

This is the author's accepted manuscript. The final publication is available at Springer via <http://dx.doi.org/10.3103/S0891416816040054>.

The full details of the published version of the article are as follows:

TITLE: The genomes and history of domestic animals

AUTHORS: Larkin, D M; Yudin, N S

JOURNAL TITLE: Molecular Genetics, Microbiology and Virology

PUBLICATION DATE: October 2016

PUBLISHER: Springer Verlag

DOI: 10.3103/S0891416816040054

D.M. Larkin^{1,2}, N.S. Yudin^{1,3}

The Genomes and History of Domestic Animals

¹The Federal Research Center Institute of Cytology and Genetics The Siberian Branch of the Russian Academy of Sciences, 630090, Novosibirsk;²The Royal Veterinary College, University of London, NW1 0TU, London, United Kingdom; ³Novosibirsk State University, 630090, Novosibirsk

Larkin Denis Mikhailovich, dmlarkin@gmail.com

Yudin Nikolay Serafimovich, yudin@bionet.nsc.ru

Summary

This paper reviews how mammalian genomes are utilized in modern genetics for the detection of genes and polymorphisms (mutations) within domesticated animal (mostly livestock) genomes that are linked to traits of economic importance to humans. Examples are given of how genetic analysis allows to determine key genes associated with the quality and quantity of milk in cattle and key genes for meat production. Various questions are reviewed, such as how contemporary methods of genome sequencing allow to maximise the effective detection of genic and regulatory DNA polymorphisms within the genomes of major domesticated mammals (cattle, sheep and pigs) and the history of their formation from the standpoint of genetics.

The Structure of the Genome

Within mammalian genomes, up to 40% of the DNA is comprised of simple repeats, DNA- and retro-transposons, SINE and LINE elements, most of which do not encode for proteins (Fig. 1). Out of 30% of DNA that belongs to genes, only 2% is coding DNA (exons). Heterochromatin DNA, repetitive and unique noncoding DNA sequences comprise a further 30% of the genome. Variation within DNA sequences in the genome as a result of errors during replication is the basis for natural (during evolution) or artificial (animal breeding) selection.

The main sources of variation of DNA in genomes include: a) point (nucleotide) mutations within genes (originally resulting from mistakes during DNA replication). Point mutations can lead to changes in the protein if the encoded amino-acid undergoes changes or lead to regulatory modification if the regulatory sequences of the gene are disturbed. It is important to note, that the majority of point mutations don't lead to changes of amino-acids. Most of them are neutral (synonymous) and don't have an immediate effect on the protein. The structure and regulation of the gene can be effected by: b) insertions of

transposons inside or near genes; c) segmental duplications of DNA and chromosomal rearrangements mediated by them, which likely to provide a “gentler” influence, changing the positions of a gene(s) within the chromosome, thus affecting their regulation. Segmental duplications can also lead to a higher number of copies, if genes are duplicated, thus leading to a higher level of gene expression. Traditional ways of searching for genes that affect economically important traits in domesticated animals use the fact that there is a high probability that the gene that controls a trait in one species (for example, a human), will have a similar function in other species of animals. Approximately 12,000 genes within the genome of a placental mammal are also represented by one copy within genomes of other placental mammals; i.e. one gene, for example, in a mouse will have a single ortholog with the similar structure and function in a human, a cow or a sheep. About 4,000 genes have multiple copies in different species of animals. Approximately 500 or more genes are unique to every species [1]. The existence of orthologous genes with similar functions in mammals allows us to use information about the functions of genes from well-studied species (e.g. the human or the mouse) for the search of “candidate genes” that have the same function within the genes of animals that have not been as carefully studied; like the genomes of livestock species.

Genes and mutations affecting economically-important traits in domesticated animals

The principle of “candidate gene” search is presented in Fig. 2. Through the use of linkage mapping techniques, it can be determined that a specific chromosomal locus controls an economically important trait in livestock. Using the methods of comparative mapping, one can find the homeologous regions of human or mouse chromosomes corresponding to this locus. The genes that are present in human or mouse homeologous intervals are screened for ones performing a role similar to the one which is studied in the

livestock animal. Gene(s) that control similar traits in a human or mouse, are then cloned and sequenced in the animal. The mutations (e.g. point mutations) are determined in these genes in livestock animal populations, and an associative analysis is conducted, which reveals the relationship (or lack thereof) between the animal's phenotype and specific nucleotide substitution(s). If an association of this sort is observed, then the gene that influences a particular trait in the animal's genome was identified correctly. Obviously, this approach has a number of drawbacks, namely: a) it "works" only for genes with similar functions in different species, and; b) generally, mutations in regulatory regions of genes are not detected by this approach because they might be located outside the cloned region. Nevertheless, with the help of the candidate gene search technique, important genes have been identified as such the gene influencing the increase in muscle mass in cattle [2]. On the linkage map of the cattle, this trait is mapped (among other locations) to a locus on chromosome 2. This locus corresponds to region 2q31-22 in the human genome, which, among other genes, contains *MSTN* gene. The knock out of the *Mstn* gene in the mouse genome significantly increases muscle mass. As a result, *MSTN* was chosen as a "candidate gene". After sequencing of the *MSTN* gene in cattle, mutations segregating with the muscle mass trait had been found.

In cases where the trait is controlled by multiple genes, the situation becomes more complicated, since often the contribution of each of these genes to the phenotype is quite small and often it is not possible to find one or two major genes/mutations that would significantly change phenotype in the desired direction. As shown in Table 1, most of the economically important traits of livestock species are "quantitative" traits.

An example of a successful identification of a mutation that has had considerable effect on economically important quantitative trait through candidate gene approach had been the detection of a mutation affecting the fat percentage in cattle milk. Among other chromosomes that trait had been mapped in the centromeric region of cattle chromosome

14 [3]. The locus influencing this polygenic trait was in a region homologous to a chromosomal interval of the mouse genome containing the gene *Dgat1*. The knockout of *Dgat1* in mice affects lactation. After cloning and sequencing of the homologous gene in cattle, several mutations had been found and their segregation with fat percentage of milk from several breeds had been confirmed.

Further progress in identifying genes and mutations associated with important traits in animals has become possible with the advent of high-throughput, low-cost methods of DNA sequencing. It is now possible to not only add the coding regions of genes to the analysis, but also regulatory regions, and introns. A good example of the effectiveness of full genome analysis may be finding a mutation associated with the appearance of early graying in horses. White color in horses has aesthetic value to humans. This coloring is the result of early graying in some horses as a result of a recessive mutation. Identifying a specific mutation would lead to the creation of genetic markers for the analysis of the population and early detection of carriers. The trait has been mapped in the area of the horse genome that contains genes *NR4A3*, *TXNDC4*, *INVS*, *STX17*. "Candidate gene" approach produced no results. However, high-throughput sequencing of the entire area of white horses and horses of ordinary color types identified a duplication (likely with a regulatory role) in the intron of the *STX17* gene, which was present in the genomes of all white and absent in all non-graying individuals.

As shown above, methods of genetic analysis allow for the identification of individual genes and mutations that have a desired effect on a number of important traits in domesticated animals. However, the traits are often affected by multiple genes, each gene only having a small effect. In this case, new approaches are needed, to allow simultaneously to find regions of chromosomes, genes and mutations in them, which together have the desired effect on the phenotype in the population. We have developed one of the variants of such approach, which allows both to identify areas subjected to

selection and the mutations in them, if the population originates from a limited number of founders and the DNA of these founders is available, along with the DNA of several generations of their descendants, which had been subjected to intensive selection. The US population of Holstein-Friesian cattle satisfies such conditions. We had sequenced genomes of the two founders of this population and tracked the frequency of their alleles for seven generations of their direct descendants. It was expected that the population frequency of "favorable" alleles will gradually increase in areas of the genome which are acted upon by selection [5]. Two of the sequenced genomes belonged to two related bulls: Walkway Chief Mark, who was born in 1978, and Pawnee Farm Arlinda Chief, who was born in 1968. Chief was Mark's sire. The data collected on the US Holstein-Friesian pedigrees show that ~12% of the alleles in the modern population were present in the genome of Chief. Having sequenced the genomes of Mark and Chief, we not only found a number of single nucleotide polymorphisms, but also were able to recover haplotypes of Chief's chromosomes. We tracked the fate of Chief's haplotypes in a population of more than 200 of his descendants through genotyping over 50,000 polymorphic markers on the Illumina Bovine 50K array. Afterwards, we identified alleles and haplotypes with a frequency in the population that was significantly higher than expected basing on the simulation of random mating. Such alleles/haplotypes have been found in 49 chromosomal intervals of cattle, representing the most likely genomic regions that have been affected by artificial selection in the US Holstein-Friesian population for the past 40 years.

After we looked at how with the use of techniques of genetic analysis we can identify genes and mutations that are beneficial to the process of artificial selection because of their impact on economic, social or aesthetically important features of domesticated animals, we may move on to how the domestication and artificial selection influences the genomes of domesticated species.

The History of Origin and Breeding of Cattle

In the world, there are more than 1,000 breeds of cattle. These breeds originate from humpless (*Bos taurus taurus*) and humped (*Bos taurus indicus*) cattle. The genomes of these subspecies do not have significant differences with the exception of their mitochondrial DNA and Y chromosome, yet breeds that descended from *Bos taurus taurus* are better adapted to moderate climate, and descendants of *Bos taurus indicus* - to hot climates. The domestication of cattle by man occurred approximately 8,000-10,000 years ago as a result of two independent events.

As shown by the structural genetic analysis of the genomes of 134 breeds of cattle, collected all over the world, in North and South America, Asia and Africa, humpless and humped cattle were crossed to produce hybrids which would be most adapted to local climates and breeding conditions [6]. The history of Asian cattle includes domestication and the mixing in of a few wild species. African humpless cattle originated through the introduction of humpless cattle from Asia Minor, and its subsequent hybridization with wild African taur which apparently made hybrids resistant to trypanosomiasis, transmitted by tsetse flies. The genetic trail of humpless African cattle can clearly be seen in the breeds farmed in Mediterranean countries. The lack of genetic material of humped cattle breeds in most European humpless cattle, along with its presence in some of the Italian breeds, corresponds to at least two waves of migration of livestock to Europe. The first wave - from the Middle East with the Romans, includes humpless cattle that had earlier been crossed with humped cattle; The second wave - from West Africa to Spain, without the introgression of humped cattle. The distribution of the second wave from Spain to the Mediterranean led to the introduction of genetic material of humpless African cattle breeds in Europe. Domestication of cattle was likely to require screening for non-aggressive animals of relatively small size, that could resist numerous pests and adapt to man-made environmental conditions. Accordingly, the signatures of this selection in the genome of

cattle were found in genes that are associated with behavior, the immune system and feed efficiency [7]. For example, one of the regions of DNA with signatures of selection contains the *R3HDM1* gene, which is related to the feed efficiency and intramuscular fat content. It is possible that mutations in cattle in the area of *R3HDM1* gene could also affect the rate of energy conversion. Among others, the signatures of directed selection were found on chromosomes 2, 6 and 14. In particular, meat breeds were subjected to selection near the *MSTN* gene on chromosome 2, which we have already mentioned above. In most dairy breeds, fixation of a mutation occurred at the *ABCG2* gene on chromosome 6, which affects the concentration of fat and protein in milk [8]. On chromosome 14, mutations in the thyroglobulin gene (*TG*) are associated with the distribution of intramuscular fat ("marbled meat") [9].

The History of Origin and Breeding of Domestic Sheep

Unlike cattle, the domestication of sheep (*Ovis aries aries*) occurred only once, about 11,000 years ago. There are a large number of sheep breeds, many of which differ both phenotypically and genetically only slightly, which makes it difficult to classify them. Genetic analysis of 74 breeds of sheep, collected from all over the world [10] showed that the majority of breeds of sheep indeed are characterized by a large effective population size, suggesting a broad genetic diversity, which hasn't been lost in the process of domestication, and an unintensive selection during the formation of breeds. This fact distinguishes the genetics of modern breeds of sheep from breeds of cattle, whose ancestor's influence is pronounced for the majority of commercial breeds. Apparently, in the breeding of modern sheep, there was either an intensive exchange of genetic material, or modern breeds are a part of the ancestral population, which migrated along with people and had to adapt to a particular environment.

Structural analysis and principal component analysis showed that European breeds of sheep are different from African and Asian breeds. In addition, African breeds influenced the formation of Caribbean and South American breeds, which is apparently associated with the direction of the slave trade from Africa to the Americas. Interestingly, primitive breeds of sheep which live on an isolated island off the coast of Scotland, are genetically different from all other breeds. Comparison of major sheep breed genotypes suggests the historical migration from South-West Asia and the Mediterranean through Central Europe to the British Isles and Scandinavia.

A more detailed study of local breeds is necessary to determine their origin and the identification of local migration routes. So, using the genome-wide genotyping for 50000 nucleotide loci, we carried out a reconstruction of the history of the formation of 18 Welsh sheep breeds in the UK [11]. As expected, the local sheep have the genetic similarity to the breeds of continental Europe, America and Australia, reflecting the exchange of animals between these regions (Fig. 3). Welsh breeds have a high level of genetic similarity with breeds from France, Spain, Portugal and Italy, which is apparently, due to the migration of people and animals across the Channel before and during the Roman Empire times. One of the smaller (in size) breeds, Black Welsh, turned out to be close to the old Norwegian Spaelsau breed, which most likely reflects the transportation of animals from Scandinavia to Wales during the Viking conquests.

Despite the overall genetic diversity, scanning the genomes of sheep revealed multiple genome areas with signatures of selection, which contain genes controlling the color, morphology of the skeleton, body size, growth, and reproductive traits [10]. A strong signal is expressed by the *RXFP2* gene which controls the rate of bone mass and puberty. Variants in this gene exist between horned and hornless breeds. Perhaps the selection on this gene contributed to the formation of hornless breeds (Fig. 4). Traces of selective breeding have been found in the *MITF* and *KIT* genes, which control coat color in many

mammals. An example of breed-specific selection could be the aforementioned *MSTN* gene in the Texel meat sheep breed.

The History of Origin and Breeding of Domestic Pigs

Whole genome sequencing of domestic pig (*Sus scrofa domestica*) and wild boar revealed that the ancestor of the domestic pig originated millions of years ago in Southeast Asia and then spread through the whole territory of Eurasia [12]. Phylogenetic analysis revealed clear differences between European and Asian swine breeds, which speaks in favor of the hypothesis that the pig was domesticated independently in several areas of Eurasia (Anatolia, China, India, etc.) [13].

It is believed that the process of domestication covered only a small number of common ancestors, which have passed through a "bottleneck" and held in conditions of reproductive isolation between wild and domesticated animals. However, the analysis of the genome sequence of domestic pigs and wild boars in Europe and Asia has shown that this traditional model of domestication was inapplicable in relation to domestic pigs [14]. Apparently, even during domestication, pigs kept quite freely interbreed with wild relatives. Therefore, the genetic diversity of modern and ancient pig populations can be more completely described by a model, which includes the constant flow of genes from wild boars to domesticated pigs. Moreover, in Europe this flow includes a second, probably extinct population of wild boar, which was significantly different from the first, which served as the source of domestication. The pressure of artificial selection on behavioral and morphological characteristics which was more typical for domestic pigs, apparently counteracted the homogenizing effect of gene flow from wild boars.

To understand the molecular mechanisms of domestication of pigs, phenotypic changes need to be separated, which formed at the early stages of domestication from later changes that occurred at the stage of formation and selection of individual breeds. Modern

methods use genome-wide DNA sequencing to examine this issue in detail. Comparing complete genomes of European wild boars and domestic pigs revealed three genes: *NR6A1*, *PLAG1* and *LCORL*, which were under positive selection that is associated with the most characteristic features of the domestic pig - an increase in body length and number of vertebrae. Duplications in the *KIT* gene were characteristic only for pigs with a white coloring [15]. Genome-wide genotyping on a chip, containing more than 60,000 single-nucleotide polymorphisms showed that the greatest number of differences between the genomes of domestic pigs and wild boars are observed in a region on chromosome 7. In this area, pigs have genes *CDKN1A* and *PPARD*, which are likely to control the distribution of fat and ear shape, respectively [16]. The nucleotide sequence of the *H19* gene, which controls the formation of a long non-coding RNA that regulates the development of muscle cells, vary considerably between wild and domestic pigs, and between different breeds [17]. Moreover, a different methylation profile led to opposite changes in the expression of this gene in domesticated European and Chinese pigs. A number of researchers have identified traces of selection, characteristic for individual breeds. Thus, the study of genomes of the Berkshire pig breed has revealed a number of genes subjected to positive selection, which are associated with the regulation of lipid metabolism, distribution of intramuscular fat and muscle fiber type [18]. Analysis of the full genome sequences of Landrace and Yorkshire pig breeds showed that the two breeds were subjected to selection in various genes and biochemical pathways [19].

Conclusion

For thousands of years domesticated animals have provided humans with protein-rich food and materials for clothing, shelter, etc. Animal genomes have changed, having been adapted to the growing needs of the people and to changing environmental conditions. As a result, several high-performance, specialized breeds were formed which exhibited

outstanding qualities, when placed in the right conditions. On the other hand, local breeds, though less productive, have better adaptation to local conditions and carry the traces of migration and history of their formation within their genomes. A parallel study of the genomes of humans and domestic animals, is necessary for an understanding of our history and to identify traces of unique genetic adaptations to local conditions. In the Russian Federation, there are about 20 local cattle breeds, 40 sheep breeds and 10 breeds of pigs. Along with the study of Russian genomes (human) [20], the study of the genomes of the native Russian breeds will reveal the history of the migration of people on the territory of the Russian Federation, it will reveal traces of a parallel adaptation to local conditions (climate, disease) and, last but not least, will allow to form next generation breeds, which would be more productive and adapted to local environment.

Research has been funded by the Russian Scientific Foundation (project No. 16-14-00090)

Tables

Table 1. The link between chromosomes and economically important traits in cattle.

Trait	Chromosome
Milk	1,2,3,5,6,7,10,13, 14, 15,17,19,20,23,27,29
Carcass	2,5,6,7,13,14,15,19,23,27,29
Fertility	5,6,7,19
Resistance to diseases	1,2,13,23
Other	2,4,5,6,14,23

Figure legends

Fig. 1. The fraction of different types of DNA sequences in the human genome.

(http://commons.wikimedia.org/wiki/File:Components_of_the_Human_Genome.jpg)

Fig. 2. The principle of genetic analysis using "candidate gene" approach.

After the localization of a certain trait in the area of a genetic map of species A (e.g. livestock species), a comparative map is created of the species' genome and human and / or mouse. Genes are determined that are in homeologous regions of the genomes of human and / or mouse, which have functions similar to those of the gene that is looked for. Then, these genes are cloned in species A and a search is conducted for polymorphisms, segregating with the desired phenotype.

Fig 3. Phylogeny of 18 native Welsh sheep breeds. On the phylogenetic tree of the Welsh breeds are marked in red, the continental European breeds – in gray, Australian and New Zealand breeds - in yellow, Scottish - in blue, and Irish – in green. The numbers correspond to clusters of individual breeds, numbers highlighted in bold correspond to breeds shown in the image (from [11]).

Fig 4. The difference in frequency of alleles for the nucleotide polymorphisms in the *RXFP2* gene region of horned and hornless sheep breeds. On the X axis the coordinates of the nucleotide polymorphisms are shown, which were genotyped in horned and hornless sheep on chromosome 10. The Y axis shows the value of F_{ST} statistics between the two breeds for each polymorphism. The largest difference is observed in the *RXFP2* gene region (from [10] with modifications).

References

1. *Elsik C.G., Tellam R.L., Worley K.C., Gibbs R.A. et al.* The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 2009; 324(5926):522-8.
2. *Grobet L., Martin L.J., Poncelet D., Pirottin D. et al.* A deletion in the bovine myostatin gene causes the double-muscling phenotype in cattle. *Nat Genet.* 1997; 17(1):71-4.
3. *Grisart B., Coppieters W., Farnir F., Karim L. et al.* 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.* 2002; 12(2):222-31.
4. *Rosengren Pielberg G., Golovko A., Sundström E., Curik I. et al.* A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. *Nat Genet.* 2008; 40(8):1004-9.
5. *Larkin D.M., Daetwyler H.D., Hernandez A.G., Wright C.L. et al.* Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *Proc. Natl. Acad. Sci. USA* 2012; 109(20):7693-8.
6. *Decker J.E., McKay S.D., Rolf M.M., Kim J., Molina Alcalá A., Sonstegard T.S. et al.* Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet.* 2014; 10(3):e1004254.
7. *Bovine HapMap Consortium.* Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 2009; 324(5926):528-32.
8. *Cohen-Zinder M., Seroussi E., Larkin D.M., Looor J.J. et al.* 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.* 2005; 15(7):936-44.
9. *Zhang L., Ren H., Yang J., Gan Q., Zhao F., Gao H. et al.* Effect of thyroglobulin gene polymorphisms on growth, carcass composition and meat quality traits in Chinese beef cattle. *Mol. Biol. Rep.* 2015; 42(9):1403-7.
10. *Kijas J.W., Lenstra J.A., Hayes B., Boitard S., Porto Neto L.R., San Cristobal M. et al.* Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* 2012; 10(2):e1001258.
11. *Beynon S.E., Slavov G.T., Farré M., Sunduimijid B., Waddams K., Davies B. et al.* Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. *BMC Genet.* 2015; 16:65.
12. *Groenen M.A., Archibald A.L., Uenishi H., Tuggle C.K., Takeuchi Y., Rothschild M.F. et al.* Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 2012; 491(7424):393-8.

13. *Larson G., Liu R., Zhao X., Yuan J., Fuller D., Barton L. et al.* Patterns of East Asian pig domestication, migration, and turnover revealed by modern and ancient DNA. *Proc. Natl. Acad. Sci. USA* 2010; 107(17):7686-91.
14. *Frantz L.A., Schraiber J.G., Madsen O., Megens H.J., Cagan A., Bosse M. et al.* Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nat Genet.* 2015; 47(10):1141-8.
15. *Rubin C.J., Megens H.J., Martinez Barrio A., Maqbool K. et al.* Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. USA* 2012; 109(48):19529-36.
16. *Wilkinson S., Lu Z.H., Megens H.J., Archibald A.L. et al.* Signatures of diversifying selection in European pig breeds. *PLoS Genet.* 2013; 9(4):e1003453.
17. *Li C., Wang X., Cai H., Fu Y. et al.* Molecular microevolution and epigenetic patterns of the long non-coding gene H19 show its potential function in pig domestication and breed divergence. *BMC Evol. Biol.* 2016; 16(1):87.
18. *Jeong H., Song K.D., Seo M., Caetano-Anollés K. et al.* Exploring evidence of positive selection reveals genetic basis of meat quality traits in Berkshire pigs through whole genome sequencing. *BMC Genet.* 2015; 16:104.
19. *Wang Z., Chen Q., Yang Y., Yang H. et al.* A genome-wide scan for selection signatures in Yorkshire and Landrace pigs based on sequencing data. *Anim. Genet.* 2014; 45(6):808-16.
20. *Oleksyk T.K., Brukhin V., O'Brien S.J.* Putting Russia on the genome map. *Science* 2015; 350(6262):747.