in 2000 in US Holsteins, and more than half of this progress was genetic (Dekkers & Hospital, 2002). However, during the same period, fertility has often severely declined. In the same US Holstein population, the number of days between calving and first estrus has increased from 126 (1976) to 169 (1999)(Washburn et al., 2002). The conception rate for the first insemination dropped from 62% (1972) to 34% (1996)(Silvia, 1998). The number of inseminations required to obtain a pregnancy increased from 1.8 (1970) to 3.0 (2000), while the interval between successive calvings increased from 13.5 (1970) to 14.9 months (2000)(Lucy et al., 2001). It is generally assumed that the negative correlation that is observed between milk yield and fertility is mainly due to the negative energy balance of high-producing cows at the peak of lactation. Accordingly, the heritability of fertility (ranging from 1 to 10%) is generally assumed to have the same quasi-infinitesimal architecture as milk yield (Kemer & Goddard, 2012; Lucy et al., 2001; Sun et al., 2010).

An alternative yet complementary hypothesis is that the decrease in fertility also involves an increase in homozygosity for EL alleles resulting from the drastic decrease in effective size witnessed for several livestock populations including cattle. That this might indeed be the case is supported by the recent identification of a number such ELs, either by positional cloning of mutations that cause a genetic defect in part of homozygotes (and embryonic lethality in the remainder)(Charlier et al., 2012), or by positional cloning of QTL affecting fertility (Kumar Kadri et al., 2014), or by the identification of disruptive variants associated with haplotypes characterized by autozygosity depletion (Sonstegard et al, 2013; Fritz et al., 2013).

It was recently demonstrated that the average human is heterozygous (respectively homozygous) for ~100 (respectively ~20) loss-of-function (LoF) variants, defined as stop gain, splice site, frame-shift and large deletions in protein coding genes (MacArthur et al., 2012). What proportion of these are highly deleterious at the homozygous state, including by causing embryonic/fetal death, remains largely unknown. It was previously estimated from the increase in perinatal lethality in consanguineous marriages that human carried on average ~1.2 highly deleterious equivalents (Bittles & Neel, 1994). However, these estimates did not account for increased embryonic lethality.

We herein describe the results of a large-scale reverse genetic screen for loss-of-function variants in domestic cattle. We first demonstrate that – against expectation – the exomic nucleotide diversity is nearly twice as high in cattle when compared to out-of-Africa humans, yet that the deleterious mutation load is approximately equal. By mining available sequence data, we identify hundreds of candidate disruptive variants, which we genotypes in 2,000 to 11,000 normal animals. When considering all candidate variants jointly, we observe a significant depletion in homozygosity for the derived allele (when compared to Hardy-Weinberg expectations or frequency-matched control SNPs). This suggests that the collection of mutations comprises highly deleterious alleles possibly including ELs. For three of the most common candidates, we demonstrate significant depletion in homozygotes in matings between carrier sires and dams, hence confirming their EL nature.

Materials and Methods

Whole genome sequencing was conducted using Illumina's TruSeq DNA PCR-Free Sample Preparation Kits for library construction and Illumina's HiSeq2000 instruments for actual sequencing. Exome sequencing was conducted using Agilent's Sure Select Target Enrichment Kits for library construction and Illumina's HiSeq2000 instruments for actual sequencing. Average sequence depths were 10-15 fold for whole genome sequencing and 25-35 fold for exome sequencing. Resulting reads were aligned to reference genomes using BWA (Li and Durbin, 2009), PCR duplicates removed using SAMTOOLS (http://samtools.sourceforge.net), and variants detected using the GATK pipeline and recommended procedures (DePristo et al., 2011). Functional effects of variants were predicted using custom-made tools, SIFT (Kumar et al., 2009) and PolyPhen2 (Adzhubei et al., 2010).

Results

Higher diversity yet equivalent deleterious mutation load in *Bos Taurus* as in human. It is often assumed that artificial selection has resulted in a drastic erosion of diversity of domestic animals and plants, which might increase communal susceptibility to pathogens, and compromise future breeding options. We aimed at rigorously quantifying the level of residual genetic variation of the exome in domestic cattle, and compare it with – for instance – equivalent metrics determined in human. To that end, we sequenced the exome of 96 animals representing 10 breeds (including six *Bos Taurus* breeds). We used the Agilent SureSelect exome capture Kit, and performed sequencing on Hiseq2000 instruments. The targeted sequence depth was 30-fold. Resulting reads were aligned to the BosTau6 genome assembly with BWA, PCR duplicates removed with SAMTOOLS, and variants called using the GATK pipeline. For human, we downloaded bam files for 100 exomes from the 1,000 Genomes Project and called variants using the exact same pipeline. In addition, we produced exome data for 50 humans in house, using the Illumina TrueSeq exome capture kit. Variants considered included synonymous, non-synonymous, stop gains and splice site variants.

We identified a subset of 159,691 exons, with same length in bovine and human, and bounded by canonical exon-intron boundaries (internal exons) or by a start/stop codon and canonical boundaries (first and last exons). These represent 77.7% of the total number of exons (205,584) defined in the human genome. Olfactory receptor genes were eliminated from the analysis. Within this sequence space, we determined an individual's genotype for all variant positions that were covered by at least 20 reads with mapping quality score of at least 30.

When considering all variant types jointly, the nucleotide diversity was nearly twice as high in *Bos taurus* breeds, when compared to out-of-Africa humans, and 1.5 times as high when compared to Africans (Table 1). Thus, against expectations, domestic taurine cattle are genetically more variable than are humans.

Table 1: Comparison of exomic variation in human and cattle

Cohort	Exomic π	S/NS	n° SG het	n° SG hom	n° SS het	n° SS hom
BT	0.00056	2.12	29	7	12	1
HS - EUR	0.00031	1.55	33	6	8	2
HS - AS	0.00029	1.56	25	6	8	3
HS - YRI	0.00041	1.59	37	6	12	1

BT: *Bos taurus*; HS: *Homo sapiens*; EUR: European; AS: Asian; YRI: Yeruban from Nigeria; π: nucleotide diversity (all variants); S/NS: ratio of synonymous over non-synonymous variants; SG: stop gains; SS: splice site variants. n° : average numbers extrapolated to a complete exome based on the number of sequenced triplets and the fact that the studied exomes represent 77.7% of exome space.

However, when focusing on two types of LoF variants, namely stop gains and splice site variants, the estimated number of heterozygous as well as homozygous LoF position per complete exome are very similar between cattle and humans (of the order of 40 heterozygous sites, and 8 homozygous sites)(Table 1). These numbers are in very close agreement with previous findings (McArthur et al., 2012), and amount to a total of ~100 heterozygous and ~20 homozygous LoF sites per individual (including frame-shifts and large deletions). Also, the ratio of synonymous over non-synonymous variants is larger in cattle (2.1) than in humans (1.6). Both observations suggest that purifying selection is more effective in cattle than in human. This may be due to the recent drastic reduction in effective population size accompanying breed formation in domestic cattle in the last two centuries, with concomitant increase in "autozygosity" and hence purging of deleterious recessives.

Depletion in homozygosity for candidate LoF and deleterious missense variants in Belgian Blue Cattle (BBC). In addition to the exome sequences of 19 BBC generated as described above, we generated whole genome sequence data for 50 BBC sires at average sequence depth >12. The whole-genome data were processed using the same pipeline as the exome data. We mined all available data and selected 73 manually curated LoF variants corresponding to frame-shift, stop-gain and splice site mutations, and 258 missense (MS) variants that were predicted by SIFT and PolyPhen2 to be damaging/disruptive. In addition, (ii) neither of these variants were reported in any other breed, (iii) all were affecting a gene expressed in embryonic tissue (as judged from available RNA-Seq data generated for pituitary, cortex and liver of 60d embryos), (iv) and for none of them was any of the sequenced individuals homozygous. Custom assays to interrogate the corresponding variants were added to the Illumina BovineLD array and – at the time of writing – >2,000 normal BBC animals had been genotyped with the corresponding SNP chip.

No homozygotes were observed amongst the genotyped animals for 112 of the 331 tested LoF and MS variants. To verify whether these numbers were higher than expected for neutral variants (while accounting for the allelic frequencies of the candidate SNPs), we randomly sampled 10,000 sets of 331"control" variants (from the non-custom SNPs on the LD array) matched for allelic frequencies. We never observed 112/331 variants without homozygotes for the derived allele (mean: 80/331; max: 96/331). This strongly suggests that our list of 112 LoF and MS variants (without homozygotes for the derived allele) includes true EL (or at least highly deleterious "juvenile lethal" variants - JL). Provided that the control SNPs are comparable to the LoF/MS variants, one can roughly estimate that at least 32/112 (i.e. 28%) of the candidate LoF/MF variants are truly EL/JL.

Confirming the EL/JL nature of common LoF variants. We then aimed at confirming the embryonic lethality of the candidate LoF/MS variants for which the depletion in homozygosity in BBC was the most significant. The top variant corresponds to a frame-shift mutation in the *SNAPC4* gene coding for the *small nuclear RNA activating complex, polypeptide 4*, for which 9% of genotyped BBC animals were heterozygous (depletion in homozygotes, $p = 4x10^{-4}$). Knock-out of the corresponding gene in the mouse causes embryonic lethality. Using a custom-made Taqman assay, we identified 17 matings between carrier sires and dams. At the time of writing, three born calves were homozygous wild-type, eight were heterozygous, while none were homozygous mutant ($p = 0.04$). Three of the remaining pregnancies resulted in spontaneous abortion (between six weeks and four months of gestation), while three others are still ongoing. Taken together, these results confirm the embryonic lethality of the *SNAPC4* frame-shift mutation in BBC. Confirmation of the EL effect of other top candidate LoF and MS variants is in progress.

We applied the same pipeline for the detection of candidate LoF variants to whole-genome sequence data generated for 497 normal animals (average sequence depth: 10 fold; range 2-148 fold) from the New Zealand dairy cattle population. This yielded a list of 275 candidate LoFs. We genotyped ~11,000 normal animals from NZ for 43 candidate variants using a MassARRAY assay (Sequenom). Homozygote mutants were not observed for 11/43 SNPs. The homozygote depletion was most significant ($p = 0.0004$) for a frame-shift mutation, for which 6.4% of genotyped animals were heterozygous. Further targeted genotyping for this variant identified 61 calves born from matings between carrier sires and dams. None of these were homozygous mutant ($p < 0.0001$), clearly confirming the EL nature of this frame-shift mutation.

At least one other of the 43 candidate LoF variants was characterized by a near significant yet incomplete depletion in homozygotes in the population as well as in matings between carrier sires and dams. This variant mapped to a region associated with the Small Calf Syndrome (SCS) that was recently described in the New-Zealand dairy cattle population. Available data strongly suggest that the corresponding mutations causes EL in part of the homozygous conceptuses, SCS in others, and no striking symptoms in the remaining.

Further characterization of the remaining variants is progressing and will be presented.

Discussion and conclusions

We herein use an NGS-based reverse genetic approach to identify EL variants that affect fertility of domestic cattle. We show – against expectations – that exomic variation is higher in domestic cattle than in humans when considering all variants indiscriminately. This could either reflect the fact that the genetic bottleneck resulting from the domestication process was not as severe as one might have thought. This implies sustained genetic exchange between domestic and wild-type populations over long periods of time. An alternative explanation, which we are testing, is that the present gene pool of domestic cattle derives from the domestication of multiple *bos* subspecies with subsequent hybridization. Thus the genome of domestic cattle would be "mosaic" in the same sense as the genome of the laboratory mouse (Wade et al., 2002). Previous analyses of the *IGF2* locus in pigs (Van Laere et al., 2003) and of the *PIGR* locus in cattle (Berry et al., 2013) suggests this may indeed apply to livestock species as well.

We also show that the number of LoF variants per exome is of the same order of magnitude in cattle as in humans, i.e. ~100. We surmise that this apparent contradiction (higher overall variation, comparable deleterious mutation load) reflects more effective purifying selection in domestic cattle than in humans as a result of the reduction in effective population size (and hence increase in inbreeding) accompanying breed creation and stringent artificial selection.

The phenotypic consequences of these highly disruptive (on protein function) LoF variants remain largely unknown. However, as shown in this and other studies, at least some of the LoF variants are highly deleterious, causing either embryonic or juvenile death. The structure and dynamics of livestock populations causes LoF variants to abruptly increase in frequency and hence contribute in non-negligible ways to decrease in fertility. What proportion of LoF variants are true EL/JL remains unknown, but our study suggests that it may be substantial. Further analyses are conducted to estimate this proportion more accurately.

Our data also suggests that a proportion of LoF variants are highly deleterious, yet without being fully penetrant EL/JL. This is exemplified by the finding of a LoF variant that is apparently causing embryonic death in some, dwarfism (SMS) in others, and no overt symptoms in the remainder of the homozygotes. Along the same lines, we previously reported a LoF mutation in the *RNF11* gene that would cause juvenile death in a majority of animals as a result of unbridled inflammatory response towards infection, and stunted growth in surviving animals (Sartelet et al., 2012). We are in the process of testing the effect of our list of LoF variants on recorded phenotypes, particularly as they relate to health and disease resistance.

It is finally worth noting that balancing selection may account for the unusually high frequency of some deleterious variants. We have previously shown that this is the case for *MRC2* LoF variants that cause Crooked Tail Syndrome in homozygotes yet increased muscle mass in heterozygotes (Fasquelle et al., 2009), as well as for a 660Kb deletion that causes EL in homozygotes yet increased milk yield in heterozygotes (Kumar Kadri et al., 2014). Preliminary evidence suggests that the *SNAPC4* mutation reported in this work may be advantageous in heterozygotes as well. Moreover, there are numerous other examples of balancing selection of deleterious variants in other livestock species.

The ability to identify EL/JL variants, including by means of the reverse genetic strategy described in this work, should provide breeders with useful information to avoid at risk matings and hence decrease pregnancy failure. Moreover, the possibility to identify candidate functional variants from population level resequencing data should allow us to enrich marker panels with causative variants, thereby increasing the accuracy of genomic selection.

Literature Cited

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) Nat Methods 7:248-249.

Berry S, Coppieters W, Davis S, Burrett A, Thomas N, et al. (2013) PLoS One 8:e57219.

Bittles AH, Neel JV (1994) Nat Genet 8: 117-121.

Dekkers JC, Hospital F (2002) Nat Rev Genet 3: 22-32.

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, et al. (2011). Nature Genetics 43:491-498.

Charlier C, Agerholm JS, Coppieters W, Karlskov-Mortensen P, Li W, et al. (2012) PLoS One 7: e43085.

Fasquelle C, Sartelet A, Li W, Dive M, Tamma N, et al. (2009) PLoS Genet 5:e1000666.

Fritz S, Capitan A, Djari A, Rodriguez SC, Barbat A, et al. (2013) PLoS One 8: e65550.

Kadri NK, Sahana G, Charlier C, Iso-Touru T, Guldbrandtsen B, et al. (2014) PLoS Genet. 10:e1004049.

Kemper KE, Goddard ME (2012) Hum Mol Genet 21: R45-51.

Kumar P, Henikoff S, Ng PC (2009) Nat Protoc 4:1073-81.

Li H, Durbin R (2009) Bioinformatics 25:1754-1760.

Lucy MC (2001) J Dairy Sci 84: 1277-1293.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) Science 335: 823-828.

Sartelet A, Druet T, Michaux C, Fasquelle C, Géron S, et al. (2012) PLoS Genet 8:e1002581.

Silvia W (1998) J Dairy Sci 81(Suppl. 1): 244.

Sun C, Madsen P, Lund MS, Zhang Y, Nielsen US, et al. (2010) J Anim Sci 88: 871-878.

Sonstegard TS, Cole JB, VanRaden PM, Van Tassell CP, Null DJ, et al. (2013) PLoS One 8: e54872.

Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, et al. (2003) Nature 425:832-836.

Wade CM, Kulbokas EJ 3rd, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ (2002) Nature 420:574-578.

Washburn SP, Silvia WJ, Brown CH, McDaniel BT, McAllister AJ (2002) J Dairy Sci 85: 244-251.