UNIVERSITÀ CATTOLICA DEL SACRO CUORE
Sede di Piacenza

Scuola di Dottorato per il Sistema Agro-alimentare

Doctoral School on the Agro-Food System

cycle XXVII

S.S.D: AGR/17

High-throughput sequencing technologies and genomics: insights into the bovine species

Candidate: Marco Milanesi
Matr. n.: 4011148

Academic Year 2013/2014

UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Scuola di Dottorato per il Sistema Agro-alimentare

Doctoral School on the Agro-Food System

cycle XXVII

S.S.D: AGR/17

High-throughput sequencing technologies and genomics: insights into the bovine species

Coordinator: Ch.mo Prof. Antonio Albanese

Candidate: Marco Milanesi
Matriculation n.: 4011148

Tutor: Prof. Paolo Ajmone-Marsan
Ph.D Lorenzo Bomba
Ph.D Stefano Capomaccio
Ph.D Ezequiel Louis Nicolazzi

Academic Year 2013/2014

Alla mia famiglia

A mio nonno Mario

INDEX

CHAPTER 1

Introduction

Background

The bovine species counts almost 1.5 billion of heads classified in some 800 breeds worldwide (http://faostat3.fao.org). The species is adapted to a wide variety of different environment and husbandry systems and is often reared for multiple purposes (Taberlet et al., 2008), draft power, milk, meat, leather production and other uses (Elsik et al., 2009).

European cattle (*Bos taurus*) derive from a single domestication event of the auroch (*Bos primigenius*) occurred in the Fertile Crescent around 10,000 years ago (Taberlet et al., 2011). Following domestication, cattle followed the Neolithic expansion of agriculture and colonised Europe (see Ajmone-Marsan et al., 2010 for a review). Many thousand years of natural and artificial selection have shaped cattle phenotype and genotype to meet environmental conditions and human needs and together with demographic events and genetic drift have started the differentiation of the bovine species in diverse and locally adapted populations. This differentiation process accelerated in the last two centuries, when the breed concept was invented, so that now this species counts hundreds of different breeds with high diversity in coat colour, body size, production aptitude, adaptation to climate, tolerance to disease, etc. (Taberlet et al., 2008; Elsik et al., 2009; The Bovine HapMap Consortium, 2009).

This diversity is now being progressively lost, with the extinction of many local populations, mainly for economic reasons. Selection efforts have therefore been focussed on a few outperforming industrial breeds that are presently selected for specialised dairy or beef production and sometimes double purpose.

In these breeds breeding schemes are well developed and have been traditionally based on the genetic evaluation of sires and dams using data collected from performance and progeny testing and unbiased linear prediction models. These are founded on the Fisher infinitesimal modelling of quantitative trait variation and need no genomic information to estimate breeding values (EBVs). The genetic progress recorded using this approach has been relevant, particularly for traits having medium to high heritability. In Italian Holstein genetic trends for the main production traits (milk, protein and fat yield) are positive. However accurate estimate of breeding values has a high cost, in terms of investment and time, so that 5 to 6 years are needed to have a dairy bull progeny tested (www.anafi.it).

The advent of genomic selection (Meuwissen et al., 2001) marked a milestone in dairy cattle breeding. The idea was developed before the molecular tools were in place for implement it in practice and has become a reality only recently. In this approach progeny tested bulls are genotyped at many thousand loci and the value of their EBVs used to estimate the value of each allele at each locus genotyped. These values are then used to predict the genetic value of young animals on the basis of their genotype. As a result, animals from the new generation receive an EBV with the same accuracy of estimates obtained by progeny testing, but much earlier. This lower the generation interval and greatly increases the rate of genetic progress in dairy populations. The initial idea was so good that nowadays genomic selection is widely applied in industrial countries to the major dairy cattle breeds. With the decrease of genotyping cost, it is likely that its use will be soon expanded to animals selected for other purposes (e.g. beef cattle) and other species (e.g. pigs). However, genomic selection is an extension of traditional selection. It is extremely useful in breeding for complex traits but gives no biological information on the genes that control them. If the traditional selection views a sire genome as a big black box having a value, but no clue why, genomic selection views the same genome disassembled in many small boxes, but still black and still with no clue on the reason of their value.

It is reasonable to think that a deeper understanding of the biological mechanisms controlling traits may further increase the accuracy of genomic evaluations. Indeed, the new technologies offer new opportunities in this sense, by lowering costs and increasing precision of data production.

This thesis is exploring the use of these technologies for retrieving biological information from genomic data. It uses methods now considered “traditional” in genomics (GWAS and analysis of selection sweeps) and less traditional ones (gene centered GWAS and exome analysis). All data used here have been produced within two national projects on farm animal genomics funded by the Italian Ministry of Agriculture:

- SELMOL "Ricerca e innovazione nelle attività di miglioramento genetico animale mediante tecniche di genetica molecolare per la competitività del sistema zootecnico nazionale I" (Research and Innovation in animal breeding sheep, goat, buffalo, horse and donkey I)
- INNOVAGEN "Ricerca e innovazione nelle attività di miglioramento genetico animale mediante tecniche di genetica molecolare per la competitività del sistema zootecnico nazionale II" (Research and Innovation in animal breeding sheep, goat, buffalo, horse and donkey II)

In these projects, the research group I work with - Istituto di Zootecnica at Università Cattolica del Sacro Cuore di Piacenza - coordinated all the research activities concerning the application of genomics techniques in dairy cattle.

Contents of the thesis

The thesis is structured into a short introduction (this first chapter), 5 core chapters (Chapter 2 to 6) and a short Conclusion (Chapter 7). With the exception of Chapter 6, still in preparation, the other core chapters contain manuscripts that at time of writing are under review for their publication in international peer-reviewed journals or manuscript.

In Chapter 2 selection signature in dairy and beef cattle are investigated with the Illumina BovineSNP50 Genotyping BeadChip. Five Italian cattle breeds (Italian Holstein, Italian Brown, Italian Simmental, Marchigiana and Piedmontese) are analyzed with the rEHH method (Sabeti et al., 2002) to find signatures of recent selection. Candidate genes in the vicinity of signatures shared by either dairy or beef breeds are identified and their possible role in dairy and beef production discussed.

In Chapter 3 a GWAS analysis is run to investigate production traits in three Italian dairy cattle breeds (Italian Holstein, Italian Brown, Italian Simmental). Data are analysed with a well established method (Amin et al., 2007) implemented in the GenABEL R package (Aulchenko et al., 2007) and with the new gene-based method. Association is investigated between 50K SNPs and milk production (milk yield) and quality traits (percentage of milk protein and fat).

In Chapter 4 the gene-based method, used in the previous chapter 3, inspired by the VEGAS approach (Liu et al., 2010) is described. The software was developed and re-coded in our laboratory in order to permit the analysis of any species in addition to human. Multi-testing correction and plot representation complete the tools made available.

In Chapter 5 we test the impact of the use of different reference sequences in the imputation step. Italian Simmental data are used to test variation in genome-wide inputation accuracy using different cattle reference sequences (BTAU4.2, UMD3.1 and BTAU4.6). To increase the analyses reliability, four different inputation softwares are tested.

In Chapter 6 Holstein data are analysed combining exome sequences and 800K SNPs data to seek deleterious recessive variants. While in natural populations the destiny of these variants is to disappear, in livestock breeds they can be maintained and disseminated if carried by a highly productive sire. The intersection of deleterious mutations identified by sequencing and lack of homozygous regions identified by the analysis of the population with the HD SNP chip identified genes candidate to carry recessive deleterious variants.

Objective

The main objective of the thesis was to investigate the bovine genome with the aid new high-throughput technology and to explore new strategies for linking biology (genes and their function) to traits. This linkage, once validated may increase the accuracy of genomic prediction and find application in selection schemes, e.g. in eradicating the most deleterious variants and/or in the planning of matings.

References

Ajmone-Marsan, P., Garcia, J.F., Lenstra, J.A., 2010. On the origin of cattle: How aurochs became cattle and colonized the world. Evol. Anthropol. 19, 148–157. doi:10.1002/evan.20267

Amin, N., van Duijn, C.M., Aulchenko, Y.S., 2007. A Genomic Background Based Method for Association Analysis in Related Individuals. PLoS ONE 2, e1274. doi:10.1371/journal.pone.0001274

Aulchenko, Y.S., Koning, D.-J. de, Haley, C., 2007. Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. Genetics 177, 577–585. doi:10.1534/genetics.107.075614

Elsik, C.G., Tellam, R.L., Worley, K.C., 2009. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science 324, 522–528. doi:10.1126/science.1169588

Liu, J.Z., Mcrae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Hayward, N.K., Montgomery, G.W., Visscher, P.M., Martin, N.G., Macgregor, S., 2010. A Versatile Gene-Based Test for Genome-wide Association Studies. The American Journal of Human Genetics 87, 139–145. doi:10.1016/j.ajhg.2010.06.009

Meuwissen, T.H., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

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., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E.S., 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832–837. doi:10.1038/nature01140

Taberlet, P., Coissac, E., Pansu, J., Pompanon, F., 2011. Conservation genetics of cattle, sheep, and goats. C. R. Biol. 334, 247–254. doi:10.1016/j.crvi.2010.12.007

Taberlet, P., Valentini, A., Rezaei, H.R., Naderi, S., Pompanon, F., Negrini, R., Ajmone-Marsan, P., 2008. Are cattle, sheep, and goats endangered species? Mol. Ecol. 17, 275–284. doi:10.1111/j.1365-294X.2007.03475.x

The Bovine HapMap Consortium, 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324, 528–532. doi:10.1126/science.1167936

CHAPTER 2

Relative extended haplotype homozygosity (rEHH) signals across breeds reveal dairy and beef specific signatures of selection

L. Bomba[1]*, E.L. Nicolazzi[2]*, M. Milanesi[1], et al.

[1] Istituto di Zootecnica, UCSC, via Emilia Parmense 84, 29122 Piacenza, Italy;
[2] Fondazione Parco Tecnologico Padano, Via Einstein, Loc. Cascina Codazza, 26900, Lodi, Italy;

*These authors contributed equally to this work

Submitted to *Genetic Selection Evolution*

Abstract

Background

Nowadays, many methods have been implemented to scan for signatures of selection at genome-scale level within and across breeds. Extended haplotype homozygosity is a proficient approach to detect genome regions under recent selective pressure within breeds. The objective of this study is to identify common regions under positive recent selection shared by most represented dairy and beef Italian cattle breeds.

Results

A total of 3,220 animals from Italian Holstein (2,179), Italian Brown (775), Simmental (493), Marchigiana (485) and Piedmontese (379) breeds were genotyped with the Illumina BovineSNP50 BeadChip v.1 in the frame of the Italian livestock genomic projects "SelMol" and "Prozoo". After standard cleaning procedures, genotypes were phased and core haplotypes identified. The decay of Linkage Disequilibrium (LD) for each core haplotype was assessed measuring the extended haplotype homozygosity (EHH). To correct for the lack of good estimates of local recombination rates, the relative EHH (rEHH) was calculated for each core haplotype. Genomic regions under recent positive selection were identified as those highly frequent core haplotypes with significant rEHH values. Significant independent core haplotypes were aligned across breeds to identify signals of recent selection shared by dairy and shared by beef breeds. Overall, 82 and 87 common regions under selection were detected among dairy and beef cattle breeds, respectively. Bioinformatics analysis identified 244 and 232 genes mapping in these common genomic regions. Pathway and network analysis of significant genes revealed molecular functions biologically related to signal of selections detected in milk or meat production types.

Conclusions

Results suggest that the multi-breed approach is a method to identify selection signatures in cattle breeds under selection for the same production goal,

allowing to better pinpoint the genomic regions of interest in dairy and beef production.

Background

The advent of genomic technologies and the consequent availability of single nucleotide polymorphism (SNP) data have enabled the study of selection events on the cattle genome (Hayes et al., 2008; Qanbari et al., 2010). Such selection signals arise from environmental or anthropogenic pressures, and help to understand the causes that led to breed formation. These studies are usually addressed in a "top-down" approach (Shahzad and Loor, 2012), from genotype to phenotype, in which genomic data is statistically evaluated to detect directional selection. The phenotype needed to run selection signatures is intended in a very general sense: a breed, a production aptitude or even a geographic area of origin. This approach holds the potential to investigate traits as adaptation to extreme climates and different feeding and husbandry systems, resilience to diseases, etc., that are very expensive, difficult and sometimes impossible to study with classic GWAS (Genome-Wide Association Study) approaches. Therefore, it is expected that genomic information retrieved in these studies only partially overlaps that identified by GWAS on specific traits. Searching for these "signatures of selection" is of interest for understanding molecular mechanisms underlying important biological processes and providing new insights to functional biological information (*e.g.* disease and/or important economic traits (Nielsen et al., 2007)). To date, many methods have been implemented to scan these selection signatures at the genomic level. Methods differ with respect to several properties (Lenstra et al., 2012): there are methods that perform analyses within-breed or across-breeds, that compare allele or haplotype frequencies and size, that detect recent or ancient selection events either still segregating or fixated into populations (Lenstra et al., 2012; O'Brien et al., 2014; Utsunomiya et al., 2013). Different methods have different sensitivity and robustness, *e.g.* they may be influenced at a different extent by marker ascertainment bias and uneven distribution of recombination hotspots along the genome.

Extended haplotype homozygosity (EHH) is a method that has been used in many animal species, including cattle (Pan et al., 2013; Qanbari et al., 2010; Zhang et al., 2006). The EHH method was first developed by Sabeti *et al.* (2002), for human population analysis. The main objective of this methodology is to identify long-range haplotypes inherited without recombination events. More in detail, under a neutral evolution model, changes in allele frequencies are driven only by genetic drift. In this scenario, a new variant would require a long time to reach high frequency in the population, and the Linkage Disequilibrium (LD) surrounding it would decay due to recombination events (Kimura, 1984). In the case of positive selection, a quick rise in frequency of a beneficial mutation in a relatively short time will preserve the original haplotype structure, as the number of recombination events would be limited. Therefore, a signature of positive selection is defined as a region with an allele in long-range LD, located in an uncommonly high frequency haplotype.

One of the most interesting features of this method is that it is able to detect genomic regions that are under recent selection. These recent selection signatures can be, therefore, used to identify the genomic regions involved in breed formation. In addition, EHH method does not require any definition of an ancestral allele (as needed in the integrated haplotype score - iHS (Voight et al., 2006) and is suited to be applied to SNP data, because it is less sensitive to ascertainment bias than other methods (Nielsen et al., 2007). However, the EHH method generates a large number of false positive and negative results, due to heterogeneous recombination rates along the genome (Qanbari et al., 2010). Moreover, another drawback not specific of EHH but shared by all selection signature method, is the lack of robust inferences able to discern true signals from those generated by chance (Kemper et al., 2014).

To partially account for these limitations, Sabeti *et al.* (2002) developed the relative extended haplotype homozygosity (rEHH), including an empirical methodology to obtain significant signals. The rEHH of a putative core haplotype compares its original EHH value with that of other haplotypes at that specific locus, using all other haplotypes as control for local variation in the recombination rate. Therefore it identifies genomic regions carrying variants under selection that are still segregating in the population analysed. Although these methods were conceived for human populations, they were successfully applied to livestock species, such as pig (Ma et al., 2012) and cattle (Qanbari et al., 2010).

After domestication - about 10,000 years ago in the fertile crescent - taurine cattle have colonized Europe and Africa and have been selected for different human needs (Taberlet et al., 2011). Particularly, in the last centuries, the anthropic pressure led to the formation of hundreds of specialized breeds adapted to different environmental conditions and linked to local tradition constituting a gene pool deserving attention for conservation (Ajmone-Marsan et al., 2010). Some of these breeds have been selected specifically for dairy, beef, or both production types, following strong artificial selection for these traits (Hayes et al., 2009). The present study aims at identifying signals of recent directional selection using rEHH method in dairy and beef production types, using data from five Italian dairy, beef and dual purpose cattle breeds. We focused on the significant core haplotypes scanned by rEHH that are shared among breeds of the same production type. Then, we used a bioinformatics approach to identify positional candidate genes lying within the genomic regions under selection and investigate their biological role.

Methods

Sample Animals and Genotyping

A total of 4,311 bulls of five Italian dairy, beef, and dual purpose breeds were genotyped with the Illumina BovineSNP50 BeadChip v.1 (Illumina, San Diego, CA), joining the genotyping efforts of two Italian projects (namely, “SelMol” and “Prozoo”). The dataset included 101 replicates and 773 sire-sons pairs, used for downstream quality check of data produced. In detail, genotypes of 2,954 dairy (2,179 Italian Holstein and 775 Italian Brown), 864 beef (485 Marchigiana and 379 Piedmontese) and 493 dual purpose (Italian Simmental) bulls were available. Data quality control (QC) was performed in two stages: a first stage on animals, independently for each breed, applying the same methods and thresholds, and a second stage applied on markers, across all the individuals of the dataset. The first stage excluded individuals with unexpectedly high (≥ 0.2%) Mendelian errors, for father-son couples, and with low call rates (≤ 95%). The second stage excluded: i) SNPs with ≤ 2.5% missing values in the whole dataset or completely missing in one breed ii) SNPs with minor allele frequency ≤ 5%; and iii) SNPs in sexual chromosomes and with unassigned chromosome or physical position.

Estimation of rEHH

Haplotypes were obtained using fastPHASE software considering default options (Scheet and Stephens, 2006), and were run breed-wise and chromosome-wise in each breed. Pedigree information for all bulls were provided by their respective breeder associations, and were used to filter out direct relatives (in father-son pairs the son was maintained in the dataset and the father discarded) and over-represented families (a maximum of 5 randomly chosen individuals per half-sib family was allowed). The final dataset containing these "less related" animals will be henceforth called "non redundant" dataset. This non-redundant dataset was used also to calculate within breed pair-wise whole-genome Linkage Disequilibrium (LD). The r2 statistic for all pairs of markers was obtained using PLINK software v.1.0.7 (Purcell et al., 2007). The decay of LD was assessed by averaging r2 values, up to 1Mb distance.

To test if the population structure influenced the detection of rEHH, we replicated all analyses on the whole Italian Holstein dataset, without any population correction (*i.e.* "redundant" dataset), focusing on genes or gene clusters that are well known to be under recent selection in cattle (*i.e.* "candidate regions"). In particular we focused on the casein, polled gene cluster and coat colour genes (*MC1R* and *KIT* (Kemper et al., 2014; Qanbari et al., 2010)).

The EHH and rEHH calculations were performed using Sweep v.1.1 software (Sabeti et al., 2002). Some default program settings had to be modified to adapt these analyses to the bovine genome, as the software was originally developed for human genetic analyses. Specifically, local recombination rates between SNPs were approximated to 1cM per Mb. EHH and rEHH calculations were performed breed-wise and chromosome-wise, using an automatic core selection with default options (*i.e.* considering longest non-overlapping cores and limiting cores to at least 3 and no more than 20 SNPs) as in Qanbari *et al.* (2010). Although EHH and rEHH values were obtained for all core haplotypes, only those with frequency ≥ 25% were retained for further analyses. The (empirical) rEHH significance threshold was obtained by dividing the rEHH values into 20 bins of 5% frequency range each, log-transforming within-bin values, to achieve normality. Core haplotypes with *p*-values $\leq$ 0.05 were considered significant. Detection of common selection signatures was obtained comparing rEHH values across genomic regions across breeds.

Breed grouping according to production type

To examine the signals common to each of the two production types (*i.e.* dairy and beef), core haplotypes that shared one or more SNP in at least two breeds from the same production type were selected. The dual purpose Italian Simmental was included in both dairy (Italian Holstein and Italian Brown) and beef groups (Piedmontese and Marchigiana), since potentially it possesses haplotypes selected for both production types. All downstream analyses were performed separately for the dairy and the beef production types.

Detection and annotation of candidate genes

Genome annotation was performed using Biomart (http://www.biomart.org/index.html), a comprehensive source of gene annotation for many species provided by Ensembl (www.ensembl.org/). The list of shared haplotypes regions (in bp) for dairy and beef was used as input file in the Biomart web interface. The set of genes retrieved by Biomart was then used as input for the Canonical pathway and regulatory network analyses. Ingenuity Pathway Analysis tool version 8.0 (IPA; Ingenuity® Systems, Inc, Redwood City, CA; http://www.ingenuity.com), and a substantial examination of published literature was used to examine the functional relationships among the resulting genes. The IPA operates with a proprietary knowledge database, providing complementary pathway analysis for several species, including cattle. In the IPA analyses, the significance of each biological function identified was estimated using Fisher's exact test. A Benjamini and Hochberg correction for multiple testing was calculated for function and canonical pathways. For each Network, IPA computed a score (*p*-score=-$\log_{10}$(*p*-value)) fitting the set studied genes and a list of biological functions present in the IPA knowledge database. The score considered the number of genes in the network and the size of the network to estimate how relevant this network was, according to the list of genes provided. Then, the score was used to rank the networks.

Results

Dataset QC

The repeatability obtained from the 101 replicates present in the whole dataset was higher than 99.8%. After the two QC stages, 105 individuals and 9,730 SNPs were discarded. Moreover, after phasing, 1,292 individuals were removed to reduce the large number of sib-families present in the redundant dataset. After quality control procedures, the final dataset had 44,271 SNPs and 1,132, 514, 393, 410 and 364 individuals from Italian Holstein, Italian Brown, Italian Simmental, Marchigiana and Piedmontese bulls, respectively (Table 1).

Table 1 - Number of animals genotyped before and after QC
Total number of animal genotyped and relative number of animal discarded after QC analysis.

Breed	Total Genotyped	ED- 5% misAN	ED-REPL	ED-MEND	Cleaned
HOL	2179	40	31	5	2093
BRW	775	6	16	4	749
SIM	493	6	6	2	479
MAR	485	37	38	-	410
PIE	379	5	10	-	364

Detection of selection signatures

Sweep v1.1. software detected rEHH signals with a frequency higher than 25% overlapping the test candidate regions in both datasets, corrected or not for population structure. In the redundant dataset, we found only one significant rEHH signal (casein cluster), whereas in the non-redundant dataset we identified 5 significant rEHH signals overlapping all the test candidate regions considered (casein cluster, polled cluster, *MC1R*) (Table2).

Table 2 - Comparison of candidate region rEHH in non-redundant and redundant dataset

rEHH overlapping candidate regions in both corrected (non-redundant) and not corrected (redundant) datasets for population

Candidate gene	BTA[1]	Closest SNP (bp)	C.H.[2] range	Redundant		Non Redundant	
				C.H.[2]Frequency	rEHH - log(pval)	C.H.[2]Frequency	rEHH - log(pval)
POLLED Cluster	1	1981154	1897418-1981154	H1:0.79	0.93/0.37	H1:0.78	1.67*/1.74*
MC1R	18	13657912	13317720-14007505	H1:0.69	1.20/0.96	H1:0.69	0.77/1.67*
KIT_BOVIN	6	72821175	72504921-72821175	H1:0.54	0.30/0.30	H1:0.54	0.33/0.48
Casein Cluster	6	88427760	88350098-88452835	H1:0.47	0.94/1.39*	H1:0.46	1.48*/1.33*
Casein Cluster	6	88427760	88350098-88452835	H2:0.32	0.02/0.03	H2:0.33	0.02/0.03

In total 5,526, 5,678, 4,772, 4,388, and 4,049 core haplotypes with a frequency higher than 25% were detected on Italian Holstein, Italian Brown, Italian Simmental, Marchigiana and Piedmontese breeds, respectively. A total of 838, 866, 740, 692 and 613 core haplotypes were found significant outliers (see Materials and Methods) in the aforementioned breeds. Table 3 shows the distribution of the total and significant core haplotypes per chromosome and breed.

Table 3 - Core haplotypes distribution

Distribution of total and significant core haplotypes per chromosome and breed.

	Italian Holstein		Italian Brown		Italian Simmental		Marchigiana		Piedmontese	
BTA[1]	**H > 25%[2]**	**Sign H[3]**	**H > 25%[2]**	**Sign H[3]**	**H > 25%[2]**	**Sign H[3]**	**H > 25%[2]**	**Sign H[3]**	**H > 25%[2]**	**Sign H[3]**
1	389	66	389	68	335	49	310	51	288	46
2	312	47	314	52	267	44	257	42	238	33
3	292	48	313	62	264	46	236	34	227	34
4	268	46	279	50	245	41	237	39	229	39
5	223	32	231	29	200	30	183	27	171	22
6	291	51	307	40	265	45	252	38	233	42
7	249	37	253	46	214	39	195	33	204	27
8	268	39	270	38	235	35	223	32	209	31
9	227	37	228	41	210	37	165	22	187	30
10	219	36	234	34	192	25	196	26	168	25
11	248	39	246	34	209	30	205	33	187	25
12	162	24	176	18	148	21	148	25	135	22
13	205	32	200	19	155	18	178	35	135	19
14	175	17	189	25	168	29	143	22	129	19
15	180	31	185	29	148	23	139	13	130	19
16	176	28	192	29	160	25	149	19	127	15
17	170	27	183	29	168	31	135	22	123	16
18	133	18	153	19	119	14	108	10	82	12
19	137	22	130	23	118	16	99	18	92	15
20	184	35	172	33	145	23	124	23	124	17
21	145	23	147	30	113	18	110	21	94	13
22	140	24	145	19	115	21	112	18	104	13
23	106	13	100	16	67	10	64	9	57	8
24	129	11	140	20	124	17	97	23	101	16
25	91	12	98	11	68	10	77	12	70	6
26	125	7	114	14	93	12	96	15	90	16
27	91	12	91	15	75	11	84	12	60	10
28	88	9	96	10	70	10	66	9	55	11
29	103	15	103	13	82	10	72	9	67	12
TOT	5526	838	5678	866	4772	740	4388	692	4049	613

[1] Bos taurus Autosomes.

[2] Detected core haplotypes with a frequency in the breed higher than 25%.

[3] Significant core haplotypes at $p \leq 0.05$

Comparison to previously reported data

A number of studies have searched selection sweeps in Holstein (Qanbari et al., 2010), Brown (Qanbari et al., 2011) and Simmental (Fan et al., 2014) breeds. Since different methods are expected to identify signatures having different characteristics, the comparison with previous studies is limited to those studies using the same method and the same breed(s) analysed in our study. Only one study analysed (German) Holstein-Friesian cattle with rEHH (Qanbari et al., 2010). Qanbari *et al.* (2010) reported candidate genes and gene clusters under recent positive selection that we compared with our rEHH in dairy breeds dataset (Table 4).

Table 4 - Core haplotypes in candidate genes following Qanbari et al. 2011

			HOLSTEIN			BROWN			SIMMENTAL		
Candidate gene	**BTA[1]**	**Closest SNP (bp)**	**C.H.[2] range**	**C.H.[2] Frequency**	**rEHH -log(p)**	**C.H.[2] range**	**C.H.[2] Frequency**	**rEHH -log(p)**	**C.H.[2] range**	**C.H.[2] Frequency**	**rEHH -log(p)**
DGAT1	14	444,963	236,653 - 443,936	H1: 0.57	-/0.11	443,936 - 763,332	H1: 0.42	0.13/0.007	-	-	-
Casein Cluster	6	88,391,612	88,350,098 - 88,452,835	H1: 0.46	1.48*/1.33*	88,326,012 - 88,452,835	H2: 0.68	0.80/1.09	88,350,098 - 88,452,835	H1: 0.44	0.37/0.22
			88,350,098 - 88,452,835	H2: 0.33	0.18/0.30	-	-	-	-	-	-
			-	-	-	-	-	-	88,350,098 - 88,452,835	H3: 0.34	0.22/0.45
GH	19	49,652,377	-	-	-	-	-	-	-	-	-
GHR	20	33,908,597	-	-	-	-	-	-	-	-	-
SST	1	81,376,956	81,283,585 - 81,376,961	H1: 0.36	2.00**/1.89**	-	-	-	81,318,451 - 81,376,961	H1: 0.42	1.26/0.53
			81,283,585 - 81,376,961	H2: 0.29	0.063/0.084	-	-	-	81,318,451 - 81,376,961	H2: 0.42	0.06/0.27
IGF-1	5	71,169,823	-	-	-	-	-	-	-	-	-
ABCG2	6	37,374,911	-	-	-	37,317,020 - 38,256,889	H1: 0.44	0.31/0.27	-	-	-
			-	-	-	37,317,020 - 38,256,889	H1:0.40	0.29/0.31	-	-	-
Leptin	4	95,715,500	-	-	-	-	-	-	-	-	-
LPR	3	85,569,203	85,497,108 - 85,594,551	H1: 0.47	0.91/0.72	85,497,108 - 85,794,693	H1: 0.68	0.80/0.63	85,497,108 - 85,794,693	H1: 0.63	0.02/0.05
			85,497,108 - 85,594,551	H2: 0.41	0.27/0.25	-	-	-	-	-	-
PIT-1	1	35,756,434	-	-	-	-	-	-	-	-	-

[1] Bos Taurus Chromosome
[2] Core haplotype

The number of core haplotypes found in our study in the candidate regions was lower previously reported. When considering Holsteins, the two most significant candidate regions in both studies match (casein cluster and the somatostatin *SST* gene), although it is impossible to determine if the haplotype under selection is the same, as this information was not provided in Qanbari et al. (2010). However, other genes considered significant in Qanbari et al. (with *p*-values ranging from 0.04 to 0.10) were not significant in our study. When using the same loose significant threshold (*p*-values ≤_0.10), "significant" signals were found also in Italian Brown for the Casein Cluster (-log$_{10}$ *p*-value = 1.09) and in Italian Simmental for the *SST* gene (-log$_{10}$ *p*-value = 1.26).

Signatures shared by dairy and by beef breeds

Significant core haplotypes were aligned across breeds to identify those shared among dairy or beef production types. Since breeds can be considered independent sets of observations, shared signatures are more likely to represent real effects rather than false positives. A total of 123 and 142 significant core haplotypes were shared between at least two dairy or beef breeds, respectively (File S1). Only 82 and 87 of those significant core haplotypes contained genes considered positional candidates under positive selection. In total, 2.2% and 1.7% of the genome was under selection for dairy and beef production types, respectively. The rEHH average lengths were 216,932 and 190,994 bp for dairy and beef, respectively.

Gene set annotation and network/pathway analysis

A total of 244 and 232 annotated genes fell within the selected regions for dairy and beef production types, respectively (Table 5).

Table 5 - Statistics on shared significant core haplotypes in dairy and beef breeds

Statistics on shared significant core haplotypes in dairy and beef breeds.

	Dairy Cattle				Beef Cattle			
BTA	**#Sel.Sign.**	**#Genes[1]**	**Sum Sel.Sign size (bp)**	**Mean Sel.Sign size (bp)**	**#Sel.Sign.**	**#Genes[1]**	**Sum Sel.Sign size (bp)**	**Mean Sel.Sign size (bp)**
1	7	11	1521608	138328	7	13	2263213	174093
2	5	9	2216685	246298	8	11	2658549	241686
3	2	5	1619336	323867	7	21	3755847	178850
4	7	21	5956755	283655	6	11	1761288	160117
5	4	17	4506119	265066	5	11	1251127	113739
6	2	4	1467738	366934	5	12	5149692	429141
7	6	30	4842969	161432	3	28	11889191	424614
8	0	0	0	0	4	12	3512215	292685
9	4	7	1554863	222123	2	6	1106768	184461
10	4	17	8839762	519986	2	3	573138	191046
11	6	8	1841802	230225	1	1	100439	100439
12	2	8	4244133	530517	1	1	536461	536461
13	3	8	1933026	241628	2	8	985376	123172
14	0	0	0	0	3	9	692397	76933
15	5	9	1415405	157267	2	2	189094	94547
16	3	6	1302506	217084	4	6	905719	150953
17	3	4	471046	117762	2	9	1551010	172334
18	0	0	0	0	3	23	2961718	128770
19	3	23	5744143	249745	2	2	116938	58469
20	1	2	148886	74443	2	4	627971	156993
21	3	7	1930206	275744	1	5	1150760	230152
22	3	5	620674	124135	3	10	2411751	241175
23	0	0	0	0	1	1	68421	68421
24	2	18	9166004	509222	2	4	803770	200942
25	2	11	2016602	183327	1	3	329199	109733
26	2	4	466134	116534	4	7	776080	110869
27	1	1	631792	631792	2	3	1004717	334906
28	0	0	0	0	1	1	93464	93464
29	2	9	935209	103912	1	5	798300	159660
TOT	82	244	65393403	216932	87	232	50024613	190994

[1] Genes identified

Genes in the regions shared by all three dairy breeds (8 genes) or all three beef breeds (11 genes) for a more detailed analysis were selected (Fig 1).

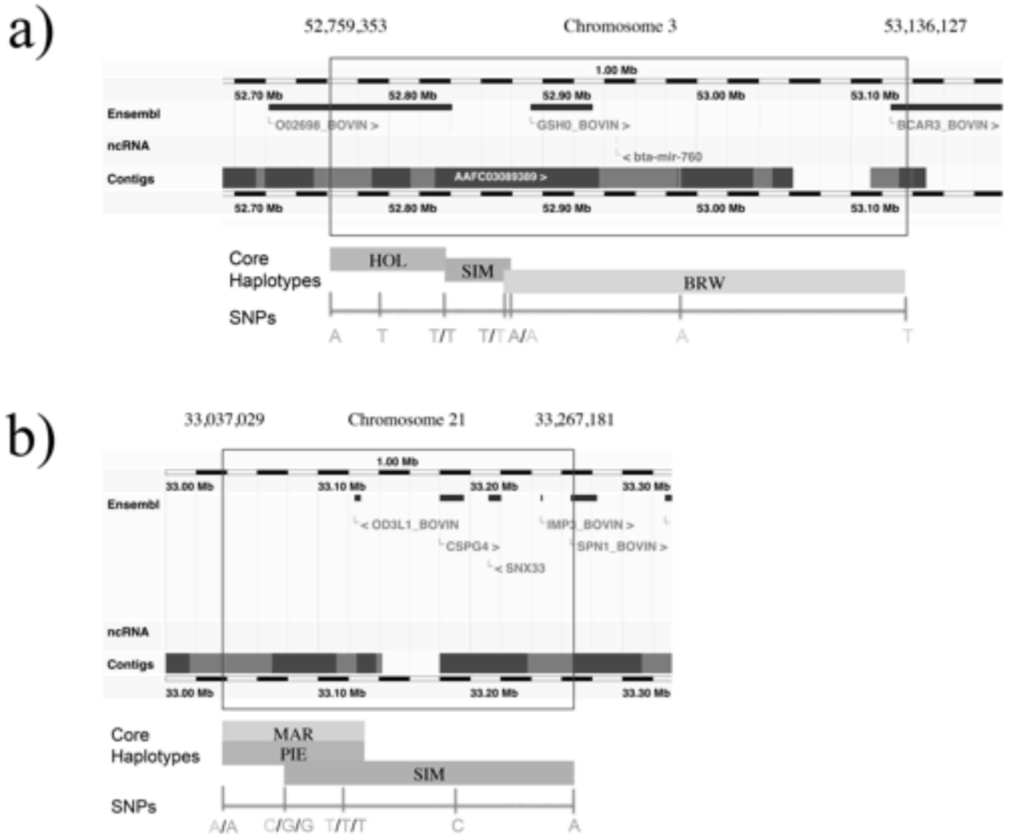

Figure 1 - Genomic location of the selection signatures shared among studied breeds

a) Genes in ensemble tracks are displayed as a red box; core haplotypes and SNPs are coloured in orange (Marchigiana; MAR), in violet (Piedmontese; PIE) and pink (Simmental; SIM). b) Genes in ensemble tracks are displayed as a red box; core haplotypes and SNPs are coloured in blue (Holstein; HOL), in green (Italian Brown; BRW) and pink (Simmental; SIM).

All the genes identified were submitted to downstream pathway/network analyses (Table 5; Figure 1; File S1). Interestingly, genes in the regions shared by all three dairy or beef breeds were manifested in the most likely networks for the two production aptitudes. The IPA analysis detected a total of 12 and 15 networks in dairy and beef breeds, respectively.

In dairy breeds, network scores ranged from 46 to 2 (File S2). Eight networks obtained scores higher than 20 and included more than 14 molecules each. They were associated to: i) gene expression, RNA damage and repair, protein synthesis and post-translational modification; ii) cell assembly, organization, function and maintenance; iii) cell cycle proliferation and cancer; iv) carbohydrate metabolism; v) gastrointestinal and neurological disease, developmental and endocrine system disorders. A total of 23 molecules were associated with the highest ranked network (score = 46). In this, the immunoglobulins, mitogenactivated protein kinases and NFkB complexes form a central hub, linking genes involved in cell-to-cell signaling and interaction, hematological system development and function and immune cell trafficking functions (Figure 2).

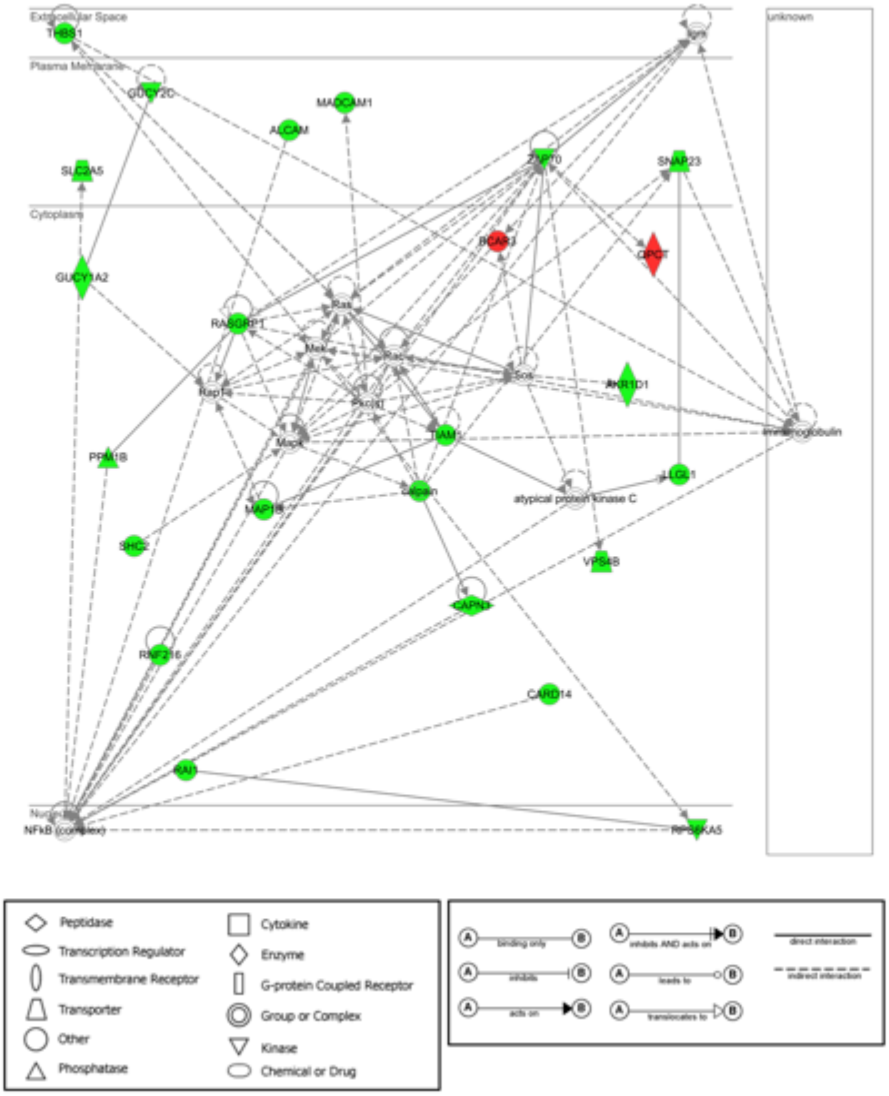

Figure 2 - The most likely IPA network in dairy cattle breeds

The functions associated to the network are cell-to-cell signalling and interaction, hematological system development and function, immune cell trafficking. Nodes in red correspond to genes identified in core haplotypes overlapping in all the three breeds of each production type, whereas those in green depict overlapping core haplotypes in at least two of those breeds.

Breast cancer anti-estrogen resistance gene 3 (*BARC3*) and *pituitary glutaminyl cyclase gene* (*QPCT*) genes were common to all three dairy breeds and are directly connected with mammary gland metabolisms (Ezura et al., 2004; G. Sun et al., 2012). The *solute carrier family 2, member 5* (*SLC2A5*) gene facilitates glucose/fructose transport (Zhao et al., 1998), and the *zeta-chain (TCR)* *associated protein kinase 70kDa* (*ZAP70*) gene plays a critical role in T-cell signaling (Bonnefont et al., 2011). Calpain is another important complex that, together with *calpain-3* (*CAPN3),* mediates epithelial-cell death during mammary gland involution (Wilde et al., 1997). Furthermore, RAS guanyl nucleotide-releasing protein (RASGRP1) activates the Erk/MAP kinase cascade, regulates T-cells and B-cells development, homeostasis and differentiation and is involved in the regulation of breast cancer cell (Madani et al., 2004; Yasuda et al., 2007; Wickramasinghe et al., 2009). Genes involved in the other networks are listed in File S2.

The score of the beef cattle breeds network ranged between 46 and 2 and the number of molecules present in each network ranged from 23 to 1 (File S3). As in dairy breeds, eight networks obtained a score higher than 20 and 13 or more molecules each. These networks were associated to: i) carbohydrate, lipid and nucleic acids metabolism, energy production and small molecule biochemistry; ii) cardiovascular system development and function; iii) cellular and tissue development, cell assembly and organization, function and maintenance and cell and organ morphology; iv) cell cycle; cell-to-cell signaling and interaction ; v) DNA replication, recombination, and repair and RNA post-transcriptional modification; vi) dermatological diseases and conditions, infectious disease, organismal injury and abnormalities. In the network with the highest score (score=46), *chondroitin sulfate proteoglycan 4* (*CSPG4*) and *snurportin-1* (*SNUPN*) genes were shared among all beef cattle breeds investigated. *CSPG4* gene is related to meat tenderness, while *SNUPN* is an imprinted gene important in the embryo development and involved in human muscle atrophy (Narayanan et al., 2002) (Figure 3). All the genes included in this network were involved in carbohydrate metabolism, small molecules biochemistry and cell cycle. The other networks were related to important signaling and metabolic process essential for animal growth (File S3).

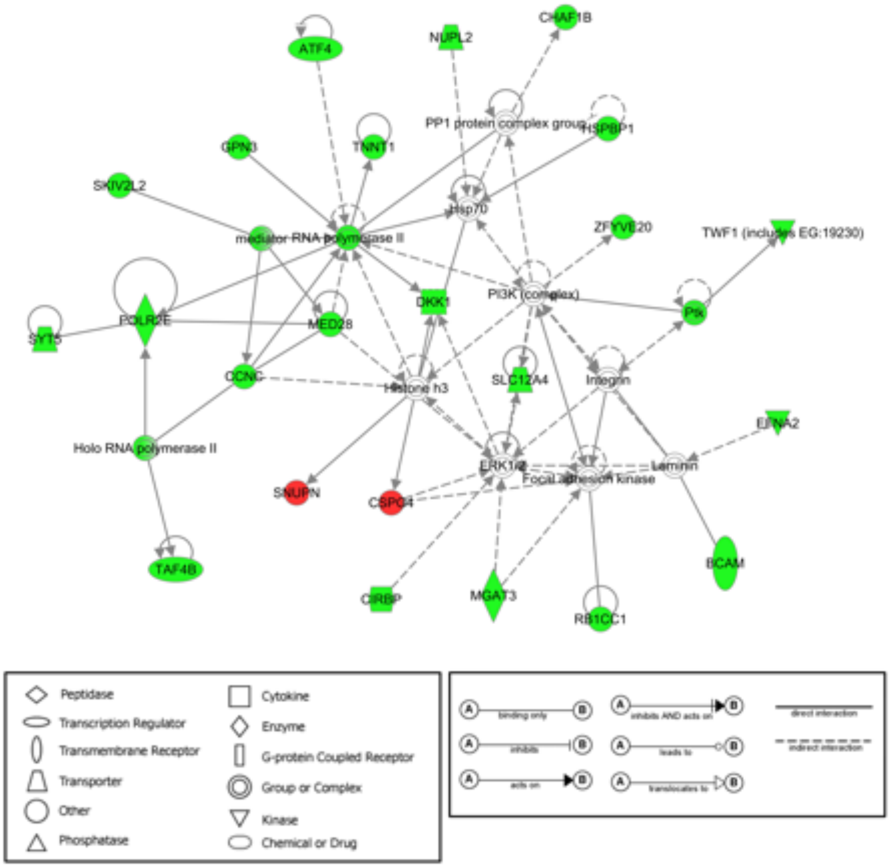

Figure 3 - The most likely IPA network in beef cattle breeds

The functions associated to the network are carbohydrate metabolism, small molecule biochemistry and cell cycle. Nodes in red correspond to genes identified in core haplotypes overlapping in all the three breeds of each production type, whereas those in green depict overlapping core haplotypes in at least two of those breeds.

A total of 6 and 9 statistically significant Canonical Pathways (FDR ≤ 0.05; $-\log_{10}$(FDR) ≥ 1.3) were identified using IPA for dairy and beef breeds, respectively (Figure 4).

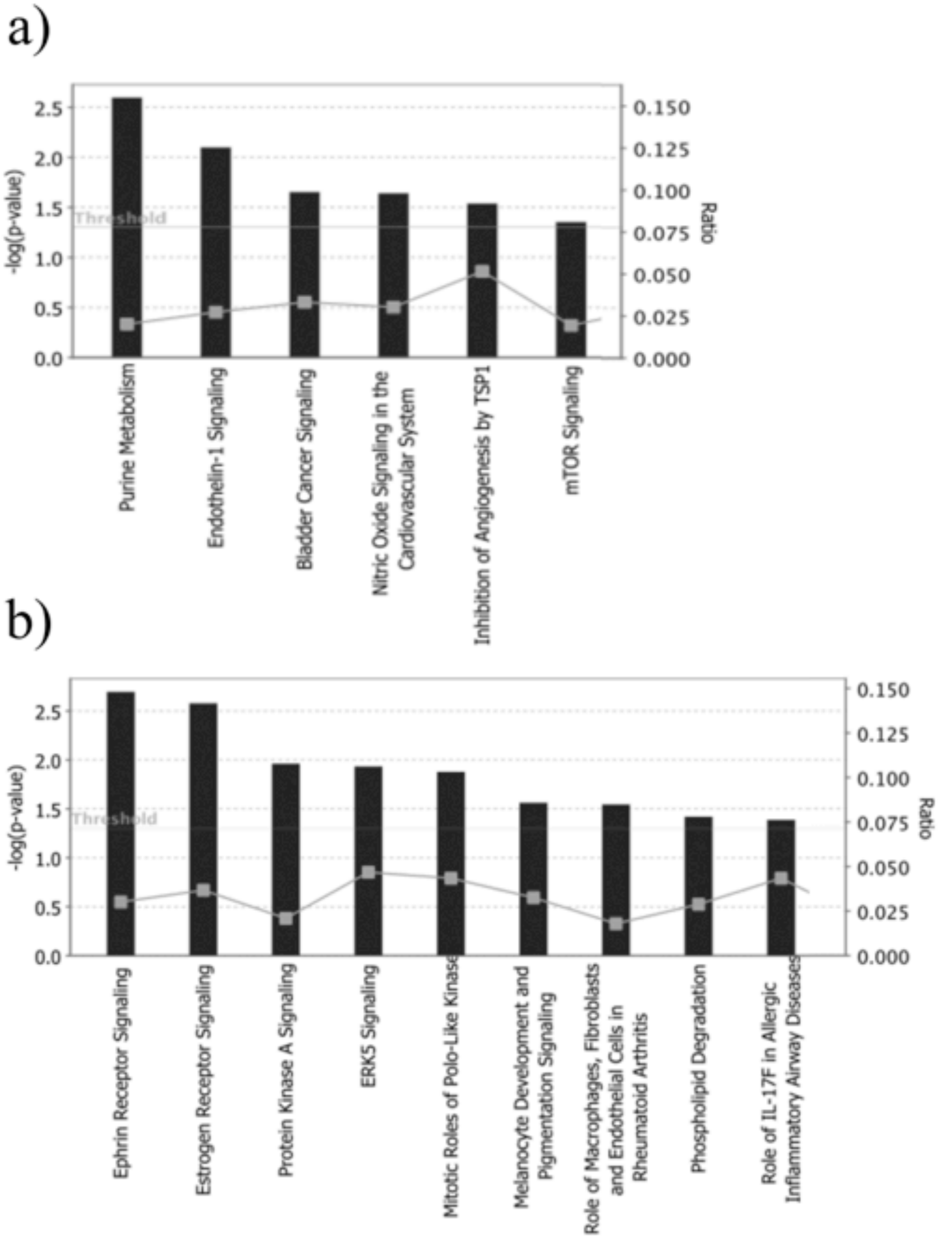

Figure 4 - Bar plot of statistically significant Canonical Pathways

P-values were corrected for multiple testing using Benjamini-Hochberg method and are presented in the graph as -log(*p*-value). The bar represents the percentage of genes in a given pathway that meet cut-off criteria within the total number of molecules that belong to the function. a) Bar plot of statistically significant Canonical Pathways in dairy cattle breeds b) Bar plot of statistically significant Canonical Pathways in beef cattle breeds.

The most significant Canonical Pathway in dairy was for purine metabolism (-$\log_{10}$(FDR) = 2.6), supporting the highly synthetic processes in the mammary epithelium (Figure 5).

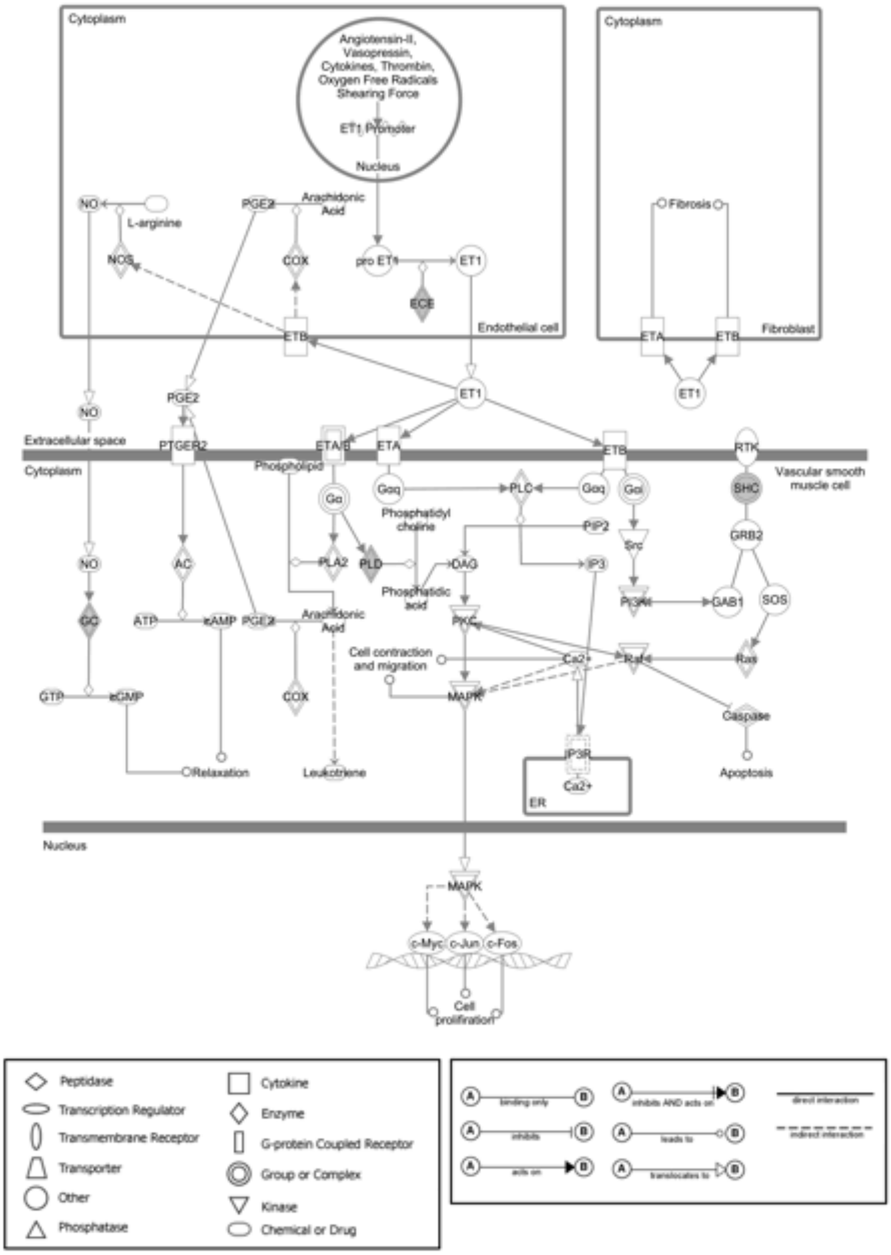

Figure 5 - Genes detected under recent positive selection in dairy cattle involved in purine metabolism canonical pathway

Nodes in red correspond to genes identified in core haplotypes overlapping in all the three breeds of each production type, whereas those in green depict overlapping core haplotypes in at least two of those breeds.

In beef production, the *ephrin receptor* signal (-log10(FDR) = 2.7) was the most significant Canonical Pathway (Figure 6). Among other functions, *ephrin* is known to promote muscle progenitor cell migration before mitotic activation (Li and Johnson, 2013). All the other Canonical Pathways are reported in File S4.

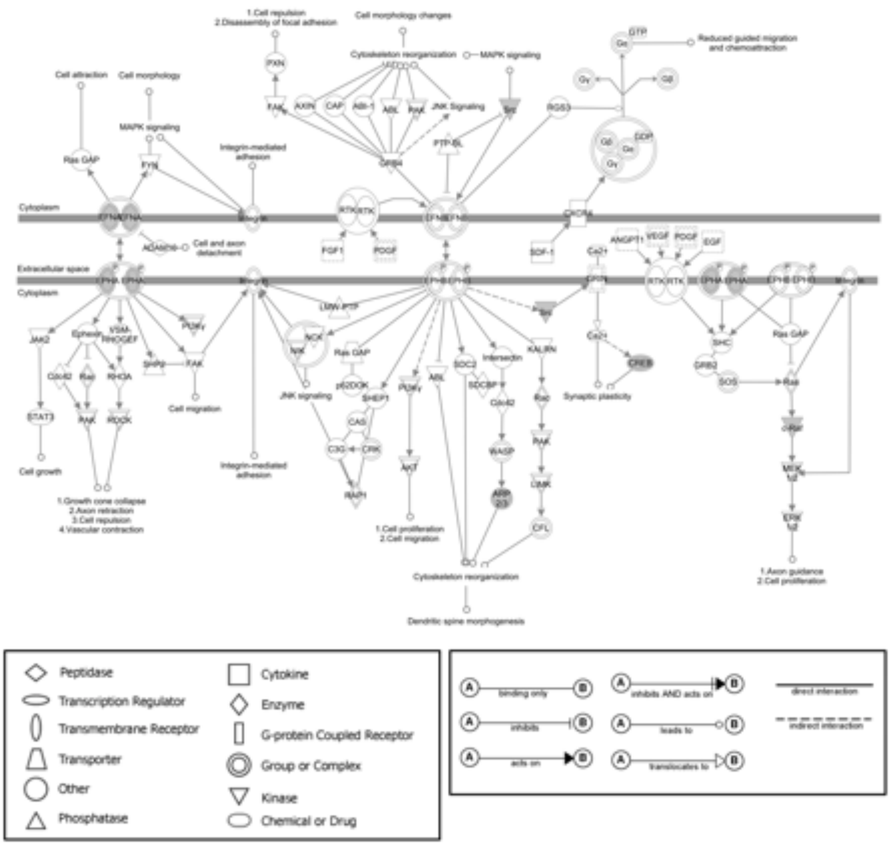

Figure 6 - Genes detected under recent positive selection in beef cattle involved in Ephrin metabolism canonical pathway

Nodes in red correspond to genes identified in core haplotypes overlapping in all the three breeds of each production type, whereas those in green depict overlapping core haplotypes in at least two of those breeds.

Discussion

In this work, more than 4,000 genotyped bulls of five dairy, beef and dual-purpose Italian breeds were screened, searching for selection signatures in dairy and beef production types. A strict data quality control was applied to reduce possible sources of bias, such as technical genotyping problems and population structure, that could have large impact on rEHH results. In fact, since the impact of population structure on rEHH values is still unexplored, we replicated part of our analyses with and without close relatives, in an attempt to study the effect of population structure on this selection signature method. In principle, population stratification should lead to an over-representation of certain haplotypes, simply due to the presence of tight pedigree links (*e.g.* sires passing half of their genetic material to their sons) rather than actual selection processes. For this reason, for the full analyses, all sire-sons couples were removed after haplotype phasing (retaining only the younger animal), and half-sib families were restricted to a maximum of five randomly chosen individuals, in an attempt to reduce over-representation. This assessment was restricted to four regions known to be under selection in Italian Holstein, as the casein cluster, the polled gene cluster, *MC1R* and *KIT*. Italian Holstein was selected for two reasons: i) it is a highly structured breed according to our data (about half of individuals removed were close relatives); ii) allowed us to compare our results with a previous study. Although the analyses conducted on both redundant and non-redundant datasets were able to identify rEHH signals in these regions, the non-redundant dataset produced five significant rEHH signals, rather than just one in the redundant dataset (Table 2). This result highlights the confounding effect of the presence of close relatives in the dataset and, consequently, the improved ability to detected significant selection signature while correcting for population structure. Interestingly, our results only partially overlap those found by Quanbari *et al.* (2010). These authors detected significant signals at the casein cluster and somatostatin *SST,* as we did, but also at a higher number of candidate regions. These inconsistencies may be the effect of different sires included in the analyses, of the different dataset size and to the close relative trimming procedure we adopted to decrease the effect of population structure and consequent bias (Table 4). However, poor correlation across investigations is often observed also in the deeply investigated human species. This is due to the use of different within and between breed statistics that identify selection signatures having different characteristics (ancient/recent, segregating/fixated, under directional/balancing selection), to the likely high rate of false positives/negatives results, and also to the different consideration of population structure and background selection (Enard et al., 2014).

The number of total and significant core haplotypes identified by the Sweep software was the highest in Italian Brown and the lowest in Piedmontese breed. Since EHH is heavily based on population LD, the average LD level over distance was calculated in each breed (Figure 7).

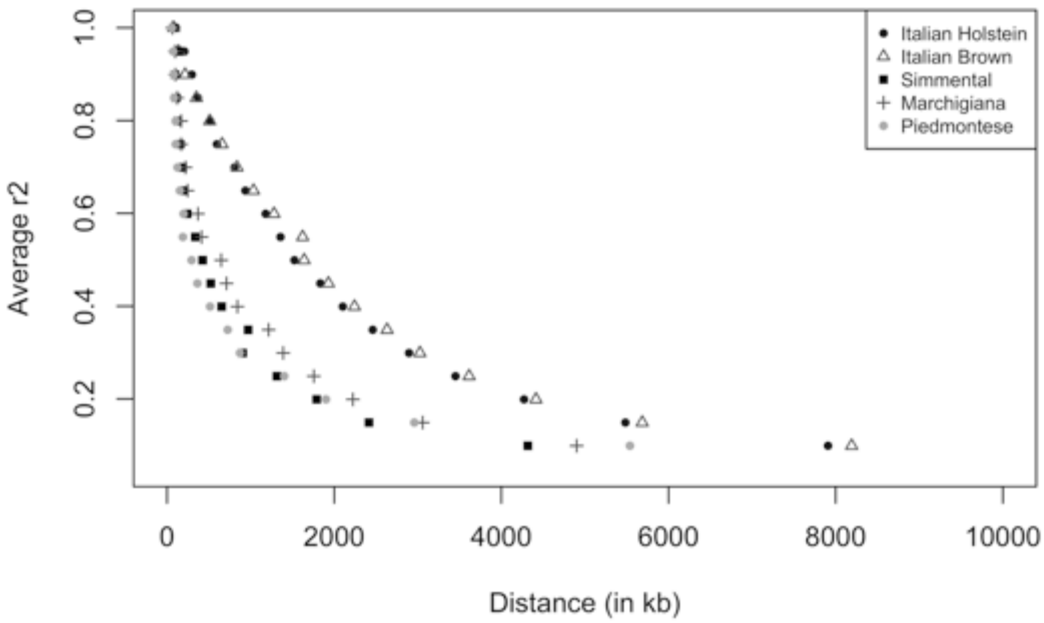

Figure 7 - Multi-breed average linkage disequilibrium against physical distance (in kb)
Marchigiana, Piedmontese and Italian Simmental breeds present a lower persistence of LD al larger distance than Italian Holstein and Italian Brown breeds.

This analysis highlighted a general positive correlation between the level of LD over distance and the number of total and significant core haplotypes found. However, considering that rEHH is a relative measure, it is more likely that the higher number of significant core haplotypes identified in dairy breeds was due mostly to a higher selective pressure in dairy compared to beef breeds.

Significance tests used in detecting selection signatures should measure the probability of a statistic to be an outlier compared to its expected distribution under a neutral model. However, no reliable neutral model has so far been developed in cattle because of the complexity of the demographic history of this species (Kemper et al., 2014). As a consequence, empiric rather than model based significance test are generally used in the detection of selection signatures. Accordingly, we considered as outliers values as those simply falling in the 5% plus variant tail of the rEHH distribution. We kept the within breed significance threshold loose, non correcting it for multiple testing, but considered only signals either exceeding this threshold by at least two orders of magnitude or shared by two or more independent breeds.

Parallel comparison of results from independent analyses of different breeds allowed the achievement of a double objective: i) to identify putative regions under (recent) selection in breeds belonging to two production types, which was the main objective of this study, and; ii) reducing false positives, since multi-breed analyses served as internal control. Since the rEHH method does not consider phenotypic information, a significant signal might arise because: i) the core haplotype is actually under a selective pressure; ii) the result is a false positive, *i.e.* caused by chance, population structure or any other driving force. However, even considering an unrealistic scenario with no false positives, a proportion of the total amount of signals would actually be selection signatures due to a selection pressure not on dairy or beef traits. This because dairy and beef breeds investigated are only a few and share a number of traits not directly related to dairy/beef production. Even considering this limitation, to our knowledge, this is the first dairy and beef multi-breed study applying this strategy to reduce false positives, at the cost of possible loss of information due to high false negative rates. Significant signal shared by dairy and by beef breeds were used for downstream network analyses on positional candidate genes, to search for a biological justification to what was obtained studying the genomic information. Only the top network for dairy and the top network for beef production are discussed in detail.

Dairy breeds

The most significant network detected in dairy breeds is associated with Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, and Immune Cell Trafficking functions (Figure 2). Immunoglobulin and NF-kB act as pivotal genes within this network. Interestingly, *BARC3* and *QPCT* genes were shared among all the three dairy breeds. Both genes have not yet been studied in cattle although, in human studies, these genes are well known to be linked to the mammary gland metobolism and the calcium regulation. The former is involved in integrin-mediated cell adhesions and signaling, which is required for mammary gland development and functions (G. Sun et al., 2012). The latter is associated with low radial body mineral density (BMD) in adult women (Ezura et al., 2004). Another important player on this network is *SLC2A5* gene, also known as *GLUT5*. *SLC2A5* gene acts as fructose transporter in the intestine and has a significant role in energy balance of dairy cows (Burant et al., 1992). Finding this gene in a dairy cattle specific network seems surprising, since there should be little need for transporting glucose and/or fructose in the ruminant intestinal tract, as simple carbohydrates are degraded into volatile fatty acids (VFA) in the rumen (Iqbal et al., 2009). However, it is often the case that large amount of starch bypass the rumen in cows fed with diets rich in cereal grains (Zhao et al., 1998). This bypassed starch needs to be digested in the small intestine to avoid high levels of glucose concentration in the large intestine.

We detected *calpain complex* and *calpain-3* (*CAPN3*) genes under positive recent selection in the dairy group, as reported also by Utsunomiya *et al.* (2013). Altough Calpain is known to be involved in postmortem meat tendernization, it is also related to dairy metabolism, as the muscle breakdown promoted by the *calpain* gene, provides an energy source for milk production especially at the beginning of lactation (Kuhla et al., 2011). In addition, as reported by Wilde *et al.* (1997), *calpain-calpastatin* system is related to programmed cell death of alveolar secretory epithelial cells during lactation. *Zap70* gene is a cytoplasmic protein tyrosine kinase. It is related to the immune system, as a component of the T-cell receptor, playing a central role in T-cell responses (Wang et al., 2010). Bonnefont *et al.* (2011) found that *Zap70* gene was up regulated (Fold-Change (FC) = 8.2; FDR = 2.77E-12) on milk somatic cells of sheep infected by *S. aureus* and *S. epidermidis*, showing an association with mastitis resistance.

Purine metabolism was the most significant canonical pathway in dairy breeds. In a study conducted in human mammary epthelium, Maningat *et al.* (2008), conducted a gene expression analysis on breast milk fat globules, identifying the Purine metabolism as the most significant pathway. Synthesis and breakdown of purine is essential in the metabolism of the tissues of many organisms, particularly of the mammary gland during lactation.

Another interesting canonical pathway was endothelin signaling (Figure 5). endothelin functions as vasocontrictor and is secreted by endothelial cells

(Nussdorfer et al., 1999). Acosta *et al.* (1999) reported that in cattle endothelins are involved in the follicular production of prostaglandins and the regulation of steroidogenesis in the mature follicle. In a recent study, Puglisi *et al.* (2013) confirmed the implication of endothelin, in particular *EDNRA* and *endothelin convertin enzyme 1*, in reproductive disorder in bovine. They classified the *EDNRA* as potential biomarker for fertility in cow.

Beef breeds

In beef cattle breeds, the *focal adhesion kinase* (*FAK*) together with *integrin* and *ERK1/2* genes play a central role as hubs in the most significant network (Figure 3). They are involved in cellular adhesion and cell motility processes, particularly during skeletal myogenesis (Nguyen et al., 2014). Moreover, the overexpression of *FAK* can determine a reduced apoptosis and an increase risk of metastatic cancers (Mehlen and Puisieux, 2006). These hubs are not directly involved in our study, but share many interactors that are located in region found to be under selction, such as *CSPG4, RB1CC1, MGAT3 , CIRPB*. *CSPG4* gene belongs to chondroitin sulfate proteoglycans (CSPGs) family. CSPGs are proteoglycans consisting of a protein core and a chondroitin sulfate side chain. They are known to be structural components of a variety of tissues, including muscle, and play key roles in neural development and glial scar formation. They are involved in cellular processes, such as cell adhesion, cell growth, receptor binding, cell migration, and interactions with other extracellular matrix constituents.

Many studies report the role of proteoglycans in meat texture of several muscles from the bovine species (Nishimura et al., 1996). Dubost *et al.* (2013) highlight a direct involvement of proteoglycans with respect to cooked meat juiciness. *RB1CC1* gene plays a crucial role in muscular differentiation and its activation is a essential for myogenic differentiation (Watanabe et al., 2005, p. 1). *Monoacylglycerol acyltransferase* (*MGAT3*) gene cause the synthesis of diacylglycerol (DAG) using 2-monoacylglycerol and fatty acyl coenzyme A. This enzymatic reaction is fundamental for the absorption of dietary fat in the small intestine. Sun et al. (2012, p. 3) reported in a study on five Chinese cattle breeds that *MGAT3* gene is associated with growth traits. *CIRPB* gene may be part of a compensatory mechanism in muscles undergoing atrophy. It preserves muscle tissue mass during cold-shock responses, aging and disuse (Dupont-Versteegden et al., 2008). *SNUPN* is an imprinted gene and is expressed monoallelically depending on its parental origin. *SNUPN* play important roles in embryo survival and postnatal growth regulation (Smith et al., 2012; Wang et al., 2012). *Ephrin receptor* signaling was the top canonical pathway produce by IPA and has also an interesting biological interpretation for meat production (Figure 6). Indeed, this pathaway is important for the muscle tissue growth and regeneration, participating in correct positioning and formation of the neural muscular junction (Li and Johnson, 2013).

Conclusions

In this study we analysed the selection signatures at genome-wide level in 5 Italian cattle breeds. Then, we used a multi-breed approach to identify the genomic regions shared among cattle breeds selected for dairy or beef purpose. This approach increased the power to pinpoint regions of the genome that play important roles in economically relevant traits in both dairy and beef cattle. Moreover, network and pathways analyses were used to display a detailed functional characterization of genes mapping in the regions detected under recent positive selection.

Particularly, in dairy cattle, genes under directional selection are related to feeding adaptation (increasing level of starch in the diet), mammary gland metabolism and resistence to mastitis. The biological features under selection in beef cattle breeds are involved in animal growth, meat texture and juiceness. These results have a logical biological explanation. However, the frequent approach of interpreting selection signatures on the basis of the function of genes located nearby the peak signal of the statistic used is presently challenged. Novel information in humans suggests that most selected variation is not located within genes and coding regions, but in regulatory sites identified by the ENCODE project (Enard et al., 2014). These may control the expression of entire genomic regions or of genes located at a relavant distance from the selected site, making biological interpretation more complex.

In any case, further studies using denser SNP chips or whole-genome sequencing, producing information not subjected to ascertainment bias (Qanbari et al., 2014), may increase the resolution of our analysis and together with new knowledge arising on the control of gene expression validate or correct our results.

Acknowledgements

ANAFI (Italian Association of Holstein Friesian Breeders), ANAPRI (Italian Association Simmental Breeders), ANABORAPI (National Association of Piedmontese Breeders), ANABIC (Association of Italian Beef Cattle Breeders) and ANARB (Italian Brown Cattle Breeders Association) are acknowledged for providing the biological samples and the necessary information for this work. The Authors are also grateful to "SelMol" project (MiPAAF, Ministero delle Politiche Agricole, Alimentari e Fortestali) and by the "ProZoo" project (Regione Lombardia–DG Agricoltura ,Fondazione Cariplo, Fondazione Banca Popolare di Lodi) for funding.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Acosta, T.J., Berisha, B., Ozawa, T., Sato, K., Schams, D., Miyamoto, A., 1999. Evidence for a Local Endothelin-Angiotensin-Atrial Natriuretic Peptide Systemin Bovine Mature Follicles In Vitro: Effects on Steroid Hormones and Prostaglandin Secretion. Biol. Reprod. 61, 1419–1425. doi:10.1095/biolreprod61.6.1419

Ajmone-Marsan, P., Garcia, J.F., Lenstra, J.A., 2010. On the origin of cattle: how aurochs became cattle and colonized the world. Evol. Anthropol. Issues News Rev. 19, 148–157.

Bonnefont, C., Toufeer, M., Caubet, C., Foulon, E., Tasca, C., Aurel, M.-R., Bergonier, D., Boullier, S., Robert-Granié, C., Foucras, G., 2011. Transcriptomic analysis of milk somatic cells in mastitis resistant and susceptible sheep upon challenge with Staphylococcus epidermidis and Staphylococcus aureus. BMC Genomics 12, 208.

Burant, C.F., Takeda, J., Brot-Laroche, E., Bell, G.I., Davidson, N.O., 1992. Fructose transporter in human spermatozoa and small intestine is GLUT5. J. Biol. Chem. 267, 14523–14526.

Dubost, A., Micol, D., Picard, B., Lethias, C., Andueza, D., Bauchart, D., Listrat, A., 2013. Structural and biochemical characteristics of bovine intramuscular connective tissue and beef quality. Meat Sci. 95, 555–561. doi:10.1016/j.meatsci.2013.05.040

Dupont-Versteegden, E.E., Nagarajan, R., Beggs, M.L., Bearden, E.D., Simpson, P.M., Peterson, C.A., 2008. Identification of cold-shock protein RBM3 as a possible regulator of skeletal muscle size through expression profiling. AJP Regul. Integr. Comp. Physiol. 295, R1263–R1273. doi:10.1152/ajpregu.90455.2008

Enard, D., Messer, P.W., Petrov, D.A., 2014. Genome-wide signals of positive selection in human evolution. Genome Res. gr.164822.113. doi:10.1101/gr.164822.113

Ezura, Y., Kajita, M., Ishida, R., Yoshida, S., Yoshida, H., Suzuki, T., Hosoi, T., Inoue, S., Shiraki, M., Orimo, H., Emi, M., 2004. Association of multiple nucleotide variations in the pituitary glutaminyl cyclase gene (QPCT) with low radial BMD in adult women. J Bone Min. Res 19, 1296–301.

Fan, H., Wu, Y., Qi, X., Zhang, J., Li, J., Gao, X., Zhang, L., Li, J., Gao, H., 2014. Genome-wide detection of selective signatures in Simmental cattle. J. Appl. Genet. 1–9. doi:10.1007/s13353-014-0200-6

Hayes, B.J., Chamberlain, A.J., Maceachern, S., Savin, K., McPartlan, H., MacLeod, I., Sethuraman, L., Goddard, M.E., 2009. A genome map of divergent artificial selection between *Bos taurus* dairy cattle and *Bos taurus* beef cattle. Anim. Genet. 40, 176–184. doi:10.1111/j.1365-2052.2008.01815.x

Hayes, B.J., Lien, S., Nilsen, H., Olsen, H.G., Berg, P., Maceachern, S., Potter, S., Meuwissen, T.H.E., 2008. The origin of selection signatures on bovine chromosome 6. Anim. Genet. 39, 105–111. doi:10.1111/j.1365-2052.2007.01683.x

Iqbal, S., Zebeli, Q., Mazzolari, A., Bertoni, G., Dunn, S.M., Yang, W.Z., Ametaj, B.N., 2009. Feeding barley grain steeped in lactic acid modulates rumen fermentation patterns and increases milk fat content in dairy cows. J. Dairy Sci. 92, 6023–6032. doi:10.3168/jds.2009-2380

Kemper, K.E., Saxton, S.J., Bolormaa, S., Hayes, B.J., Goddard, M.E., 2014. Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics 15, 246. doi:10.1186/1471-2164-15-246

Kimura, M., 1984. The neutral theory of molecular evolution.

Kuhla, B., Nürnberg, G., Albrecht, D., Görs, S., Hammon, H.M., Metges, C.C., 2011. Involvement of Skeletal Muscle Protein, Glycogen, and Fat Metabolism in the Adaptation on Early Lactation of Dairy Cows. J. Proteome Res. 10, 4252–4262. doi:10.1021/pr200425h

Lenstra, J.A., Groeneveld, L.F., Eding, H., Kantanen, J., Williams, J.L., Taberlet, P., Nicolazzi, E.L., Solkner, J., Simianer, H., Ciani, E., Garcia, J.F., Bruford, M.W., Ajmone-Marsan, P., Weigend, S., 2012. Molecular tools and analytical approaches for the characterization of farm animal genetic diversity. Anim Genet 43, 483–502.

Li, J., Johnson, S.E., 2013. Ephrin-A5 promotes bovine muscle progenitor cell migration before mitotic activation. J. Anim. Sci. 91, 1086–1093. doi:10.2527/jas.2012-5728

Ma, Y.-L., Zhang, Q., Ding, X.-D., 2012. [Detecting selection signatures on X chromosome in pig through high density SNPs]. Yi Chuan Hered. Zhongguo Yi Chuan Xue Hui Bian Ji 34, 1251–1260.

Madani, S., Hichami, A., Charkaoui-Malki, M., Khan, N.A., 2004. Diacylglycerols Containing Omega 3 and Omega 6 Fatty Acids Bind to RasGRP and Modulate MAP Kinase Activation. J. Biol. Chem. 279, 1176–1183. doi:10.1074/jbc.M306252200

Maningat, P.D., Sen, P., Rijnkels, M., Sunehag, A.L., Hadsell, D.L., Bray, M., Haymond, M.W., 2008. Gene expression in the human mammary epithelium during lactation: the milk fat globule transcriptome. Physiol. Genomics 37, 12–22. doi:10.1152/physiolgenomics.90341.2008

Mehlen, P., Puisieux, A., 2006. Metastasis: a question of life or death. Nat. Rev. Cancer 6, 449–458. doi:10.1038/nrc1886

Narayanan, U., Ospina, J.K., Frey, M.R., Hebert, M.D., Matera, A.G., 2002. SMN, the Spinal Muscular Atrophy Protein, Forms a Pre-Import Snrnp Complex with Snurportin1 and Importin ? Hum. Mol. Genet. 11, 1785–1795.

Nguyen, N., Yi, J.-S., Park, H., Lee, J.-S., Ko, Y.-G., 2014. Mitsugumin 53 (MG53) Ligase Ubiquitinates Focal Adhesion Kinase during Skeletal Myogenesis. J. Biol. Chem. 289, 3209–3216. doi:10.1074/jbc.M113.525154

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., Clark, A.G., 2007. Recent and ongoing selection in the human genome. Nat. Rev. Genet. 8, 857–868. doi:10.1038/nrg2187

Nishimura, T., Hattori, A., Takahashi, K., 1996. Relationship between degradation of proteoglycans and weakening of the intramuscular connective tissue during post-mortem ageing of beef. Meat Sci. 42, 251–260.

Nussdorfer, G.G., Rossi, G.P., Malendowicz, L.K., Mazzocchi, G., 1999. Autocrine-Paracrine Endothelin System in the Physiology and Pathology of Steroid-Secreting Tissues. Pharmacol. Rev. 51, 403–438.

O'Brien, A.M.P., Utsunomiya, Y.T., Mészáros, G., Bickhart, D.M., Liu, G.E., Tassell, C.P.V., Sonstegard, T.S., Silva, M.V.D., Garcia, J.F., Sölkner, J., 2014. Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. Genet. Sel. Evol. 46, 19. doi:10.1186/1297-9686-46-19

Pan, D., Zhang, S., Jiang, J., Jiang, L., Zhang, Q., Liu, J., 2013. Genome-Wide Detection of Selective Signature in Chinese Holstein. PLoS ONE 8, e60440. doi:10.1371/journal.pone.0060440

Puglisi, R., Cambuli, C., Capoferri, R., Giannino, L., Lukaj, A., Duchi, R., Lazzari, G., Galli, C., Feligini, M., Galli, A., Bongioni, G., 2013. Differential gene expression in cumulus oocyte complexes collected by ovum pick up from repeat breeder and normally fertile Holstein Friesian heifers. Anim. Reprod. Sci. 141, 26–33. doi:10.1016/j.anireprosci.2013.07.003

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. doi:10.1086/519795

Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., Moore, S., Thaller, G., Simianer, H., 2011. Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. BMC Genomics 12, 318. doi:10.1186/1471-2164-12-318

Qanbari, S., Pausch, H., Jansen, S., Somel, M., Strom, T.M., Fries, R., Nielsen, R., Simianer, H., 2014. Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. PLoS Genet. 10. doi:10.1371/journal.pgen.1004148

Qanbari, S., Pimentel, E.C.G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R., Simianer, H., 2010. A genome-wide scan for signatures of recent selection in Holstein cattle. Anim. Genet. doi:10.1111/j.1365-2052.2009.02016.x

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E.S., 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832–7.

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

Shahzad, K., Loor, J.J., 2012. Application of top-down and bottom-up systems approaches in ruminant physiology and metabolism. Curr. Genomics 13, 379.

Smith, L., Suzuki, J., Goff, A., Filion, F., Therrien, J., Murphy, B., Kohan-Ghadr, H., Lefebvre, R., Brisville, A., Buczinski, S., Fecteau, G., Perecin, F., Meirelles, F., 2012. Developmental and Epigenetic Anomalies in Cloned Cattle. Reprod. Domest. Anim. 47, 107–114. doi:10.1111/j.1439-0531.2012.02063.x

Sun, G., Cheng, S.Y.S., Chen, M., Lim, C.J., Pallen, C.J., 2012. Protein Tyrosine Phosphatase Phosphotyrosyl-789 Binds BCAR3 To Position Cas for Activation at Integrin-Mediated Focal Adhesions. Mol. Cell. Biol. 32, 3776–3789. doi:10.1128/MCB.00214-12

Sun, J., Zhang, C., Lan, X., Lei, C., Chen, H., 2012. Exploring polymorphisms and associations of the bovine MOGAT3 gene with growth traits. Genome Natl.

Res. Counc. Can. Génome Cons. Natl. Rech. Can. 55, 56–62. doi:10.1139/g11-077

Taberlet, P., Coissac, E., Pansu, J., Pompanon, F., 2011. Conservation genetics of cattle, sheep, and goats. C. R. Biol., On the trail of domestications, migrations and invasions in agriculture Dominique Job, Georges Pelletier, Jean-Claude Pernollet 334, 247–254. doi:10.1016/j.crvi.2010.12.007

Utsunomiya, Y.T., Perez O'Brien, A.M., Sonstegard, T.S., Van Tassell, C.P., do Carmo, A.S., Meszaros, G., Solkner, J., Garcia, J.F., 2013. Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. PLoS One 8, e64280.

Voight, B.F., Kudaravalli, S., Wen, X., Pritchard, J.K., 2006. A Map of Recent Positive Selection in the Human Genome. PLoS Biol. 4, e72. doi:10.1371/journal.pbio.0040072

Wang, H., Kadlecek, T.A., Au-Yeung, B.B., Goodfellow, H.E.S., Hsu, L.-Y., Freedman, T.S., Weiss, A., 2010. ZAP-70: An Essential Kinase in T-cell Signaling. Cold Spring Harb. Perspect. Biol. 2. doi:10.1101/cshperspect.a002279

Wang, M., Zhang, X., Kang, L., Jiang, C., Jiang, Y., 2012. Molecular characterization of porcine NECD, SNRPN and UBE3A genes and imprinting status in the skeletal muscle of neonate pigs. Mol. Biol. Rep. 39, 9415–9422. doi:10.1007/s11033-012-1806-6

Watanabe, R., Chano, T., Inoue, H., Isono, T., Koiwai, O., Okabe, H., 2005. Rb1cc1 is critical for myoblast differentiation through Rb1 regulation. Virchows Arch. Int. J. Pathol. 447, 643–648. doi:10.1007/s00428-004-1183-1

Wickramasinghe, N.S., Manavalan, T.T., Dougherty, S.M., Riggs, K.A., Li, Y., Klinge, C.M., 2009. Estradiol downregulates miR-21 expression and increases miR-21 target gene expression in MCF-7 breast cancer cells. Nucleic Acids Res. 37, 2584–2595. doi:10.1093/nar/gkp117

Wilde, C.J., Addey, C.V., Li, P., Fernig, D.G., 1997. Programmed cell death in bovine mammary tissue during lactation and involution. Exp. Physiol. 82, 943–953.

Yasuda, S., Stevens, R.L., Terada, T., Takeda, M., Hashimoto, T., Fukae, J., Horita, T., Kataoka, H., Atsumi, T., Koike, T., 2007. Defective Expression of Ras

Guanyl Nucleotide-Releasing Protein 1 in a Subset of Patients with Systemic Lupus Erythematosus. J. Immunol. 179, 4890–4900.

Zhang, C., Bailey, D.K., Awad, T., Liu, G., Xing, G., Cao, M., Valmeekam, V., Retief, J., Matsuzaki, H., Taub, M., Seielstad, M., Kennedy, G.C., 2006. A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. Bioinformatics 22, 2122–8.

Zhao, F.-Q., Okine, E.K., Cheeseman, C.I., Shirazi-Beechey, S.P., Kennelly, J.J., 1998. Glucose transporter gene expression in lactating bovine gastrointestinal tract. J. Anim. Sci. 76, 2921–2929.

Supplementary files

You can find Chapter 2 supplementary files at:

- DocTa (Doctoral Thesis Archive) (http://tesionline.unicatt.it/)
- Google Drive (http://tinyurl.com/PhDMMChapter02) or scan the QR code

Supplementary file S1 – Significant core haplotypes and genes shared in dairy and beef

Significant core haplotypes (*p*-value ≤ 0.05; haplotype frequency ≥0.25) shared in dairy and beef cattle breeds and relative genes intersecting with them. The file is divided into 4 tables: The first 2 sheets (sheet 1 and 2) shown the significant core haplotypes for dairy and beef. The last 2 sheets (sheet 3 and 4) shown genes intersecting with the significant core haplotypes for dairy and beef.

Supplementary file S2 – Networks ranking in dairy breeds

Networks ranking in dairy breeds with relative molecules symbols list, score value, number of molecules belonging to the dataset (focus molecules) and top function for each network.

Supplementary file S3 – Networks ranking in beef breeds

Networks ranking in beef breeds with relative molecules symbols list, score value, number of molecules belonging to the dataset (focus molecules) and top function for each network.

Supplementary file S4 – Canonical pathways ranking in dairy or beef cattle breeds

Canonical pathways ranking in dairy (sheet 1) or beef (sheet 2) cattle breeds with relative molecules symbols list, ratio and –log10 of the *p*-values for each canonical pathway.

CHAPTER 3

Searching new signals for productive traits through gene based association analysis in three Italian cattle breeds

S. Capomaccio[1*], M. Milanesi[1*], et al.

[1] Istituto di Zootecnica, UCSC, via Emilia Parmense 84, 29122 Piacenza, Italy;

*These authors contributed equally to this work

Submitted to *Animal Genetics*

Abstract

Genome wide association studies (GWAS) have been widely applied to disentangle the genetic basis of complex traits. In cattle breeds, classical GWAS approach with medium density markers panels are far to be conclusive, especially for complex traits. This is due to the intrinsic limitations of GWAS and to the choices that are made to step from the association's signals to the functional units.

Here, we applied a gene-based association strategy to prioritize genotype-phenotype associations found with classical approaches in three Italian dairy cattle breeds with different sample sizes (Italian Holstein n=2058; Italian Brown n=745; Italian Simmental n=477) for milk production and quality traits.

While classical single marker regression failed to reveal genome-wide significant genotype-phenotype association, except for Italian Holstein, the gene based approach found specific genes in each breed that are associated with milk physiology and mammary gland development.

Since no established method is yet implemented to step from variation to functional units, our strategy may contribute to reveal new genes that play significant roles in complex traits like those here investigated, condensing low signals of association in a gene-centric fashion approach.

Background

Modern cattle breeds have no more than 200 years of history (Taberlet et al., 2008), a time span where selection, reproductive isolation and the consequent genetic drift shaped allele frequencies and linkage disequilibrium throughout the genome. This is particularly true for industrial dairy breeds, where the joint application of reproductive technologies and modern genetic evaluation methods have imposed a high selection pressure on a limited number of traits (Taberlet et al., 2008) involved in milk quantity and quality.

Indeed, the milk production industry has been historically ready to acquire technical improvements from different fields: statistics, biology, and engineering.

For example, BLUP based strategies and high-density genotyping through SNP-chips are now routinely included in bull evaluations having significant impact on selection efficiency and population structure.

These techniques have dramatically improved milk yield and composition without knowing the biological basis of the traits under selection. To date, despite many efforts and many QTL mapped, only a few genes have been effectively associated with milk yield and composition (Grisart et al., 2002; Blott et al., 2003; Cohen-Zinder et al., 2005).

In this context, Genome Wide Association Studies (GWAS) is the most common method for dissecting the biology that underlies complex traits (McCarthy et al., 2008).

The underlying idea of GWAS is simple: find marker-trait associations exploiting the linkage disequilibrium (LD) that exists between the causative mutation – which we ignore – and one or more analyzed markers, which we well know. In this way one can pinpoint genomic regions carrying causal variants for any trait.

Although a number of GWAS have been conducted in cattle productive and morphological traits (some examples (Jiang et al., 2010; Cole et al., 2011; Meredith et al., 2012)) and found several genomic regions associated to these traits, none focused their attention directly on genes involved in milk biology. Moreover, one of the major drawbacks of GWAS is related to the scarce results portability across populations due to a variety of confounding factors: population structure, differential LD levels, breed specific selection targets and even SNPs ascertainment bias (Clark et al., 2005). For instance, significant associations on fat percentage were identified in many regions of the *Bos taurus* genome in different breeds. Except for genes of major effect, there is little agreement on which regions and - more importantly - genes are actually involved in the control of the trait investigated. This is somehow not surprising: one favorable allele may segregate in one breed and be fixated in a different breed, the same allele segregates in both breeds but alleles may differ or the segregates in both breeds but the genetic background masks the effect of segregation.

Another limitation in GWAS is the stringent significance threshold often applied to correct for multiple testing. As a result, a large proportion of genes with low effect are disregarded, with consequent overestimation of the high effect genes variance (Xu, 2003). This bias is particularly worrying when investigating traits controlled by a large number of genes with small effect and in high LD as milk yield (GIANOLA ET AL., 2013).

To overcome this kind of limitation, different strategies have been invented using prior knowledge. Interesting "candidate pathway" approaches have been conducted (Raven et al., 2013) testing for association between a phenotype and genes in pathways that are likely to be involved in the control of the trait. This approach reduces the number of markers to be tested, increases the statistical power of the experiment restricting the analysis to existing knowledge.

Other methods based on statistical manipulations or particular experimental designs have been developed to prioritize or validate genotype-phenotype associations. For example, gene-based association strategies, while restricting GWA study only to genes and neighboring genomic regions, may identify new candidate genes deserving post-GWAS analysis (Akula et al., 2011; Cantor et al., 2010; Liu et al., 2010). Moreover, it is true that classical GWAS downstream bioinformatics analyses concentrate their efforts in finding meaningful results using genes as final target.

The aim of our work was to find new candidate genes and novel information associated to complex traits as milk production (yiled) and quality (fat and protein percentage), using a gene-centric genome wide association in three different dairy breeds: Italian Holstein, Italian Brown and Italian Simmental.

Methods

Biological Samples

Bulls from the three most represented dairy breeds in Italy (Italian Holstein, Italian Brown and Italian Simmental) were selected from Artificial Insemination (AI) bulls progeny tested in Italy until 2010 and for which biological samples were available. Italian Holstein bulls were chosen according to following criteria: i) having the national Selection Index namely production functionality type (PFT) with reliability > 75% and ii) having the lowest possible relationship with the other sires in the set. This was achieved by maximizing the variability within the group. On the contrary, considering the low number of available Italian Brown and Italian Simmental bulls, they were all included in the analyses. After this procedure, a total of 2139 Italian Holstein, 775 Italian Brown and 486 Italian Simmental samples (including quality control replicates) were genotyped with the BovineSNP50 BeadChip v.1 (Illumina, San Diego, CA).

Traits considered in the analysis

Italian Holstein (ANAFI), Italian Brown (ANARB) and Italian Simmental (ANAPRI) breeders associations provided deregressed proofs (DRP) or daughter yield deviations (DYD), hereafter referred as phenotype, for all traits analyzed. Traits analyzed were milk yield, fat percentage and protein percentage.

Genotyping and quality control

Quality controls (QC) were run independently for each breed, using the same pipeline and thresholds. In detail, QC excluded: i) the replicate sample with the higher number of missing genotypes ii) samples with unexpectedly high (≥100) mendelian errors (for father-son pairs); iii) samples with more than 5% missing genotypes; iv) SNPs with ≥ 2.5% missing data; v) SNPs with minor allele frequency (MAF) ≤ 5%; vi) SNPs with mendelian errors ≥ 2.5% and; vii) SNPs out of Hardy-Weinberg equilibrium tested by Fisher exact test (p ≤ 0.001, Bonferroni corrected).

Analysis of population structure

Multi Dimension Scaling (MDS) was performed using PLINK 1.07 (Purcell et al., 2007) and results plotted with R scripts. This analysis, shown in supplementary material (Figure S1), explored population substructure and verified the genetic homogeneity of the dataset prior to further analyses.

Pairwise LD was calculated using PLINK 1.07 (Purcell et al., 2007). LD decay and LD structure on specific regions were calculated with in house R scripts and snp.plotter R package (Luna and Nicodemus, 2007).

GWAS

Genome-wide association analysis was made with the GenABEL package in R (Team, 2014) using GRAMMAR-CG (Genome wide Association using Mixed Model and Regression - Genomic Control) approach (Amin et al., 2007; Aulchenko et al., 2007).

Basing on prior knowledge (Minozzi et al., 2013), in Italian Holstein *the DGAT1* region was also treated as fixed effect in GenABEL analysis of the three traits. Results using this model are referred to as "HolsteinDGAT".

Manhattan and quantile-quantile plots for each trait were produced to allow a thorough evaluation of the GWAS results. All calculated *p*-values from the GWAS analysis were retained for the downstream prioritization with the gene-based association.

Working on genome coordinates

Coordinates of the markers were lifted over from BTA4.0 (Elsik et al., 2009) to UMD3.1 assembly (Zimin et al., 2009) using SnpChimp v.1 (Nicolazzi et al., 2014) and gene coordinates were retrieved using BioMart@Ensembl.

QTL coordinates were downloaded from the Animal QTLdb (Hu et al., 2013) in bed format.

Operations on genomic intervals were performed using the bedtools suite ver. 2.16.2 (Quinlan and Hall, 2010).

Gene Based Association analysis

To reduce the antagonism between significance threshold and false positive rate of a classical GWA study (Gianola et al., 2013), and to facilitate selection signature comparison across the three breeds investigated, we used a gene-centric statistic that collapses in a single value all the *p*-values from SNPs related to a gene, correcting for LD information.

Our approach derives directly from the software VEGAS (Versatile Gene-Based Test for Genome-wide Association Studies) (Liu et al., 2010) adapted to pursue meaningful results in any species beyond humans.

Briefly, the gene-wise test statistic is the sum of the n upper tail chi-squared values of a SNP subset with one degree of freedom (where n is the number of SNP mapping in the given gene). With perfect linkage equilibrium, the gene-wise p-value would be the one-side p-value of a chi-square distribution with n degrees of freedom. Otherwise, more likely, the distribution of the p-values has to be evaluated (with simulations) and weighted (with LD values) (Liu et al., 2010).

The VEGAS species-free procedure was implemented in a UNIX environment using Bash and R languages, using Ensembl annotation as input files, PLINK data type and p-values for each SNP from the previously cited GRAMMAR procedure.

We mapped SNPs on *Bos taurus* genes (UMD 3.1) and their boundaries (100,000 bp up- and down-stream) to account for regulatory regions. Only genes with at least 2 SNPs mapped were retained. Significant entries from the gene-based association were mapped in the cattle QTL database with bedtools.

Results and Discussion

In this study, a gene based association strategy was applied to identify new genomic regions boosting GWAS analysis results for milk production and quality from three Italian dairy breeds.

Complex traits like those here evaluated are affected by a few major genes with large effects and many others with moderate to low effects; the latter are not easily identified by genome wide scans in modern cattle breeds due mainly to sample size limitation, while signal from the formers are lost in genome scans - with some exceptions - due rapid fixations.

In this scenario, with medium density SNP panels like the one used in this study, association based on single SNP regression or other “classical” methods could be inconclusive in terms of results.

After the QC process, 745, 2058 and 477 samples and 35759, 38535 and 38574 SNPs were retained for Italian Brown, Italian Holstein and Italian Simmental bulls, respectively (Table S1). HolsteinDGAT analysis retained 38534 markers, one SNP less than the other Holstein panel.

Descriptive statistics on genotypes for each breed (histogram and barplot for heterozygosis and call rate per individual, Figure S7 and multi dimensional scaling, Figure S1) revealed absence of confounding effects.

With these datasets, at the nominal genome-wide p-value significance threshold of 0.05 ($p \leq 1.25 \times 10^{-6}$, after Bonferroni correction (Burton et al., 2007; Dudbridge and Gusnanto, 2008)), significant associations were found only in Italian Holstein between milk yield and fat percentage and 26 SNPs mapping in the proximal region of BTA14, nearby *DGAT1* (Table 1).

Table 1. P-values of the SNPs overcoming the genome wide significance threshold in the Holstein breed for milk yield and fat percentage (not accounting for *DGAT1* effect). SNPs are sorted by positions in the UMD 3.1 assembly.

SNP ID	BTA	Position (Bp)	P-value (Fat)	P-Value (Milk)
Hapmap30383-BTC-005848	14	1489496	1.12E-12	3.61E-12
ARS-BFGL-NGS-57820	14	1651311	2.47E-46	1.61E-16
ARS-BFGL-NGS-34135	14	1675278	8.57E-17	
ARS-BFGL-NGS-94706	14	1696470	6.65E-16	
ARS-BFGL-NGS-4939	14	1801116	5.34E-49	1.06E-20
Hapmap52798-ss46526455	14	1923292	9.15E-21	
UA-IFASA-6878	14	2002873	3.82E-24	
ARS-BFGL-NGS-107379	14	2054457	1.18E-33	6.64E-15
ARS-BFGL-NGS-18365	14	2117455	2.50E-16	
Hapmap30922-BTC-002021	14	2138926	4.39E-13	
Hapmap25384-BTC-001997	14	2217163	3.90E-10	6.97E-09
Hapmap24715-BTC-001973	14	2239085	1.08E-09	3.69E-09
BTA-35941-no-rs	14	2276443	4.02E-13	
Hapmap30374-BTC-002159	14	2468020	5.28E-12	
Hapmap30086-BTC-002066	14	2524432	7.98E-11	
ARS-BFGL-NGS-74378	14	3640788	1.61E-08	
ARS-BFGL-BAC-1511	14	3765019	3.75E-09	
UA-IFASA-9288	14	3956956	8.11E-09	
ARS-BFGL-NGS-100480	14	4364952	1.33E-09	
UA-IFASA-5306	14	4468478	4.58E-08	

This is likely due to the effect of the *K232A DGAT1* mutation that is segregating in this breed (Fontanesi et al., 2014). No other SNP was significantly associated to any other trait in the 3 breeds investigated.

Some SNPs were under a suggestive significance threshold of $p \leq 5 \times 10^{-5}$. Details on these markers are in supplementary material (Table S3), but they are not discussed here. Plots describing at glance the association results, revealing a low signal to noise ratio, are available in the online supplementary material (Figure S2, S3, S4 and S5 respectively for Italian Brown, Italian Holstein, HolsteinDGAT and Italian Simmental).

As per the gene centric analysis we mapped a total of 19237, 19899, 19894 and 19844 genes on 24616 available on Ensembl (release version 73) in Italian Brown, Italian Holstein, HolsteinDGAT and Italian Simmental respectively. Genes with a gene-wise p-value at least 10^{-5} were considered significantly associated with the studied trait. The significance appeared to be adequate to the high rate of redundancy of the annotation features and to figure out the top associations with a reasonable level of confidence. Results on these associations per breed and per trait are summarized in Table 2, Table 3 and in supplementary materials (Table S2 and Table S4). The majority of significant genes found with the gene-centric analysis mapped in previously characterized QTL that are associated to milk production or quality. These are summarized in the Table S5.

Table 2. Number of significantly associated genes for each breed in each trait in the gene based association test.

	Fat %	Milk yield	Protein %
Brown	-	3	1
Holstein	92	80	1
HolsteinDGAT	1	4	1
Simmental	1	13	-

Table 3. Features significantly associated per breed/trait. Holstein analysis not accounting for *DGAT1* is shown in Table S3.

Breed	Trait	Ensembl ID	GeneWise PValue	SNPsInGene	BestSNP	BestSNP_PValue	BTA	Start (bp)	End (bp)	Gene Name
BROWN	*FAT*	No significant association found								
	MILK	ENSBTAG00000019237	5.00E-05	4	Hapmap53770-ss46526325	2.04E-04	23	47333403	47348212	*TXNDC5*
		ENSBTAG00000043918	5.00E-05	4	Hapmap53770-ss46526325	2.04E-04	23	47337650	47337787	*5S_rRNA*
		ENSBTAG00000046839	5.00E-05	4	Hapmap53770-ss46526325	2.04E-04	23	47328131	47352621	*Uncharacterized Protein*
	PROTEIN	ENSBTAG00000001260	7.00E-05	3	ARS-BFGL-NGS-116222	4.41E-03	18	52120387	52124418	*PINLYP*
HOLSTEINDGAT	*FAT*	ENSBTAG00000000369	1.00E-05	3	Hapmap53294-rs29016908	3.33E-05	5	94673755	94749093	*EPS8*
	MILK	ENSBTAG00000026896	1.30E-05	5	ARS-BFGL-NGS-104949	3.85E-04	23	51549721	51553563	*FOXF2*
		ENSBTAG00000005260	4.00E-05	5	Hapmap33288-BTC-033751	1.20E-03	6	38120578	38127577	*SPP1*
		ENSBTAG00000006579	5.00E-05	4	ARS-BFGL-BAC-2207	6.32E-05	15	54418559	54459500	*P4HA3*
		ENSBTAG00000007013	7.00E-05	6	UA-IFASA-7665	3.96E-05	19	5221507	5283308	*TOM1L1*
	PROTEIN	ENSBTAG00000006638	6.00E-05	3	ARS-BFGL-NGS-104774	1.76E-04	18	56523439	56530499	*BCL2L12*
SIMMENTHAL	*FAT*	ENSBTAG00000037649	9.00E-05	6	Hapmap30001-BTA-142716	5.29E-05	4	120427915	120494453	*VIPR2*
	MILK	ENSBTAG00000017538	1.00E-05	6	ARS-BFGL-BAC-13093	5.87E-06	12	85630428	85630912	*Pseudogene*
		ENSBTAG00000000856	2.00E-05	2	ARS-BFGL-NGS-4939	4.13E-06	14	1766767	1769754	*FBXL6*
		ENSBTAG00000000857	2.00E-05	2	ARS-BFGL-NGS-4939	4.13E-06	14	1763994	1766621	*GPR172B*
		ENSBTAG00000026356	2.00E-05	2	ARS-BFGL-NGS-4939	4.13E-06	14	1795351	1804562	*DGAT1*
		ENSBTAG00000035158	2.00E-05	2	ARS-BFGL-NGS-4939	4.13E-06	14	1770660	1772329	*TMEM249*
		ENSBTAG00000046208	2.00E-05	2	ARS-BFGL-NGS-4939	4.13E-06	14	1782901	1788087	*SCRT1*
		ENSBTAG00000009811	4.00E-05	2	ARS-BFGL-NGS-4939	4.13E-06	14	1825982	1844587	*Uncharacterized Protein*
		ENSBTAG00000009816	4.00E-05	2	ARS-BFGL-NGS-4939	4.13E-06	14	1838135	1840081	*Uncharacterized Protein*
		ENSBTAG00000014458	4.00E-05	2	ARS-BFGL-NGS-4939	4.13E-06	14	1844664	1894424	*MROH1*
		ENSBTAG00000020751	4.00E-05	2	ARS-BFGL-NGS-4939	4.13E-06	14	1806081	1825793	*HSF1*
		ENSBTAG00000044406	4.00E-05	2	ARS-BFGL-NGS-4939	4.13E-06	14	1883849	1883971	*SCARNA15*
		ENSBTAG00000046889	4.00E-05	5	ARS-BFGL-BAC-13093	5.87E-06	12	85610279	85610762	*Uncharacterized Protein*
		ENSBTAG00000020117	8.00E-05	3	BTA-78494-no-rs	1.88E-04	7	21189879	21196263	*ZBTB7A*
	PROTEIN	No significant association found								

1. Italian Holstein

The large number of significant signals was observed in the Italian Holstein, especially for milk yield and fat percentage. As shown in Table 2, the strong effect of *DGAT1* in Italian Holsteins observed with individual SNPs, was confirmed by the gene-centric approach. In this region, almost the totality of the gene-wise p-values under the chosen threshold appeared to be influenced by this major gene and by the high level of linkage disequilibrium that characterizes cosmopolitan bovine breeds (see Figure S6). Within this pool of genes, it is worth to mention *PLEC* (*plectin*) a cell surface receptor gene in the *RNASE5* pathway (Angiogenin encoded by angiogenin, ribonuclease, RNase A family 5). Proteins encoded by these genes, that are present in milk, are involved in protein synthesis and act as growth factors in vitro. Moreover, *RNASE5* seems to have other functions, such as regulation of stress response and apoptosis and even angiogenesis, through which it may also contribute to the regulate milk production (Raven et al., 2013).

The exclusion of *DGAT1* effect (HolsteinDGAT results) highlighted that all the significant effects on fat percentage found for genes located within approximately 2.5 Mb up or downstream from the excluded SNP (*ARS-BFGL-NGS-4939*) were redundant signals of *DGAT1* (Table3 and Table S2 and S3). This is also observed for milk yield due to the genetic correlations between fat and milk. Also, *PLEC* was not significant in our analyses when taking into account the *DGAT1* effect, contradicting previous findings in Holstein, that report association of *PLEC* gene with production traits either accounting or not for the effect of the diglyceride acyltransferase (Raven et al., 2013). A possible explanation for the different results between these studies is a difference between genetic structures of the populations analyzed. To reduce false positive signals, we only evaluated the genes based on HolsteinDGAT results for the Italian Holstein.

1.1 Milk yield

We found genes associated with this trait in Holstein, in particular *SPP1*, *P4HA3, FOXF2* and *TOM1L1* genes.

Among these, the osteopontin *SPP1* (secreted phosphoprotein 1) is directly related to milk production (de Mello et al., 2012). It is located in BTA6, nearby

ABCG2, and there is evidence that polymorphisms in this region are strongly correlated with high production levels of milk in various breeds (Cohen-Zinder et al., 2005). Osteopontin *SPP1,* is a secreted glycophosphoprotein that plays a crucial role in mammary gland development, having an essential role also in mammary gland differentiation and branching of the mammary epithelial ductal system. This gene has also been implicated in many types of cancer, including breast cancer, for the potentiality of proliferation and alveologenesis (Hubbard et al., 2013). It is worth mentioning that, the most significant SNPs within this gene was *Hapmap33288-BTC-033751 with* a *p*-value of 0.0012, largely over the genome-wide significance threshold. This signal would be ignored in the single SNP approach.

Another gene revealed by the gene-based association is *P4HA3*, encoding for the prolyl 4-hydroxylase, α polypeptide III, involved in the correct folding of nascent procollagen chains. It is known that stromal-epithelial interactions are critical for mammary gland development, and defective deposition or reorganization of extracellular matrix can lead to incorrect mammary epithelial morphogenesis (Gillette et al., 2013) highlighting the role of this gene in the healthy breast tissue.

FOXF2 gene (Forkhead Box F2) is a member of the forkhead box transcription factors, known to be involved in tissue development as part of the "Hedgehog signaling pathway". This key cascade is essential for the correct segmentation of tissues during embryonic development but has been recently demonstrated that is active in adult stem cells proliferation in mammary tissue (Liu et al., 2006). In particular, *FOXF2*, plays a key role in extracellular matrix synthesis and epithelial-mesenchymal interactions, and could therefore affect the proper development of the mammary stroma. In addition, low levels of transcription of this gene are correlated to early onset metastasis in breast cancer (Kong et al., 2013), underlining a specific role in the mammary gland.

Lastly, *TOM1L1* (Target Of Myb1 Chicken-Like 1) is another gene potentially involved with the milk physiology as is known its implication with the early endosome sorting coupled with *EGFR* (epidermal growth factor receptor), another important player in lactation (Fowler et al., 1995). TOM1L1 belongs to the VHS domain containing proteins, that are involved in the post-Golgi trafficking and signaling (Wang et al., 2010).

1.2 Fat percentage

In Italian Holstein breed we found a significant association, with or without *DGAT1* effect correction, to the *EPS8* gene located on BTA5, at around 95 Mbp. This gene encodes for a substrate of the already cited *EGFR* and is involved in the fatty acid metabolism (Fazioli et al., 1993). A recent study on Holstein Friesian (Wang et al., 2012) found strong association with a variant located in the second intron of the epidermal growth factor receptor pathway substrate, enforcing the hypothesis of a true involvement of this gene on fat percentage.

1.3 Protein percentage

The *BCL2L12* gene (Bcl-2-like protein 12), associated with protein percentage in Italian Holstein breed, represents another interesting result. The product of this gene is an important oncoprotein in two fundamental cell cycle pathways, namely *p53* and cytoplasmic caspase signaling. BCL2L12 resides in both cytoplasm and nucleus, inhibiting caspases 3 and 7 and forming a complex with p53 thereby inhibiting p53-directed transcriptomic changes upon DNA damage (Stegh and DePinho, 2011). The members of BCL2 family are considered anti-apoptotic genes and it is obvious that dysfunctional programmed cell death plays an important role in the pathogenesis and in the progression of cancer. Many members of this family were found to be modulated in various malignancies and *BCL2L12* in particular was expressed in mammary gland and proposed as prognostic cancer biomarker (Thomadaki et al., 2007). In dairy cows, the members of BCL2 family are involved in the key events in the mammary tissue development and remodeling during lactation and involution (Stefanon et al., 2002, p. 200). Members of BCL2 family regulate the turnover of cell population in the mammary gland during lactation and its overexpression was found to be linearly related to milk production in Holstein cow (Colitti et al., 2004).

2. Italian Simmental

In the Italian Simmental breed, we found one association on BTA4 for fat percentage (*VIPR2)* and two associations for milk yield: one in BTA7 (*ZBTB7A*) and the other in BTA14 related to the *DGAT1*. No association was found for protein percentage.

2.1 Fat percentage

Among the genes found significantly associated with production traits in Italian Simmental, *VIPR2* a gene encoding a receptor for a small vasoactive intestinal neuropeptide, was the single signal in BTA4 for fat percentage. Vasoactive intestinal peptide (VIP) is synthesized in the pituitary gland and, among other functions, stimulates prolactin secretion and is involved in proliferation, survival, and differentiation in human breast cancer cells (Valdehita et al., 2010). An interesting novel finding is the effect of immunosuppression (through direct effect of *IL-6*) on prolactin cells and its ability to regulate prolactin by acting on pituitary VIP through VIP receptors (Blanco et al., 2013). These evidences suggest that *VIPR2* could be an important candidate for further investigations through post GWAS approaches such as re-sequencing or fine mapping.

2.2 Milk yield

Surprisingly, the *DGAT1* gene was significantly associated with milk yield but not with fat percentage in Italian Simmental breed. Indeed, the most significant SNP for the milk yield trait in this breed was the *ARS-BFGL-NGS-4939*, with a p-value of 4,13E-06. The importance of this gene in the lactation is largely known (Grisart et al., 2002).

To investigate *DGAT1* signal in Simmental, we analyzed the fine LD structure genome-wide and in BTA14 proximal region in all the three breeds, revealing different patterns.

Observed LD decay on the three breeds genomes revealed similar behavior for Italian Holstein and Italian Brown, with $r^2 > 0.10$ until around 250 Kbp, while Italian Simmental showed lower LD. Distances higher that 400 Kbp revealed half level of LD in Simmental (Figure S6).

In *DGAT1* region, LD behavior is very similar between Italian Holstein and Italian Brown (Figure S8) while association values are really different because *ARS-BFGL-NGS-4939* is fixed in Italian Brown and purged by QC process. Interestingly, association patterns become similar when *DGAT1* effect is removed from the Holstein panel (HolsteinDGAT, Figure S8). In Simmental we found a different situation: very low LD level and significant association levels with *ARS-BFGL-NGS-4939* with the milk trait. This is probably due to the low frequency of the p.232K allele in Italian Simmental (Scotti et al., 2010) and to the different selection strategies applied in these breeds.

The relatively high *p*-value of this SNP is responsible for the statistical significance of many of the 13 genes found significant in this breed for this trait as shown in Table 3, therefore reducing the gene set to three locations: two overlapping features indicating a pseudogene in BTA12 and a zinc finger protein *ZBTB7A* in BTA7. *DGAT1* role and function will be not discussed here as already done elsewhere (Grisart et al., 2002).

ZBTB7A is part of zinc finger and BTB proteins (Broad complex, Tramtrack, and Bricabrac), transcription factors with emerging importance for their involvement in oncogenesis and cell differentiation by repressing key genes of cell cycle regulation as retinoblastoma (Rb) and *p53* ("Jeon et al. - 2008 - Proto-oncogene FBI-1 (PokemonZBTB7A) Represses Tr.pdf," n.d.). *ZBTB7A* over expression is shown in primary and recurrent breast cancer tissues leading to disease progression and poor survival (Zu et al., 2011). Link between this gene polymorphism and milk production might be in the proper cellular development of the mammary gland. This transcription factor is also known as a master regulator of B and T-cell lineage (Lee and Maeda, 2012) enforcing the link between milk production and immune protection, as already outlined by of Lemay and colleagues with proteomic studies, where the "activation of the innate immune system" resulted as one of the most represented gene ontology categories (Lemay et al., 2009).

3. Italian Brown

In the Italian Brown breed, we found one association on BTA23 (*TXNDC5)* for milk yield and one in BTA18 for protein percentage (*PINLYP*). The three features listed in Table 3 for milk yield in Italian Brown were collapsed due to overlap.

3.1 Milk yield

Thioredoxin domain-containing protein 5 (*TXNDC5*) belongs to thioredoxin family, small ubiquitous proteins containing disulphide domain affecting protein folding with chaperone activity, preventing protein mis-folding within the lumen of the endoplasmic reticulum. This multifunctional organelle have a key role in several processes, including lipid catabolism and protein synthesis (Ramírez-Torres et al., 2012). Because of its disulphide bonds, thioredoxin facilitate the reduction of other proteins by cysteine thiol disulfide exchange. This mechanism is essential for detossification.

During lactation, indeed, an enhanced detoxification level is functional to preserve milk quality against increased metabolic activities. Higher metabolic levels lead to free radicals production, making necessary the maintenance of redox homeostasis that include, among others, the reduction and oxidation of peroxiredoxin and thioredoxin (Lu and Holmgren, 2014).

3.2 Protein percentage

For protein percentage in Italian Brown there is only one associated gene. *PINLYP* is encoding a phospholipase A2 inhibitor, but this protein product is not yet well studied. Many types of phospholipase A2 were characterized and it is known that they are mediator of inflammation, atherosclerosis and cancer in mammals and that plays an important role in the innate immune response (Qiu and Lai, 2013). It may be reasonable to imagine a role for *PINLYP* in the lactation in light of previously discussed results that link milk to innate immune response (Lemay et al., 2009).

Although limited by the size of the study and the number of the included markers, these results represent both a confirmation and a step forward towards the comprehension of the biological bases of milk yield and quality.

New genes potentially involved with production traits in dairy cows were tracked when GWAS signals resulted too weak to be considered in a classical SNP based mapping with the "closest gene to peak" or the window-based strategy.

It is worth mentioning that none of the analyzed SNPs (not considering the Italian Holstein with *ARS-BFGL-NGS-4939* included in the analysis) were above the genome-wide significance threshold in single SNP analysis, meaning that important information is often lost in such studies.

In addition to the newly identified loci, we observed a confirmation of existing associations with genes related to milk production, not only dissecting gene specific functions but from mapping these genes in previously characterized regions (Table S5). This is not trivial when using new data analysis techniques, especially in high noise to signal ratio datasets like those here presented. We believe that this approach will benefit of data from denser marker panels (e.g. BovineHD) to exploit local linkage disequilibrium and increase the number of mapped genes with more than two SNPs.

Results obtained with the gene centric approach, although having intrinsic limitations, are not influenced by a number of subjective choices often made in GWAS. For instance, when "window mapping" is used to decrease the noise to signal ratio e.g. due to nearby markers having different allele frequencies, windows size is to be decided. This can be based on equal number of markers, equal number of base pairs or equal map distance in cM per window. The three choices do not coincide, because markers are not equally spread along the chromosomes and recombination differs in different genomic regions. In addition windows can be chosen non-overlapping or with different degree of overlap. Finally, when stepping from significant SNPs or windows to candidate genes, the size of the flanking region from which picking up candidate genes is to be decided as well. Sometimes, the gene closer to the signal peak is selected, in other cases all genes within certain boundaries are included in further analyses. All these choices may generate misleading results. For example, not truly associated genes may be retained for further analyses, disregarding the correct ones. In addition, a large number of false positives can be included, to be later interpreted with the help of network and pathway analysis. As a consequence, these decisions have a direct impact in terms of production and interpretation of results, leading to conclusions heavily based on the enrichment analysis and pre-existing knowledge.

With the gene based association, we also noticed some advantages in the downstream data mining. First, genes are functional units and the golden targets of the "classical post-GWAS" bioinformatics pipelines. Hence, the gene based association leads to genes with an embedded statistically robust framework. This is a "plus" in studying complex traits because one can concentrate on specific signals rather than cherry picking meaningful features from a list constructed with the previously cited approaches.

Moreover, it is easier to track back gene clusters to major gene effects - like for example *DGAT1*: in a gene rich genome chunk, many genes may result significant due to the same SNP (Table S2 and Table 3). With the VEGAS method, finding the "true" association is facilitated by following the "best SNP" in the region.

In addition, we found phenotype coherent signals in such situation, where only a few SNPs - if none – are above an acceptable statistical genome-wide threshold using a single SNP regression method.

The gene centric approach is a first and slight improvement in livestock GWAS analyses, since we still have a partial genome annotation. In addition, it is now known that the definition of gene is to be revised; many genes do not code for proteins, have alternative starting sites, have antisense transcripts, host regulatory RNA, are subjected to alternative splicing and may be regulated by sequences quite far away from their coding regions (Consortium, 2012). Other limitations are the need for larger number of animals in these type of studies, as in this study, and the still relatively high cost of sequencing that cannot be used routinely in large experiments, at least not yet.

Conclusions

This study underlines the importance of the implementation of new methods to analyze complex traits with genomic data. No golden path(s) is discovered yet to walk from variation to functional units. The gene-centric analysis may contribute to reveal new signals throughout the bovine genome, condensing weak association signals in chromosome bits having functional significance.

Acknowledgements

The Authors would like to thank Dr. Jimmy Liu for his advices in building the software. AIA, ANAFI, ANARB and ANAPRI breeders' associations for their help and support during the data analysis. SelMol and ProZoo projects that provide samples. The Authors are also grateful to MIPAAF and Cariplo Foundation for funding.

References

Akula, N., Baranova, A., Seto, D., Solka, J., Nalls, M.A., Singleton, A., Ferrucci, L., Tanaka, T., Bandinelli, S., Cho, Y.S., Kim, Y.J., Lee, J.-Y., Han, B.-G., Bipolar Disorder Genome Study (BiGS) Consortium, The Wellcome Trust Case-Control Consortium, McMahon, F.J., 2011. A Network-Based Approach to Prioritize Results from Genome-Wide Association Studies. PLoS ONE 6, e24220. doi:10.1371/journal.pone.0024220

Amin, N., van Duijn, C.M., Aulchenko, Y.S., 2007. A Genomic Background Based Method for Association Analysis in Related Individuals. PLoS ONE 2, e1274. doi:10.1371/journal.pone.0001274

Aulchenko, Y.S., de Koning, D.-J., Haley, C., 2007. Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. Genetics 177, 577–585. doi:10.1534/genetics.107.075614

Blanco, E.J., Carretero-Hernández, M., García-Barrado, J., Iglesias-Osma, M.C., Carretero, M., Herrero, J.J., Rubio, M., Riesco, J.M., Carretero, J., 2013. The activity and proliferation of pituitary prolactin-positive cells and pituitary

VIP-positive cells are regulated by interleukin 6. Histol. Histopathol. 28, 1595–1604.

Blott, S., Kim, J.-J., Moisio, S., Schmidt-Kuntzel, A., Cornet, A., Berzi, P., Cambisano, N., Ford, C., Grisart, B., Johnson, D., Karim, L., Simon, P., Snell, R., Spelman, R., Wong, J., Vilkki, J., Georges, M., Farnir, F., Coppieters, W., 2003. Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 163, 253–266.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., Todd, J.A., Donnelly, P., Barrett, J.C., Burton, P.R., Davison, D., Donnelly, P., Easton, D., Evans, D., Leung, H.-T., Marchini, J.L., Morris, A.P., Spencer, C.C.A., Tobin, M.D., Cardon, L.R., Clayton, D.G., Attwood, A.P., Boorman, J.P., Cant, B., Everson, U., Hussey, J.M., Jolley, J.D., Knight, A.S., Koch, K., Meech, E., Nutland, S., Prowse, C.V., Stevens, H.E., Taylor, N.C., Walters, G.R., Walker, N.M., Watkins, N.A., Winzer, T., Todd, J.A., Ouwehand, W.H., Jones, R.W., McArdle, W.L., Ring, S.M., Strachan, D.P., Pembrey, M., Breen, G., Clair, D.S., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E.K., Grozeva, D., Hamshere, M.L., Holmans, P.A., Jones, I.R., Kirov, G., Moskvina, V., Nikolov, I., O’Donovan, M.C., Owen, M.J., Craddock, N., Collier, D.A., Elkin, A., Farmer, A., Williamson, R., McGuffin, P., Young, A.H., Ferrier, I.N., Ball, S.G., Balmforth, A.J., Barrett, J.H., Bishop, D.T., Iles, M.M., Maqbool, A., Yuldasheva, N., Hall, A.S., Braund, P.S., Burton, P.R., Dixon, R.J., Mangino, M., Stevens, S., Tobin, M.D., Thompson, J.R., Samani, N.J., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C.W., Nimmo, E.R., Satsangi, J., Fisher, S.A., Forbes, A., Lewis, C.M., Onnie, C.M., Prescott, N.J., Sanderson, J., Mathew, C.G., Barbour, J., Mohiuddin, M.K., Todhunter, C.E., Mansfield, J.C., Ahmad, T., Cummings, F.R., Jewell, D.P., Webster, J., Brown, M.J., Clayton, D.G., Lathrop, G.M., Connell, J., Dominiczak, A., Samani, N.J., Marcano, C.A.B., Burke, B., Dobson, R., Gungadoo, J., Lee, K.L., Munroe, P.B., Newhouse, S.J., Onipinla, A., Wallace, C., Xue, M., Caulfield, M., Farrall, M., Barton, A., (braggs), T.B. in R.G. and G., Bruce, I.N., Donovan, H., Eyre, S., Gilbert, P.D., Hider, S.L., Hinks, A.M., John, S.L., Potter, C., Silman, A.J., Symmons, D.P.M., Thomson, W., Worthington, J., Clayton, D.G., Dunger, D.B., Nutland, S., Stevens, H.E., Walker, N.M., Widmer, B., Todd, J.A., Frayling, T.M., Freathy, R.M., Lango, H., Perry, J.R.B., Shields, B.M., Weedon, M.N., Hattersley, A.T., Hitman, G.A., Walker, M., Elliott, K.S., Groves, C.J., Lindgren, C.M., Rayner, N.W., Timpson, N.J., Zeggini, E., McCarthy, M.I., Newport, M., Sirugo, G., Lyons, E., Vannberg, F., Hill, A.V.S., Bradbury, L.A., Farrar, C., Pointon, J.J., Wordsworth, P., Brown, M.A., Franklyn, J.A., Heward, J.M., Simmonds, M.J., Gough, S.C.L., Seal, S., (uk), B.C.S.C., Stratton, M.R., Rahman, N., Ban, M., Goris, A., Sawcer, S.J., Compston, A., Conway, D., Jallow, M., Newport, M., Sirugo, G., Rockett, K.A., Kwiatkowski, D.P., Bumpstead, S.J., Chaney, A., Downes, K., Ghori, M.J.R., Gwilliam, R., Hunt, S.E., Inouye, M., Keniry, A., King, E., McGinnis, R., Potter, S., Ravindrarajah, R., Whittaker, P., Widden, C., Withers, D., Deloukas, P., Leung, H.-T., Nutland, S., Stevens, H.E., Walker, N.M., Todd, J.A., Easton, D., Clayton, D.G., Burton, P.R., Tobin, M.D., Barrett, J.C., Evans, D., Morris, A.P., Cardon, L.R., Cardin, N.J., Davison, D., Ferreira, T., Pereira-Gale, J., Hallgrimsdóttir, I.B., Howie, B.N., Marchini, J.L., Spencer, C.C.A., Su, Z., Teo, Y.Y., Vukcevic, D., Donnelly, P., Bentley, D., Brown, M.A., Cardon, L.R., Caulfield, M., Clayton, D.G., Compston, A., Craddock, N., Deloukas, P., Donnelly, P., Farrall, M., Gough, S.C.L., Hall, A.S., Hattersley, A.T., Hill, A.V.S., Kwiatkowski, D.P., Mathew, C.G., McCarthy, M.I., Ouwehand, W.H., Parkes, M., Pembrey, M., Rahman, N., Samani, N.J., Stratton, M.R., Todd, J.A., Worthington, J., 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678. doi:10.1038/nature05911

Cantor, R.M., Lange, K., Sinsheimer, J.S., 2010. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am. J. Hum. Genet. 86, 6–22. doi:10.1016/j.ajhg.2009.11.017

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., Nielsen, R., 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 15, 1496–1502. doi:10.1101/gr.4107905

Cohen-Zinder, M., Seroussi, E., Larkin, D.M., Loor, J.J., Wind, A.E. der, Lee, J.-H., Drackley, J.K., Band, M.R., Hernandez, A.G., Shani, M., Lewin, H.A., Weller, J.I., Ron, M., 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res. 15, 936–944. doi:10.1101/gr.3806705

Cole, J.B., Wiggans, G.R., Ma, L., Sonstegard, T.S., Lawlor, T.J., Crooker, B.A., Tassell, C.P.V., Yang, J., Wang, S., Matukumalli, L.K., Da, Y., 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics 12, 408. doi:10.1186/1471-2164-12-408

Colitti, M., Wilde, C.J., Stefanon, B., 2004. Functional expression of bcl-2 protein family and AIF in bovine mammary tissue in early lactation. J. Dairy Res. 71, 20–27.

Consortium, T.E.P., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. doi:10.1038/nature11247

De Mello, F., Cobuci, J.A., Martins, M.F., Silva, M.V.G.B., Neto, J.B., 2012. Association of the polymorphism g.8514C>T in the osteopontin gene (SPP1) with milk yield in the dairy cattle breed Girolando. Anim. Genet. 43, 647–648. doi:10.1111/j.1365-2052.2011.02312.x

Dudbridge, F., Gusnanto, A., 2008. Estimation of significance thresholds for genomewide association scans. Genet. Epidemiol. 32, 227–234. doi:10.1002/gepi.20297

Elsik, C.G., Tellam, R.L., Worley, K.C., 2009. The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution. Science 324, 522–528. doi:10.1126/science.1169588

Fazioli, F., Minichiello, L., Matoska, V., Castagnino, P., Miki, T., Wong, W.T., Di Fiore, P.P., 1993. Eps8, a substrate for the epidermal growth factor receptor kinase, enhances EGF-dependent mitogenic signals. EMBO J. 12, 3799–3808.

Fontanesi, L., Calò, D.G., Galimberti, G., Negrini, R., Marino, R., Nardone, A., Ajmone-Marsan, P., Russo, V., 2014. A candidate gene association study for nine economically important traits in Italian Holstein cattle. Anim. Genet. n/a–n/a. doi:10.1111/age.12164

Fowler, K.J., Walker, F., Alexander, W., Hibbs, M.L., Nice, E.C., Bohmer, R.M., Mann, G.B., Thumwood, C., Maglitto, R., Danks, J.A., 1995. A mutation in the epidermal growth factor receptor in waved-2 mice has a profound effect on receptor biochemistry that results in impaired lactation. Proc. Natl. Acad. Sci. 92, 1465–1469. doi:10.1073/pnas.92.5.1465

Gianola, D., Hospital, F., Verrier, E., 2013. Contribution of an additive locus to genetic variance when inheritance is multi-factorial with implications on interpretation of GWAS. Theor. Appl. Genet. 126, 1457–1472. doi:10.1007/s00122-013-2064-2

Gillette, M., Bray, K., Blumenthaler, A., Vargo-Gogola, T., 2013. P190B RhoGAP Overexpression in the Developing Mammary Epithelium Induces TGFβ-dependent Fibroblast Activation. PLoS ONE 8, e65105. doi:10.1371/journal.pone.0065105

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M., Snell, R., 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. doi:10.1101/gr.224202

Hubbard, N.E., Chen, Q.J., Sickafoose, L.K., Wood, M.B., Gregg, J.P., Abrahamsson, N.M., Engelberg, J.A., Walls, J.E., Borowsky, A.D., 2013. Transgenic Mammary Epithelial Osteopontin (Spp1) Expression Induces Proliferation and Alveologenesis. Genes Cancer 4, 201–212. doi:10.1177/1947601913496813

Hu, Z.-L., Park, C.A., Wu, X.-L., Reecy, J.M., 2013. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res. 41, D871–D879. doi:10.1093/nar/gks1150

Jeon et al. - 2008 - Proto-oncogene FBI-1 (PokemonZBTB7A) Represses Tr.pdf, n.d.

Jiang, L., Liu, J., Sun, D., Ma, P., Ding, X., Yu, Y., Zhang, Q., 2010. Genome Wide Association Studies for Milk Production Traits in Chinese Holstein Population. PLoS ONE 5, e13661. doi:10.1371/journal.pone.0013661

Kong, P.-Z., Yang, F., Li, L., Li, X.-Q., Feng, Y.-M., 2013. Decreased FOXF2 mRNA Expression Indicates Early-Onset Metastasis and Poor Prognosis for Breast Cancer Patients with Histological Grade II Tumor. PLoS ONE 8, e61591. doi:10.1371/journal.pone.0061591

Lee, S.-U., Maeda, T., 2012. POK/ZBTB proteins: an emerging family of proteins that regulate lymphoid development and function. Immunol. Rev. 247, 107–119.

Lemay, D.G., Lynn, D.J., Martin, W.F., Neville, M.C., Casey, T.M., Rincon, G., Kriventseva, E.V., Barris, W.C., Hinrichs, A.S., Molenaar, A.J., 2009. The bovine lactation genome: insights into the evolution of mammalian milk. Genome Biol. 10, R43.

Liu, J.Z., Mcrae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., 2010. A versatile gene-based test for genome-wide association studies. Am. J. Hum. Genet. 87, 139.

Liu, S., Dontu, G., Mantle, I.D., Patel, S., Ahn, N., Jackson, K.W., Suri, P., Wicha, M.S., 2006. Hedgehog Signaling and Bmi-1 Regulate Self-renewal of Normal and Malignant Human Mammary Stem Cells. Cancer Res. 66, 6063–6071. doi:10.1158/0008-5472.CAN-06-0054

Lu, J., Holmgren, A., 2014. The thioredoxin superfamily in oxidative protein folding. Antioxid. Redox Signal. 140202091634000. doi:10.1089/ars.2014.5849

Luna, A., Nicodemus, K.K., 2007. snp.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package. Bioinforma. Oxf. Engl. 23, 774–776. doi:10.1093/bioinformatics/btl657

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., Hirschhorn, J.N., 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9, 356–369. doi:10.1038/nrg2344

Meredith, B.K., Kearney, F.J., Finlay, E.K., Bradley, D.G., Fahey, A.G., Berry, D.P., Lynn, D.J., 2012. Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland. BMC Genet. 13, 21. doi:10.1186/1471-2156-13-21

Minozzi, G., Nicolazzi, E.L., Stella, A., Biffani, S., Negrini, R., Lazzari, B., Ajmone-Marsan, P., Williams, J.L.., 2013. Genome Wide Analysis of Fertility and Production Traits in Italian Holstein Cattle. PLoS ONE 8, e80219. doi:10.1371/journal.pone.0080219

Nicolazzi, E.L., Picciolini, M., Strozzi, F., Schnabel, R.D., Lawley, C., Pirani, A., Brew, F., Stella, A., 2014. SNPchiMp: a database to disentangle the SNPchip jungle in bovine livestock. BMC Genomics 15, 123. doi:10.1186/1471-2164-15-123

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. doi:10.1086/519795

Qiu, S., Lai, L., 2013. Antibacterial properties of recombinant human non-pancreatic secretory phospholipase A2. Biochem. Biophys. Res. Commun. 441, 453–456. doi:10.1016/j.bbrc.2013.10.092

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. doi:10.1093/bioinformatics/btq033

Ramírez-Torres, A., Barceló-Batllori, S., Martínez-Beamonte, R., Navarro, M.A., Surra, J.C., Arnal, C., Guillén, N., Acín, S., Osada, J., 2012. Proteomics and gene expression analyses of squalene-supplemented mice identify microsomal thioredoxin domain-containing protein 5 changes associated with hepatic steatosis. J. Proteomics 77, 27–39. doi:10.1016/j.jprot.2012.07.001

Raven, L.-A., Cocks, B.G., Pryce, J.E., Cottrell, J.J., Hayes, B.J., 2013. Genes of the RNASE5 pathway contain SNP associated with milk production traits in dairy cattle. Genet. Sel. Evol. 45, 25. doi:10.1186/1297-9686-45-25

Scotti, E., Fontanesi, L., Schiavini, F., La Mattina, V., Bagnato, A., Russo, V., 2010. DGAT1 p.K232A polymorphism in dairy and dual purpose Italian cattle breeds. Ital. J. Anim. Sci. 9. doi:10.4081/ijas.2010.e16

Stefanon, B., Colitti, M., Gabai, G., Knight, C.H., Wilde, C.J., 2002. Mammary apoptosis and lactation persistency in dairy animals. J. Dairy Res. 69, 37–52.

Stegh, A.H., DePinho, R.A., 2011. Beyond effector caspase inhibition: Bcl2L12 neutralizes p53 signaling in glioblastoma. Cell Cycle 10, 33–38. doi:10.4161/cc.10.1.14365

Taberlet, P., Valentini, A., Rezaei, H.R., Naderi, S., Pompanon, F., Negrini, R., Ajmone-Marsan, P., 2008. Are cattle, sheep, and goats endangered species? Mol. Ecol. 17, 275–284. doi:10.1111/j.1365-294X.2007.03475.x

Team, R.C., 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Thomadaki, H., Talieri, M., Scorilas, A., 2007. Prognostic value of the apoptosis related genes BCL2 and BCL2L12 in breast cancer. Cancer Lett. 247, 48–55. doi:10.1016/j.canlet.2006.03.016

Valdehita, A., Bajo, A.M., Fernández-Martínez, A.B., Arenas, M.I., Vacas, E., Valenzuela, P., Ruíz-Villaespesa, A., Prieto, J.C., Carmena, M.J., 2010. Nuclear localization of vasoactive intestinal peptide (VIP) receptors in human breast cancer. Peptides 31, 2035–2045. doi:10.1016/j.peptides.2010.07.024

Wang, T., Liu, N.S., Seet, L.-F., Hong, W., 2010. The Emerging Role of VHS Domain-Containing Tom1, Tom1L1 and Tom1L2 in Membrane Trafficking. Traffic 11, 1119–1128. doi:10.1111/j.1600-0854.2010.01098.x

Wang, X., Wurmser, C., Pausch, H., Jung, S., Reinhardt, F., Tetens, J., Thaller, G., Fries, R., 2012. Identification and Dissection of Four Major QTL Affecting Milk Fat Content in the German Holstein-Friesian Population. PLoS ONE 7, e40711. doi:10.1371/journal.pone.0040711

Xu, S., 2003. Theoretical Basis of the Beavis Effect. Genetics 165, 2259–2268.

Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Tassell, C.P.V., Sonstegard, T.S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J.A., Salzberg, S.L., 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 10, R42. doi:10.1186/gb-2009-10-4-r42

Zu, X., Ma, J., Liu, H., Liu, F., Tan, C., Yu, L., Wang, J., Xie, Z., Cao, D., Jiang, Y., 2011. Pro-oncogene Pokemon promotes breast cancer progression by upregulating survivin expression. Breast Cancer Res. 13, R26. doi:10.1186/bcr2843

Supplementary files

You can find Chapter 3 supplementary files at:

- DocTa (Doctoral Thesis Archive) (http://tesionline.unicatt.it/)
- Google Drive (http://tinyurl.com/PhDMMChapter03) or scan the QR code

Supplementary Figure S1: Multidimensional scaling

Spatial distribution of genetic structure in the analyzed breeds.

Supplementary Figure S2: GWAS plot

Manhattan plots and QQ-plots respectively for Fat percentage, Milk yield and Protein percentage in the Italian Brown

Supplementary Figure S3: GWAS plot

Manhattan plots and QQ-plots respectively for Fat percentage, Milk yield and Protein percentage in the Italian Holstein

Supplementary Figure S4: GWAS plot

Manhattan plots and QQ-plots respectively for Fat percentage, Milk yield and Protein percentage in the HolsteinDGAT

Supplementary Figure S5: GWAS plot

Manhattan plots and QQ-plots respectively for Fat percentage, Milk yield and Protein percentage in the Italian Simmental

Supplementary Figure S6: LD decay

Genome-wide LD decay for the three breeds (Holstein, Simmental, Brown)

Supplementary Figure S7: Descriptive plots

Descriptive statistics for the Italian Brown, Italian Holstein and Italian Simmental.

Supplementary Figure S8: DGAT1 LD

Focus on the LD of the DGAT1 region in all the datasets (Brown, Holstein, HolsteinDGAT, Simmental)

Supplementary Table S1: SNP fact

Number of subjects and SNPs after QC

Supplementary Table S2: Gene Based Association results

Full results of the significant genes associated with the traits in all the datasets.

Sheet 1: Brown – fat percentage
Sheet 2: Brown – milk yield
Sheet 3: Brown – protein percentage
Sheet 4: Holstein – fat percentage
Sheet 5: Holstein – milk yield
Sheet 6: Holstein – protein percentage
Sheet 7: HolsteinDGAT – fat percentage
Sheet 8: HolsteinDGAT – milk yield
Sheet 9: HolsteinDGAT – protein percentage
Sheet 10: Simmental – fat percentage
Sheet 11: Simmental – milk yield
Sheet 12: Simmental – protein percentage

Supplementary Table S3: Single SNP Association results

Full results of the SNP that overcome the suggestive threshold.

Sheet 1: Brown – fat percentage

Sheet 2: Brown – milk yield

Sheet 3: Brown – protein percentage

Sheet 4: Holstein – fat percentage

Sheet 5: Holstein – milk yield

Sheet 6: Holstein – protein percentage

Sheet 7: HolsteinDGAT – fat percentage

Sheet 8: HolsteinDGAT – milk yield

Sheet 9: HolsteinDGAT – protein percentage

Sheet 10: Simmental – fat percentage

Sheet 11: Simmental – milk yield

Sheet 12: Simmental – protein percentage

Supplementary Table S4: Genes and traits

GeneWise p-values of the discussed genes in all the analysis. In grey are evidenced the p-values that overcome the significance threshold. Genes are alphabetically listed.

Supplementary Table S5: Genes in QTLs

Genes from the gene-based association mapping into known QTLs.

CHAPTER 4

MUGBAS: a species free gene based association suite

S. Capomaccio[1,*], M. Milanesi[1,*], L. Bomba[1,*] et al.

[1] *Istituto di Zootecnica, Università Cattolica del Sacro Cuore, via Emilia Parmense 84, Piacenza, 29122, Italy*

*These authors contributed equally to this work

Application note submitted to *Bioinformatics*

Abstract

The association studies between genome-wide markers and phenotypes (GWAS) are now widely used in model and non-model species. However, tools that merge single marker results and their probability of being associated to functional units (typically genes) are seldom developed for species other than human.

Here we present MUGBAS, a suite that calculates a gene/region-based p-value from a given annotation, single marker GWA result and genotype data, with the objective to ease post-GWAS analysis. The software is species and annotation independent, fast, highly parallelized, and ready for high-density marker studies.

Availability and implementation: *https://bitbucket.org/capemaster/mugbas*

Introduction

With the availability of high-density SNP panels, Genome Wide Association Studies (GWAS) have become the gold standard for dissecting the biology of complex traits (McCarthy et al., 2008). This is true for humans, and other species. The underlying idea of GWAS is simple: find marker-trait associations exploiting the linkage disequilibrium (LD) that exists between the causative mutation – which we ignore – and one or more markers of known chromosomal position.

While a number of well established approaches can be easily used in a standard GWAS (Aulchenko et al., 2007; Kang et al., 2010; Purcell et al., 2007) to find marker(s) associated to a particular phenotype, several issues have to be accounted for during the up- and down-stream analyses.

Firstly, genome wide studies have to take into account multiple testing. If stringent statistical thresholds are applied, a proportion of moderate or small effects actually associated with the trait will be disregarded. Therefore, reducing the number of tests while maintaining all the information provided by high-density panels would be a clear advantage. In addition, once a significant signal is detected, different strategies may be used to tag the candidate causative gene. Sometimes, the gene closer to the peak signal is selected, in other cases all genes within a flanking region of an arbitrary size of the significant SNP(s) are included in further analyses. These choices may end up in bias results, either retaining a large number of not truly associated genes or disregarding the correct ones.

Some methods have recently been proposed to overcome these limitations using *a priori* knowledge, as the "candidate pathway" analysis (Raven et al., 2013), and the "gene-based" association strategies (Akula et al., 2011; Cantor et al., 2010; Liu et al., 2010). These approaches restrict GWA studies only to known genes or gene subsets, resizing the multiple testing problem and increasing the power of the analyses.

Usually, bioinformatics pipelines are developed for the analysis of human and model organisms, and are not adapted to non-model organism. This is the case of the VEGAS software (Liu *et al.* 2010, Yang *et al.*, 2014), developed only for humans and only for gene-based analyses.

Here, we introduce MUGBAS [MUlti species Gene-Based Association Suite], a software based on VEGAS that uses GWAS data and a given annotation (typically genes but any region may be suitable) to estimate gene-based and region-based association p-values. The suite is species and annotation free, ready to analyze high-density data from any experimental designs.

Methods

MUGBAS is a suite of scripts built in Bash, R, Python and Perl. The suite requires several pre-requisites that can be easily fulfilled in a Linux/Unix/Mac environment with a few command lines. Further information on how to meet these requirements are detailed in the online repository.

MUGBAS suite can be invoked launching the Bash wrapper that checks dependencies and stores user choices.

Required input files are: MAP/PED files in PLINK format with genotype data from (at least 200 individuals are recommended for a proper LD calculation), an annotation file in BED format (chromosome, starting position, ending position, name and user custom fields) of the desired genome feature and the GWAS result file with mandatory headers "SNPNAME" and "PVALUE". Multiple association results can be tested at the same time, if provided in separate files in the same directory. A trial dataset is downloadable along with the software release. Please note that for simplicity we provide a gene annotation, but any BED file with any desired feature can be used

The program starts the analysis assigning SNPs to any feature of the given annotation using bedtools (Quinlan and Hall, 2010). This step is customizable by user defined boundaries (up- and down-stream) in order to catch regulatory regions that could be in high LD with the current gene/region.

In order to speed up the analysis, the program subsamples 200 individuals from the input dataset. We then implemented two levels of parallelization: one for the LD and one for the "gene-wise" or "region-wise" p-value calculation. The latter is particularly useful analyzing more than one GWA results from the same dataset.

Briefly, MUGBAS first splits the LD calculation in *n* cores (as defined by the user) using the *foreach* / *doParallel* packages (Analytics and Weston, 2014a, 2014b) and then distributes traits across cores using GNU Parallel and *foreach* / *doParallel*.

Since LD calculation is a slow process, VEGAS benefits of HapMap phase 2 data to speed up the process while preserving the option for a user defined (human) population using PLINK. In order to expand these analysis to any species, LD calculation in MUGBAS is implemented with the *r2fast()* function from *GenABEL* package (Aulchenko et al., 2007), while the gene-wise p-value is calculated as in the VEGAS approach. Briefly, the gene-wise test statistic is the sum of the *n* upper tail chi-squared values of a SNP subset with one degree of freedom (where *n* is the number of SNP mapping in the given gene). With perfect linkage equilibrium, the gene-wise p-value would be the one-tailed p-value of a chi-square distribution with *n* degrees of freedom. Otherwise, more likely, the distribution of the p-values has to be evaluated (with simulations) and weighted (with LD values) (Liu et al., 2010).

Once every gene/region has its "gene-wise" p-value, a FDR q-value (False Discovery Rate) statistics is calculated with the base R function *p.adjust()*. Results are stored and enriched with "manhattan" and "underground" plots of the "gene-wise" p-value and the q-value , respectively.

Results

MUGBAS output is gene/region-oriented with useful information about the location of the analyzed feature, the “Best SNP” within that feature with its genomic coordinates and the original p-value, the gene- or region-wise statistics (p-value, number of simulations and q-value), the custom annotation of that particular entry and the position of the SNP with respect to the feature and the decided boundaries. All this information are useful to track the effect of highly significant markers that could map into multiple region of interest and effectively pinpoint the most probable location to focus on with data mining processes.

A typical analysis, performed with a high-density panel (>600K markers), distributed in 10 mid-performance cores (Intel Xeon X5675 @ 3.07 GHz), on three traits simultaneously takes about four hours.

Discussion

Gene/region-based methods are now widely used and accepted in genomic research. However, non model-species suffer for the lack of availability of specific tools to enhance discovery from genome-wide data. MUGBAS provides a robust and accepted statistics to researchers dealing with species other than human to easily jump from associated markers to the desired functional units.

Acknowledgements

We would like to thank Dr Jimmy Liu for his advices in building the suite.

Reference

Akula, N., Baranova, A., Seto, D., Solka, J., Nalls, M.A., Singleton, A., Ferrucci, L., Tanaka, T., Bandinelli, S., Cho, Y.S., Kim, Y.J., Lee, J.-Y., Han, B.-G., Bipolar Disorder Genome Study (BiGS) Consortium, The Wellcome Trust Case-

Control Consortium, McMahon, F.J., 2011. A Network-Based Approach to Prioritize Results from Genome-Wide Association Studies. PLoS ONE 6, e24220. doi:10.1371/journal.pone.0024220

Analytics, R., Weston, S., 2014a. doParallel: Foreach parallel adaptor for the parallel package.

Analytics, R., Weston, S., 2014b. foreach: Foreach looping construct for R.

Aulchenko, Y.S., de Koning, D.-J., Haley, C., 2007. Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. Genetics 177, 577–585. doi:10.1534/genetics.107.075614

Cantor, R.M., Lange, K., Sinsheimer, J.S., 2010. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am. J. Hum. Genet. 86, 6–22. doi:10.1016/j.ajhg.2009.11.017

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S., Freimer, N.B., Sabatti, C., Eskin, E., 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42, 348–354. doi:10.1038/ng.548

Liu, J.Z., Mcrae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., 2010. A versatile gene-based test for genome-wide association studies. Am. J. Hum. Genet. 87, 139.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., Hirschhorn, J.N., 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9, 356–369. doi:10.1038/nrg2344

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. doi:10.1086/519795

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. doi:10.1093/bioinformatics/btq033

Raven, L.-A., Cocks, B.G., Pryce, J.E., Cottrell, J.J., Hayes, B.J., 2013. Genes of the RNASE5 pathway contain SNP associated with milk production traits in dairy cattle. Genet. Sel. Evol. 45, 25. doi:10.1186/1297-9686-45-25

CHAPTER 5

Imputation accuracy is robust to cattle reference genome updates

M. Milanesi[1], et al.

[1] Istituto di Zootecnica, Università Cattolica del Sacro Cuore, via Emilia Parmense 84, Piacenza, 29122, Italy

Short Communication accepted in *Animal Genetics*

Summary

Genotype imputation is routinely applied in a large number of cattle breeds. Imputation has become a need, due to the large number of SNP arrays with variable density (currently, from 2,900 to 777,962 SNPs). Although many authors have studied the effect of different statistical methods on imputation accuracy, the impact of a (likely) change in the reference genome assembly on imputation from lower to higher density has not been determined so far. In this work, 1,021 Italian Simmental SNP genotypes were re-mapped on the three most recent reference genome assemblies. Four imputation methods were used to assess the impact of an update in the reference genome. As expected, the four methods behaved differently, with large differences in terms of accuracy. Updating SNP coordinates on the three tested cattle reference genome assemblies determined only a slight variation on imputation results within method.

Keywords:

Imputation, reference genome assembly, SNP, bovine

Main text

From the publication of the first bovine single nucleotide polymorphism (SNP) array, the whole genomic scenario of this species evolved rapidly. Nowadays several SNP panels have been designed, with a range of densities, mapping on different reference genome assemblies and produced by different manufacturers. When running analyses with SNP arrays of different densities, or from different commercial companies (*e.g.* Illumina and Affymetrix) an imputation step is often required. The impact of different statistical methods on imputation accuracy has already been assessed (Pei et al., 2008); (Marchini and Howie, 2010); (Ma et al., 2013). Surprisingly, up to now, there are not available studies investigating the impact of an update on the reference genome assembly on imputation accuracy. This assessment would be of fundamental importance, considering that re-mapping SNPs determines a rearrangement of the haplotype structure of the genotypes (in some cases SNPs are re-mapped to a different chromosome), and that the bovine reference assembly is likely to be repeatedly updated over time. This work was aimed at assessing the effect of different reference genome assemblies in cattle. As test case, we compared four different imputation methods from low (Illumina BovineLDv.1.0 Beadchip) to medium/high (Illumina BovineSNP50v.2 Beadchip) density in the Italian Simmental population. For each method we assessed the impact on imputation accuracy of mapping SNPs on the BTAU4.2, BTAU4.6 (www.hgsc.bcm.edu/other-mammals/bovine-genome-project) or UMD3.1 (Zimin *et al.* 2009) cattle reference genome assemblies.

A total of 900 and 121 Simmental bulls were genotyped with the Illumina BovineSNP50v.1 and v.2 Beadchips (Illumina Inc., San Diego, CA), respectively. Illumina BovineSNP50v.2 Beadchip (hereinafter named 54k), the official reference chip for the Italian Simmental breeders association (ANAPRI), was considered as reference SNP panel. The SNPchiMp v.1 tool (Nicolazzi et al., 2014) was used to obtain SNP coordinates on BTAU4.2, UMD3.1 and BTAU4.6 genome assemblies. For each assembly, SNPs not contained in the 54k, unmapped or in the sexual chromosomes, were discarded. The final dataset contained 51,189, 52,886 and 51,290 SNPs in BTAU4.2, UMD3.1 and BTAU4.6, respectively.

Four imputation methods were considered: PedImpute (Nicolazzi et al., 2013), FindHap v.2 (VanRaden et al., 2011), FImpute (Sargolzaei *et al.* 2014) and Beagle v.3.3.2 (Browning and Browning, 2007). The first three use pedigree and population (*i.e.* linkage disequilibrium; LD) information, whereas the fourth uses only LD information. Performance of the four methods tested across genomic assemblies was compared using wrongly imputed alleles ("%errors") and allelic squared correlation ("allelic-R2") statistics. Allelic-R2 was defined as the squared correlation between imputed and true (*e.g.* genotyped) alleles multiplied by 100, as in Browning & Browning (2009). The 878 oldest sires were used as training population, whereas the 143 youngest bulls were used as test population. SNPs not in common with the low density panel were set to missing in the test population. The test population was split in four scenarios: individuals with sire and maternal grandsire genotyped in the reference population (scenario A - 84 animals), individuals with only the sire genotyped (scenario B - 34 animals), individuals with only the maternal grandsire genotyped (scenario C - 13 animals) or none of close relatives genotyped (scenario D - 12 animals). The population used was highly structured, with nearly half of the bulls present in the dataset (490) sired by 115 genotyped sires. In addition, 35 out of these 115 genotyped sires were sired by 22 bulls (*e.g.* grandsires), also present in the dataset.

Differences across assemblies are not homogeneous. From BTAU4.2 to BTAU4.6, 46 SNPs were mapped to different chromosomes and the average difference in physical position within chromosomes was nearly 0.25 Mbp. However, there was no substantial difference in the order of markers. In fact, not considering unmapped SNPs on either of the aforementioned assemblies, a difference in the sequential order was observed for only 134 single (*i.e.* spread throughout the genome) SNPs. On the contrary, from UMD3.1 to BTAU4.6, 222 SNPs were mapped to different chromosomes, and a much larger average difference in SNPs physical position was observed (*i.e.* more than 2Mbp). Also the number of SNPs with different sequential order was much larger (*i.e.* 1893 SNPs). However, more than 50% of the rearrangements involved blocks from 3 to nearly 100 markers. As a result, slight differences in overall imputation accuracy from BTA4.6 to UMD3.1, rather than between the two BTAU assemblies, were expected across methods. On the contrary, if rearrangements are due to issues in the assembly of scaffolds, they might have a direct (negative) impact on local imputation performances. In fact, a preliminary analysis on BTA1, considering local %error across assemblies and methods, suggest at least two such SNPs clusters at 44 and 138Mbp on UMD3.1 (Figures S1, S2 and S3).

Considering all methods and scenarios, a maximum average variation of 0.2 in %errors (Table 1) and of 0.9 in allelic-R2 (Table 2) was observed across reference assemblies. The standard deviation (SD) of the imputation accuracy was stable across reference assemblies within methods. On the contrary, a larger variability was observed across methods.

PedImpute and FindHap showed a relatively large average allelic-R2 variability across scenarios, evaluated over the same reference assembly. Only PedImpute showed a large variation in SD across scenarios in %errors and allelic-R2, with high variability on scenarios C and D, showing a low performance of this method when there are no close relatives (*e.g.* sire/dam) genotyped in the reference dataset. FImpute, instead, achieved the highest accuracy across all reference genomes and scenarios, surprisingly obtaining its best results - % errors of 1.4 (0.6 SD), and allelic-R2 of 95.5 (2.6 SD) - in scenario C, where only one relatively close relative is genotyped (maternal grand-sire). FImpute appear to be able to detect shorter haplotypes from distant relatives more accurately than any other method tested in this study; thus, it is less dependent on the availability of pedigree information than the other two pedigree-based methods (Sargolzaei *et al.* 2010).

The large variability across methods within assembly was expected, although the differences obtained among PedImpute, FindHap and Beagle were larger in this study compared to a similar study in Italian Holstein (Nicolazzi *et al.* 2013). This could depend on a number of factors. First, the sample size of the test population is small (especially for groups C and D, where less than 15 individuals were present). Second, the interaction between population structure and imputation methods may have a role. PedImpute and FindHap rely much more heavily on pedigree information than FImpute, and Beagle assumes all individuals are unrelated. However, the latter two methods indirectly benefit greatly of from highly structured populations with large number of relatives – close or distant – present in the reference population (Sargolzaei *et al*. 2014). Another interesting factor that influences the imputation accuracy may be the persistence of LD at long distances: this is much lower in the Italian Simmental than in the Italian Holstein population (Figure S4). Information on relatives (either pedigree information or genotyped relatives in the reference population) is expected to play a bigger role if LD is conserved at shorter distances. This hypothesis is supported by the fact that, although there is high variability across methods, better performances of the pedigree-based methods are still observed in scenarios A and B (and C for FImpute).

The slight differences in imputation accuracy observed across reference assemblies may be explained by the aforementioned small variation in the SNP blocks across reference assemblies, irrespectively of the rearrangements of single SNPs. Considering that SNP blocks are mostly maintained, we expect that breeds with larger LD persistence at longer distances (*i.e.* Holstein, Brown) will obtain even better imputation accuracy across assemblies, compared to what reported in this study in Italian Simmental. However, a still relatively unexplored aspect of imputation performance is the distribution of imputation errors across the genome. Not considering genotyping errors (which are expected to be randomly distributed in the genome), a cluster of consecutive markers with high percentage of imputation error across methods are most probably identifying miss-assembled regions in the genome (Figures S1,S2 and S3). Although the impact on overall imputation accuracy is limited, as results in this study show, these errors might have a large impact on any analyses relying on specific genomic coordinates (e.g. genome-wide association studies). More comprehensive results, including annotation studies and whole genome sequencing data are required to confirm and investigate these observations.

To conclude, only a small effect of updating the reference genome assembly on imputation accuracy was observed in Italian Simmental. On the other hand, as already reported by other studies, the variability of imputation accuracy estimates is much larger across imputation methods. The choice of the most suitable imputation method, based on the breed genetic background, its characteristics, and on the available data, is therefore critical.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 289592 –

Gene2Farm. Part of the genotypic data used in this study was provided by project "SelMol" funded by the Italian Ministry of Agriculture (MiPAAF). MM was supported by the Doctoral School on the Agro-Food System (Agrisystem) of the Università Cattolica del Sacro Cuore (Italy). FB was supported by the Marie Curie European Reintegration Grant "NEUTRADAPT" (FP7-PEOPLE-2009-RG).

References

Browning, S.R., Browning, B.L., 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet 81, 1084–1097.

Ma, P., Brøndum, R.F., Zhang, Q., Lund, M.S., Su, G., 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. J. Dairy Sci. 96, 4666–4677. doi:10.3168/jds.2012-6316

Marchini, J., Howie, B., 2010. Genotype imputation for genome-wide association studies. Nat Rev Genet 11, 499–511. doi:10.1038/nrg2796

Nicolazzi, E.L., Biffani, S., Jansen, G., 2013. Short communication: Imputing genotypes using PedImpute fast algorithm combining pedigree and population information. Journal of Dairy Science 96, 2649–2653. doi:10.3168/jds.2012-6062

Nicolazzi, E.L., Picciolini, M., Strozzi, F., Schnabel, R.D., Lawley, C., Pirani, A., Brew, F., Stella, A., 2014. SNPchiMp: a database to disentangle the SNPchip jungle in bovine livestock. BMC Genomics 15, 123. doi:10.1186/1471-2164-15-123

Pei, Y., Li, J., Zhang, L., Papasian, C.J., Deng, H., 2008. Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE 3551.

VanRaden, P.M., O'Connell, J.R., Wiggans, G.R., Weigel, K.A., 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol 43.

Table 1. Imputation accuracy across assemblies for PedImpute, FindHap, FImpute and Beagle: percentage of wrongly imputed alleles.

		PedImpute						FindHap						FImpute						Beagle					
		BTAU 4.2		UMD 3.1		BTAU 4.6		BTAU 4.2		UMD 3.1		BTAU 4.6		BTAU 4.2		UMD 3.1		BTAU 4.6		BTAU 4.2		UMD 3.1		BTAU 4.6	
Scenario	N	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A[1]	84	2,0	0,7	2,1	0,7	2,1	0,7	3,2	0,9	3,2	0,9	3,2	0,9	1,5	0,9	1,5	0,9	1,5	0,9	2,4	1,1	2,5	1,1	2,5	1,1
B[2]	34	2,3	0,8	2,3	0,9	2,2	0,8	3,6	1,1	3,6	1,2	3,6	1,1	1,5	0,7	1,5	0,9	1,4	0,6	2,3	0,8	2,3	0,8	2,3	0,8
C[3]	13	9,8	4,1	9,8	4,1	9,8	4,1	5,1	1,0	5,2	1,0	5,1	1,0	1,4	0,6	1,4	0,6	1,4	0,6	2,3	0,6	2,3	0,6	2,4	0,6
D[4]	12	8,8	4,9	8,9	4,9	8,8	4,9	5,4	1,1	5,6	1,2	5,4	1,1	1,6	1,1	1,7	1,2	1,6	1,1	2,6	1,5	2,6	1,5	2,6	1,4

[1] Sire and maternal grandsire genotyped

[2] Sire genotyped, maternal grandsire not genotyped

[3] Sire not genotyped, maternal grandsire genotyped.

[4] Sire and maternal grandsire not genotyped.

Table 2. Imputation accuracy across assemblies for PedImpute, FindHap, FImpute and Beagle: allelic-R2.

		PedImpute						FindHap						FImpute						Beagle					
		BTAU 4.2		UMD 3.1		BTAU 4.6		BTAU 4.2		UMD 3.1		BTAU 4.6		BTAU 4.2		UMD 3.1		BTAU 4.6		BTAU 4.2		UMD 3.1		BTAU 4.6	
Scenario	N	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A[1]	84	94,1	2,0	94,0	2,0	94,1	2,0	90,7	2,4	90,8	2,5	90,7	2,4	95,2	2,6	95,2	2,7	95,2	2,7	92,5	3,3	92,3	3,3	92,5	3,3
B[2]	34	93,5	2,4	93,4	2,5	93,5	2,4	89,6	3,0	89,6	3,4	89,5	3,2	95,4	2,0	95,4	2,1	95,4	2,0	92,9	2,4	92,8	2,5	92,8	2,4
C[3]	13	75,9	8,9	76,1	8,8	75,8	8,7	84,8	3,3	84,5	3,0	84,8	3,2	95,5	1,8	95,5	1,9	95,5	1,9	93,1	1,9	92,8	1,9	92,7	1,9
D[4]	12	78,2	11,3	77,7	11,3	78,1	11,2	84,1	3,2	83,2	3,5	83,9	3,3	94,9	3,4	94,7	3,8	94,9	3,5	92,1	4,4	91,9	4,5	92,1	4,4

[1] Sire and maternal grandsire genotyped

[2] Sire genotyped, maternal grandsire not genotyped

[3] Sire not genotyped, maternal grandsire genotyped.

[4] Sire and maternal grandsire not genotyped.

Supplementary files

You can find Chapter 5 supplementary files at:

- DocTa (Doctoral Thesis Archive) (http://tesionline.unicatt.it/)
- Google Drive (http://tinyurl.com/PhDMMChapter05) or scan the QR code

Supplementary Figure S1: Error percentage across methods for BTAU4.2 on BTA1

Supplementary Figure S2: Error percentage across methods for UMD3.1 on BTA1

Supplementary Figure S3: Error percentage across methods for BTAU4.6 on BTA1

Supplementary Figure S4: Linkage disequilibrium (LD) decay over distance in Italian Simmental and Holstein.
The dotted line with black dots indicates LD decay over distance in Italian Holstein, whereas the dashed line with white diamonds indicates the same values for Italian Simmental. Italian Holstein genotypes were a subset of the data used in Nicolazzi et al. (2013).

CHAPTER 6

Quest for deleterious recessive variants in Italian Holstein bulls combining exome sequencing and high density SNP analysis

M. Milanesi [1], et al.

[1] Istituto di Zootecnica, UCSC, via Emilia Parmense 84, 29122 Piacenza, Italy;

Manuscript

Abstract

Outbreed species carry a genetic load of deleterious recessive variants. Those that are lethal survive only in the heterozygous state in the population while others simply reduce the fitness of individuals homozygous for the variant. The long-term destiny of these mutations is to decrease in frequency due to negative selection and eventually disappear because of genetic drift. However, they may be maintained in livestock populations and even increase in frequency when in linkage with favourable alleles for traits under selection or occurring in the genome of sires having high genetic value for productive traits. We searched for these variants combining exome sequence data from 20 Italian Holstein bulls sampled from opposite tails of the male and female fertility EBV distribution, the High Density (HD) SNP genotyping of 1009 progeny tested Italian Holstein bulls and information existing in public databases. Exome sequencing detected 202,956 variants. Among these, 5,167 were classified as "deleterious" when annotated with SIFT and Annovar. These were filtered by picking those mapping in regions out of equilibrium and lack of homozygotes in the progeny tested Holstein bulls or having an asymmetric distribution in bulls plus-variant and minus-variant for fertility EBV. The resulting 193 variants in 163 genes were further assessed by comparative genomics, haplotype analysis and function. For each assessment a positive or negative score was given to evidence in favour or against the variant to be deleterious. A total of 35 variants had a positive total score. These genes are involved in development, cell motility and immune response. Some of them are known to be expressed in the testicles and appear important for sperm fertility in model species; others are responsible for genetic defects in human and mouse carrying mutations similar to the ones detected in cattle. These gene variants are candidate to be among those that constitute the "genetic load" of deleterious recessive genes in the Holstein population.

Introduction

The presence of deleterious recessive variants is well tolerated by outbred species. When these variants appear they may spread and survive a long time in the population in the heterozygous state (Simmons and Crow, 1977). It is estimated that outbred individuals, human included, carry on average a genetic load of more than 100 loss-of-function variants in their genome (MacArthur et al., 2012).

The effect of deleterious recessive variants appears when demographic events, e.g. genetic bottlenecks, reduce the genetic variability of the population or when related individuals are intentionally mated for breeding purposes. In these cases, recessive deleterious reduce the fitness of individuals carrying them in the homozygous state and in extreme case induce very severe genetic defects, premature death and abortion (McClure et al., 2014; Sahana et al., 2013; VanRaden et al., 2011). They therefore contribute to explain the phenomenon of inbreeding depression and its negative effect on fertility (Charlesworth and Willis, 2009).

The search for deleterious recessive variants is particularly interesting in Holstein cattle. This breed is experiencing a negative genetic trend for fertility traits (Jamrozik et al., 2005); it is worth noting that this is occurring while inbreeding is under strict monitoring and is theoretically increasing at a very low pace (about 0.05 per year - Mrode et al., 2009). In addition, mating is planned to minimise relationship between animals and the additive genetic variance for most traits – including milk production - is apparently not decreasing, in spite of the selection pressure applied by modern schemes (Biffani et al., 2002), indicating that overall diversity is only slightly decreasing.

Selection to increase fertility is challenging. Selection criteria generally adopted to improve this trait are far from the biological mechanisms that govern reproduction. In addition, these mechanisms are complex and very sensitive to environmental (management) conditions (Lane et al., 2013; Wathes et al., 2014). As a result, fertility traits have low or very low heritability (Pryce et al., 1997) and therefore low response to selection and slow genetic progress.

For these reasons, the quest for molecular variants influencing fertility traits is on-going since the availability of molecular tools for DNA analysis. A number of investigations tested the effect of functional candidate genes running genome wide association analyses to find QTLs for fertility traits using hundreds microsatellite markers in structured families (Ashwell et al., 2004) or tens of thousand SNP panels in whole populations (Aston and Carrell, 2009). The availability of medium density SNP panels allowed the search for deleterious recessives even without the use of phenotypic traits, by scanning the genome in search of regions lacking one of the two homozygous classes (Sahana et al., 2013; VanRaden et al., 2011). This approach has highlighted a number of genomic regions likely carrying deleterious recessives. This information has immediate application in molecular assisted breeding and a relevant scientific interest in the search of causative variants and assessment of their functions.

The development of new sequencing technologies has greatly decreased the cost of sequencing, so that nowadays the sequencing of few target individuals is becoming the gold standard to seek for causal mutations for rare monogenic defects in humans and other species (Chun and Fay, 2009). Often only the exome, the protein coding part of the genome, is sequenced searching for those mutations that severely alter protein structure and/or function (Bamshad et al., 2011; Li et al., 2013). A typical mammalian exome consists in a few tens of million nucleotides (around 2% of the reference genome), compared to the few billion of the whole genome. Therefore, with the same number of reads, exome sequencing has the advantage of a more accurate genotyping due to the higher sequencing depth, an easier data management, lower computation load and easier interpretation (Biesecker et al., 2011; Majewski et al., 2011). Indeed the number variants found in exomes are one-two orders of magnitude lower than that detected in whole genomes and mutations are detected in annotated genes or immediately nearby. On the other hand, the exome approach can only identify variants on annotated genes. Variants in unknown exons or that occur in non-coding regulatory features will be missed (Bamshad et al., 2011). In humans these are turning out to be of outmost importance for gene regulation and downstream phenotypic effects (ENCODE Consortium, 2012), but they are still very poorly annotated in livestock.

Here we sequenced the exomes of 20 Holstein bulls progeny tested selected from the tails of male and female fertility EBVs (Estimated Breeding Value). Since livestock fertility is a complex trait and causative mutations cannot be sought by the approach used in monogenic diseases, we integrated the sequence information with high density SNP chip data obtained on a much larger population of Italian Holstein bulls to detect genes carrying alleles candidate to be deleterious recessives and confirm their likely deleterious effect by comparative genomics.

Materials and methods

1. Sampling and DNA extraction

Semen (straws) and blood samples were provided by Italian certificated artificial insemination centres (UE Directive 88/407/CEE), thus ethics committee approval for this study is not required.

A total of 20 Italian Holstein progeny tested bulls were sampled for complete exome sequencing. Among them, 19 were chosen from the tails of female and male fertility EBV (Estimated Breeding Value) distribution (Table 1), to have 5 animals from each tail. One animal was to be in the minus-variant tail of both traits. Bulls were selected as unrelated as possible. Sampling was possible thanks to the collaboration with ANAFI (Italian Holstein Breeder Association).

ANAFI developed a genetic evaluation for fertility combining different traits, as explained by Biffani and colleagues (Biffani et al., 2005). The traits are: “days from calving to first insemination”, “calving interval”, “first-service non return rate to 56 days”, “angularity” and “mature equivalent milk yield at 305 days”.

Male fertility is expressed as an Estimated Relative Conception Rate (ERCR) and is a measure of conception rate of a service sire relative to service sires of herd mates (Soggiu et al., 2013).

Table 1. Female and male fertility EBVs of the 20 animals sequenced

Animal	EBV Female Fertility	EBV Male Fertility	Group
fri-01	97	-1,11	NA
fri-02	98	-1,66	MalL
fri-03	115	0	FemH
fri-04	114	0	FemH
fri-05	114	0	FemH
fri-06	86	0	FemL
fri-07	98	1,95	MalH
fri-08	105	2,13	MalH
fri-09	95	-1,54	MalL
fri-10	95	1,49	MalH
fri-11	102	2,64	MalH
fri-12	91	-4,75	FemL; MalL
fri-13	105	-1,79	MalL
fri-14	102	2,07	MalH
fri-15	99	-4,05	MalL
fri-16	85	0	FemL
fri-17	87	0	FemL
fri-18	121	0	FemH
fri-19	117	0	FemH
fri-20	88	0	FemL

A total of 1009 Italian Holstein bulls were chosen from the entire population balancing their presence in the tails of the EBV for kg protein (PROT), somatic cell count (SCC), fertility (FERT) and functional longevity (LONG), for high density genotyping (Table S1).

DNA was isolated from whole blood and semen straws samples in LGS laboratory (Cremona, Italy, www.lgscr.it/) using the Genomix kit, according to LGS internal protocol. DNAs for exome capture and sequencing were extracted independently, to satisfy the requirement indicated by IGA technologies (Udine, Italy): at least 12mg of DNA at a concentration 100 ng/ul with R260/280 higher than 1.8 and R260/230 near 1.9.

2. Molecular data production and quality checking

2.1 Exome capture and sequencing

Exomes were captured using the “SureSelected All exon Bovine” (Agilent Technologies) kit, in early version. Probes coordinates are based on UMD3.1 bovine reference sequence (Zimin et al., 2009) and designed against NCBI exons as well as predicted exons, UTRs, and miRNAs. Number and distribution of probes per chromosome are detailed in Table S2.

”Ovation Ultralow Library Systems” protocol was followed for library preparation. After quantification with Qubit (Life technologies) and quality control with 2100 Bioanalyzer (Agilent Technologies), DNA was sonicated with Bioruptor (Diagenode) to obtain fragments between 150 to 200 bp in length, blunt end-repaired, adenylated, ligated with paired-end adaptors and purified following the protocol provided by the manufacturer.

Libraries were hybridized with “SureSelected All exon Bovine” probes using the Encore Target Capture Module (Nugen), SureSelect Target Enrichment and Herculase II Fusion Enzyme with dNTP Combo (Agilent Technology) kit. Then, captured DNA was amplified and tagged with library specific indexes. Quality of captured samples was tested with Qubit and 2100 Bioanalyzer.

Paired-end 2X100bp sequencing was performed on Illumina HiSeq2500 (Illumina, San Diego, CA) at an expected coverage of 50X.

2.2 Sequence alignment

Adapters and reads shorter than 36bp were trimmed using Trimmomatic (version 0.32) (Bolger et al., 2014). For each Fastq file, the alignment and pairing of the reads to the UMD3.1 reference sequence was performed with Burrows-Wheeler Aligner (BWA) (version 0.7.7-r441) (Li and Durbin, 2009), applying also trimming option (-q 5). SAM files resulted were converted into BAM file using SAMtools (version 0.1.18) (Li et al., 2009). For each individual, BAM files were merged in a single one with Picard (version 1.111) (https://github.com/broadinstitute/picard). Duplicate reads were identified and marked with Picard. Finally individual BAM files were indexed with SAMtools.

Sequence depth was evaluated simultaneously on all 20 sequenced animals and only on regions targeted by probes using *DepthOfCoverage* of GATK (McKenna et al., 2010)

2.3 Exome Variant calling and quality check

Single nucleotide polymorphisms (SNPs), insertions and deletions (InDels), were called simultaneously on all 20 sequenced animals and only on regions targeted by probes using *UnifiedGenotyper* of GATK (DePristo et al., 2011).

All variants were filtered using *VariantFiltration* of GATK to remove ones with low mapping and/or sequencing quality. The choice of filters and thresholds was made in accordance to "GATK best practice" for exome sequencing and distribution of filter values:

- "SNP cluster": excluded variant if more than 3 variants were found in 10 bp;
- "Base Quality Rank Sum Test": excluded if it's lower then -5 or more than 5;
- "Mapping Quality Rank Sum Test": excluded if it's lower then -7 or more than 6;
- "Read Position Rank Sum Test": excluded if it's lower then -4 or more than 4;
- "Fisher Strand": excluded if it's more than 30;
- "Depth": excluded if it's lower then 60X (i.e. a minimum value of 3X per animals) or more than 3500X;
- "Quality": excluded if it's lower then 40;
- "Mapping quality": excluded if it's lower then 30.

Filters were applied to all variants but only bi-allelic SNPs were used in the following analyses. Genotypes with "Genotype Depth" lower than 5 and "Genotype Quality" lower than 20 were discarded. The remaining ones were converted into "AB" code (A = reference allele and B = alternative allele). Then data were transformed in PLINK (Purcell et al., 2007) format, using a home-made python script. To prepare the working dataset the final quality control removed variants with more than 25% missing data (that is with more than 5 animals out of 20 without genotypes), monomorphic (minor allele frequency, MAF < 1%) or not mapping in autosomes. Basic population genetics parameters were calculated to evaluate the presence of cryptic structure in the data that may influence the outcome of the downstream analyses.

In total 24,390 SNPs identified in the exomes were also included in the HD SNPchip panel. HD genotype from nineteen out to twenty of the sequenced animals sequenced was used to compare technologies. Genotypes at common locations were compared across technologies, to evaluate sequence quality and the effect of sequence depth on genotype calling.

2.4. Bovine HD Genotyping BeadChip data and quality check

All animals were genotyped using the BovineHD Genotyping BeadChip (Illumina, San Diego, CA). Genotypes were quality controlled and filtered with standard threshold. Markers with more than 5% missing data and MAF < 1% were excluded. Animals with more than 5% missing data were also excluded. Only autosomal SNPs were retained. Thereafter genotypes were phased and the remaining missing data imputed using Beagle v.3.3.2 (Browning and Browning, 2007).

Of the 20 animals sequenced, 12 had also Bovine HD Genotyping BeadChip and 7 BovineSNP50 Genotyping BeadChip version 2 genotypes available. For concordance and haplotype analysis we imputed the 50K data to 800K as previously described above.

3. Strategy for detecting candidate deleterious variants

We searched for candidate deleterious variants following an integrated approach that joined the analysis of: i) exomes of a few Italian Holstein bulls having plusvariant and minusvariant EBVs for a target trait; ii) HD genotyping of a population of 1009 progeny tested Italian Holstein bulls; iii) the comparative genomic analysis of 100 vertebrate species. The strategy followed for the analysis is shown in Figure 1. In the first step (Search) we looked for candidate deleterious variants in the exomes. In step 2 (Filter) we filtered candidates by assessing which of these variants either had a non-random distribution among animals having extreme EBVs or co-mapped with genomic regions carrying outlier signals in the large population of progeny tested bulls. In step 3 (Asses) candidates were further evaluated by: i) assessing their conservation across a panel of 100 vertebrate species, ii) searching for associated haplotypes and evaluating their frequencies in the population of progeny tested bulls; iii) evaluating their association with genetic defects in human and model species; iv) examining their expression profile and function, when known. Methods are below described in detail.

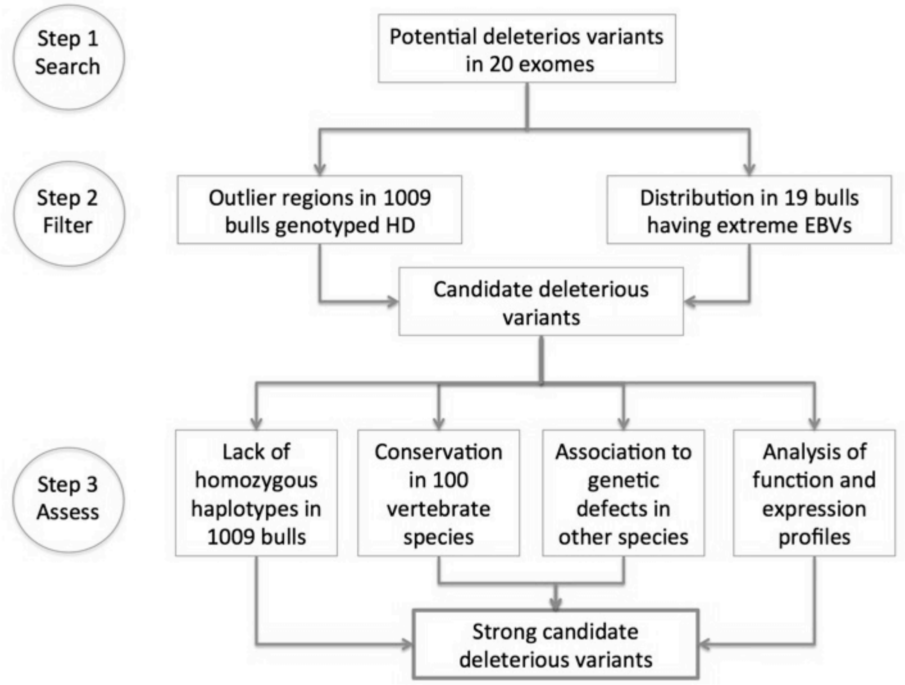
Step 1
Search
Potential deleterios variants in 20 exomes
Step 2
Filter
Outlier regions in 1009 bulls genotyped HD
Distribution in 19 bulls having extreme EBVs
Candidate deleterious variants
Step 3
Assess
Lack of homozygous haplotypes in 1009 bulls
Conservation in 100 vertebrate species
Association to genetic defects in other species
Analysis of function and expression profiles
Strong candidate deleterious variants

Figure 1. Strategy followed to identify deleterious variants

3.1 Step 1: Search for candidate deleterious variants

Quest for candidate deleterious variants in exome sequence data by variant annotation

We annotated all variants using ANNOVAR (Wang et al., 2010) and VEP (Variant Effect Predictor) (McClure et al., 2014) software. They use the Ensembl gene annotation database (UMD3.1 reference sequence, release 74) to investigate location and potential effect on coding sequences of all variants. VEP integrates the SIFT algorithm (Kumar et al., 2009) to predict the effect of amino acid substitution on protein function. For each sequence change, sequence homology and physic-chemical similarity between the reference amino acid and the alternative is considered providing a score that evaluates the mutation tolerability. Variants with score lower than 0.05 were classified as "deleterious", in according to software recommendations.

SIFT also compared all variants found in the exomes with those present in the dbSNP database (version 140), to identify those that were novel.

3.2 Step 2: Filtering of candidate deleterious variants

2a. Analysis of deleterious variant distribution among animals having extremes EBV values for male and female fertility traits

Association between variants and the classification in the tails of the EBV distribution was estimated by *Fisher exact test* which compares frequencies of alleles (a vs. A) in the two tails not assuming H-W equilibrium; *genotypic test*, that compares the frequency of genotypes (aa vs. Aa vs. AA) and the *recessive gene action test*, which tests for the distribution of one homozygous class versus the others (aa vs aA, AA). All tests are implemented in PLINK functions. The 5% significance threshold was set empirically by permutation (N=100,000) either point-wise, that is not corrected for multiple testing (EMP1 = Empirical Probability 1) and thereafter family-wise, corrected for multiple testing (EMP2 = Empirical Probability 2).

Genes exceeding EMP1 were considered suggestive and those exceeding EMP2 significant candidates to be deleterious recessives and their characteristics conservation across species and function further investigated.

2b. Quest for regions candidate to carry deleterious variants in the population analysed with the HD BeadChip.

Basic population genetics parameters and Multi Dimensional Scaling were calculated with GenABEL R package (Aulchenko et al., 2007) to evaluate the presence of cryptic structure in the data that may influence the outcome of the downstream analyses.

Expected and observed allelic and genotypic frequencies and Fisher exact test of Hardy-Weinberg proportions (Janis E Wigginton, 2005) were calculated with PLINK on the HD SNP dataset. Values were averaged in sliding windows of 10 markers. This value (10 markers) was chosen after testing different window size as represent a good compromise between noise reduction and resolution (not shown).

Candidate regions containing deleterious mutations were identified following two different approaches.

In one approach, hereafter refereed to as "Out of Hardy-Weinberg" approach (OHW), markers significantly deviating from of Hardy-Weinberg equilibrium at $P \leq 8.47 * 10^{-8}$ (nominal value $P \leq 0.05$ Bonferroni corrected for multiple testing) were first identified. Windows with at least three significant markers below this threshold and having a homozygote deficit were considered candidates to carry deleterious mutations. In this OHW approach we applied very stringent constraints to reduce false positive discoveries due to the Wahlund effect induced by population substructure (see results and Figure S5).

In the second approach, hereafter referred to as "Lack Of Homozygotes" (LOH) approach, the differences between observed and expected frequencies of homozygotes for the minor allele were computed and averaged across markers included in each sliding window. Values were standardized and only regions carrying three overlapping windows below -3Z score value were considered candidates to carry deleterious mutations. Thereafter, significant windows carrying at least one marker in common were joined to define region boundaries.

Exome sequencing knowledge and HD BeadChip data were crossed in order to identify genes carrying severe variants located within significant regions or their proximity (50Kb upstream or downstream). These genes were considered candidates to be deleterious recessives and their characteristics conservation across species and function was further investigated.

3.3. Step 3. Assessment of candidate deleterious variants

Genes carrying variations classified as deleterious by SIFT or ANNOVAR and either as suggestive by fertility EBV tails analysis or as significant by OHW e LOH analyses of the Holstein bull population were submitted to comparative genomics, population genetics and functional analyses. In order to prioritize genes/variants discovered, we scored each evaluation using a subjective scale ranging from -3 to 3 to measuring the evidence in favour (scores 1 to 3) or against (scores -1 to -3) the deleterious effect of the SNPs. The score 0 in all cases indicated the absence of evidence or that a specific assessment failed. The complete score scale therefore ranged from -9 to 9.

3a. Comparative genomics

For comparative genomics purposes, the position of each mutation was "lifted over" from UMD3.1 bovine coordinates to hg19 human coordinates, using the specific tool at UCSC (Kent et al., 2002 - http://genome.ucsc.edu/). Orthologous proteins were aligned using the UCSC genome browser and amino acid conservation assessed in a set of 100 vertebrate species that included human, primates (11 species), euarchontoglires (e.g. mouse, 14 species), laurasiatheria (e.g. sheep; 25 species); afrotheira (e.g. elephant; 6 species); other mammals (e.g. platypus; 5 species); birds (e.g. chicken; 14 species); sarcopterygii (e.g. turtle; 8 species); fish (e.g. zebrafish; 16 species). Results of amino acid conservation analysis were recorded only if homologous protein from at least 50 species could be retrieved and aligned. Thereafter amino acid conservation was scored 3 if the amino acid was conserved in all vertebrates, 2 if conserved in all mammals and 1 if conserved in the large majority (at least 90%) of mammals. A score of -1 was given when the amino acid was not conserved among species and a score -3 when both variants observed in cattle were carried by other species. Finally when either the lift over process or the protein alignment failed, the comparison was considered non informative and given a score 0.

3b. Search of haplotypes associated to deleterious variants and analysis of HD population data

To acquire further evidence, we searched for the haplotypes that carry the deleterious variations and estimated their frequency in the homozygous/heterozygous status within the bull population genotyped with the HD BeadChip.

We used a multistep approach to find haplotypes carrying deleterious variants: i) from the phased HD BeadChip data we extracted the haplotypes of the 20 sequenced animals in the region of the variant (100 markers on each side); ii) in these 20 animals we searched the shorter haplotype able to distinguish homozygous and/or heterozygous animals for the deleterious variant from all other haplotypes; iii) we assessed the frequency of homozygotes and heterozygotes for the haplotype in the whole population of bulls characterised with the HD BeadChip.

Scores in this case were -3 in the case of unique haplotypes showing a number of homozygotes non significantly different from the number expected from Hardy-Weinberg proportions; 3 in the case of unique haplotypes showing a number of homozygotes significantly lower than expected; 1 in the case of unique haplotypes never homozygous and rare; 0 whenever no unique haplotype could be found.

3c. Functional analyses

With PantherDB (Mi et al., 2010) resources, we classified our gene list into functional categories according to the Gene Ontology classification. All suggestive and candidate Ensembl Gene ID submitted to the bulk analysis were recognized and annotated. Beyond the GO term characterization, in order to understand potential detrimental effects of mutations and also pinpoint potential candidates for fertility traits, we retrieved information for each gene according to the following parameters: i) the involvement in known Mendelian or complex diseases from OMIM database ii) the availability of knock-out mice iii) and the eventual expression in reproductive tissues from the Gene Expression Atlas (http://www.ebi.ac.uk/gxa/) (Table 4).

Also for these data we used the same scale attributing a range from -3 points to genes having no relation with fertility, viability and genetic defects and no relevant phenotypic effects in knock-out mouse models, to 3 to gene closely related to fertility and viability, associated to genetic defects in human and affecting fertility or survival when knocked-out. In particular a score ranging -1 to +1 was attributed to genes associated to human syndromes (-1=no association recorded in OMIM database, 0=no information in OMIM, 1=association in OMIM)**.** A score with same range to information collected from knock-out mouse phenotypes (-1=knock-out mouse shows no impaired phenotype; 0=no knockout mouse has been produced; 1= knock-out mouse shows impaired phenotype). The same score was attributed to molecular functions (-1=function unrelated to fertility and viability; 0=function indirectly related, as immunity and metabolism; 1=function directly related, as fertility and development).

Results

Exome capture and sequencing

A total of 225414 probes were used for capturing the cattle exome (53.55 Mb, about 2% of UMD3.1 reference genomes). Mitochondrial (MT) and Y chromosome were not included in probe design. The number of probes per chromosome spanned from 2,500 on BTA27 to almost 12,000 in BTA3 (Table S2).

The average exome sequencing depth was 40.58, slightly lower than the expected 50X. All animals had at least 24X average genomes sequenced but in a same animal the sequencing depth varied across chromosomes and regions within chromosomes. The lowest chromosome average value observed was 10X in BTA25 of the shallower sequenced animal. Average coverage per animal and chromosome and overall means are shown in Figures S1.

In total, 1,613,461,402 raw 100bp pair-ends sequence reads were produced. A total of 1,425,514,112 100bp reads were mapped and properly paired to the reference genome. Paired reads per animal ranged between 41,973,510 and 138,840,392.

In synthesis, sequencing depth was variable, as expected, and almost all animals had some regions of scarce or no sequence depth, but overall all exome regions were sufficiently covered to permit a comprehensive analysis of molecular variants of the animals investigated.

Step 1. Search for candidate deleterious variants

Variant detection and annotation

The sequencing depth of all variants called on the alignment of all reads across animals followed a Chi-square distribution. Average depth was 796.94, spanning from 2X to 5000X, with peak frequency class at 150-250X (Figure S2).

A multistep approach was followed to clean the exome sequence dataset. Cleaning process steps and the number on markers retained in each step are described in Table 2.

Table 2. Type and number of markers detected and retained after dataset cleaning.

Type of variation	Number of variations (N.)		
	Raw data	After variant cleaning	After genotype cleaning (only autosomal)
Complex	1213	963	0
Deletion *	6938	5031	0
Insertion *	5510	3656	0
Multiallelic Complex	5710	3995	0
Multiallelic SNP	803	512	0
SNP	242988	203079	164277
Total	263162	217236	164277

* when compared to the reference genome

A total of 263,162 variants were identified. Among these 217,236 (82,55%) were retained after applying the variant filters (see materials and methods section).

According to GATK classification, the vast majority (93.48%) of variations identified were biallelic SNPs, followed by insertions and deletions (InDels; 4.0%), more complex variations and multiallelic SNPs (2.52%). Indels and complex variations, although less frequent than SNPs, affected a relevant portion of the exome (49,589 bp). The number of variants detected per chromosome was significantly correlated to the number of probes used (r=0.8761 P=2.28x10^{-10}) and of bases captured (r=0.8424 P=5.3x10^{-19}) per chromosome, however BTA23, BTA27 BTAX had an outlier behaviour, the former two containing higher variation (4.50 and 4.45 variants per Kb) and the latter lower (0.84 variants per Kb) than the average (3.22 variants per Kb).

Bialleic SNPs were further filtered for genotype quality (sequence depth > 5X and sequence quality > 20) and BTAX markers. The final autosomal SNP dataset counted 164,277 SNPs. Among these 17,829 biallelic SNPs (10.9%) described for the first time have been submitted to dbSNP.

Variants' annotation is shown in Table 3a. Variants were detected in different gene domains (exons, introns, 3' untranslated regions) as well as upstream and downstream genes. More than 31% of the variations were detected in exons and among them more than 41% have the potential to have an effect on gene function by affecting splicing or alter protein structure (Table 3b).

Table 3: A) number of autosomal SNPs annotated in each gene region; B) number and classification of autosomal SNPs altering gene function

A)

Gene region	N. of SNPs (autosomes)
Downstream	1170
Exonic	51293
Intergenic	49901
Intronic	58116
ncRNA_exonic	160
ncRNA_intronic	3
ncRNA_splicing	2
Splicing	184
Upstream	1445
Upstream;downstream	54
UTR3	1345
UTR5	604
Total	164277

B)

Effect of exonic SNPs	N. of SNPs (autosomes)
Nonsynonymous single nucleotide variation	20839
Stopgain single nucleotide variation *	213
Stoploss single nucleotide variation *	10
Synonymous single nucleotide variation	30226
Unknown	5
Total	51293

* when compared to the reference genome

Among the 51,293 exonic SNPs, 4,166 in 2,978 genes were classified "deleterious" by SIFT. Among these, 3,210 in 2,356 genes were observed at least twice. ANNOVAR identified 223 stop gain (213) or loss (10) variants in 217 genes.

The average concordance at 24,390 SNPs in common between exomes and the HD Beadchip was 97.16% (Table S3). Discordance was due to either alleles detected by sequencing and not detected by genotyping (exome=heterozygote, Beadchip=homozygote) and vice-versa (exome=homozygote, Beadchip=heterozygote). Discordances were randomly distributed among markers and inversely correlated to sequencing depth (Figure S3), indicating that most errors are likely occurring during sequencing, even if occasional

Beadchip failure in detecting some allele cannot be excluded. Two animals had an outlier behaviour, one having 84.6% and the second 71.9% concordance. Both animals were retained for variant discovery, while they were excluded from the analyses of EBV tails. Without these two outlier animals the mean concordance increases to 99.38%.

Multidimensional scaling confirmed the absence of a substructure in animals sequenced (Figure S4).

Step 2. Filtering of candidate deleterious variants

Analysis of EBV tails

Fisher exact test is generally used to test for the association between a variant and a Mendelian disease. Samples are divided in cases and controls and the association between alleles or genotypes and the disease is tested under different inheritance models. Here we used the same method to test for association between alleles/genotypes and animals from opposite tails of the EBV distribution for i) male fertility (N=4 high + 4 low); ii) female fertility (N=5 high + 4 low) and iii) male and female fertility (N=9 high + 8 low).

Genotypes were tested under all inheritance models implemented in PLINK. Given the low number of samples, no variant was significant after correction for multiple comparisons (EMP2) and only suggestive variants could be found (significant at 5% at EMP1). We considered suggestive of carrying deleterious alleles a list of 101 genes carrying 112 variants (Table S4).

Detection of regions carrying deleterious mutation from 800K SNP data

After applying quality control parameters, 26,653 and 146,991 markers were removed from the original dataset for excessive missing data and MAF, respectively. In addition 17 individuals removed for low genotyping rate. As with exomes, only autosomal SNPs were retained for the analyses, thus obtaining a working dataset of 590,680 SNPs and 992 animals.

MDS (Figure S5) analysis indicates that the dataset has some substructure that might affect downstream analyses. We accounted for the risk of finding markers out of Hardy-Weinberg proportions due to the Wahlund effect, by choosing very stringent significance threshold in the OHW analyses.

To detect regions candidate to carry deleterious recessive mutations we used two approaches, the "Outside Hardy Weinberg" (OHW) and the "Lack Of Homozygotes" (LOH) approaches (see materials and methods). The two approaches complement each other in the detection of regions carrying less homozygotes than expected: OHW identifies regions in which a few markers have extreme homozygote deficit, while LOH detects regions in which many markers have a moderate deficit.

In the SNP dataset, 1884 markers were significantly out of Hardy-Weinberg equilibrium ($P \leq 8.47 * 10^{-8}$). A total of 485 windows were identified with at least 3 markers out of 10 exceeding this threshold. Among these, 84 contained markers significant because of homozygote deficit. These 84 windows identified 11 OHW candidate regions in 9 chromosomes (Table S5).

The LOH approach identified 2335 windows below -3 Z score of the standardised difference between observed and expected frequency of homozygotes for MAF. These identified 326 LOH candidate regions spread on all autosomes (Table S6). Only regions formed by, at least, 3 single windows were retained reducing LOH candidate regions to 240.

Deleterious variants (SIFT deleterious and stop codon gained/lost) were intersected with significant OHW and LOH region. Seven of these mapped within OHW sliding windows, 36 within LOH sliding windows. Among these 5 were located in regions significant in both OHW and LOH. Other 51 deleterious variants mapped in the immediate vicinity (+/-50Kb) of significant windows. These 83 variants affect 66 unique genes candidate to carry deleterious recessive alleles (Table S7).

Five regions were common to OHW and LOH: one on BTA4 (LOH=74,959,610-75,151,417; OHW=74,970,653-75,151,417), two on BTA7 (LOH=9.626,435-10,099.512; OHW=9,628,735-10,099,512; LOH=10,575,691-11,171,132;

OHW=10,486,424-11,087,441), one on BTA15 (LOH=80,299,924-80,928,311; OHW=80,299,924-80,921,274) and one on BTA18 (LOH=28,122,373-28,185,525; OHW= 28,106,022-28,164,693).

Step 3. Assessing candidate deleterious variants

After the two filtering step previously described a total of 191 SNPs were retained out of the 3433 classified as deleterious by SIFT and Annovar, and analysed further. These are carried by 163 genes. Eighteen genes carried 2 candidate SNPs, three genes 3 SNPs and one gene 5 SNPs. In addition 4 SNPs in four genes were identified as potential candidates in both the analysis of EBV tails and of bull population.

A total of 12 SNPs introduced stop codons while the remaining induced amino acid substitutions or splicing variants. In the filtered dataset the proportion of stop codons on total variation (6.3%) was only slightly higher than the proportion found in the entire dataset (5.1%: 225 stop codon out of 4391 deleterious SNPs).

Comparative genomics

Protein multiple alignment was successful in 146 out of 191 comparisons. In 26 cases the most frequent amino acid in cattle was very highly conserved either in all vertebrates (N=9) or in all mammals (N=17). In other 23 cases the amino acid was conserved in at least 90% of mammalian species. In the last class were generally included amino acids that changed in mammalian species phylogenetically very distant from cattle (e.g. platypus and related species). In 97 instances the amino acid could change quite freely and occasionally both variants observed in cattle were carried by other species.

Haplotype analysis of HD population data

The search for haplotypes associated to the 191 variants for further testing in the population identified 101 unique and invariant haplotypes common to all carriers of a target variant and different from all those found in non carriers. In additional 36 cases a non completely invariant haplotype could be associated to the mutation. Conversely in 54 instances a reliable haplotype could not be found (table S8). In total 38 haplotypes had no homozygous individuals in the population. This was expected in the case of 33 rather rare haplotypes (fewer than 100 heterozygotes observed in the population). Conversely 5 haplotypes were rather frequent (from 170 to almost 300 heterozygotes observed in the population) and the expectation was to find 7.8 to 22.2 homozygous individuals in the population instead of the none found. Another variant had a remarkable high frequency (35%) and a clear outlier behaviour. Only 11 homozygous individuals were observed out of the 121 expected.

Functional analysis of candidate genes

To assess the potential deleterious effects of the variants identified we also retrieved functional information for each gene according to the following parameters: i) association to known Mendelian or complex diseases listed in the OMIM database ii) availability and effect of gene knock-out in mice iii) expression profiles, with particular attention in reproductive tissues, from the Gene Expression Atlas (Table 4).

Twenty genes out of 163 have a recognized phenotype in humans. These are mostly severe syndromes. Unfortunately only a few genes have a null-mice line produced with phenotypic data available

According to the expression data available from Ensembl, these genes are generally (with some exceptions) expressed in a variety of tissues with a clear tendency to be highly expressed in testis.

Table 4. Information on 163 genes analyzed concerning eventual phenotypes in man or knock-outs in mouse. *Header codes:* A: brain; B: colon; C: heart; D: kidney; E: liver; F: lung; G: skeletal muscle; H: spleen; I: testis. Colours represent the gene level of expression in different tissue: from low (white/grey) to high (blue). *NC* in *Function* column means "Not Classified".

Gene	Gene name	Human defect	Knock-Outs Mouse	Mouse Phenotype	A	B	C	D	E	F	G	H	I	Function
ENSBTAG00000000079	CCSAP		Cells		9	1	0.3	1	0.7	2	0.6	2	2	NC
ENSBTAG00000000149	USP40		Cells		6	7	2	7	5	7	2	8	9	fertility
ENSBTAG00000000414	FUT5		No match		-	37	-	1	-	-	-	-	-	cellular process
ENSBTAG00000000918	ERMARD	Periventricular nodular heterotopia 6; X-linked periventricular heterotopia; 6q terminal deletion syndrome; Periventricular nodular heterotopia;	No match		5	4	4	9	6	7	4	9	17	cellular process
ENSBTAG00000000998	SLC39A6		Cells		6	9	2	7	4	10	1	10	9	cellular process
ENSBTAG00000001133	VWA8		Mice		2	3	6	7	10	3	3	4	4	blood metabolism
ENSBTAG00000001140	IAH1		https://www.mousephenotype.org/data/genes/MGI:1914982#section-associations	BOTH SEX:adipose tissue, behaviour/neurological , growth/size/body . FEMALE: homeostasis/metabolism . MALE: limbs/digits/tail , skeleton	9	18	30	60	23	19	12	17	30	cellular process
ENSBTAG00000001262	IRGQ		Cells		6	1	5	3	0.6	2	5	0.4	2	immunity
ENSBTAG00000001286	ELMO3		https://www.mousephenotype.org/data/genes/MGI:2679007#section-associations	Decreased startle reflex	0.2	27	-	14	7	7	-	0.2	1	fertility
ENSBTAG00000001291	OR8H3		No match		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000001500	FIGNL1		Cells		0.8	2	0.1	0.5	0.7	0.8	0.4	2	3	cellular process
ENSBTAG00000001505	GRK4		Cells		10	7	0.9	4	6	3	2	6	218	signalling
ENSBTAG00000002220	SPTLC1	Hereditary sensory and autonomic neuropathy type 1	No		17	27	3	28	12	28	5	19	6	metabolism
ENSBTAG00000002288	NT5DC4		No match		-	-	-	0.2	-	-	-	-	-	cellular process
ENSBTAG00000002289	NLRP14		No		-	-	-	-	-	-	-	-	0.4	cellular process
ENSBTAG00000002660	CCDC11	Situs inversus totalis	No match		3	1	0.6	2	0.7	3	0.3	2	54	fertility

Gene	Gene name	Human defect	Knock-Outs Mouse	Mouse Phenotype	A	B	C	D	E	F	G	H	I	Function
ENSBTAG00000002809	CAPN10	Autosomal Recessive Mental Retardation; DIABETES MELLITUS NONINSULIN-DEPENDENT 1	Cells		2	1	3	4	1	3	3	2	9	immunity
ENSBTAG00000002966	DNAJC13	Young adult-onset Parkinsonism	No		5	5	2	9	4	6	2	8	4	fertility
ENSBTAG00000003231	LIM2	Pulverulent cataract	Cells		-	0.1	-	0.2	0.1	-	-	-	0.2	cellular process
ENSBTAG00000003508	CFAP69		https://www.mousephenotype.org/data/genes/MGI:2443778#section-associations	Abnormal circulating insulin level, Male Infertility, Decreased mean platelet volume, Increased lean body mass	2	0.3	-	0.7	-	1	-	0.1	7	NC
ENSBTAG00000003523	UGT2B15		No match		-	-	-	-	34	-	-	-	-	metabolism
ENSBTAG00000004081	FAT3		No		2	0.3	-	-	-	0.1	-	0.1	0.5	signalling
ENSBTAG00000004555	LRP2	DONNAI-BARROW SYNDROME: Intellectual disability	Cells		-	-	-	56	-	3	-	-	0.1	disease
ENSBTAG00000004772	THEM4		Cells		8	16	2	95	27	5	2	6	3	metabolism
ENSBTAG00000004860	SLC27A6		Cells		6	0.2	-	10	6	0.2	2	0.7	0.1	immunity
ENSBTAG00000005021	SEMA5A	Monosomy 5p	Cells		9	0.4	8	9	1	3	0.8	0.8	6	immunity
ENSBTAG00000005170	GPR114		Mice		-	1	0.1	0.2	0.1	0.9	-	6	0.1	fertility
ENSBTAG00000005248	OFCC1		No		-	0.1	0.1	0.3	0.1	0.1	0.1	0.3	-	NC
ENSBTAG00000005340	SERGEF		Cells		7	10	3	5	1	5	5	8	23	cellular process
ENSBTAG00000005357	VPS11		Mice		14	13	6	19	7	17	8	15	15	cellular process
ENSBTAG00000005466	C8H8orf58		No match		-	0.5	0.1	0.8	0.2	1	0.3	0.2	0.4	NC
ENSBTAG00000005514	TRIO		Cells		17	6	6	10	2	10	8	8	7	immunity
ENSBTAG00000005633	RGNEF		https://www.mousephenotype.org/data/genes/MGI:1346016#section-associations	Vertebral fusion	3	6	2	14	0.5	8	0.1	0.8	1	cellular process
ENSBTAG00000006021	CEP250		Mice		3	2	2	4	3	4	1	8	8	metabolism
ENSBTAG00000006027	USP34		Mice		11	6	3	7	4	8	8	9	12	fertility
ENSBTAG00000006532	CDK5RAP2	Autosomal recessive primary microcephaly	https://www.mousephenotype.org/data/genes/MGI:2384875#section-associations	skeleton, immune system, growth/size/body, hematopoietic system	5	3	4	9	2	7	4	7	84	development
ENSBTAG00000007001	SLC26A1		No		0.3	0.1	0.2	32	9	0.3	0.1	-	0.7	metabolism
ENSBTAG00000007340	LOC101906349		No match		-	-	-	-	-	-	-	-	-	NC
ENSBTAG00000007568	MEPE		Cells		-	-	-	-	-	-	-	-	-	cellular process
ENSBTAG00000007910	EXOC7		Cells		13	24	23	22	12	22	15	22	28	cellular process

Gene	Gene name	Human defect	Knock-Outs Mouse	Mouse Phenotype	A	B	C	D	E	F	G	H	I	Function
ENSBTAG00000008650	CAMK1D		Cells		24	2	0.1	0.6	1	2	0.3	7	3	signalling
ENSBTAG00000008797	Uncharacterized		No match		-	0.9	-	0.4	0.7	0.4	-	-	-	NC
ENSBTAG00000008871	DDX4		Cells		0.2	-	-	-	-	0.1	0.1	-	132	cellular process
ENSBTAG00000008996	SFI1		Mice		2	1	0.6	3	0.4	3	-	2	5	NC
ENSBTAG00000009014	UPK1B		Mice		2	2	2	5	4	4	2	7	3	fertility
ENSBTAG00000009030	NOX5		No match		-	0.4	-	0.1	-	8	-	-	0.1	immunity
ENSBTAG00000009033	TEX37		https://www.mousephenotype.org/data/genes/MGI:1921471#section-associations	Abnormal startle reflex	-	-	-	-	-	-	0.3	-	83	NC
ENSBTAG00000009064	CSF2RB	Congenital pulmonary alveolar proteinosis	Mice		0.1	2	0.4	0.6	2	23	0.3	17	0.5	signalling
ENSBTAG00000009144	BPIFA2A		No match		-	6	-	-	-	0.6	-	-	0.5	metabolism
ENSBTAG00000009337	PNMAL1		Cells		8	0.5	0.1	3	0.2	0.3	0.3	0.4	3	Immunity
ENSBTAG00000009444	SLC15A5		Cells		-	-	-	-	-	-	-	-	0.1	cellular process
ENSBTAG00000009568	KANK2		Cells		1	17	30	9	3	31	9	10	3	cellular process
ENSBTAG00000009569	DOCK6	Adams-Oliver syndrome	Cells		2	5	10	12	5	15	5	6	21	metabolism
ENSBTAG00000009836	CHGA		Mice		140	555	0.1	2	-	0.2	0.1	0.7	1	metabolism
ENSBTAG00000010522	Uncharacterized		No match		0.1	-	-	0.2	-	0.2	0.3	-	2	metabolism
ENSBTAG00000010601	DOLK	Familial isolated dilated cardiomyopathy	No		4	8	6	22	9	6	5	5	12	metabolism
ENSBTAG00000010847	FOXN4		Cells		-	-	-	-	-	0.1	-	-	0.1	metabolism
ENSBTAG00000010954	ART3		Cells		0.2	1	55	0.1	0.8	2	10	0.6	71	metabolism
ENSBTAG00000011011	SSH2		No		3	0.9	5	4	2	2	6	7	10	cellular process
ENSBTAG00000011387	NAT9		Cells		6	7	3	8	6	8	4	6	6	metabolism
ENSBTAG00000011420	CA9		Cells		0.4	3	0.2	2	0.3	1	0.1	2	56	metabolism
ENSBTAG00000011433	RGP1		https://www.mousephenotype.org/data/genes/MGI:1915956#section-associations	Complete pre-weaning lethality	7	19	11	18	8	12	9	8	25	cellular process
ENSBTAG00000011622	C18orf21		Mice		9	10	6	6	2	10	5	10	34	NC
ENSBTAG00000011683	NUMB		Cells		21	25	8	35	17	47	6	8	19	immunity
ENSBTAG00000011684	RAB11FIP1		Cells		-	6	0.1	8	0.9	11	0.2	1	4	NC
ENSBTAG00000012053	ACCSL		No		0.3	-	-	-	-	-	-	-	0.1	cellular process
ENSBTAG00000012314	LDLR	Homozygous familial hypercholesterolemia	Cells		8	33	11	21	17	24	2	15	2	cellular process

Gene	Gene name	Human defect	Knock-Outs Mouse	Mouse Phenotype	A	B	C	D	E	F	G	H	I	Function
ENSBTAG00000012326	LOC101906218		No match		-	11	0.1	0.1	-	0.9	-	0.1	-	NC
ENSBTAG00000012458	PSAPL1		Cells		-	-	-	0.1	-	-	-	-	-	NC
ENSBTAG00000012565	PCNX		No		3	3	2	3	2	3	3	4	8	development
ENSBTAG00000012833	ATG2B		No		4	2	1	2	1	2	1	2	9	NC
ENSBTAG00000012921	KIF24		Mice		0.5	0.7	0.3	0.8	0.2	2	0.2	2	14	metabolism
ENSBTAG00000013568	UEVLD		Cells		5	4	1	5	3	4	2	2	14	metabolism
ENSBTAG00000013685	KRT31		Mice		-	-	-	-	-	-	-	-	-	cellular process
ENSBTAG00000013753	DDRGK1		Mice		22	32	18	56	55	25	14	22	37	NC
ENSBTAG00000013789	OR6K6		No		-	-	-	-	-	-	-	-	-	NC
ENSBTAG00000013838	ALLC		Mice		0.2	-	-	-	-	-	0.1	-	78	metabolism
ENSBTAG00000013916	HERC2	Mental retardation autosomal recessive 38; SKIN/HAIR/EYE PIGMENTATION VARIATION IN 1; Developmental delay with autism spectrum disorder and gait instability;	No		8	7	4	8	3	7	4	6	5	metabolism
ENSBTAG00000014001	MAP1A		Cells		136	23	22	2	0.2	4	3	3	52	cellular process
ENSBTAG00000014058	LDB2		No		28	3	3	12	8	25	4	10	3	cellular process
ENSBTAG00000014284	ALPK2		No		-	0.1	10	-	-	0.1	3	0.4	0.2	cellular process
ENSBTAG00000014750	EPB41L4A		https://www.mousephenotype.org/data/genes/MGI:103007#section-associations	Decreased circulating glucose level, Abnormal cone electrophysiology, Immune system s	1	2	2	20	0.7	12	1	3	43	NC
ENSBTAG00000014752	AKAP11		https://www.mousephenotype.org/data/genes/MGI:2684060#section-associations	No pheno	24	7	2	6	3	3	0.6	6	8	cellular process
ENSBTAG00000014768	ZNF786		Cells		3	2	3	4	2	4	8	3	3	cellular process
ENSBTAG00000014794	SCYL3		Cells		6	10	5	9	6	13	5	16	26	cellular process
ENSBTAG00000015027	Clusterin		https://www.mousephenotype.org/data/genes/MGI:88423#section-associations	Aggression, vertebral fusion	0.5	0.3	-	0.3	0.2	0.5	-	-	3	NC

Gene	Gene name	Human defect	Knock-Outs Mouse	Mouse Phenotype	A	B	C	D	E	F	G	H	I	Function
ENSBTAG00000015210	LOXHD1	DEAFNESS AUTOSOMAL RECESSIVE 77; Autosomal recessive nonsyndromic sensorineu; Dominant late-onset Fuchs corneal dystrophy; deafness autosomal recessive type 77;	No		3	-	-	-	-	-	-	-	0.3	NC
ENSBTAG00000015294	COX10	Leigh syndrome; Mitochondrial complex IV deficiency (MT-C4D)	Cells		4	6	24	5	2	3	14	3	16	cellular process
ENSBTAG00000015335	BAI3		No		38	0.3	8	-	-	-	0.1	0.1	2	signalling
ENSBTAG00000015490	HS1BP3		Cells		1	10	5	9	4	8	2	1	37	cellular process
ENSBTAG00000015602	C7H5orf45		No match		-	-	-	-	-	-	-	-	10	cellular process
ENSBTAG00000015839	MAP4		No		27	85	104	35	11	100	82	46	61	cellular process
ENSBTAG00000015912	DMKN		No		-	0.1	-	0.4	-	-	-	-	0.3	NC
ENSBTAG00000015980	FASN	Autosomal Recessive Mental Retardation	Cells		19	29	7	12	6	39	6	7	22	metabolism
ENSBTAG00000015985	GPR20		https://www.mousephenotype.org/data/genes/MGI:2441803#section-associations	Anatomical defects various tissues.	-	-	0.3	0.2	-	0.8	0.1	-	0.9	signalling
ENSBTAG00000016495	LINS	Autosomal Recessive Mental Retardation	Cells		6	3	1	5	3	3	1	5	12	NC
ENSBTAG00000016691	LACC1		Mice		2	9	1	9	9	8	0.8	7	27	NC
ENSBTAG00000017096	ERICH1		Cells		6	4	2	4	2	5	2	5	13	metabolism
ENSBTAG00000017165	MATN2		No		2	7	2	20	0.8	3	2	2	7	cellular process
ENSBTAG00000017418	RRP1B		Cells		6	4	4	5	3	7	3	12	159	cellular process
ENSBTAG00000017436	ALS2CR11		Cells		0.1	-	-	-	-	-	-	-	37	disease
ENSBTAG00000017505	PAXIP1		Cells		2	3	2	4	2	4	2	3	12	signalling
ENSBTAG00000017576	GPR144		No match		-	-	-	-	-	-	-	-	0.2	signalling
ENSBTAG00000018528	USP45		No		1	1	0.4	1	0.5	0.6	0.2	2	0.4	cellular process
ENSBTAG00000018810	THBS2	INTERVERTEBRAL DISC DISEASE	Cells		1	2	0.4	0.9	1	0.9	0.4	1	2	fertility
ENSBTAG00000019072	PSD4		Cells		-	7	8	1	1	7	0.2	6	4	cellular process

Gene	Gene name	Human defect	Knock-Outs Mouse	Mouse Phenotype	A	B	C	D	E	F	G	H	I	Function
ENSBTAG00000019265	AGK	Sengers syndrome; Congenital cataract - hypertrophic cardiomyopathy - mitochondrial myopathy	Cells		7	3	6	4	3	5	2	6	5	metabolism
ENSBTAG00000019752	BPIFA2B		No match		-	0.1	-	-	-	-	-	-	1	metabolism
ENSBTAG00000019799	FCAMR		Cells		-	0.7	-	0.1	0.2	0.1	-	2	7	immunity
ENSBTAG00000019961	ATP4B		Cells		-	-	12	-	-	0.3	-	1	0.5	cellular process
ENSBTAG00000020414	DHX37		No		3	3	3	7	3	7	3	4	13	cellular process
ENSBTAG00000021073	KIAA1549	PILOCYTIC ASTROCYTOMA SOMATIC	https://www.mousephenotype.org/data/genes/MGI:2669829#section-associations	behaviour/neurological , homeostasis/metabolism , hematopoietic system	3	0.4	2	3	0.1	2	1	0.4	3	NC
ENSBTAG00000021233	OR5B17		No		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000021375	OR10Q1		No		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000021643	EFCAB12		Cells		-	-	-	0.1	-	1	-	-	9	cellular process
ENSBTAG00000021645	MBD4		Cells		8	4	3	8	2	5	1	10	6	cellular process
ENSBTAG00000022509	DNAH9		No		0.2	-	-	-	-	5	-	0.5	5	metabolism
ENSBTAG00000022858	LOC783561		No match		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000026247	PKHD1L1		Cells		-	-	-	-	-	0.2	-	-	0.1	cellular process
ENSBTAG00000026323	LYSB		No match		-	68	-	-	-	-	-	-	-	metabolism
ENSBTAG00000026350	SPATC1		Cells		0.5	-	-	0.6	-	0.1	0.1	0.1	39	NC
ENSBTAG00000027390	RTFDC1		Mice		28	19	21	20	9	22	13	20	49	NC
ENSBTAG00000031115	Uncharacterized		No match		-	-	-	-	-	-	-	-	0.1	cellular process
ENSBTAG00000031718	OGFR		Mice		6	17	6	12	5	17	3	9	1	fertility
ENSBTAG00000032264	Uncharacterized		No match		52	-	-	4	-	-	-	-	22	NC
ENSBTAG00000032481	DAPL1		No		-	1	-	6	-	0.7	0.1	13	10	cellular process

Gene	Gene name	Human defect	Knock-Outs Mouse	Mouse Phenotype	A	B	C	D	E	F	G	H	I	Function
ENSBTAG00000032527	ERCC6	CEREBROOCULOFACIOSKELETAL SYNDROME 1; COCKAYNE SYNDROME B; De Sanctis-Cacchione syndrome; MACULAR DEGENERATION AGE-RELATED 5; UV-SENSITIVE SYNDROME 1; COFS syndrome; Cockayne syndrome type 1; Cockayne syndrome type 2; Cockayne syndrome type 3; UV-sensitive syndrome; Cockayne syndrome type B (CSB); De Sanctis-Cacchione syndrome (DSC); UV-sensitive syndrome (UVS); cerebro-oculo-facio-skeletal syndrome type 1 (COFS1);	Mice		3	2	0.5	2	2	2	0.7	4	2	cellular process
ENSBTAG00000032821	SCEL		No		0.1	-	-	-	0.2	17	-	-	0.4	NC
ENSBTAG00000033954	PRPS1L1		No		-	-	-	-	-	-	-	-	35	metabolism
ENSBTAG00000034991	SLX4IP		Cells		6	2	1	3	2	2	0.7	3	3	NC
ENSBTAG00000035226	TOR1AIP1		Cells		9	12	5	13	8	16	7	16	83	NC
ENSBTAG00000035309	OR4F15		No		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000035708	LOC527795		No match		-	-	-	-	0.1	-	0.1	-	23	metabolism
ENSBTAG00000035985	LOC101909743/ OR8K1		No		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000036019	RIPPLY3		https://www.mousephenotype.org/data/genes/MGI:2181192#section-associations	hemopoietic system	0.1	-	-	0.4	0.1	9	-	0.1	-	malformation
ENSBTAG00000037399	Uncharacterized		No match		2	9	2	4	-	2	4	0.5	5	signalling
ENSBTAG00000038606	DBDR		No		0.2	-	-	-	0.1	-	-	-	-	signalling
ENSBTAG00000038831	PHYHD1		No		2	18	21	23	30	18	16	15	1	metabolism
ENSBTAG00000039110	OR8U1		No		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000039117	FAM179B		Cells		14	2	0.6	3	1	3	0.8	3	3	cellular process
ENSBTAG00000039520	SIRPB1		Cells		4	0.8	3	2	5	16	0.5	12	0.1	signalling

Gene	Gene name	Human defect	Knock-Outs Mouse	Mouse Phenotype	A	B	C	D	E	F	G	H	I	Function
ENSBTAG00000040568	Uncharacterized		No match		4	3	0.7	5	4	3	3	2	5	cellular process
ENSBTAG00000045558	LOC100299084		No match		1	2	2	1	10	10	0.9	5	2	fertility
ENSBTAG00000045588	LOC100298356		No match		-	-	-	-	-	-	-	-	-	metabolism
ENSBTAG00000045709	LOC514057		No match		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000046175	Uncharacterized		No match		0.1	-	-	-	-	-	-	0.1	-	cellular process
ENSBTAG00000046228	Olfactory receptor		No match		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000046239	C10orf12		No		0.3	1	0.4	0.7	1	0.8	0.6	2	0.9	cellular process
ENSBTAG00000046285	OR8I2		No		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000046360	Uncharacterized		No match		-	-	-	-	-	-	-	-	-	NC
ENSBTAG00000046423	LOC618007		No match		-	-	-	-	-	-	-	-	-	NC
ENSBTAG00000046527	LOC100139830		No match		-	-	-	-	-	-	-	-	-	fertility
ENSBTAG00000046590	FAM181A		Cells		0.9	-	-	-	-	-	-	-	44	NC
ENSBTAG00000046593	Uncharacterized		No match		-	-	-	-	-	-	-	-	20	metabolism
ENSBTAG00000046727	Olfactory receptor		No match		-	-	-	-	-	-	-	-	0.4	fertility
ENSBTAG00000047150	RTEL		Cells		4	2	3	6	4	5	5	6	16	cellular process
ENSBTAG00000047166	SH2D7		Mice		0.3	1	0.1	0.1	-	1	0.2	0.2	2	immunity
ENSBTAG00000047174	Uncharacterized		No match		0.3	7	105	0.2	0.1	2	71	0.5	0.3	NC
ENSBTAG00000047317	CL43		No match		-	-	-	0.1	120	-	-	-	0.1	immunity
ENSBTAG00000047587	Uncharacterized		No match		22	0.9	0.9	35	0.3	4	-	0.6	2	cellular process
ENSBTAG00000047661	FABP12		https://www.mousephenotype.org/data/genes/MGI:1922747#section-associations	Growth/size/body	0.1	-	-	-	-	-	0.2	-	69	metabolism
ENSBTAG00000048021	PNMAL2		No		0.4	0.3	-	2	-	0.3	0.3	-	0.2	NC
ENSBTAG00000048153	Uncharacterized		No match		2	-	-	0.5	-	0.1	-	0.2	0.1	cellular process

Discussion

Deleterious recessive variants are naturally present in outbred animal species. They do not create problems as long as they remain at low frequency in random mating populations. Under these conditions these variants remain in the heterozygous state and do not produce phenotypic effects. However, as soon as their frequency increases and mating occurs between related individuals, their deleterious effects become apparent and highly affect the welfare and viability of animals carrying them in the homozygous state.

In industrial dairy cattle breeds crossbreeding is generally avoided and breeds are in practice genetic isolates in which artificial insemination (AI) is widely used for rapidly spreading the genetic progress in the population. As a consequence elite sires have many thousand and sometimes hundreds of thousand daughters. Under these conditions mating between relatives becomes practically unavoidable in the long term. In fact, in spite of the close control of inbreeding and of the mating system, genetic defects regularly appear to the surface. Some examples are the Complex Vertebral Malformation (CVM) and the Bovine Leucocyte Adhesion Deficiency (BLAD) in Holstein (Schütz et al., 2008) and Weaver disease in Brown (McClure et al., 2013). In these cases the effects of the recessive deleterious alleles are clear and molecular diagnostic tools have been developed to avoid their further spreading and to initiate the eradication process.

However, recessive deleterious alleles may produce their effects in a more deceitful way, for example by affecting fertility, inducing early lethality or influence disease susceptibility. All these effects are difficult to track with the current methods of phenotype measurements, but have an impact on the economy of the livestock sector and on animal welfare.

In this paper we sought for deleterious recessive alleles in Italian Holstein cattle. This breed is recently experiencing a decrease in the genetic trend of fertility traits (Pryce et al., 2014), therefore we started our search from the exome sequences of 20 sires, 19 of which having extreme EBV values for male and female fertility.

The choice of sequencing exomes was a compromise between completeness of information, accuracy of variant and genotype detection and cost. Exomes clearly give incomplete information, since they miss all variation in regulatory sites outside gene boundaries (ENCODE Consortium, 2012). These may be very relevant for the control of trait variation, but they are still very poorly annotated in livestock. On the other hand, sequencing exomes yields millions instead of billions base pairs, a deeper sequencing of target regions and a consequent more reliable variant and genotype calling (Figure S1 and S2). The analysis of exome data is also less demanding in terms of computer time (thousands instead of millions variants are to be analysed) and easier in terms of biological interpretation. Exome sequencing revealed to be very effective in the identification of a number of mutations affecting human health (Bamshad et al., 2011; Yang et al., 2013). All these were defects caused by mutations in single genes producing clear phenotypic effects having Mendelian inheritance. These conditions do not hold in our investigation, therefore we had to devise strategies to integrate exome data with other information collected from the Holstein population and other species.

The overall procedure for identify the candidate deleterious recessive was based on the rationale that deleterious recessive alleles in coding sequences are expected to be: i) caused by severe alteration of the proteins encoded by genes; ii) mostly, if not exclusively, carried at the heterozygous state in the population; iii) unevenly distributed in animals of high and low fertility; iv) likely conserved in different species because of selection constraint; v) likely associated to genetic defects in any species, when severely damaged.

A stepwise procedure was set up for reducing candidates to a number manageable for functional analysis (Figure 1). In the first step mutations were discovered, in the second filtered and in third assessed.

This procedure identifies only some of the deleterious recessives existing in the population analysed. Since we sequenced only exomes and not whole genomes, our analyses is limited to coding sequences and at the moment to SNP variation. We therefore miss regulatory variations outside genes and the effects of indels and more complex variations.

The number of animals sequenced is limited. Deleterious recessives are expected to be rare in the population and we have sampled only 20 animals from a subset of sires authorized to AI in Italy. Although we have taken among them those in the minus tail of fertility trait, that might have a higher genetic load, it's likely that we have missed other variants existing in the population. On the other hand, variants carried by AI authorized bulls are the ones that will have a larger impact on the population if widely spread by elite bulls and therefore they are the most interesting from a breeding point of view.

Most deleterious variants are expected to be very rare, and therefore not detected by the approach here adopted or detected and afterwards discarded during the filtering processes.

Sequencing is prone to errors. Some real variants may have been discarded during QC of sequence data, due to poor sequence quality or poor coverage of the region.

Some of the variations classified as "non deleterious" by the software used may indeed have an effect. In fact synonymous mutation are non-neutral whenever the different tRNAs for a same amino acid have different relative abundance and translation rates (Goymer, 2007).

1. Searching candidate deleterious variants in exomes

Exomes produced a fair amount of data. All together 53.55 Mb were sequenced. These represent 99.98% of the exome sequences targeted by probes. Among these only 3.5% had a sequencing depth across animals <60X. Sequencing depth is a key factor for variant detection and even more so when the genotype of sequenced animals is to be inferred. Investigating the concordance between the genotypes detected at more than 20K SNP by both Beadchip and exome sequencing, we observed a clear inverse relationship between sequence depth and number of discordances (Table S3 and Figure S3). The average sequencing depth expected (50X) and realised in practice (40.58) in this investigation, yielded rather accurate genotypes, since the average concordance observed was over 97% and increased to over 99% when two outlier animals were removed from the dataset. The reason for the outlier behaviour of these samples was likely a suboptimal quality of DNA submitted to sequencing. Overall, almost 215,000 variants passed quality controls and among them 192,440 autosomal SNPs have been investigated in detail. As found in other investigations, most variants were rare (Peloso et al., 2014). For example the MAF distribution observed in 46904 SNPs classified neutral has a mean of 0.197 and a median of 0.167 values (Figure S6).

The first step in our research was to identify SNPs causing severe alterations in the proteins coded by genes. These were relevant changes in conformation, alterations of splicing sites or truncations of proteins. To accomplish this task we relied on two software, SIFT and ANNOVAR that identified 4,902 such variations in 3423 genes. Interestingly the subset of deleterious variants has significantly lower mean (0.156) and median (0.10) compared to neutral variations (Figure S6). This subset is therefore at least enriched in variants under purifying selection having a deleterious effect at the population level.

The number of deleterious alleles detected is high and it is unlikely they all have a relevant phenotypic effect, even if most variants are rare. In fact, variants deleterious for a protein are not automatically deleterious for an organism, as the protein change induced by the variant, although relevant, may not affect protein function; variation may occur in one member of a gene family, so that its function may be compensated by other family members; the protein function may be compensated by other proteins or the affected metabolic pathway compensated by alternative pathways. The challenge was therefore to filter a subset of variants likely having a phenotypic effect to be proposed for validation at a larger scale. To accomplish this task we sought for additional information in exome sequences and in a large population of Holstein sires genotyped with the bovine HD Beadchip.

2. Filtering deleterious variants

The second step consisted in filtering the several thousand deleterious SNPs identified in exomes to search for SNPs and genes that worth deeper investigation. We used two strategies to accomplish this task.

2.1. Distribution of deleterious variants in fertility EBV plus-variant and minus-variant animals

Fertility traits are controlled by a high number of genes and are highly genetically heterogeneous. Even using a selected subset of variants identified from exome sequencing, classic association analyses would need sample sizes much larger that the one available in this investigation to have a reasonable statistical power (Stitziel et al., 2011). The first approach we applied, was the analysis of SNP allelic and genotypic distribution between the minus and plus-variant tails of male and female fertility EBVs. Given the very low number of animals sequenced, we decided to use a simple model to exclude from further analyses all markers showing no evidence of association with the EBV tails, rather than to find significant associations.

We divided samples in cases (minus-variant animals for fertility EBV) and controls (plus-variant animals) and tested all possible models of inheritance implemented in the Plink package for tackling simple genetic defects. We then used an empirical point-wise significant threshold based on permutations to highlight SNPs having alleles or genotypes showing an uneven distribution between cases and controls and considered these as worth further investigation. A total of 112 SNPs in 101 genes passed this filtering process.

2.2 Detection of regions carrying deleterious mutation from 800K SNP data

The analysis of exomes alone was therefore insufficient to tag recessive deleterious genes in Italian Holstein and we sought for additional information in 1009 Italian Holstein bulls genotyped at high density that are part of the genomic selection program run by ANAFI. The rationale of the approach is that deleterious recessives are expected to be absent or rare in the homozygous status in the population. Hence, they may be tagged searching for genome regions having markers or haplotypes lacking or having significantly less homozygotes than expected under a neutral model.

We used two complementary approaches to identify these candidate regions. One (OHW) searched for very strong signals produced by single markers out of Hardy-Weinberg equilibrium at $P \leq 8.47 \times 10^{-8}$. A minimum of three significant markers in a sliding windows of 10 was set to reach significance, to capture more reliable signals, avoid excessive single marker noise and account for a possible Wahlund effect due to presence of the genetic structure in the population, difficult to avoid in the set of progeny tested bulls analysed. The second approach (LOH) captured more moderate signals but from multiple markers in larger regions. Both approaches were designed to detect signals in regions in which deleterious variants are still in segregation and are not excessively rare.

As expected, the two approaches identified regions that only partially overlapped (Tables S5, S6 e S7). In particular, 5 regions out of the 11 identified by OHW were in common to the 240 identified by LOH. Two of these regions overlap with previous investigations that used the BovineSNP50 Beadchip. Sahana et al. (2013) found a region candidate to carry potential lethal haplotypes in Nordic Holstein on BTA7 in the region 6,708,007-10,907,840. This interval overlaps with both BTA7 regions detected in this study. VanRaden et al. (2011) detected in U.S. Holstein a large region on BTA15 spanning 76Mb-82M (on BTA 3.0) that included *LRP4* gene, responsible for the mulefoot defect. Our analyses with the HD Beadchip indicates on the one side that the large region detected by Sahana et al. (2013) may consist in two nearby regions. On the other side the signal we detected on BTA 15 does not include the *LRP4* locus that is completely fixed in the Italian bulls population analysed.

In total 83 mutations in 66 genes fell in either or both OHW and LOH regions. These were joined to the SNPs output of the EBV tail analyses for further assessment.

3. Assessing filtered deleterious variants

Different characteristics of the filtered deleterious variants were assessed to identify among them strong candidates to submit to a large-scale validation study. The assessment phase permitted to gain additional information in favour or against SNP and gene candidature to be deleterious. We quantified this evidence using a subjective score.

3.1 Comparative genomics analysis

In the case of comparative genomics, the score quantified the level of conservation of the alternative amino acid residues induced by SNP variants in Italian Holstein cattle. Comparison was run against a set 100 vertebrate species that included 62 mammals. The level of conservation is a good proxy for the expected severity of a mutation. In the case of the 46 SNPs that induced the change of an amino acid highly conserved in the species investigated we may hypothesize a strong selection pressure in favour of the maintenance or the invariant amino acid at that position. Therefore the amino acid change or the stop codon we observed in cattle may affect not only the protein structure and function but also the phenotype of homozygous animals. Interestingly, in all these cases, the SNP allele coding for the non conserved variants was the minor allele.

Conversely, in all cases of high variability of the target amino acid across species (as many as seven different amino acids have sometimes been observed) poorly supports the presence of phenotypic effects due to its change. Sometimes the information collected is against the presence of a deleterious mutation, as in the case of the presence in different species of both alleles identified by exome sequencing. This suggests that neither of these is greatly affecting the viability and fitness of animals carrying them.

We also observed a number of failures (N=45) in the lift over of SNP coordinates from cattle to human or in protein alignment. This happened for example in the case of uncharacterized cattle proteins that have no orthologous in the human genome, and of some olfactory receptors that belong to a highly variable gene family. Even if the failure of alignment is an indication of poor protein conservation across species, the conservation of the target amino acid could not be assesses and in these cases we preferred to consider the information missing and to score it accordingly.

3.2 Haplotype analysis

A second assessment of the filtered candidates was on their behaviour in the sire population. The ideal test would be the direct genotyping of the SNPs in the population and the evaluation of their allelic and genotypic proportions. In the absence of this information, we first inferred the HD haplotype associated to SNP alleles and assessed the behaviour of the haplotype in the population, assuming this as a surrogate of the SNP behaviour. Since deleterious alleles are expected to be rare and this is also what we observed from comparative genomic analyses, we investigate the frequency only of the haplotype carrying the minor SNP allele against the expectation to find the corresponding haplotype never homozygous or a number of times much lower than expected from its frequency in the population. The strategy yielded some interesting confirmation but was only partially successful.

First of all the identification of a unique haplotype carrying the rare variant was not always possible or straightforward. In some cases carriers of a same rare variant did not share exactly the same haplotype. We accepted some variation if the haplotype remained clearly distinguishable from those of animals not carrying the rare variant, but at the same time reduced the stringency of the test. In quite a number of cases a same haplotype carried both candidate deleterious SNP variants. This may occur for example in the case of recent mutations. In these last cases the test could not be carried out. Finally, most rare variants were associated to rare haplotypes that are not likely to have homozygous individuals in the population independently of the SNP effect.

This assessment was therefore only significant for a few haplotypes that had a large number of heterozygous individuals and no or only a very few homozygotes and for those that had a large number of homozygotes.

3.3 Functional analysis

The final assessment was the analysis of the function of genes carrying the candidate deleterious SNP. This has some limitation, since knowledge of gene function is only partial and mainly gained from investigations on species different from cattle. In any case, we considered the association to genetic defects in human and the induction of severe phenotypes in mouse knock out as a positive indication of the likely phenotypic effect of the severe protein alteration we observed in cattle. In addition gene function and expression profile was evaluated in search for evidence of gene involvement in fertility, development or processes fundamental for cell and organism viability.

Among genes identified 22 are associated to human syndromes (Table 4). Eleven syndromes are neurological diseases affecting brain development and inducing mental retardation (e.g. Donnay-Barrow Syndrome and Autosomal recessive primary microcephaly), the others influence development (e.g. Cerebrooculofacioskeletal Syndrome 1 and Adams-Oliver syndrome) or organ function (e.g. Congenital pulmonary alveolar proteinosis).

Most of the same defects translated into Holstein would induce physical and behavioural anomalies likely causing young bulls to be excluded from progeny testing. It should in fact be considered that the animals we analysed are on the one side those that will most influence the genetic make of future generation but are not a random sample of Holstein animals. These bulls have been progeny tested and authorised to be used in artificial insemination in Italy. Candidates to progeny testing derive from the mating of the best 1% sires and best 2% dams of the Italian Holstein population. At the age of 4 months they are taken to the Italian Holstein breeder association (ANAFI) test station where they are subjected to performance test, sanitary controls and behavioural tests. Thereafter the quality of their semen is checked, before authorization is granted to enter in progeny test. In this sample of animals it is therefore expected to find never at the homozygous state mutations causing lethality, physical malformation, abnormal behaviour and semen infertility. Also those having a major effect on semen quality and high susceptibility to diseases are expected to be rarely carried at the homozygous status.

Association with human syndromes is a strong evidence of the potential phenotypic effect of mutations altering the gene structure. However it should be considered that variants found in humans are different from variants found in cattle and that the severity of the phenotype induced by a deleterious gene is often variable even within species. Therefore we evaluated this information together with the other indirect evidences of variants to be deleterious.

We also enquired the website of the International Mouse Phenotype Consortium (https://www.mousephenotype.org/) to search for the effects of null alleles in mice. The few genes for which null-mice line is available are involved in basic functions for the organism that span from vision (*IAH1* and *EPB41L4A*), to aging (*RGP1*) and tissues development (*FABP12, RIPPLY3, RGNEF*) or fertility (*CFAP96*). As with genes associated to human syndromes, most of the abnormal phenotypes observed in knock mouse would hamper young bulls to enter progeny testing.

Gene Ontology (GO) was evaluated on the entire gene set. No enrichment on particular categories/GO terms was observed. The most represented categories in Biological Processes were as expected "metabolic process" and "cellular process", however, lower level terms "reproduction" and "developmental process" are represented with meaningful hits. The analysis of Molecular Function indicates that many of the hits are enzymes ("catalytic activity") and have regulatory roles ("binding" and "enzyme regulator activity"). This result can be interpreted in different ways. On the one side deleterious alleles may exert their deleterious effect in a number of different cellular and developmental processes: the correct process is one, while there is many options to make it wrong. Also, the number of genes investigated is relatively low and for many of them there is little or no information available (e.g. 11 genes code for a completely uncharacterized protein). In addition some of them may turn out not to be really deleterious and add noise to the GO analysis.

As a final step all the previous information (GO, expression patterns, known function) have been used to estimate a score measuring the importance of the gene function for cattle survival and reproduction. The scoring of function was particularly difficult. GO categories were general and all functions tagged have a relevant importance for animal survival and production. Therefore in this case we didn’t use the whole score range planned (from -1 to 1) and gave no negative scores. A positive score of 1 was attributed only to genes playing an important role in fertility and development.

3.4 Strong candidate deleterious variants

The scoring process yielded 21 variants with a total score = 0, 35 variants with positive and 135 with negative values (Figure 2 and Table 5). Although the process was not error free, as it relied on incomplete information and had some degree of subjectivity, particularly in the choice of relative weights, it identified a set of variants that may be considered strong candidates to have deleterious effects in the Italian Holstein breed. Variants were evaluated individually and those in a same genes had sometimes different scores. For example the five variants carried by *ALPK2* had scores ranging from -7 to -2.

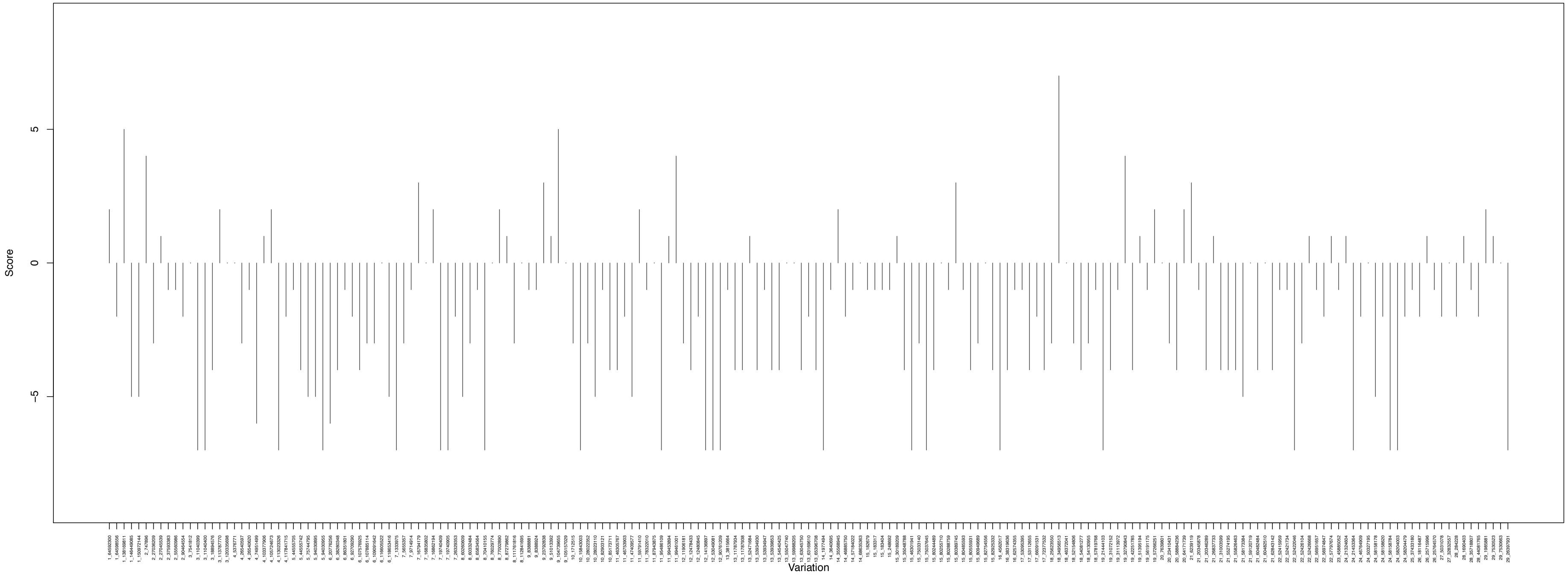

Figure 2. Graphical representation of total score for each variants.

Table 5. Single analyses score and total for each candidate variants. *Chr*: chromosome; *Pos*: physical position;

Mutation				Score					
Chr	**Pos**	**Gene**	**Gene name**	**Haplotypes**	**Conservation**	**Human defect**	**Mouse model**	**Function**	**Total Score**
1	64592300	ENSBTAG00000009014	UPK1B	0	2	-1	0	1	2
1	64598556	ENSBTAG00000009014	UPK1B	0	-1	-1	0	0	-2
1	138169811	ENSBTAG00000002966	DNAJC13	0	3	1	0	1	5
1	146449085	ENSBTAG00000017418	RRP1B	-3	-1	-1	0	0	-5
1	150972144	ENSBTAG00000036019	RIPPLY3	-3	-1	-1	-1	1	-5
2	747896	ENSBTAG00000013916	HERC2	3	0	1	0	0	4
2	27036209	ENSBTAG00000004555	LRP2	-3	-1	1	0	0	-3
2	27045539	ENSBTAG00000004555	LRP2	-3	3	1	0	0	1
2	37593383	ENSBTAG00000032481	DAPL1	0	0	-1	0	0	-1
2	55560986	ENSBTAG00000032264	Uncharacterized	3	-3	-1	0	0	-1
2	90464554	ENSBTAG00000017436	ALS2CR11	0	-1	-1	0	0	-2
3	7541812	ENSBTAG00000046175	Uncharacterized	0	0	0	0	0	0
3	11040399	ENSBTAG00000013789	OR6K6	-3	-3	-1	0	0	-7
3	11040400	ENSBTAG00000013789	OR6K6	-3	-3	-1	0	0	-7
3	18894576	ENSBTAG00000004772	THEM4	0	-3	-1	0	0	-4
3	113787770	ENSBTAG00000000149	USP40	1	1	-1	0	1	2
3	120335698	ENSBTAG00000002809	CAPN10	-3	2	1	0	0	0
4	5376771	ENSBTAG00000001500	FIGNL1	0	1	-1	0	0	0
4	26540597	ENSBTAG00000033954	PRPS1L1	1	-3	-1	0	0	-3
4	26540620	ENSBTAG00000033954	PRPS1L1	1	-1	-1	0	0	-1
4	74951499	ENSBTAG00000003508	CFAP69	-3	-3	-1	1	0	-6
4	103377906	ENSBTAG00000021073	KIAA1549	0	-1	1	1	0	1
4	105724673	ENSBTAG00000019265	AGK	0	1	1	0	0	2

Mutation				Score					
Chr	Pos	Gene	Gene name	Haplotypes	Conservation	Human defect	Mouse model	Function	Total Score
4	113023326	ENSBTAG00000014768	ZNF786	-3	-3	-1	0	0	-7
4	117841715	ENSBTAG00000017505	PAXIP1	0	-1	-1	0	0	-2
5	44555705	ENSBTAG00000026323	LYSB	-3	3	-1	0	0	-1
5	44555742	ENSBTAG00000026323	LYSB	-3	0	-1	0	0	-4
5	75744795	ENSBTAG00000009064	CSF2RB	-3	-3	1	0	0	-5
5	94030895	ENSBTAG00000009444	SLC15A5	-3	-1	-1	0	0	-5
5	94030955	ENSBTAG00000009444	SLC15A5	-3	-3	-1	0	0	-7
6	20776256	ENSBTAG00000008797	Uncharacterized	-3	-3	0	0	0	-6
6	38280348	ENSBTAG00000007568	MEPE	0	-3	-1	0	0	-4
6	86351801	ENSBTAG00000003523	UGT2B15	1	-1	-1	0	0	-1
6	92709290	ENSBTAG00000010954	ART3	-3	2	-1	0	0	-2
6	107576925	ENSBTAG00000038606	DBDR	0	-3	-1	0	0	-4
6	107885114	ENSBTAG00000001505	GRK4	-3	1	-1	0	0	-3
6	109091642	ENSBTAG00000007001	SLC26A1	1	-3	-1	0	0	-3
6	116605522	ENSBTAG00000014058	LDB2	0	1	-1	0	0	0
6	118653416	ENSBTAG00000012458	PSAPL1	-3	-1	-1	0	0	-5
7	1333975	ENSBTAG00000015602	C7H5orf45	-3	-3	-1	0	0	-7
7	5655357	ENSBTAG00000045588	LOC100298356	1	-3	-1	0	0	-3
7	9714914	ENSBTAG00000046423	LOC618007	0	0	-1	0	0	-1
7	16794179	ENSBTAG00000012314	LDLR	0	2	1	0	0	3
7	16835802	ENSBTAG00000009568	KANK2	0	1	-1	0	0	0
7	16862194	ENSBTAG00000009569	DOCK6	0	1	1	0	0	2
7	19740409	ENSBTAG00000000414	FUT5	-3	-3	-1	0	0	-7
7	19740905	ENSBTAG00000000414	FUT5	-3	-3	-1	0	0	-7

Mutation				Score					
Chr	**Pos**	**Gene**	**Gene name**	**Haplotypes**	**Conservation**	**Human defect**	**Mouse model**	**Function**	**Total Score**
7	26329353	ENSBTAG00000004860	SLC27A6	0	-1	-1	0	0	-2
8	60260609	ENSBTAG00000011420	CA9	-3	-1	-1	0	0	-5
8	60332484	ENSBTAG00000011433	RGP1	-3	0	-1	1	0	-3
8	65834564	ENSBTAG00000035708	LOC527795	0	0	-1	0	0	-1
8	70410155	ENSBTAG00000005466	C8H8orf58	-3	-3	-1	0	0	-7
8	76029774	ENSBTAG00000015027	Clusterin	0	0	-1	1	0	0
8	77002890	ENSBTAG00000012921	KIF24	1	2	-1	0	0	2
8	87279862	ENSBTAG00000002220	SPTLC1	0	0	1	0	0	1
8	111761816	ENSBTAG00000006532	CDK5RAP2	-3	-3	1	1	1	-3
8	112841695	ENSBTAG00000013838	ALLC	0	1	-1	0	0	0
9	8388881	ENSBTAG00000015335	BAI3	0	0	-1	0	0	-1
9	8388924	ENSBTAG00000015335	BAI3	0	0	-1	0	0	-1
9	23792838	ENSBTAG00000046593	Uncharacterized	0	3	0	0	0	3
9	51013392	ENSBTAG00000018528	USP45	0	2	-1	0	0	1
9	104739655	ENSBTAG00000018810	THBS2	0	3	1	0	1	5
9	105157029	ENSBTAG00000000918	ERMARD	-3	2	1	0	0	0
10	1712515	ENSBTAG00000014750	EPB41L4A	0	-1	-1	-1	0	-3
10	15843003	ENSBTAG00000009030	NOX5	-3	-3	-1	0	0	-7
10	28022352	ENSBTAG00000035309	OR4F15	-3	1	-1	0	0	-3
10	28023110	ENSBTAG00000035309	OR4F15	-3	-1	-1	0	0	-5
10	82903121	ENSBTAG00000012565	PCNX	-3	2	-1	0	1	-1
10	85173711	ENSBTAG00000011683	NUMB	-3	0	-1	0	0	-4
11	46305767	ENSBTAG00000002288	NT5DC4	0	-3	-1	0	0	-4
11	46753903	ENSBTAG00000019072	PSD4	0	-1	-1	0	0	-2

Mutation				Score					
Chr	Pos	Gene	Gene name	Haplotypes	Conservation	Human defect	Mouse model	Function	Total Score
11	47409577	ENSBTAG00000009033	TEX37	-3	-2	-1	1	0	-5
11	59791410	ENSBTAG00000006027	USP34	0	2	-1	0	1	2
11	78322010	ENSBTAG00000015490	HS1BP3	0	0	-1	0	0	-1
11	87943875	ENSBTAG00000001140	IAH1	-3	3	-1	1	0	0
11	95486109	ENSBTAG00000017576	GPR144	-3	-3	-1	0	0	-7
11	99453894	ENSBTAG00000038831	PHYHD1	1	1	-1	0	0	1
11	99461001	ENSBTAG00000010601	DOLK	1	2	1	0	0	4
12	11906181	ENSBTAG00000001133	VWA8	-3	1	-1	0	0	-3
12	12478433	ENSBTAG00000014752	AKAP11	1	-3	-1	-1	0	-4
12	12480845	ENSBTAG00000014752	AKAP11	1	-1	-1	-1	0	-2
12	14136897	ENSBTAG00000016691	LACC1	-3	-3	-1	0	0	-7
12	53049081	ENSBTAG00000032821	SCEL	-3	-3	-1	0	0	-7
12	90761209	ENSBTAG00000019961	ATP4B	-3	-3	-1	0	0	-7
13	3815664	ENSBTAG00000034991	SLX4IP	0	0	-1	0	0	-1
13	11787934	ENSBTAG00000008650	CAMK1D	-3	0	-1	0	0	-4
13	11787938	ENSBTAG00000008650	CAMK1D	-3	0	-1	0	0	-4
13	52471684	ENSBTAG00000013753	DDRGK1	0	2	-1	0	0	1
13	53934930	ENSBTAG00000039520	SIRPB1	0	-3	-1	0	0	-4
13	53934947	ENSBTAG00000039520	SIRPB1	0	0	-1	0	0	-1
13	53939853	ENSBTAG00000039520	SIRPB1	0	-3	-1	0	0	-4
13	54540425	ENSBTAG00000047150	RTEL	-3	0	-1	0	0	-4
13	55047740	ENSBTAG00000031718	OGFR	0	1	-1	0	0	0
13	59988205	ENSBTAG00000027390	RTFDC1	0	1	-1	0	0	0
13	63045750	ENSBTAG00000009144	BPIFA2A	0	-3	-1	0	0	-4

Mutation				Score					
Chr	Pos	Gene	Gene name	Haplotypes	Conservation	Human defect	Mouse model	Function	Total Score
13	63169610	ENSBTAG00000019752	BPIFA2B	0	-1	-1	0	0	-2
13	65396708	ENSBTAG00000006021	CEP250	0	-3	-1	0	0	-4
14	1977494	ENSBTAG00000026350	SPATC1	-3	-3	-1	0	0	-7
14	3640695	ENSBTAG00000015985	GPR20	0	0	-1	0	0	-1
14	35568945	ENSBTAG00000037399	Uncharacterized	3	0	-1	0	0	2
14	46885750	ENSBTAG00000047661	FABP12	0	-1	-1	0	0	-2
14	57184022	ENSBTAG00000026247	PKHD1L1	-3	3	-1	0	0	-1
14	68635363	ENSBTAG00000017165	MATN2	0	1	-1	0	0	0
15	182679	ENSBTAG00000046228	Olfactory receptor	0	0	-1	0	o	-1
15	183317	ENSBTAG00000046228	Olfactory receptor	0	0	-1	0	0	-1
15	183404	ENSBTAG00000046228	Olfactory receptor	0	0	-1	0	0	-1
15	248692	ENSBTAG00000045558	LOC100299084	0	0	-1	0	o	-1
15	30189209	ENSBTAG00000005357	VPS11	0	2	-1	0	0	1
15	35048788	ENSBTAG00000005340	SERGEF	0	-3	-1	0	0	-4
15	46301941	ENSBTAG00000002289	NLRP14	-3	-3	-1	0	0	-7
15	75033140	ENSBTAG00000012053	ACCSL	1	-3	-1	0	0	-3
15	75037645	ENSBTAG00000012053	ACCSL	-3	-3	-1	0	0	-7
15	80244489	ENSBTAG00000046527	LOC100139830	-3	0	-1	0	0	-4
15	80255710	ENSBTAG00000046285	OR8I2	1	0	-1	0	0	0
15	80288759	ENSBTAG00000001291	OR8H3	1	-1	-1	0	0	-1
15	80389745	ENSBTAG00000045709	LOC514057	1	3	-1	0	0	3
15	80485593	ENSBTAG00000035985	LOC101909743/OR8K1	1	-1	-1	0	0	-1
15	80555931	ENSBTAG00000022858	LOC783561	-3	0	-1	0	0	-4
15	80949689	ENSBTAG00000039110	OR8U1	1	-3	-1	0	0	-3

Mutation				Score					
Chr	Pos	Gene	Gene name	Haplotypes	Conservation	Human defect	Mouse model	Function	Total Score
15	82754556	ENSBTAG00000021375	OR10Q1	0	1	-1	0	0	0
15	82925332	ENSBTAG00000021233	OR5B17	-3	0	-1	0	0	-4
16	4562017	ENSBTAG00000019799	FCAMR	-3	-3	-1	0	0	-7
16	38319636	ENSBTAG00000014794	SCYL3	-3	0	-1	0	0	-4
16	62574355	ENSBTAG00000035226	TOR1AIP1	0	0	-1	0	0	-1
17	53095395	ENSBTAG00000020414	DHX37	1	-1	-1	0	0	-1
17	53112655	ENSBTAG00000020414	DHX37	0	-3	-1	0	0	-4
17	66091531	ENSBTAG00000010847	FOXN4	0	-1	-1	0	0	-2
17	72377532	ENSBTAG00000008996	SFI1	0	-3	-1	0	0	-4
18	25623550	ENSBTAG00000005170	GPR114	0	-3	-1	0	1	-3
18	34956513	ENSBTAG00000001286	ELMO3	3	3	-1	1	1	7
18	46372041	ENSBTAG00000015912	DMKN	0	1	-1	0	0	0
18	52154806	ENSBTAG00000001262	IRGQ	-3	1	-1	0	0	-3
18	54081277	ENSBTAG00000009337	PNMAL1	0	-3	-1	0	0	-4
18	54130955	ENSBTAG00000048021	PNMAL2	-3	1	-1	0	0	-3
18	57819768	ENSBTAG00000003231	LIM2	1	-3	1	0	0	-1
19	21444103	ENSBTAG00000011011	SSH2	-3	-3	-1	0	0	-7
19	31072152	ENSBTAG00000022509	DNAH9	-3	0	-1	0	0	-4
19	31113972	ENSBTAG00000022509	DNAH9	0	0	-1	0	0	-1
19	32790843	ENSBTAG00000015294	COX10	1	2	1	0	0	4
19	42251785	ENSBTAG00000013685	KRT31	-3	0	-1	0	0	-4
19	51395194	ENSBTAG00000015980	FASN	1	-1	1	0	0	1
19	56191175	ENSBTAG00000007910	EXOC7	3	-3	-1	0	0	-1
19	57266251	ENSBTAG00000011387	NAT9	1	2	-1	0	0	2

Mutation				Score					
Chr	Pos	Gene	Gene name	Haplotypes	Conservation	Human defect	Mouse model	Function	Total Score
20	7639801	ENSBTAG00000005633	RGNEF	1	-1	-1	1	0	0
20	23410431	ENSBTAG00000008871	DDX4	1	-3	-1	0	0	-3
20	58884236	ENSBTAG00000005514	TRIO	-3	0	-1	0	0	-4
20	64171739	ENSBTAG00000005021	SEMA5A	0	1	1	0	0	2
21	6239119	ENSBTAG00000016495	LINS	0	2	1	0	0	3
21	20345878	ENSBTAG00000012326	LOC101906218	0	0	-1	0	0	-1
21	20346289	ENSBTAG00000012326	LOC101906218	-3	0	-1	0	0	-4
21	26837733	ENSBTAG00000047587	Uncharacterized	1	0	0	0	0	1
21	31003599	ENSBTAG00000047166	SH2D7	0	-3	-1	0	0	-4
21	55274195	ENSBTAG00000039117	FAM179B	0	-3	-1	0	0	-4
21	55826463	ENSBTAG00000014001	MAP1A	0	-3	-1	0	0	-4
21	58173384	ENSBTAG00000009836	CHGA	-3	-1	-1	0	0	-5
21	59120719	ENSBTAG00000046590	FAM181A	0	1	-1	0	0	0
21	60482484	ENSBTAG00000017096	ERICH1	0	-3	-1	0	0	-4
21	60482510	ENSBTAG00000017096	ERICH1	0	1	-1	0	0	0
21	62843142	ENSBTAG00000012833	ATG2B	0	-3	-1	0	0	-4
22	52415959	ENSBTAG00000015839	MAP4	0	0	-1	0	0	-1
22	52421734	ENSBTAG00000015839	MAP4	1	-1	-1	0	0	-1
22	52422046	ENSBTAG00000015839	MAP4	-3	-3	-1	0	0	-7
22	52426104	ENSBTAG00000047174	Uncharacterized	-3	0	0	0	0	-3
22	52426668	ENSBTAG00000047174	Uncharacterized	1	0	0	0	0	1
22	56951657	ENSBTAG00000021645	MBD4	1	-1	-1	0	0	-1
22	56974847	ENSBTAG00000021643	EFCAB12	0	-1	-1	0	0	-2
22	57797674	ENSBTAG00000031115	Uncharacterized	0	1	0	0	0	1

Mutation				Score					
Chr	Pos	Gene	Gene name	Haplotypes	Conservation	Human defect	Mouse model	Function	Total Score
23	45885052	ENSBTAG00000005248	OFCC1	0	0	-1	0	0	-1
24	21324604	ENSBTAG00000000998	SLC39A6	0	2	-1	0	0	1
24	21453364	ENSBTAG00000011622	C18orf21	-3	-3	-1	0	0	-7
24	46764909	ENSBTAG00000015210	LOXHD1	0	-3	1	0	0	-2
24	50327195	ENSBTAG00000002660	CCDC11	-3	1	1	0	1	0
24	58158178	ENSBTAG00000014284	ALPK2	-3	-1	-1	0	0	-5
24	58158620	ENSBTAG00000014284	ALPK2	0	-1	-1	0	0	-2
24	58158784	ENSBTAG00000014284	ALPK2	-3	-3	-1	0	0	-7
24	58204003	ENSBTAG00000014284	ALPK2	-3	-3	-1	0	0	-7
24	58204479	ENSBTAG00000014284	ALPK2	-3	2	-1	0	0	-2
25	37433180	ENSBTAG00000040568	Uncharacterized	0	0	-1	0	0	-1
26	18116487	ENSBTAG00000046239	C10orf12	0	-1	-1	0	0	-2
26	25715996	ENSBTAG00000010522	Uncharacterized	1	0	0	0	0	1
26	25764570	ENSBTAG00000046727	Olfactory receptor	0	0	-1	0	0	-1
27	5031078	ENSBTAG00000007340	LOC101906349	0	-1	-1	0	0	-2
27	32832557	ENSBTAG00000011684	RAB11FIP1	1	0	-1	0	0	0
28	284228	ENSBTAG00000000079	CCSAP	0	-1	-1	0	0	-2
28	1690403	ENSBTAG00000048153	Uncharacterized	1	0	0	0	0	1
28	35718807	ENSBTAG00000047317	CL43	1	-1	-1	0	0	-1
28	44081765	ENSBTAG00000032527	ERCC6	-3	0	1	0	0	-2
29	1985881	ENSBTAG00000004081	FAT3	3	0	-1	0	0	2
29	7530523	ENSBTAG00000046360	Uncharacterized	0	1	0	0	0	1
29	7530663	ENSBTAG00000046360	Uncharacterized	1	-1	0	0	0	0
29	26397931	ENSBTAG00000013568	UEVLD	-3	-3	-1	0	0	-7

Interestingly, 13 of the 22 genes associated to human syndromes carry variants appearing in the short list of the 35 positives and only 6 in that of 135 with negative values. This in spite of the low weight (from -1 to 1) given to this parameter. Eight of the 35 positive variants are in proteins still uncharacterized.

Six variants had a score of 4 or higher. The highest score (total score=7) was for *ELMO3* (Engulfment and cell motility gene 3). This gene is involved in cell motility and migration. *ELMO3* is also a homolog of *C. elegans* Ced12 that is required for many developmental processes included gonadal morphogenesis. Mutant for this gene in *C. elegans* suffer from abnormal development, usually with lethal consequences or sub-lethal developmental defects, including small embryos, delayed development, and reduced fertility (Gumienny et al., 2001). In human it is expressed in testis, prostate and ovary (http://www.genecards.org/cgi-bin/carddisp.pl?gene=ELMO3).

The gene *DNAJC13* (DnaJ (Hsp40) homolog, subfamily C, member 13) had a total score of 5. This gene is involved in the onset of Parkinson disease. *DNAJC13* regulates the dynamics of clathrin coats on early endosomes. Cellular analysis shows that the mutation Arg855Ser identified by Vilariño-Güell et al. (2014) confers a gain-of-function that impairs endosomal transport. A possible role of *DNAJC13* gene in Tourette syndrome/chronic tic disorder is under investigation (Sundaram et al., 2011).

A second gene had a score of 5, *THBS2* (Thrombospondin 2). It belongs to the thrombospondin family and is a powerful inhibitor of tumour growth and angiogenesis. It is highly expressed in testis, in fetal Leydig cell. Hirose et al. (2008) found a significant association between an intronic SNP in the THBS2 gene and lumbar disc herniation. Thrombospondin-2 is also a matricellular protein found in human serum. Deletion of TSP-2 causes age-dependent dilated cardiomyopathy (Hanatani et al., 2014). Mutant mice for this gene display a variety of mutant phenotype; Kyriakides et al., 1998 suggest that *THBS2* modulates the cell surface properties of mesenchymal cells, affecting cell functions such as adhesion and migration.

Three genes had a score of 4, all three are associated to human syndromes. Mutation in *HERC2* (HECT and RLD domain containing E3 ubiquitin protein ligase 2) is involved in autosomal recessive mental retardation-38 (Puffenberger et al., 2012). Ji et al. (2000) identified in mutant mouse neuromuscular secretory vesicle defects and sperm acrosome defects. Moreover genetic variations in this gene are associated with skin/hair/eye pigmentation variability. Mutations in *DOLK* gene (dolichol kinase) are associated with dolichol kinase deficiency (Lefeber et al., 2011). *COX10* (cytochrome c oxidase assembly homolog 10) gene is associated to mitochondrial complex IV deficiency causing Leigh syndrome (Antonicka et al., 2003).

Conclusions

Exome sequencing therefore provided valuable information on coding sequence variants and on their potential effect on protein function. As a group, variants classified as deleterious were rarer than variants classified as neutral. Since this is an expected effect of purifying selection, this set of variant was likely enriched for variants having a phenotypic effect.

The inclusion in the analyses of plus and minus variant animals for fertility traits likely permitted to sequence the most important variants influencing these traits but the low number of animals sequenced combined with the complexity of the trait permitted to find only a suggestive association between SNPs and traits. It's likely that a much larger number of individuals is to be sequenced if genome wide association is the strategy ones want to pursue.

We explored an alternative strategy to confirm suggestive deleterious variants based on indirect evidence of the variant effect by assessing: i) evolutionary constraints associated to the amino acid residue coded by the variant; ii) variant behaviour in a large population. In the absence of direct genotyping of the candidate variants, we used haplotype information as surrogate; iii) function and effects in other species.

We attempted to quantify these criteria through a score subjective and only indicative, but permitted to highlighted 35 genes so far neglected in cattle, but deserving further attention and potentially having deleterious effects in Holstein and other cattle breeds. These may be part of the recessive deleterious set that constitute the genetic load of Italian Holstein.

References

Antonicka, H., Leary, S.C., Guercin, G.-H., Agar, J.N., Horvath, R., Kennaway, N.G., Harding, C.O., Jaksch, M., Shoubridge, E.A., 2003. Mutations in COX10 result in a defect in mitochondrial heme A biosynthesis and account for multiple, early-onset clinical phenotypes associated with isolated COX deficiency. Hum. Mol. Genet. 12, 2693–2702. doi:10.1093/hmg/ddg284

Ashwell, M.S., Heyen, D.W., Sonstegard, T.S., Van Tassell, C.P., Da, Y., VanRaden, P.M., Ron, M., Weller, J.I., Lewin, H.A., 2004. Detection of Quantitative Trait Loci Affecting Milk Production, Health, and Reproductive Traits in Holstein Cattle. Journal of Dairy Science 87, 468–475. doi:10.3168/jds.S0022-0302(04)73186-0

Aston, K.I., Carrell, D.T., 2009. Genome-Wide Study of Single-Nucleotide Polymorphisms Associated With Azoospermia and Severe Oligozoospermia. Journal of Andrology 30, 711–725. doi:10.2164/jandrol.109.007971

Aulchenko, Y.S., Ripke, S., Isaacs, A., Duijn, C.M. van, 2007. GenABEL: an R library for genome-wide association analysis. Bioinformatics 23, 1294–1296. doi:10.1093/bioinformatics/btm108

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., Shendure, J., 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12, 745–755. doi:10.1038/nrg3031

Biesecker, L.G., Shianna, K.V., Mullikin, J.C., 2011. Exome sequencing: the expert view. Genome Biology 12, 128. doi:10.1186/gb-2011-12-9-128

Biffani, S., Marusi, M., Biscarini, F., Canavesi, F., 2005. Developing a genetic evaluation for fertility using angularity and milk yield as correlated traits. Interbull Bulletin 0, 63.

Biffani, S., Samoré, A.B., Canavesi, F., 2002. Inbreeding depression for production, reproduction and functional traits in Italian Holstein cattle. Institut National de la Recherche Agronomique (INRA), pp. 0–4.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics btu170. doi:10.1093/bioinformatics/btu170

Browning, S.R., Browning, B.L., 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet 81, 1084–1097.

Charlesworth, D., Willis, J.H., 2009. The genetics of inbreeding depression. Nat Rev Genet 10, 783–796. doi:10.1038/nrg2664

Chun, S., Fay, J.C., 2009. Identification of deleterious mutations within three human genomes. Genome Res 19, 1553–1561. doi:10.1101/gr.092619.109

Consortium, T.E.P., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. doi:10.1038/nature11247

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491–498. doi:10.1038/ng.806

Goymer, P., 2007. Synonymous mutations break their silence. Nat Rev Genet 8, 92–92. doi:10.1038/nrg2056

Gumienny, T.L., Brugnera, E., Tosello-Trampont, A.-C., Kinchen, J.M., Haney, L.B., Nishiwaki, K., Walk, S.F., Nemergut, M.E., Macara, I.G., Francis, R., Schedl, T., Qin, Y., Van Aelst, L., Hengartner, M.O., Ravichandran, K.S., 2001. CED-12/ELMO, a Novel Member of the CrkII/Dock180/Rac Pathway, Is Required for Phagocytosis and Cell Migration. Cell 107, 27–41. doi:10.1016/S0092-8674(01)00520-7

Hanatani, S., Izumiya, Y., Takashio, S., Kimura, Y., Araki, S., Rokutanda, T., Tsujita, K., Yamamoto, E., Tanaka, T., Yamamuro, M., Kojima, S., Tayama, S., Kaikita, K., Hokimoto, S., Ogawa, H., 2014. Circulating thrombospondin-2 reflects disease severity and predicts outcome of heart failure with reduced ejection fraction. Circ. J. 78, 903–910.

Hirose, Y., Chiba, K., Karasugi, T., Nakajima, M., Kawaguchi, Y., Mikami, Y., Furuichi, T., Mio, F., Miyake, A., Miyamoto, T., Ozaki, K., Takahashi, A., Mizuta, H., Kubo, T., Kimura, T., Tanaka, T., Toyama, Y., Ikegawa, S., 2008. A functional polymorphism in THBS2 that affects alternative splicing and MMP binding is associated with lumbar-disc herniation. Am. J. Hum. Genet. 82, 1122–1129. doi:10.1016/j.ajhg.2008.03.013

https://genome.ucsc.edu/index.html [WWW Document], n.d.

Jamrozik, J., Fatehi, J., Kistemaker, G.J., Schaeffer, L.R., 2005. Estimates of Genetic Parameters for Canadian Holstein Female Reproduction Traits. Journal of Dairy Science 88, 2199–2208. doi:10.3168/jds.S0022-0302(05)72895-2

Janis E Wigginton, D.J.C., 2005. A note on exact tests of Hardy-Weinberg equilibrium. American journal of human genetics 76, 887–93. doi:10.1086/429864

Ji, Y., Rebert, N.A., Joslin, J.M., Higgins, M.J., Schultz, R.A., Nicholls, R.D., 2000. Structure of the Highly Conserved HERC2 Gene and of Multiple Partially Duplicated Paralogs in Human. Genome Res 10, 319–329.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, and D., 2002. The Human Genome Browser at UCSC. Genome Res. 12, 996–1006. doi:10.1101/gr.229102

Kumar, P., Henikoff, S., Ng, P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4, 1073–1081. doi:10.1038/nprot.2009.86

Kyriakides, T.R., Zhu, Y.H., Smith, L.T., Bain, S.D., Yang, Z., Lin, M.T., Danielson, K.G., Iozzo, R.V., LaMarca, M., McKinney, C.E., Ginns, E.I., Bornstein, P., 1998. Mice that lack thrombospondin 2 display connective tissue abnormalities that are associated with disordered collagen fibrillogenesis, an increased vascular density, and a bleeding diathesis. J. Cell Biol. 140, 419–430.

Lane, E.A., Crowe, M.A., Beltman, M.E., More, S.J., 2013. The influence of cow and management factors on reproductive performance of Irish seasonal calving dairy cows. Anim. Reprod. Sci. 141, 34–41. doi:10.1016/j.anireprosci.2013.06.019

Lefeber, D.J., de Brouwer, A.P.M., Morava, E., Riemersma, M., Schuurs-Hoeijmakers, J.H.M., Absmanner, B., Verrijp, K., van den Akker, W.M.R., Huijben, K., Steenbergen, G., van Reeuwijk, J., Jozwiak, A., Zucker, N., Lorber, A., Lammens, M., Knopf, C., van Bokhoven, H., Grünewald, S., Lehle, L., Kapusta, L., Mandel, H., Wevers, R.A., 2011. Autosomal Recessive Dilated Cardiomyopathy due to DOLK Mutations Results from Abnormal Dystroglycan O-Mannosylation. PLoS Genet 7, e1002427. doi:10.1371/journal.pgen.1002427

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, M.-X., Kwan, J.S.H., Bao, S.-Y., Yang, W., Ho, S.-L., Song, Y.-Q., Sham, P.C., 2013. Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. PLoS Genet 9, e1003143. doi:10.1371/journal.pgen.1003143

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., Albers, C.A., Zhang, Z., Conrad, D.F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M.A., Banks, E., Hu, M., Handsaker, R.E., Rosenfeld, J., Fromer, M., Jin, M., Mu, X.J., Khurana, E., Ye, K., Kay, M., Saunders, G.I., Suner, M.-M., Hunt, T., Barnes, I.H.A., Amid, C., Carvalho-Silva, D.R., Bignell, A.H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D.N., Xue, Y., Romero, I.G., Wang, J., Li, Y., Gibbs, R.A., McCarroll, S.A., Dermitzakis, E.T., Pritchard, J.K., Barrett, J.C., Harrow, J., Hurles, M.E., Gerstein, M.B., Tyler-Smith, C., 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335, 823–828. doi:10.1126/science.1215040

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A., Jabado, N., 2011. What can exome sequencing do for you? J Med Genet 48, 580–589. doi:10.1136/jmedgenet-2011-100223

McClure, M.C., Bickhart, D., Null, D., VanRaden, P., Xu, L., Wiggans, G., Liu, G., Schroeder, S., Glasscock, J., Armstrong, J., Cole, J.B., Van Tassell, C.P., Sonstegard, T.S., 2014. Bovine Exome Sequence Analysis and Targeted SNP Genotyping of Recessive Fertility Defects BH1, HH2, and HH3 Reveal a Putative Causative Mutation in SMC2 for HH3. PLoS ONE 9, e92769. doi:10.1371/journal.pone.0092769

McClure, M., Kim, E., Bickhart, D., Null, D., Cooper, T., Cole, J., Wiggans, G., Ajmone-Marsan, P., Colli, L., Santus, E., Liu, G.E., Schroeder, S., Matukumalli, L., Van Tassell, C., Sonstegard, T., 2013. Fine Mapping for Weaver Syndrome in Brown Swiss Cattle and the Identification of 41 Concordant Mutations across NRCAM, PNPLA8 and CTTNBP2. PLoS ONE 8, e59251. doi:10.1371/journal.pone.0059251

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. doi:10.1101/gr.107524.110

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., Thomas, P.D., 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucl. Acids Res. 38, D204–D210. doi:10.1093/nar/gkp1019

Mrode, R., Kearney, J.F., Biffani, S., Coffey, M., Canavesi, F., 2009. Short communication: Genetic relationships between the Holstein cow populations of three European dairy countries. Journal of Dairy Science 92, 5760–5764. doi:10.3168/jds.2008-1931

Peloso, G.M., Auer, P.L., Bis, J.C., Voorman, A., Morrison, A.C., Stitziel, N.O., Brody, J.A., Khetarpal, S.A., Crosby, J.R., Fornage, M., Isaacs, A., Jakobsdottir, J., Feitosa, M.F., Davies, G., Huffman, J.E., Manichaikul, A., Davis, B., Lohman, K., Joon, A.Y., Smith, A.V., Grove, M.L., Zanoni, P., Redon, V., Demissie, S., Lawson, K., Peters, U., Carlson, C., Jackson, R.D., Ryckman, K.K., Mackey, R.H., Robinson, J.G., Siscovick, D.S., Schreiner, P.J., Mychaleckyj, J.C., Pankow, J.S., Hofman, A., Uitterlinden, A.G., Harris, T.B., Taylor, K.D., Stafford, J.M., Reynolds, L.M., Marioni, R.E., Dehghan, A., Franco, O.H., Patel, A.P., Lu, Y., Hindy, G., Gottesman, O., Bottinger, E.P., Melander, O., Orho-Melander, M., Loos, R.J.F., Duga, S., Merlini, P.A., Farrall, M., Goel, A., Asselta, R., Girelli, D., Martinelli, N., Shah, S.H., Kraus, W.E., Li, M., Rader, D.J., Reilly, M.P., McPherson, R., Watkins, H., Ardissino, D., Zhang, Q., Wang, J., Tsai, M.Y., Taylor, H.A., Correa, A., Griswold, M.E., Lange, L.A., Starr, J.M., Rudan, I., Eiriksdottir, G., Launer, L.J., Ordovas, J.M., Levy, D., Chen, Y.-D.I., Reiner, A.P., Hayward, C., Polasek, O., Deary, I.J., Borecki, I.B., Liu, Y., Gudnason, V., Wilson, J.G., van Duijn, C.M., Kooperberg, C., Rich, S.S., Psaty, B.M., Rotter, J.I., O'Donnell, C.J., Rice, K., Boerwinkle, E., Kathiresan, S., Cupples, L.A., 2014. Association of Low-Frequency and Rare Coding-Sequence Variants with Blood Lipids and Coronary Heart Disease in 56,000 Whites and Blacks. The American Journal of Human Genetics 94, 223–232. doi:10.1016/j.ajhg.2014.01.009

Pryce, J.E., Veerkamp, R.F., Thompson, R., Hill, W.G., Simm, G., 1997. Genetic aspects of common health disorders and measures of fertility in Holstein Friesian dairy cattle. Animal Science 65, 353–360. doi:10.1017/S1357729800008559

Pryce, J.E., Woolaston, R., Berry, D.P., Wall, E., Winters, M., Butler, R., Shaffer, M., 2014. World Trends in Dairy Cow Fertility, in: 10 Th World Congress of Genetics Applied to Livestock Production.

Puffenberger, E.G., Jinks, R.N., Wang, H., Xin, B., Fiorentini, C., Sherman, E.A., Degrazio, D., Shaw, C., Sougnez, C., Cibulskis, K., Gabriel, S., Kelley, R.I., Morton, D.H., Strauss, K.A., 2012. A homozygous missense mutation in HERC2 associated with global developmental delay and autism spectrum disorder. Hum. Mutat. 33, 1639–1646. doi:10.1002/humu.22237

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A

Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81, 559–575.

Sahana, G., Nielsen, U.S., Aamand, G.P., Lund, M.S., Guldbrandtsen, B., 2013. Novel Harmful Recessive Haplotypes Identified for Fertility Traits in Nordic Holstein Cattle. PLoS ONE 8, e82909. doi:10.1371/journal.pone.0082909

Schütz, E., Scharfenstein, M., Brenig, B., 2008. Implication of complex vertebral malformation and bovine leukocyte adhesion deficiency DNA-based testing on disease frequency in the Holstein population. J. Dairy Sci. 91, 4854–4859. doi:10.3168/jds.2008-1154

Simmons, M.J., Crow, J.F., 1977. Mutations Affecting Fitness in Drosophila Populations. Annual Review of Genetics 11, 49–78. doi:10.1146/annurev.ge.11.120177.000405

Soggiu, A., Piras, C., Hussein, H.A., Canio, M.D., Gaviraghi, A., Galli, A., Urbani, A., Bonizzi, L., Roncada, P., 2013. Unravelling the bull fertility proteome. Mol. BioSyst. 9, 1188–1195. doi:10.1039/C3MB25494A

Stitziel, N.O., Kiezun, A., Sunyaev, S., 2011. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biology 12, 227. doi:10.1186/gb-2011-12-9-227

Sundaram, S.K., Huq, A.M., Sun, Z., Yu, W., Bennett, L., Wilson, B.J., Behen, M.E., Chugani, H.T., 2011. Exome sequencing of a pedigree with Tourette syndrome or chronic tic disorder. Ann. Neurol. 69, 901–904. doi:10.1002/ana.22398

VanRaden, P.M., Olson, K.M., Null, D.J., Hutchison, J.L., 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. Journal of Dairy Science 94, 6153–6161. doi:10.3168/jds.2011-4624

Vilariño-Güell, C., Rajput, A., Milnerwood, A.J., Shah, B., Szu-Tu, C., Trinh, J., Yu, I., Encarnacion, M., Munsie, L.N., Tapia, L., Gustavsson, E.K., Chou, P., Tatarnikov, I., Evans, D.M., Pishotta, F.T., Volta, M., Beccano-Kelly, D., Thompson, C., Lin, M.K., Sherman, H.E., Han, H.J., Guenther, B.L., Wasserman, W.W., Bernard, V., Ross, C.J., Appel-Cresswell, S., Stoessl, A.J., Robinson, C.A., Dickson, D.W., Ross, O.A., Wszolek, Z.K., Aasly, J.O., Wu, R.-M., Hentati, F., Gibson, R.A., McPherson, P.S., Girard, M., Rajput, M., Rajput, A.H., Farrer, M.J., 2014. DNAJC13 mutations in Parkinson disease. Hum. Mol. Genet. 23, 1794–1801. doi:10.1093/hmg/ddt570

Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucl. Acids Res. 38, e164–e164. doi:10.1093/nar/gkq603

Wathes, D.C., Pollott, G.E., Johnson, K.F., Richardson, H., Cooke, J.S., 2014. Heifer fertility and carry over consequences for life time production in dairy and beef cattle. Animal 8 Suppl 1, 91–104. doi:10.1017/S1751731114000755

Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., Hardison, M., Person, R., Bekheirnia, M.R., Leduc, M.S., Kirby, A., Pham, P., Scull, J., Wang, M., Ding, Y., Plon, S.E., Lupski, J.R., Beaudet, A.L., Gibbs, R.A., Eng, C.M., 2013. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. New England Journal of Medicine 369, 1502–1511. doi:10.1056/NEJMoa1306555

Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Tassell, C.P.V., Sonstegard, T.S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J.A., Salzberg, S.L., 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 10, R42. doi:10.1186/gb-2009-10-4-r42

Supplementary files

You can find Chapter 6 supplementary files at:

- DocTa (Doctoral Thesis Archive) (http://tesionline.unicatt.it/)
- Google Drive (http://tinyurl.com/PhDMMChapter06) or scan the QR code

Supplementary Figure S1. Coverage analyses

A) Average coverage per animals

B) Boxplot of chromosomes coverage per animal

C) Boxplot of animal coverage per chromosome. Blue line represent the expected 50X coverage

Supplementary Figure S2. Distribution of variant sequencing depth

Supplementary Figure S3. Correlation between sequencing depth and number of discordant genotypes. In red values for two outlier animals having concordance <84.6%.

Supplementary Figure S4. Multidimensional scaling of Italian Holstein bulls sequenced.

Supplementary Figure S5. Multidimensional scaling of Italian Holstein bulls analysed with the 800K SNPchip

Supplementary Figure S6 MAF distribution on neutral (blue) and deleterious (red) variants

Supplementary Table S1. EBV distribution on Italian Holstein population and SNPchip dataset

Sheet 1: EBV threshold value to define plus and minus tail for PROT (kg protein), SCC (somatic cell count), FERT (fertility) and LONG (functional longevity). * All Italian Holstein FA bulls were include in this analysis

Sheet 2: Number of dataset animals in each tail for each trait.

Supplementary Table S2. Exome probes statistics: number and bp covered per chromosomes

Supplementary Table S3: Concordance between animal genotypes in sequencing and SNPchip data

Error1: HD data: homozygotes – Exome data: heterozygotes; Error2: HD data: heterozygotes – Exome data: homozygotes

Supplementary Table S4. List of suggestive genes deriving from the analysis of exomes of animals in the tails of male and female fertility EBV distributions.

Header code:

- Bfreq_tot: Frequency of B allele in entire dataset;
- Bfreq_MH: Frequency of B allele in high male fertility dataset;
- Bfreq_ML: Frequency of B allele in high female fertility dataset;
- Bfreq_FH: Frequency of B allele in low male fertility dataset;
- Bfreq_FL: Frequency of B allele in low male fertility dataset;
- CallRate_tot: SNP call rate – entire dataset;
- CallRate_MH: SNP call rate – high male fertility dataset;
- CallRate_ML: SNP call rate – low male fertility dataset;
- CallRate_FH: SNP call rate – high female fertility dataset;
- CallRate_FL: SNP call rate – high male fertility dataset;
- HWeq: Hardy-Weinberg equilibrium p-value;
- Geno_HomoMinor: Number of animals with homozygote genotype for minor allele;
- Geno_Het: Number of animals with heterozygote genotype;

- Geno_HomoMajor: Number of animals with homozygote genotype for major allele;
- LOW_MinorAllele: minor allele for low fertility (male and female) dataset;
- LOW_MAF: minor allele frequency from low fertility (male and female) dataset;
- LOW_CR: SNP call rate from low fertility (male and female) dataset;
- LOW_HomoMinor: Number of animals from low fertility (male and female) dataset with homozygote genotype for minor allele;
- LOW_Het: Number of animals from low fertility (male and female) dataset with heterozygote genotype;
- LOW_HomoMajor: Number of animals from low fertility (male and female) dataset with homozygote genotype for major allele;
- LOW_HWeqPvalue: Hardy-Weinberg equilibrium p-value from low fertility (male and female) dataset;
- HIGH_MinorAllele: minor allele from high fertility (male and female) dataset;
- HIGH_MAF: minor allele frequency from high fertility (male and female) dataset;
- HIGH_CR: SNP call rate from high fertility (male and female) dataset;
- HIGH_HomoMinor: Number of animals from high fertility (male and female) dataset with homozygote genotype for minor allele;
- HIGH_Het: Number of animals from high fertility (male and female) dataset with heterozygote genotype;
- HIGH_HomoMajor: Number of animals from high fertility (male and female) dataset with homozygote genotype for major allele;
- HIGH_HWeqPvalue: Hardy-Weinberg equilibrium p-value from high fertility (male and female) dataset;
- Male_TREND: inheritance models implemented in PLINK, trend test, on male fertility dataset;
- Male_ GENOTYPES: inheritance models implemented in PLINK, genotypic test, on male fertility dataset;
- Male_ RECESSIVE: inheritance models implemented in PLINK, recessive test, on male fertility dataset;
- Female_TREND: inheritance models implemented in PLINK, trend test, on female fertility dataset;
- Female_GENOTYPES: inheritance models implemented in PLINK, genotypic test, on female fertility dataset;
- Female_RECESSIVE: inheritance models implemented in PLINK, recessive test, on female fertility dataset;
- Male&Female_TREND: inheritance models implemented in PLINK, trend test, on male and female fertility dataset;
- Male&Female_GENOTYPES: inheritance models implemented in PLINK, genotypic test, on male and female fertility dataset;

- Male&Female_RECESSIVE: inheritance models implemented in PLINK, recessive test, on male and female fertility dataset;
- EMP1_Male: Empirical p-value on male fertility dataset;
- EMP1_Female: Empirical p-value on female fertility dataset;
- EMP1_Male&Female: Empirical p-value on male and female fertility dataset.

Supplementary Table S5. Regions candidate to carry deleterious variants identified by OHW approach.

Supplementary Table S6. Regions candidate to carry deleterious variants identify by LOH approach.

In grey regions removed in further analysis.

Supplementary Table S7. List of candidate deleterious variants mapping in OHW and LOH regions

Header code:

- Bfreq_tot: Frequency of B allele in entire dataset;
- Bfreq_MH: Frequency of B allele in high male fertility dataset;
- Bfreq_ML: Frequency of B allele in high female fertility dataset;
- Bfreq_FH: Frequency of B allele in low male fertility dataset;
- Bfreq_FL: Frequency of B allele in low male fertility dataset;
- CallRate_tot: SNP call rate – entire dataset;
- CallRate_MH: SNP call rate – high male fertility dataset;
- CallRate_ML: SNP call rate – low male fertility dataset;
- CallRate_FH: SNP call rate – high female fertility dataset;
- CallRate_FL: SNP call rate – high male fertility dataset;
- HWeq: Hardy-Weinberg equilibrium p-value;
- Geno_HomoMinor: Number of animals with homozygote genotype for minor allele;
- Geno_Het: Number of animals with heterozygote genotype;
- Geno_HomoMajor: Number of animals with homozygote genotype for major allele;
- HW_inside: Deleterious mutation/s inside "Out of Hardy-Weinberg" region;
- HW_outside: Deleterious mutation/s outside "Out of Hardy-Weinberg" region;
- DH_inside: Deleterious mutation/s inside "Lack of Homozygosity" region;
- DH_outside: Deleterious mutation/s outside "Lack of Homozygosity" region;

- GeneClass: Classifications of gene - 1: deleterious mutation/s inside OHW and LOH region, 2: deleterious mutation/s inside OHW or LOH region;
- 2: deleterious mutation/s outside OHW and/or LOH region;
- LOW_MinorAllele: minor allele for low fertility (male and female) dataset;
- LOW_MAF: minor allele frequency from low fertility (male and female) dataset;
- LOW_CR: SNP call rate from low fertility (male and female) dataset;
- LOW_HomoMinor: Number of animals from low fertility (male and female) dataset with homozygote genotype for minor allele;
- LOW_Het: Number of animals from low fertility (male and female) dataset with heterozygote genotype;
- LOW_HomoMajor: Number of animals from low fertility (male and female) dataset with homozygote genotype for major allele;
- LOW_HWeqPvalue: Hardy-Weinberg equilibrium p-value from low fertility (male and female) dataset;
- HIGH_MinorAllele: minor allele from high fertility (male and female) dataset;
- HIGH_MAF: minor allele frequency from high fertility (male and female) dataset;
- HIGH_CR: SNP call rate from high fertility (male and female) dataset;
- HIGH_HomoMinor: Number of animals from high fertility (male and female) dataset with homozygote genotype for minor allele;
- HIGH_Het: Number of animals from high fertility (male and female) dataset with heterozygote genotype;
- HIGH_HomoMajor: Number of animals from high fertility (male and female) dataset with homozygote genotype for major allele;
- HIGH_HWeqPvalue: Hardy-Weinberg equilibrium p-value from high fertility (male and female) dataset;
- Male_TREND: inheritance models implemented in PLINK, trend test, on male fertility dataset;
- Male_ GENOTYPES: inheritance models implemented in PLINK, genotypic test, on male fertility dataset;
- Male_ RECESSIVE: inheritance models implemented in PLINK, recessive test, on male fertility dataset;
- Female_TREND: inheritance models implemented in PLINK, trend test, on female fertility dataset;
- Female_GENOTYPES: inheritance models implemented in PLINK, genotypic test, on female fertility dataset;
- Female_RECESSIVE: inheritance models implemented in PLINK, recessive test, on female fertility dataset;
- Male&Female_TREND: inheritance models implemented in PLINK, trend test, on male and female fertility dataset;

- Male&Female_GENOTYPES: inheritance models implemented in PLINK, genotypic test, on male and female fertility dataset;
- Male&Female_RECESSIVE: inheritance models implemented in PLINK, recessive test, on male and female fertility dataset;
- EMP1_Male: Empirical p-value on male fertility dataset;
- EMP1_Female: Empirical p-value on female fertility dataset;
- EMP1_Male&Female: Empirical p-value on male and female fertility dataset.

Supplementary Table S8. Results of haplotype analysis.

Header code:

- Chrom: chromosome
- Pos: physical position on UMD 3.1 genome
- Exome_homo: number of homozygote, for deleterious mutation, sequenced animals
- Exome_het: number of heterozygote sequenced animals for deleterious mutation
- Outcome: *NotFound*: It was impossible to associated a haplotype to the mutation; *NotOpt*: Non completely invariant haplotype could be associated to the mutation; *OK*: Putative haplotype associate to mutation was found. In few case alternative haplotype was reported.
- Len: length of haplotype, in number of SNPs
- Pop_num: numbers of selected haplotype detected in Holstein population
- Pop_homo: numbers of homozygote animals in Holstein population for selected haplotype
- Pop_het: numbers of heterozygotes animals in Holstein population for selected haplotype
- Exp: Expected number of homozygote animal
- Alternative: if only one heterozygote sequenced animals for deleterious mutation was found all two possible haplotypes where tested

Supplementary Table S9. Descriptive results for the Biological Process (BP)

Number of genes belonging to Biological Process GO terms.

Supplementary Table S10. Descriptive results for the Molecular Function (MF)

Number of genes belonging to Molecular Function GO terms.

CHAPTER 7

Conclusion

Since the domestication event, occurred in the Fertile Crescent around 10,000 years ago, the bovine species was selected by humankind to fulfil his needs. The creation of breeds, occurred around 200 years ago, intensified the differentiation between populations and promoted the development of intensive selection schemes. Traditional- and nowadays genomic- selections are producing positive genetic trends in many productive traits, despite the still poor knowledge about phenotype biology. A deeper understanding of genome organization and gene biology would increase the accuracy of genomic evaluation by incorporating prior knowledge.

This thesis explores the bovine genome by high-throughput technologies - such as Next Generation Sequencing and High-Density Genotyping - with established and innovative procedures, to investigate the biology of complex traits and to provide new data and molecular tools to bovine breeding.

These goals have been achieved by complementary approaches and different methods. In Chapter 2 selection signatures were investigated independently into five Italian breeds using Illumina BovineSNP50 BeadChip. Then, a multi-breed approach was applied, clustering breeds into dairy (Italian Holstein, Italian Brown and Italian Simmental) and beef (Marchigiana, Piedmontese and Italian Simmental) groups. Regions under recent positive selection shared by breeds with a same productive aptitude were investigated in further detail by retrieving genes in the region and analysing their function by ontology and pathway analysis. In the dairy group, selection signals appear nearby genes related to mammary gland (metabolism and resistance to mastitis) and feeding adaptation. In the beef group, genes in selected regions are involved in animal growth and meat quality (texture and juiciness). Hence, candidate genes identified belong to networks and pathways important in directing a breed towards either beef or dairy production.

In Chapter 3, a GWAS approach was used in dairy cattle to correlate SNP to complex phenotypes having an economic value. Three independent "classical" single-marker regressions were run in three Italian dairy cattle (Italian Holstein, Italian Brown and Italian Simmental). Using the traditional single SNP approach, only few SNPs were above the nominal genome-wide significance threshold, while applying a gene-based method identified significant candidate genes associated with milk physiology and mammary gland development in all three populations. Interestingly, genes found in different breeds have similar function but are not the same, suggesting that breed history, demography, selection criteria and stochastic events, as genetic drift, have an impact on the definition of the main molecular target of selection schemes.

Current intensive selection strategies rapidly spread the genetic gain in the population but hold the associated risk of disseminating defective genes causing genetic defects (*e.g.* mulefoot, complex vertebral malformation and others) or decreasing fertility. In Chapter 6, different technologies and approaches were combined to obtain, filter and prioritize genes candidate to be deleterious. Exome sequence data were exploited to identify deleterious variants affecting Italian Holstein cattle in a set of 20 animals extreme for fertility traits. More than 5000 candidate deleterious SNP variants were identified. To bypass the small number limitation and the lack of power of any statistic analysis on 20 animals, a stepwise approach was used to first filter and then prioritize these candidate deleterious variants. Polymorphisms were retained if having asymmetric distribution in bulls from opposite tails of fertility EBV distributions and/or mapping in genomic regions outlier for lack of homozygosity in a 1009 bull population genotyped with the Illumina BovineHD BeadChip. A total of 191 variants in 163 genes passed the processes. These were assessed by searching additional evidences in favour or against their deleterious effect through different approaches: i) checking the variant conservation across 100 vertebrate species by comparative genomics; ii) finding haplotypes associated with candidate mutations and evaluating their frequency in Holsteins; iii) integrating human and mouse functional annotations. A score was proposed to weight the result of these assessments and rank the variants: 35 of them had positive scores. These occurred in genes involved in basic biological mechanisms as fertility, development, and immunity and resulted enriched in genes associated to genetic defects reported in human. These mutations are strong candidates to be recessive deleterious variants worth to be further investigated in a larger population to better understand their biology, impact and possible application in breeding.

To increase the reliability of results based on genomic data and to reduce the impact of subjective choices in data analysis, new tools/approaches were developed (*i.e.* the multi-breed approach in Selection Signature – Chapter 2; the post-GWAS gene-based approach – Chapters 3 and 4) and the existent ones critically evaluated (imputation methods – Chapter 5).

In Chapter 4, the post-GWAS gene-based method used in Chapter 3 was implemented in the MUGBAS (MUlti species Gene-Based Association Suite) software. In contrast to existing software, MUGBAS is species and annotation free. A multiple-testing correction was included in the package, in addition to computing parallelization, for faster processing, and plot representation for better understanding. The software uses single-marker GWAS data and a given annotation to estimate gene-/region-wise association p-values. As previously mentioned in Chapter 3, the "classical" GWAS approach (single SNP) found signals only in Italian Holstein, while MUGBAS gene-based approach identified significant signals in all breeds investigated.

In Chapter 5, the impact of different bovine reference sequences on genotype imputation was investigated. In this work, Illumina BovineSNP50 BeadChip genotypes of Italian Simmental bulls were re-mapped on the three most recent reference genome assemblies. We compared four different imputation methods from low (Illumina BovineLD v.1.0 Beadchip) to high density. The test population was divided into four subpopulations on the basis of relationship and genotype data availability. Updating SNP coordinates on the three tested cattle reference genome assemblies determined only a slight variation on imputation results within methods. On the other hand, large differences in terms of accuracy were observed among the four imputation methods tested.

Genetic structure of industrial breeds was evaluated assessing Linkage Disequilibrium (LD). Since LD decay is correlated to intensity of selection we expected different behaviours in the breeds analysed. Indeed, as evident in Chapter 2 – Figure 7, breeds subjected to lower selection intensity (*i.e.*

Marchigiana, Piedmontese and Italian Simmental) display a lower persistence of LD at greater distances compared to those subjected to higher selection pressure (*i.e.* Italian Holstein and Italian Brown). Chapter 3 investigates LD in the *DGAT1* region of dairy cattle breeds. In Italian Simmental, LD is lower compared to Italian Holstein and Italian Brown, In spite of this, significant association between the *DGAT1* region and milk traits was found both in Italian Simmental and in Italian Holstein. In Italian Brown no association was found because *DGAT1* is fixed.

Data and results presented in this thesis represent a clear example of the progress of molecular technologies occurred in last few years, spanning from medium density genotyping to next generation sequencing.

The biology landscapes of some complex traits (*i.e.* milk production and quality, fertility) were improved, finding candidate genes and pathways not previously associated to the phenotype in the bovine species. Moreover, insights on deleterious mutations were inferred through a deep and integrated analysis using NGS data.

This new knowledge could be used to improve cattle breeding. For example, deleterious recessive variants could be included in bull evaluations by breeders' associations to reduce the genetic load of deleterious genes in parallel to improving productivity.

While technological progress has been impressive, we are still only scratching the surface of animal biology. New frontiers are being opened by whole genome sequencing of many animals (*e.g.* in the 1000 bull genome project) and by stepping beyond sequencing (a number of epigenomic projects in animals are following the footprints of the human ENCODE project). The trend is towards unravelling complexity and data production ability is to be flanked by data storing and processing and above all by new methods and tools for data analysis.

ACKNOWLEDGMENTS

"Non è la destinazione, ma il viaggio che conta"

Questi tre anni sono stati un gran bel viaggio, di crescita personale e ricchi di esperienze. Non è stato un viaggio solitario ma condiviso con vecchie e nuove conoscenze che mi sembra doveroso menzionare.

Innanzitutto vorrei ringraziare il prof Paolo Ajmone-Marsan che mi ha dato la possibilità di fare questa esperienza. Grazie per gli insegnamenti e per le opportunità che mi ha offerto. Spero di non aver tradito le sue aspettative.
Agli attuali colleghi, Lorenzo, Stefano, Licia, Elisa, Riccardo ed Elia, e a quelli passati, Nicola, Fatima, Raffaele, Ezequiel e Barbara, va il mio più sincero ed affettuoso ringraziamento per aver condiviso, anche solo in parte, questo mio percorso. Grazie per avermi sopportato (e non è poco), guidato (non solo nel campo lavorativo) e fatto vivere in un modo speciale l'ambiente di lavoro.

Un sincero grazie va a tutti i miei colleghi del XXVII ciclo Agrisystem. Ricorderò sempre tutti i momenti che ho potuto condividere con voi.

Un grazie anche a chi ha fatto parte della mia esperienza svedese. Grazie di cuore a Carl, Flavio, Christian, Erica, a tutti i ragazzi italiani, a tutti i colleghi "svedesi" e a tutte le altre persone che ho conosciuto. Mi avete fatto vivere appieno l'esperienza estera e fatto tornare a casa con qualche consapevolezza in più.

Un sentito ringraziamento a tutti gli altri colleghi con cui ho collaborato in questi anni.

Ringrazio tutti gli amici e conoscenti che mi sono stati vicino per avermi supportato e sopportato, ognuno a modo proprio.

Ultimo, ma non per importanza, il ringraziamento che va a tutta la mia famiglia per avermi aiutato, spronato, sostenuto sempre e in particolare in questa esperienza. Semplicemente grazie.

P.S.: A dirla tutta, anche la destinazione non è poi così male!