Current Status of Genomic Maps: Genomic Selection/GBV in Livestock

4

Agustin Blasco and R. N. Pena

Abstract

Our understanding on how the genome is structured has improved substantially since the human genome was first sequenced in 2001. The sequencing of livestock and other model animals, in addition to other organisms, has also helped to identify common genomic patterns and features, which can now be summarised in genome maps. The annotation of sequence variation in the livestock genomes has opened up the possibility of using its genomic information for improving the prediction accuracy of its genetic merit. This chapter will give a general view on the main features annotated to the livestock genomes and outline the application of molecular information in the prediction of the genetic breeding value of the animals. The advantages and limitations of implementing this methodology in distinct production systems are also discussed.

4.1 The Evolution of Genetic Maps

Before the sequence of the genome was available for most livestock and model animals, researchers used genetic maps to orderly map genes and markers in the genome. A genetic map is simply a representation of the distribution of genes and other genetic features within the genome of one species. Specific techniques were developed to respond to questions such as in which chromosome a certain gene (or

A. Blasco (✉)
Institute for Animal Science and Technology, Universitat Politècnica de València, Valencia, Spain
e-mail: ablasco@dca.upv.es

R. N. Pena
Department of Animal Science, University of Lleida – Agrotecnio Centre, Lleida, Spain
e-mail: romi.pena@prodan.udl.cat

H. Niemann, C. Wrenzycki (eds.), *Animal Biotechnology 2*,
https://doi.org/10.1007/978-3-319-92348-2_4

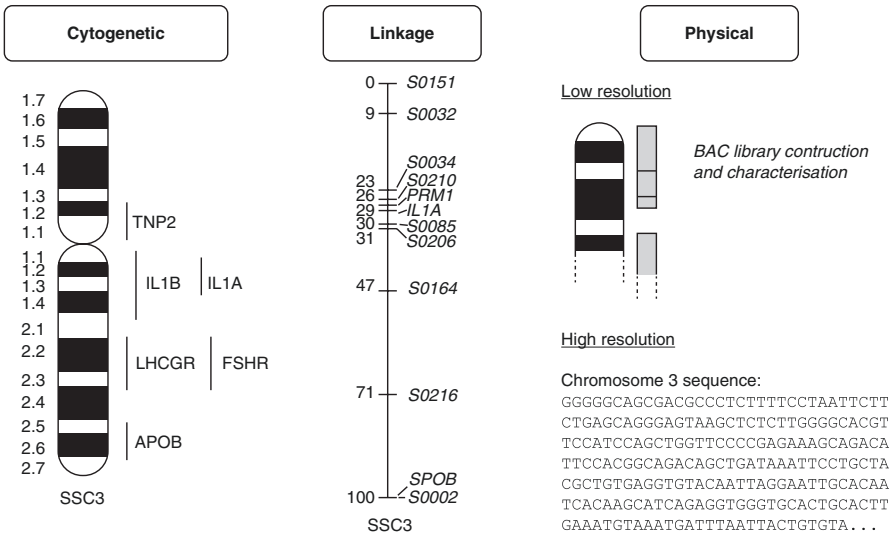

Fig. 4.1 Comparison of cytogenetic, linkage and physical maps. The three types of maps are shown for loci in pig chromosome 3 (SSC3). The cytogenetic and linkage maps extend over the entire chromosome, and loci are position in relative order and distance. Chromosome bands are indicated in the long (q) and short (p) chromosomal arms in the cytogenetic map. Linkage units are centiMorgans. The three maps can be connected to allow complementing the information from one another. Genetic maps from different species can also be compared by matching blocks of synteny (comparative mapping). Data source: Archibald et al. (1995), Groenen et al. (2011), Yerle et al. (1995)

marker) is mapped, or which were its closest genes/markers, or even in which particular order a small number of loci were mapped in a given chromosome. Thus, three distinct types of genetic maps—cytogenetic, linkage and physical—were developed to answer the questions above complementary (Fig. 4.1). Silver (1995) includes an excellent revision on genetic maps. A brief summary is presented here.

Cytogenetic maps relied on the hybridisation of a fluorescently labelled gene-specific probe (a synthetic DNA fragment) to its target gene in condensed whole-chromosome preparations (such as in karyotypes). The resolution of this type of mapping was low, but it allowed mapping a limited number of genes to the telomeric, centromeric or short (named 'p') or long (named 'q') arms of chromosomes. Complementary to these efforts, other researchers developed *linkage maps*, which were based on the frequency of recombination between two or more heterozygous loci (markers or genes) over generations. Loci that are close together in the same chromosome tend to be inherited together more often than loci that are apart. Linkage maps are generated by counting the number of offspring that receive either parental or recombinant allele combinations from a heterozygous parent. The frequency of recombination between two loci is directly related to the distance between them, measured in centiMorgans (1 cM equals a crossover rate of 1%). This measure of the linkage disequilibrium between loci allowed establishing their relative order and distance, a critical information in the pre-genomic era. Finally, the *physical maps* analysed the genomic DNA directly, usually by subcloning large DNA fragments into

DNA vectors such as BACs (bacteria artificial chromosomes) or YACs (yeast artificial chromosomes), which could be easily propagated in the lab using standard microbiology methods. These DNA fragments were usually generated by restricting targeted fractions of chromosomal DNA with several restriction enzymes to obtain overlapping fragments. By comparing the structure of these fragments, the relative position of each gene and their upstream and downstream flanking sequences could be identified. At its highest resolution, a physical map will give us the full sequence of the whole genome. Consequently, physical maps are measured in base pairs (bp) or its derived units (kbp, Mbp, Gbp). Nowadays, the genome of the main livestock species (chicken, cow, sheep, pig, horse and rabbit) has been sequenced, and efforts are being made to update and improve the information annotated to them. A summary and comparison of this information are given in the following sections.

4.2 Current State of the Livestock Genomes

While the first draft of the human genome sequence was delivered in 2001, we had to wait a number of years for the first sequence of the cow (2004), chicken (2005), horse (2007), pig (2010), rabbit (2014) and sheep (2014) genomes. Although whole genomes can be sequenced by different methods, in practice all of them result in a pool of millions of short (75–150 bp) or long (>500 bp) sequence reads. The first hurdle in describing a genome is to identify and assemble overlapping sequences into larger fragments (called contigs) to eventually reconstitute the sequence of whole chromosomes. For this, new bioinformatic programmes able to deal with these massive data had to be developed and implemented. In all species, the first genome drafts had a large number of gaps rendering incomplete chromosomes. However, these have progressively been filled in as newer versions were released. The exception is the chicken genome, which is structured in 38 autosomes, many of which are relatively small and uniform in size, often termed microchromosomes. Several properties (e.g. %GC content, gene and repeat density) contribute to the fact that some of them are not yet assembled (or only partially) even in the latest version of the genome (Warren et al. 2017). In this species, linkage groups estimated from linkage maps are still of use to study genes located in these missing regions.

The most updated version of farm animal genomes is available at www.ensembl.org. The importance of these updated versions is double: first, it is a precious material for researchers to study the structure of the genome and to investigate genes related to production traits or disease. They also provide a scaffold to assemble new whole-genome sequencing (WGS) data from other animals of the same species in a much faster and accurate way. As the costs of WGS have become more affordable, it is now feasible to describe genetic variability in a population by sequencing key genetic contributors. Sound scaffolds are critical to identify, map and compare sequence variants across these individuals.

As a result of the genome sequencing projects, we have been able to measure the *total size* of the genome, which is specific to each species. In the five farm animals analysed here, it ranges from 1.2 Gbp in chicken to 2.7 Gbp in rabbit (Table 4.1). As a reference, the human genome is slightly longer (3.1 Gbp), but the longest genome

Table 4.1 Genome size and annotation of genes, transcripts and sequence variation features in the latest assemblies available for the main livestock species

		Genome					
	Chicken	Cow	Sheep	Pig	Horse	Rabbit	Human
Assembly	Gga5.0	UMD_3.1	Oar_v3.1	Sscrofa11.1	Equ Cab 2	OryCun2.0	GRCh38.p10
Length (bp)	1,230,258,557	2,670,422,299	2,619,054,388	2,501,912,388	2,474,929,062	2,737,490,501	3,096,649,726
Genes							
Protein-coding genes	18,346	19,994	20,921	22,452	20,449	19,293	20,338
Non-coding genes	6491	3825	5843	3250	2142	3375	22,521
– Small nc-genes	1705	3650	3624	2503	1967	3059	5363
– Long nc-genes	4643	175	1858	361	–	–	14,720
– Miscellaneous nc-genes	144	797	361	386	175	316	2222
Pseudogenes	43	26,740	290	178	4400	1001	14,638
Gene transcripts	38,118	19,994	29,118	49,573	29,196	24,964	200,310
Sequence variation							
Short variants	23,873,479	102,499,615	60,323,418	64,310,125	5,217,806	–	329,465,985
Structural variants	–	10,462	2	224,038	193,747	–	5,864,995

Data from the human genome are also included as a reference. Source: www.ensembl.org

so far sequenced is the loblolly pine tree (*Pinus taeda*) which spans 23.2 Gbp (Neale et al. 2014). As we will see below, there is no linear correlation between the size of a genome and the number of genes it contains.

4.3 Gene Annotation in the Livestock Genomes

Once the sequence is established, the next step in order to build a genomic map is to annotate the genetic elements underlying each genome. This annotation step is constantly evolving as new elements are still being discovered. The first features to be mapped to the genomes were the *protein-coding genes*. By doing so, researchers realised that animal genomes were, at once, simpler and more complex than expected. Humans, farm animals, mice and simpler animals such as the earthworm *Caenorhabditis elegans* have all approximately the same number of genes, around 20,000 (Fig. 4.2 and Table 4.1). This number of genes seemed too low to explain the complexity of larger mammals. Moreover, the coding sequences only spanned a very small percentage of the total genomic sequence of farm animals, about 1.5–2%. This means 98% of the genome does not encode for proteins, the ultimate effectors of cellular functions. About a quarter of this non-coding (nc) DNA are intron sequences, that is, gene sequences that are transcribed by the RNA polymerases but that are spliced out of the mature mRNA by the spliceosome. Half of the 70% remaining genomic DNA contains repetitive DNA elements such as micro-/minisatellites or transposon-derived sequences (LINEs, SINEs, Alu, LTRs, etc.).

Strikingly, the proportion of ncDNA in the genome, unlike the total number of protein-coding genes, increases in parallel with evolutionary complexity (Fig. 4.2). Thus, in simple organisms such as prokaryotes or yeasts, 70–85% of the genome encodes proteins, while in invertebrates (earthworm, fruitful), this figure drops to 20–25% and reaches the overwhelming 1.5–2% value in humans and farm animals. The presence of ncDNA has been explained by several mechanisms. Initially, all this additional ncDNA of unknown (and unpredictable) function was thought to be an evolutionary artefact, a carry-over of non-functional (and non-damaging) DNA that had accumulated over evolution without adding any specific advantage to the species. Moreover, although there was a degree of sequence conservation in the protein-coding DNA, sequences were much more divergent in ncDNA, reinforcing the hypothesis of lack of function. In consequence, the ncDNA was often called 'junk DNA' to designate its lack of purpose. However, as it became more and more obvious that the number of protein-coding genes was not the main drive of biological evolution, the attention was turned into ncDNA.

In this context, the ENCODE project was set up to annotate functional elements in the genome of humans and model organisms. The consortia of research groups participating in this initiative designed two types of experiments: one group aiming at identifying DNA that was being transcribed into RNA and another group targeting chemical labels in the chromatin (epigenome). One of the first results reported by the ENCODE consortia was that more than 80% of the genome was being transcribed into RNA. This phenomenon was called pervasive transcription to express

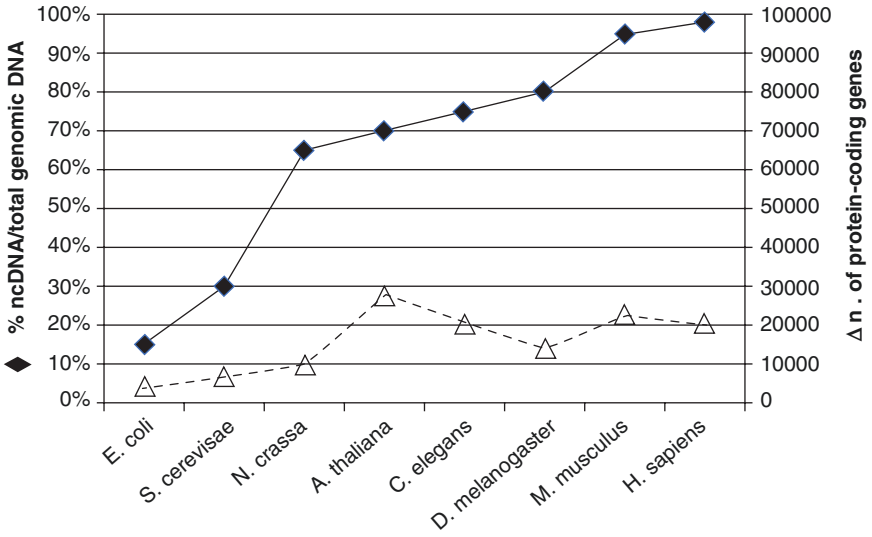

Fig. 4.2 The proportion of non-coding DNA (ncDNA) in the genome increases with developmental complexity (adapted from Mattick 2011), but the number of protein-coding genes does not (source www.ensembl.org)

the permanent state of transcription of most of the genome (Libri 2015). As protein-coding genes only span 25% of the genome (adding exons and introns together), that means most of the transcribed RNA was in fact non-coding RNA (ncRNA) molecules. Thus, a new category of *non-coding genes* was defined (Table 4.1), which, like coding genes, are also organised in exons and introns and have regulatory elements that control expression.

Currently, the annotation of protein-coding genes is almost complete in most animal genomes. The exception is again the chicken genome, where 360 genes are still missing in the current annotation, which most likely map to unassembled microchromosomes (Warren et al. 2017). In contrast, the mapping of nc-genes is still at its initial stages, particularly in farm animals. As a reference, in humans there are similar numbers of protein-coding and non-coding genes, indicating that annotation in farm animals is probably underestimated (Table 4.1). Based on the length of the transcripts, ncRNAs can be classified into small ncRNAs (usually <200 nt-long) and long ncRNA (>200 nt-long) molecules. Small ncRNA can be divided into further categories (Wright 2014), although probably the best characterised are the family of microRNA (*miRNA*) genes. These represent a group of genes that, once transcribed and processed, generate short structures of double-stranded (ds) RNA, usually ~21 nt-long. About 80% of miRNA genes map to intronic DNA, usually in polycistronic clusters from which up to ten miRNAs are co-expressed (Hausser and Zavolan 2014). This has facilitated their mapping, and they are probably the best annotated class of nc-genes in the farm animal genomes. miRNA are strong regulators of the translation rate of protein-coding mRNAs. By binding usually to the 3′ untranslated regions (3′UTR) of the mRNA, the miRNA are able to put the

translation of that miRNA on hold. This represents an additional layer of regulation of the expression of protein-coding genes, from DNA to proteins. On the other hand, long intergenic ncRNA (*lincRNA*) represents a new class of ncRNA that has brought much excitement, even though few data are yet available for most of them. In general, these genes are transcribed at very low levels (about 100- to 1000-fold lower than the average protein-coding gene) from >60% of the genome. Most lincRNA genes are active only in some cell types or at certain developmental stages and are thought to be one of the key organisers of development and probably a main evolutionary drive (Hangauer et al. 2013).

The third type of genes mapped to the genomes is *pseudogenes* (Table 4.1). These represent 'dead genes', relics from former protein-coding genes, usually generated by gene duplication, that have been inactivated in the course of evolution through accumulation of mutations. The gene graveyard is extensive in the human and cow genome (14,638 and 26,740 pseudogenes, respectively) but is probably underrepresented in chicken, pig, sheep and horse. It is not unusual for a pseudogenes to be transcribed into mRNA, but they very rarely get translated into proteins, due to unstable messengers or to accumulation of premature STOP codons (Xu and Zhang 2016).

Altogether, protein-coding genes, non-coding genes and pseudogenes generate a large number of transcripts (around 20,000–50,000 in farm animals but close to 200,000 in humans). The tenfold higher number of transcripts in humans is explained mainly by alternative splicing of exons and introns, which takes place in ~94% of the human (protein-coding and non-coding) genes. This is a process that also takes place in the animal transcripts but to a lower extent (for instance, if has been estimated to affect 21% of cow genes). Current genomic maps also include information on alternative transcripts and predicted proteins generated by each gene. Beyond question, this is a major source of functional variation that can explain the larger biological complexity of livestock animals and certainly that of humans.

4.4 Annotation of Regulatory Elements

The second set of experiments carried out in the frame of the ENCODE project had the aim to identify the regulatory elements of the genome, that is, stretches of genomic DNA that regulate (activate/inactivate) the expression of genes. The two main types of regulatory elements are promoters and enhancers (Fig. 4.3). *Promoters* are DNA sequences around the transcription start site of a gene where the proteins of the transcription machinery assemble. The transcription complex represents a runway for the RNA polymerase II to land and start transcription. *Enhancers*, on the other hand, are usually located remotely from gene promoters. They physically interact with promoters stabilising or disassembling the transcription complex. Enhancers are essential for the correct spatio-temporal activation of gene expression (Andersson 2015). For instance, an enhancer may act to increase the transcription of a gene with a possibly weak promoter or may provide essential, additional information not encoded in the gene promoter itself. Enhancer function is highly specific to cell type and state

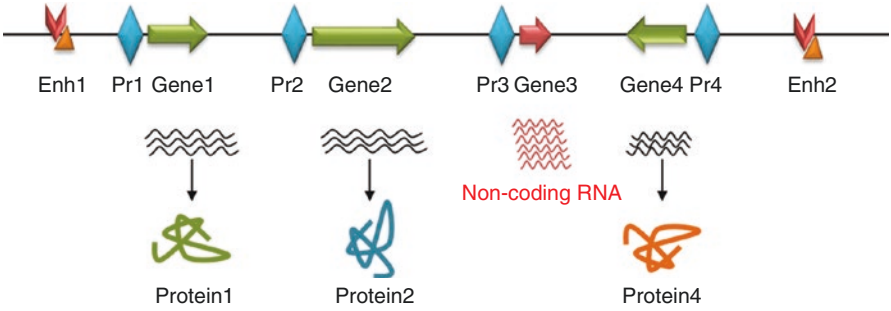

Fig. 4.3 Spatial relationship between enhancers (Enh), promoters (Pr) and genes. Promoter elements are positioned close to the transcriptional start site of both protein-coding and non-coding genes. Both types of genes are transcribed into RNA, but only the protein-encoding genes are translated into proteins. Enhancers can be located upstream, downstream and even inside the genes they are regulating

compared to protein-coding genes. Hence, a gene may be regulated by different enhancers in different cell types, at different developmental stages and in response to different signals. Enhancers can be hundreds of kbp away from the regulated genes, and it is not unusual to find several (untargeted) genes between them (Fig. 4.4). Hence, to put enhancers proximal to the correct target gene promoters in three-dimensional space, the DNA must be structured into chromatin loops (Fig. 4.5). A current hype is the elaboration of 3D dynamic genomic maps of how these loops evolve during cellular differentiation according to the required change in gene expression.

These functional elements are currently being annotated to the livestock genomes thanks to the efforts of the FAANG (Functional Annotation of the Animal Genomes) initiative. Annotations are much more advanced in humans and model animals. As a reference, there are 70,292 promoters and 399,124 enhancers in the human genome (ENCODE Project Consortium 2012), and about half of each are active in any given cell (Won et al. 2013). Regulatory elements are difficult to identify by computational analysis of the genome sequence as in general they lack evolutionary constraint, which means their sequence is not conserved across species despite having the same function. A combination of wet-lab techniques is needed to position epigenetic labels that are characteristic of silent, poised or active regulatory elements (ENCODE Project Consortium 2012). A second common feature of promoters and enhancers is that they are bidirectionally transcribed; that is, RNA is synthesised from both strands flanking the element, producing relatively short non-polyadenylated enhancer RNAs (eRNAs). Synthesis of eRNA is essential for full enhancer functionality. This explains a large part of the non-coding RNA pervasively transcribed in the genome and indicates there is an extensive overlap between transcription and regulation. Overall, the results from the ENCODE project claim a shift from the gene-centric vision of the genome to a more dynamic and holistic interpretation of genomic function.

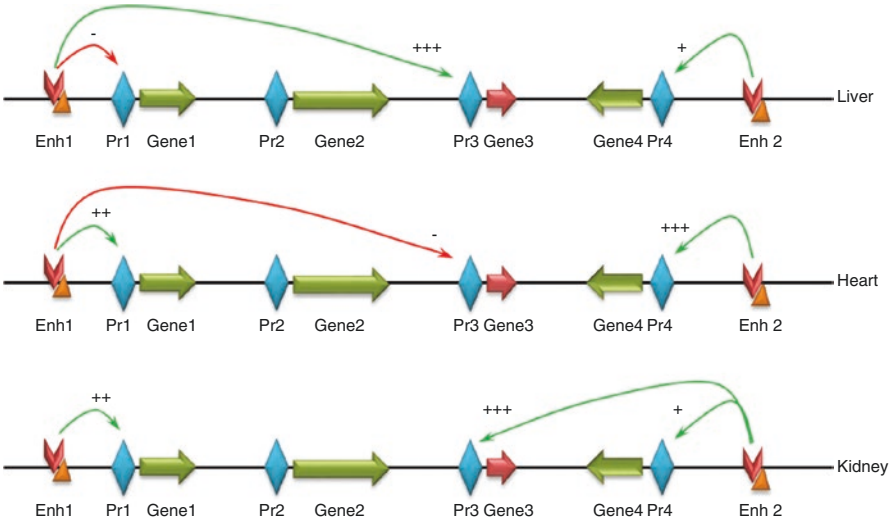

Fig. 4.4 Enhancers (Enh) regulate gene expression by interacting with gene promoters (Pr), which might be several genes away from the enhancer site. To bring enhancers and promoters together, genomic DNA needs to bend in a 3D loop. Enhancers can activate or silence transcription depending on the gene, the tissue and the stage of development. Enhancer engagement and displacement are very dynamic events as they regulate more than one gene at a time

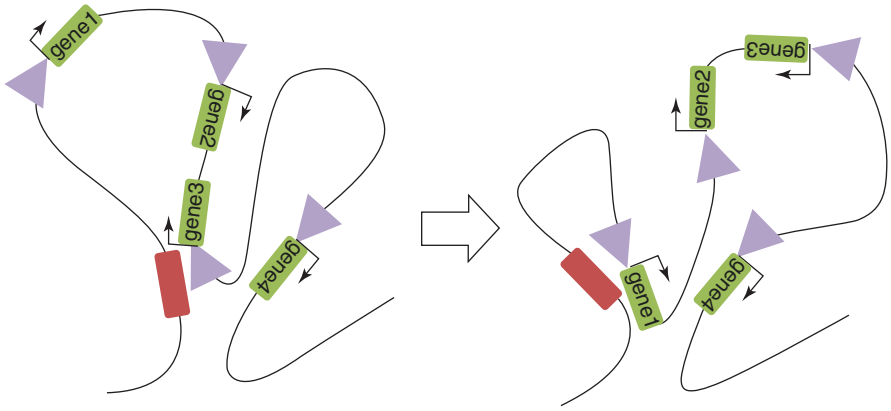

Fig. 4.5 Three-dimensional chromatin looping is necessary for the correct regulation of gene expression by distal enhancers. On the left, the enhancer (dark block) interacts only with the promoter (triangle) of gene 3. In order to regulate the expression of gene 1, a different loop needs to be formed (right). This is a dynamic process that can change rapidly in response to cellular signals

4.5 Mapping the Genomic Sequence Variation

Another objective of the genome annotation initiatives has been to catalogue the new mutations described in each genome and to map them in the genomic context. For simplification, in the annotated genomes, mutations are classified as either short variants or structural variants. *Short variants* include single nucleotide polymorphisms (SNPs) and insertion/deletions (indels) of short nucleotide runs. By large, most short variants have two possible versions, called alleles. For instance, for a given SNP, an adenine can change into a guanidine, so A and G constitute the two alternative alleles. The data to annotate these variants come from specialised databases such as the dbSNP (www.ncbi.nlm.nih.gov/snp) and from sequencing centres (e.g. the Broad Institute). Short variants are very common. Taking figures in Table 4.1, it is estimated that in farm animals there is one short variant per 30–50 bp (even more frequent in humans; 1 in 15 bp). However, although the amount of intraspecies sequence variation is disturbingly high, the numbers are expected to be much lower within a given breed or commercial line. In genomic maps, short variants are annotated over other genomic features such as genes. For variants overlapping protein-coding genes, an estimation of the effects on the final protein is also calculated and annotated on the map. Protein structure predictor programmes such as SWIFT (www.bioinfo.org.cn/swift) are routinely used for this purpose. Under *structural variants*, repetitions in larger regions of DNA, of at least 1 kb in size, are gathered. It can include inversions, balanced translocations or genomic imbalances (insertions and deletions), commonly referred to as copy number variants (CNVs). This is an area of uneven annotation across the genomes, with total numbers ranging from ~200,000 in horse and pig to none in the chicken genome. The number of possible alleles per structural variants is more variable and can go from complete deletion (zero copies) up to three to four copies of the fragment.

Variation data, particularly SNP information, have been used to build for each species dense panels of markers evenly distributed across the genome. Novel biotechnological tools have been developed to genotype these panels. Currently, two companies lead the market of genotyping platforms for livestock animals. They provide a range of SNP-based arrays (also known as *SNP-chips*) to genotype at variable densities (Table 4.2). These chips are currently used to improve the accuracy of predictions of breeding values in several species, as we will see in the following sections.

4.6 Genomic Selection

The use of genetic markers to improve the efficiency of current selection programmes was proposed 40 years ago by Moses Soller (1978). At that time, few markers were available, and the expectation was to find some gene with a substantial effect linked to the marker and increase its frequency in the population. Unfortunately, most production traits in livestock species are determined by a large number of genes with small effect, and consequently the method was inefficient.

Table 4.2 Summary of the commercial high-density genotyping chips currently available for the main livestock species

Species	Chip name	No. of SNPs	Average interval between SNPs (kb)	Average MAF across tested populations	Supplier
Chicken	Axiom Chicken array	580,961			Affymetrix
Cattle	BovineSNP50	54,001	50.6	0.26	Illumina
	BovineHD	777,000	3.43	0.25	Illumina
	BovineLDv2	7931	383	0.31	Illumina
	Bovine3K[a]	2900			Illumina
	Axiom BOS1 array	648,855			Affymetrix
Horse	Axiom Equine Array	650,000			Affymetrix
Pig	PorcineSNP60	64,232	43.4	0.28	Illumina
	GGP Porcine HD	>65,000	43.0		GeneSeek
	Axiom Porcine 650K	658.692	3.34	0.32	Affymetrix
Sheep	OvineSNP50	54,241	50.9	0.28	Illumina
	OvineHD	603,350	5	0.30	Illumina
Rabbit	Axiom OricunSNP	200,000	15–20	0.20	Affymetrix

[a]Subset of SNP50 panel for prediction of milk yield, protein % and fertility

When QTL detection started in the 1990s and it was feasible to use more markers, it seemed that alleles of medium effect could increase their frequencies by marker-assisted selection, and higher responses to selection could be obtained (Lande and Thompson 1990). However, as Blasco (2008) noticed, there was a notorious discrepancy between simulation results, relatively optimistic, and practical application of marker-assisted selection, which gave deceptively small improvements. The problem was, as Smith and Smith (1993) stressed, the lack of enough markers to cover the whole genome and capture the signals of genes with small effect on the traits. When the genome sequences of livestock species were published, first in 2004 in cattle and later in the other species, chips with a large amount of SNPs became available at an affordable cost, and its use in selection programmes was examined. The first chips of 10,000 SNPs were not well distributed along the genome and were not efficient, but in December 2007 a well-distributed 57,000 SNPs chip was for the first time commercialised. In 2008 the first genetic evaluations for dairy cattle using genomics started in several countries, and in 2009, the USDA published the first official dairy cattle genomic evaluations. The impact in dairy cattle selection programmes was dramatic, doubling the rate of improvement of total genetic merit (Wiggans et al. 2017); thus the use of genomic selection was rapidly investigated for the other livestock species. Dairy cattle has some special characteristics that permit an efficient use of genomic selection, as we will see later, but the use of genomic selection in other species is not as straightforward (Blasco and Toro 2014; Jonas and de Koning 2015). Nevertheless, genomic selection can

contribute to the efficiency of current selection programmes, if the strategies of implementation are carefully studied and the cost of genotyping is low enough. It seems that very high-density SNP chips do not lead to a much higher accuracy of prediction; for example, Van Raden et al. (2011) obtained a gain in accuracy of only 1.6% when using 500,000 markers instead of 50,000. Even SNP chips of 3000 markers, using imputation techniques that we will comment later, give good results (Berry and Kearney 2011, in cattle; Cleveland and Hickey 2013, in pigs), which permits examining scenarios less favourable than the dairy cattle one.

4.7 Predicting Breeding Values with Genomic Selection

The methods for predicting breeding values with genomic selection were developed in a seminal paper by Meuwissen et al. (2001) before we had access to the SNP chips. Predicting breeding values has two steps. First, we collect data form a set of animals, for example, 4000 animals, the 'reference population', and genotype all of them with a high-density chip with, for example, 50,000 SNPs. Now we need to prepare the prediction equation. To do this, we generate one variable $z_i$per SNP having an arbitrary value indicating whether the SNP 'i' is 'homozygous' for one base, 'heterozygous' (i.e. has different bases) or 'homozygous' for the other base. Calling 'M' and 'm' the two positions of the bases of one SNP, we have for each SNP.

SNP_i	M_iM_i	M_im_i	m_im_i
z_i	1	0	−1

The values 1, 0 and −1 are arbitrary and can be substituted by other values (e.g. 2, 1, 0). The coding is additive; it is related to the number of copies of one reference allele, 'M' in this example. The use of capital letter 'M' does not mean that we are considering dominance effects, although models that are more complex can include this possibility (Vitezica et al. 2016). We will consider in this simple example that the data are pre-corrected to make the formula simpler. The regression equation is

$$y = a_0 + a_1 z_1 + a_2 z_2 + a_3 z_3 + \cdots + a_{50,000} z_{50,000} + e$$

where a_1, a_2, $a_3 \cdots a_{50,000}$ are the coefficients of regression and z_1, z_2, $z_3 \cdots z_{50,000}$ are the variables associated to each SNP. The genetic value of the animal is

$$a = a_1 z_1 + a_2 z_2 + a_3 z_3 + \cdots + a_{50,000} z_{50,000}$$

(we can add, if we like, the intercept $a_0$of the regression equation). Now we have to estimate the coefficients, and we are faced with the problem that there are much more unknowns than the equations that we have; the equation system cannot be solved by classical procedures. However, the system has a solution using Bayesian statistics under some assumptions about the prior information on the SNPs we have (see Blasco 2017, for details). We can thus obtain the estimates of the regression coefficients $\hat{a}_1, \hat{a}_2, \hat{a}_3 \cdots \hat{a}_{50,000}$, and we are ready to predict the breeding value of new individuals.

In the second step, we will predict the genetic value of animals that may have or not have their own data. Suppose first that the animal has no phenotypic data, but it has been genotyped, and we know the values of each of the variables z_i for this animal. By substituting in the equation, we can predict its genetic value:

$$\hat{a} = \hat{a}_1 z_1 + \hat{a}_2 z_2 + \hat{a}_3 z_3 + \cdots + \hat{a}_{50,000} z_{50,000}$$

The genetic value of each new animal will be predicted using the same coefficients $\hat{a}_1, \hat{a}_2, \hat{a}_3 \cdots \hat{a}_{50,000}$ with the variables z_i of the new animal provided by its SNPs. It is important to notice that the coefficients do not indicate the importance of each SNP, since the variables z_1, z_2, $z_3 \cdots z_{50,000}$ are correlated. We have said before that SNPs close to each other are frequently associated in its genetic transmission; even if they are in different chromosomes, they can be associated in its transmission, for example, due to selection. The equation is useful to predict the whole genetic value '$\hat{a}$' of an animal, but not to detect single genes. The coefficient of a SNP in a multiple regression is not the same as the coefficient that can be found when this SNP is fit in isolation. The coefficients of the equation will also change depending on the number of SNPs considered, because the sum of all of the terms in the multiple regression equation should give the same genetic value $\hat{a}$ of the animal; thus when many SNPs are considered, they have smaller individual effects.

The different Bayesian statistical methods for solving the equations depend on different prior assumptions about the genetic determination of the traits; for example, the trait can be determined by many genes of small effect each one or by some major genes, some intermediate ones and many genes with small effects. The success of each method depends on whether the actual genetic determination of the trait reflects well what the prior information assumes, although Fernando and Garrick (2013) have noticed that in real applications, the simplest model that considers the traits determined by many genes with small effects works just as well as the more complex models and sometimes even better. This occurs probably because in practice, even if there are genes with medium-large effects, they are not in close association with only few markers, but their effect is captured by many markers.

When the animal has no data, its breeding value can be estimated by weighing the information of its relatives appropriately, a technique called selection index, in which several traits can be simultaneously used for selection weighed according to their economic importance (Falconer and Mackay 1996). Now, if the animal is genotyped, the estimated breeding value from the genomic equation can also be appropriately weighed and integrated with the breeding value provided by the selection index. The information given by the SNP chips can be used to better assess the actual relationships between individuals. For example, we know that on average full sibs share half of their genomic information, but by crossing two heterozygotes Aa × Aa, we could produce full sibs that are more similar than others. If we have three full sibs AA, AA and aa coming from this cross, the two first full sibs are more similar than the first and the third or the second and the third sib. Taking into account all SNPs, we can have a more accurate idea about the actual correlation between relatives. This allows being more accurate in the genetic evaluation.

In current breeding programmes, the correction of environmental effects (parity, season, herd, batch, etc.) is done at the same time as the genetic evaluation, using a technique called best linear unbiased prediction (BLUP, see Blasco 2017, for details). When all genomic relationships are used in the evaluation, this procedure is known as 'genomic-BLUP' (G-BLUP; see e.g. Clark and Van der Werf 2013). It can be shown that this procedure is equivalent to solving the genomic equations under a model assuming that the genetic determination of the trait depends on many genes with small effects each one (Habier et al. 2007). Nowadays there is a wide consensus about the reasons of the success of genomic selection; rather than a better assessment of the 'genetic architecture' of the trait, it is mainly related to a better determination of the actual relationships between relatives. Genomic information can be integrated with BLUP, and the evaluation is made with all data of all animals and all important traits, integrating the information provided by the genomic equations, a procedure that is called 'single step' (Legarra et al. 2009; Misztal et al. 2009).

4.8 Difficulties in Implementing Genomic Selection

Blasco and Toro (2014) and Jonas and de Koning (2015) have detailed some of the difficulties of implementing genomic selection in current breeding programmes. First, to create the equations to be used in prediction, we need a large 'reference population' of several thousand animals. This is not a problem in dairy cattle, but in other species, it can be a serious problem. In prolific species, for example, selection is performed in small nucleuses, sometimes with few hundred females. Some alternatives can be considered, for example, using sibs from multiplication farms or using animals from several generations (Chen et al. 2011) or crossbred animals (Knol et al. 2016), but the efficiency of the equations rapidly decays, thus alternative strategies should be examined with care. A second problem is the need of generating new equations every three or four generations, because due to recombination, the associations between SNPs and causal genes are lost with time. Ibañez and Blasco (2011) have shown that the accuracy of the equations is rapidly lost generation after generation, which means that new large reference populations are needed from time to time. In practice, instead of having large reference populations every few generations, phenotypes are collected every generation to update the equations. This is not a problem for routinely recorded traits (e.g. litter size), but it can be a problem for more expensive traits.

Another major problem of genomic selection is the cost of genotyping. This cost has been dramatically reduced in the last years, but it is still important for species in which the individual value of the animal is small and the generation interval is short (pigs, poultry, rabbits), which implies frequent genotyping with high cost with respect to the value of the animal. A way of facing this problem is to use low-density chips with only few hundreds or thousands SNPs, inferring the missing SNPs from high-density chips. This technique, called 'imputation', is based on that recombination which is low in a single generation and has produced efficient results (Huang et al. 2012; Cleveland and Hickey 2013). Imputation from high-density chips should

be repeated every three or four generations, because recombination leads to errors of imputation. Nowadays, instead of having a reference population, some high-density chips are used every generation for repeating imputation.

Genomic selection was considered as a possible procedure for improving 'difficult' traits. For example, meat quality traits were considered natural candidates for genomic selection, hoping that after collecting the data in a reference population, many animals could be evaluated using their genomic data without the need of collecting their phenotypic data or data from relatives. However, as the equations have to be reformulated after three or four generations, there is the need of continuous data collection to avoid reconstituting reference populations every few generations; thus genomic selection became less attractive, at least for short generation interval species such as pigs, poultry or rabbit. In dairy cattle, the index of conversion of food for milk is economically attractive but difficult to be recorded, but by collecting these records in some specialised farms, a reference population and the equations needed for genomics could be prepared. As dairy cattle, particularly the Holstein breed, constitutes a global population in which most farmers use the same bulls, genomics could be used to estimate the genetic value of animals that have not been measured for this trait. Here the problem comes from the genotype per environment interactions. Farms measuring food efficiency for milk production are good farms having the cows under a good environment. It is not clear that the best genetic animals in these farms will be the best in common farms under other environments. This has happened yet with another 'new trait' in pigs, residual feed intake, where the relationship between the breeding values of the animals in the nucleus of selection and the commercial farms was null (Knap and Wang 2012).

The difficulties in the implementation of genomic selection do not invalidate genomics for selection programmes, since genomics is a tool and how to use it efficiently is a matter of research. As we will see below, genomic selection has proved to be extremely useful in dairy cattle, but the cost of genotyping prevents its use in rabbit breeding programmes and complicates its application in pigs, lamb or poultry. Nevertheless, in pigs, poultry, lamb and beef cattle, genomic selections is, or can be, a useful complement to current selection programmes.

4.9 The Use of Genomic Selection in Breeding Programmes

Genomic selection has been applied with success in breeding programmes, with spectacular results in dairy cattle and with more modest results in other species. Nowadays there is no doubt that genomics is a useful tool for selection, but careful strategies for its implementation should be developed in most species to ensure its profitability.

Dairy cattle. Genomic selection has revolutionised the dairy cattle breeding programmes. As Schaeffer (2006) predicted before the first SNP chip was available, dairy cattle is particularly suitable for genomic selection. It has a long generation interval (6 years) due to the need of progeny test, the traits of interest (milk production and quality) cannot be measured in the sire, selection pressure has to be applied

essentially in sires because the average parities of dams is about 2.7, and the dissemination of the genetic progress is cheap and easy via artificial insemination. It was not a problem to create large reference populations and maintain a continuous recording system to update the equations, since a single bull can have many daughters and all farms constitute a global nucleus linked by artificial insemination. Moreover, sires have a high price; thus genomic cost is not an impediment for developing genomic selection compared to other species. Genomic selection was implemented in 2008 in several countries and nowadays is widely used for sire evaluation. Nowadays, as using imputation with 3000 SNPs chips has a high accuracy (up to 99%, Van Eenennaam et al. 2013), dairy cows are also being genotyped. As a result of this wide implementation, generation interval has been halved and genetic progress doubled (Wiggans et al. 2017). It is interesting to notice that other efficient programmes based on reducing the generation interval were proposed in the past, for example, MOET (multiple ovulation and embryo transfer, Nicholas and Smith 1983). However, in addition to difficulties in implementation MOET (Simianier 2016), as the accuracy of bulls evaluation was lower compared to proven bulls with 100 daughters (0.45 versus 0.95), farmers were reluctant to use them. Now, genomic bulls have still lower accuracy (around 0.8), but farmers accept the loss of accuracy and use several genomic bulls to lower the risk. Obviously, this loss of accuracy is compensated by far by the reduction of the generation interval, but the fascination for the new technique may have played a role in its rapid acceptance.

Beef cattle. The success of genomic selection in dairy cattle has moved the whole industry to consider the introduction of genomics in current breeding programmes. However, beef cattle is organised in many breeding associations, with a much lower size than the dairy cattle breeds. Moreover, beef cattle are not always well connected by artificial insemination. Because of this, it is not feasible for most beef cattle associations to have a 'training population' and a continuous recording as large as in dairy cattle. This has led to the proposal of using multibreed training populations for predictions, but the problem is that effectiveness of genomic breeding value prediction is higher when training populations are close to the animals to be predicted, otherwise the prediction is poor (Lund et al. 2014); thus the use of multibreed populations is now under discussion. Another problem is the cost of genotyping. In beef cattle, the most commonly measured traits are weights at a given age. Usually these traits have relatively high heritabilities (about 0.40), which means that the accuracy of the individual phenotype is about 0.6–0.7, and it can become higher by adding information from relatives. Therefore, genomics should improve accuracy over 0.7 when the trait of interest can be measured just using a scale, although in some extensive systems collecting samples for genomics may be easier than using a scale. Imputation may be a solution, but imputation is precise only when the low-density chip is used in animals closely related to the ones used for imputation (Rolf et al. 2014); thus multibreed low-density chips may be of little utility. Although it is true that genomics has been used by commercial companies as a marketing tool (Rolf et al. 2014), genomics could improve the accuracy for traits not directly measured, for example, when the objective is weight at slaughter but only weight at weaning is measured, or for carcass traits. Even in all these cases, a

careful study should be made taking into account the large training populations needed and the permanent cost of genotyping in relation to the benefits expected.

Sheep and goat. Lambs and goats bred for milk production have the same scheme as in dairy cattle at a much lower scale, which limits the application of genomic selection. Meat sheep shares with beef cattle most of its problems for the efficient use of genomic selection. In both cases, the low price of the animals limits the application of genomics due to the relatively high cost of genotyping. Rupp et al. (2016) have recently reviewed the application of genomics in sheep and goats. Gains in accuracy when applying genomic selection were rather modest, around 10–20%, even for milk production traits. Considering costs of genotyping, Shumbusho et al. (2016) estimated the economic advantages of using genomic selection in sheep meat to be only 15% in the best scenario. Similar results were found in Australian merino breed by Horton et al. (2015). Multibreed SNP chips have also been proposed, but they share the same problems as in beef cattle.

Pigs. In pigs, progeny test is not performed, and generation interval is consequently short (around 1 year). Selection objectives are traits expressed in males and females with the exception of litter size, dissemination of genetic progress is made through a pyramidal structure of nucleus-multiplier-commercial farms (where genetic improvement is performed only in the nucleus), and selection can be applied on dams because they are prolific animals. There is no global nucleus but several companies competing in a free market, having small nucleuses of around 25–50 males and 300–2000 females per line, and the price of selected animals is much lower than in dairy cattle. Moreover, as pigs are normally produced in a three-way cross scheme, the costs of genotyping are three times higher than when a single breed is used in production. With all of these constrains, the application of genomics has had less spectacular results than in dairy cattle; nevertheless the increment in profit when using genomic selection has been evaluated from 10% (Lillehammer et al. 2013) to 50% (Knol et al. 2016), depending on the implementation. Litter size is an obvious candidate for genomic selection because the trait is not expressed in the female when it should be selected, but heritability of litter size is very low, so large reference populations are needed, and the strategy for obtaining them is not evident; for example, information from multipliers can be used or even information from crossbred commercial females, as we mentioned before. The success of genomic selection in pigs comes from a careful study of the strategies for implementing genomic selection (see Ibáñez et al. 2014 and Knol et al. 2016 for a detailed description of some strategies). Imputation is important because genotyping is still economically relevant relatively to the price of selected animals, and the accuracy of imputation is high (around 97%, Cleveland and Hickey 2013). The success of genomic selection comes, again, from a better estimation of the relationships between animals.

Poultry. Similar constraints to pigs arise in poultry, in which four-way cross schemes are common; nucleuses are also small, although it can be found large nucleuses up to 2000 males and 10,000 females. Generation interval is also very short, females produce a large amount of eggs, and the relevant traits are expressed mainly in females in layers and in both sexes in broilers. Genomics was implemented in 2013 in both production systems. Careful imputation procedures have

obtained very good results in both layers and broilers, with accuracies of around 97% with respect to the high-density SNP chip, and due to the continuous decreasing in genotyping cost, medium density SNP chips are being used, removing the need of imputation (Wolc et al. 2016). A selection experiment in layers has evaluated the response to selection using genomics when compared with a line in which the same sort of selection was performed without genomics. The results were variable depending on the trait used in the selection index; traits like egg production number showed little advantage, but for some traits like egg weight, the use of genomic selection was much more efficient (Wolc et al. 2015). Efficiency of genomic selection in broilers has been evaluated by comparing the increasing in precision when evaluating the genetic merit of some traits (Chen et al. 2011). In the sire line, selected mainly for growth rate, the increment of precision for body weight when using genomic selection was 20%, and for ultrasound measurements of the breast, it was 17%; the dam line had better results for the same traits, but it was selected mainly for reproductive traits. In general, the best advantage of the use of genomic selection, as in the other species, comes from traits that are not available at the moment of selection (Wolc et al. 2016).

Rabbits. Genomics has not been implemented in rabbits yet, mainly due to the cost of genotyping. The rabbit chip of 200,000 SNPs appeared recently (October of 2015), and no low-density chips have been produced yet. Rabbit selection schemes are three-way crosses with the same structure as in pigs, and nucleuses are even smaller (from 20 males and 150 females per line); dam lines are mainly selected for litter size and sire lines for growth rate. Generation interval is very short (6–9 months), and the price of the animals is low, which represents the main constraint for the application of genomic selection. Research needs to be done to find the best strategy for implementing genomic selection in rabbits.

References

Andersson R (2015) Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. BioEssays 37:314–323. https://doi.org/10.1002/bies.201400162

Archibald AL et al (1995) The PiGMaP consortium linkage map of the pig (Sus scrofa). Mamm Genome 6:157–175

Berry DP, Kearney JF (2011) Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal 5:1162–1169

Blasco A (2008) The role of genetic engineering in livestock production. Livestock Sci 113:191–201

Blasco A (2017) Bayesian statitics for animal scientists. Springer, New York

Blasco A, Toro MA (2014) A short critical history of the application of genomics to animal breeding. Livstock Sci 166:4–9

Chen CY, Misztal I, Aguilar I, Tsuruta S, Meuwissen THE, Aggrey SE, Wing T, Muir WM (2011) Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: an example using broiler chickens. J Anim Sci 89:23–28

Clark SA, van der Werf J (2013) Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. In: Gondro C, van der Werf J, Hayes B (eds) Genome-wide association studies and genomic prediction. Springer, New York

Cleveland MA, Hickey JM (2013) Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J Anim Sci 91:3583–3592
ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. https://doi.org/10.1038/nature11247
Falconer D, Mackay TFC (1996) Introduction to quantitative genetics. Longman, Edinburgh
Fernando RL, Garrick D (2013) Bayesian methods applied to GWAS. In: Gondro C, van der Werf J, Hayes B (eds) Genome-wide association studies and genomic prediction. Springer, New York
Groenen MAM, Schook LB, Archibald AL (2011) Pig genomics. In: Rothschild MF, Ruvinsky A (eds) The genetics of the pig, 2nd edn. CAB International, Wallingford, UK, p 496. https://doi.org/10.1079/9781845937560.0000
Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397
Hangauer MJ, Vaughn IW, McManus MT (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet 9:e1003569. https://doi.org/10.1371/journal.pgen.1003569
Hausser J, Zavolan M (2014) Identification and consequences of miRNA-target interactions--beyond repression of gene expression. Nat Rev Genet 15:599–612. https://doi.org/10.1038/nrg3765
Horton BH, Banks R, Van der Werf JHJ (2015) Industry benefits from using genomic information in two- and three-tier sheep breeding systems. Anim Prod Sci 55:437–446
Huang Y, Hickey JM, Cleveland MA, Maltecca C (2012) Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet Sel Evol 44:25
Ibañez N, Blasco A (2011) Modifying growth curve parameters by multitrait genomic selection. J Anim Sci 89:661–668
Ibáñez-Escriche N, Forni S, Noguera JL, Varona L (2014) Genomic information in pig breeding: science meets industry needs. Livestock Sci 166:94–100
Jonas E, de Koning DJ (2015) Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. Anim Front 6:1–8
Knap PW, Wang L (2012) Pig breeding for improved feed efficiency. In: Patience JF (ed) Feed efficiency in swine. Wageningen Academic Publishers, Wageningen
Knol EF, Nielsen B, Knap PW (2016) Genomic selection in commercial pig breeding. Anim Front 6:15–22
Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756
Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. J Dairy Sci 92:4656–4663
Libri D (2015) Sleeping beauty and the beast (of pervasive transcription). RNA 21:678–679. https://doi.org/10.1261/rna.050948.115
Lillehammer M, Meuwissen THE, Sonesson AK (2013) Genomic selection for two traits in a maternal pig breeding scheme. J Anim Sci 91:3079–3087
Lund MS, Su G, Janss L, Guldbrandtsen B, Brøndum RF (2014) Genomic evaluation of cattle in a multi-breed context. Livestock Sci 166:101–110
Mattick JS (2011) The central role of RNA in human development and cognition. FEBS Lett 585:1600–1616
Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
Misztal I, Legarra A, Aguilar I (2009) Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. J Dairy Sci 92:4648–4655
Neale DB et al (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15:R59. https://doi.org/10.1186/gb-2014-15-3-r59
Nicholas FW, Smith C (1983) Increased rates of genetic change in dairy cattle by embryo transfer and splitting. Anim Prod Sci 36:341–353
Rolf MM, Decker JE, Mckay SD, Tizioto PC, Branham KA, Whitacre LK, Hoff JL, Regitano LCA, Taylor JF (2014) Genomics in the United States beef industry. Livestock Sci 166:84–93

Rupp R, Mucha S, Larroque H, McEwan J, Conington J (2016) Genomic application in sheep and goat breeding. Anim Front 6:39–44
Schaeffer LR (2006) Strategy for applying genome-wide selectionin strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123:218–223
Shumbusho F, Raoul J, Astruc JM, Palhiere I, Lemarié S, Fugeray-Scarbel A, Elsen JM (2016) Economic evaluation of genomic selection in small ruminants: a sheep meat breeding program. Animal 6:1033–1041
Silver LM (1995) Mouse genetics. Oxford University Press, Bar Harbor, Maine
Simianier H (2016) Genomic and other revolutions why some technologies are quickly adopted and others are not. Anim Front 6:53–58
Smith C, Smith DJ (1993) The need for close linkages in markers-assisted selection for economic meritin livestock. Anim Breed Abst 61:197–204
Soller M (1978) The use of loci associated with quantitative traits in dairy cattle improvement. Anim Prod 27:133–139
Van Eenennaam AL, Weigel KA, Young AE, Matthew AC, Dekkers JCM (2013) Applied animal genomics: results from the field. Annu Rev Anim Biosci 2:9.1–9.35
Van Raden PM, O'Connell JR, Wiggans GR, Weigel KA (2011) Genomic evaluations with many more genotypes. Genet Sel Evol 43:10
Vitezica ZG, Varona L, Elsen JM, Misztal I, Herring W, Legarra A (2016) Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. Genet Sel Evol 48:6
Warren WC et al (2017) A new chicken genome assembly provides insight into avian genome structure. G3 7:109–117. https://doi.org/10.1534/g3.116.035923
Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS (2017) Genomic selection in dairy cattle: the USDA experience. Annu Rev Anim Biosci 5:309–327
Wolc A, Zhao HH, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R, Stricker C, Habier D, Fernando RL, Garrick DJ, Lamont SJ, Dekkers JCM (2015) Response and inbreeding from a genomic selection experiment in layer chickens. Genet Sel Evol 47:59
Wolc A, Kranis A, Arango J, Settar P, Fulton JE, O'Sullivan NP, Avendano A, Watson KA, Hickey JM, De los Campos G, Fernando RL, Garrick DJ, Dekkers JCM (2016) Implementation of genomic selection in the poultry industry. Anim Front 6:23–31
Won KJ et al (2013) Comparative annotation of functional regions in the human genome using epigenomic data. Nucleic Acids Res 41:4423–4432. https://doi.org/10.1093/nar/gkt143
Wright MW (2014) A short guide to long non-coding RNA gene nomenclature. Hum Genomics 8:7. https://doi.org/10.1186/1479-7364-8-7
Xu J, Zhang J (2016) Are human translated pseudogenes functional? Mol Biol Evol 33:755–760. https://doi.org/10.1093/molbev/msv268
Yerle M et al (1995) The PiGMaP consortium cytogenetic map of the domestic pig (Sus scrofa domestica). Mamm Genome 6:176–186