1 Historical Overview

Introduction

Genomic selection is based on the synthesis of statistical and molecular genetics that occurred during the last three decades. In this introductory chapter we will review the landmark breakthroughs that lead to this synthesis. The first section reviews the milestones in the synthesis of Mendelian and quantitative genetics. The next section reviews the early experiments of quantitative trait locus (QTL) detection using morphological and biochemical markers, beginning with Sax's landmark experiment with beans (*Phaseolus vulgaris*). The following sections describe the development of DNA-level markers starting with restriction fragment length polymorphisms (RFLPs) to single nucleotide polymorphisms (SNPs) and copy number variations (CNV). The final sections describe QTL detection and marker-assisted selection (MAS) prior to genomic selection.

The Mendelian Theory of Genetics

Modern genetics is usually considered to have started with the rediscovery of Mendel's paper in 1900. The rediscovery of Mendel's laws led to a rapid first synthesis of genetics, statistics, and cytology. Boveri (1902) and Sutton (1903), first proposed the "chromosomal theory of inheritance" that the Mendelian factors were located on the chromosomes. Using *Drosophila*, Morgan (1910) demonstrated that Mendelian genes were linked and could be mapped into linear linkage groups of a number equal to the haploid number of chromosomes. Hardy (1908) and Weinberg (1908) independently derived their famous equation to describe the distribution of genotypes in a segregating population at equilibrium. That is, the frequencies of genotypes for a locus with two alleles with frequencies p and $1-p$ will be p^2, $2p(1-p)$, and $(1-p)^2$ for homozygotes for p-allele and heterozygotes and homozygotes for the other allele, respectively.

In 1919 Haldane derived a formula to convert recombination frequencies into additive "map units" denoted "Morgans" or "centimorgans," assuming a random distribution of events of recombination along the chromosome. The Haldane mapping function (Haldane, 1919) is based on the assumption of zero "interference" throughout the genome. That is, all events of recombination are statistically independent. In this case the number of events of recombination in any given chromosomal segment corresponds to a Poisson distribution. The map distance between

Genomic Selection in Animals, First Edition. Joel Ira Weller.

two genes in Morgans, M, which is a function of the frequency of observed recombination between them, R, is derived as follows:

$$M = -\frac{1}{2}\ln(1-2R) \tag{1.1}$$

The Mendelian Basis of Quantitative Variation

Unlike the morphological traits analyzed first by Mendel and then by Morgan, most traits of economic interest in agricultural species display continuous variation, rather than the discrete distribution associated with Mendelian genes. Despite the early synthesis between Mendelian genetics and cytogenetics, there seemed to be no apparent connection between Mendelian genetics on the one hand and quantitative variation and natural selection on the other.

Experiments by Johanssen (1903) with beans demonstrated that environmental factors are a major source of variation in quantitative traits, leading to the conclusion that the phenotype for these traits is not a reliable indicator for the genotype. Yule in 1906 first suggested that continuous variation could be explained by the cumulative action of many Mendelian genes, each with a small effect on the trait. (Many different terminologies have been employed for these genes. I will use the term "QTL" throughout.) Fisher in 1918 demonstrated that segregation of QTL in an outcrossing population would generate correlations between relatives. Payne (1918) demonstrated that the X chromosome from selected lines of *Drosophila* contains multiple factors, which influenced scutellar bristle number. Thus, by 1920, the basic theory necessary for detection of individual genes affecting quantitative traits was in place.

Detection of QTL with Morphological and Biochemical Markers

In 1923 Sax demonstrated with beans that the effect of an individual locus on a quantitative trait could be isolated through a series of crosses, resulting in randomization of the genetic background with respect to all genes not linked to the genetic markers under observation. Even though all of his markers were morphological seed markers with complete dominance, he was able to show a significant effect on seed weight associated with some of his markers.

During the next 50 years, there were relatively few successful experiments that found marker–QTL linkage in plant and animal populations, and of these even fewer were independently repeated. A major problem was the relatively small size of most experiments. In most cases in which QTL effects were not found, power was too low to find segregating QTL of a reasonable magnitude (Soller *et al.,* 1976).

In 1961 Neimann-Søressen and Robertson proposed a half-sib design for QTL detection in commercial dairy cattle populations. Although the actual results were disappointing, this was the first attempt to detect QTL in an existing segregating population. All previous studies were based on experimental populations produced specifically for QTL detection. This study was also groundbreaking in other aspects. It was the first study to use blood groups rather than morphological markers, and the proposed statistical analyses—a χ^2 (chi-squared) test, based on a squared sum of normal distributions, and ANOVA—were also unique. This was the first study that attempted to estimate the power to detect QTL and to consider the problem of multiple comparisons when several traits and markers were analyzed jointly.

Lewontin and Hubby showed in 1966 that electrophoresis could be used to disclose large quantities of naturally occurring enzyme polymorphisms in *Drosophila*. Almost all enzymes analyzed showed some polymorphism that could be detected by the speed of migration in an electric field. Studies with domestic plant and animal species found that electrophoretic polymorphisms were much less common in agricultural populations. During the 1980s there were a number of QTL detection studies in agricultural plants based on isozymes using crosses between different strains or even species in order to generate sufficient electrophoretic polymorphism (Tanksley *et al.,* 1982; Kahler and Wherhahn, 1986; Edwards *et al.,* 1987; Weller *et al.,* 1988). It was clear, though, that naturally occurring biochemical polymorphisms were insufficient for complete genome analyses in populations of interest.

DNA-Level Markers, 1974–1994

The first detected DNA-level polymorphisms were RFLPs. Grodzicker *et al.* (1974) first showed that restriction fragment band patterns could be used to detect genetic differences in viruses. Solomon and Bodmer (1979) and Botstein *et al.* (1980) proposed RFLP as a general source of polymorphism that could be used for genetic mapping. Although RFLPs are diallelic, initial theoretical studies demonstrated that they might be present throughout the genome. Beckmann and Soller (1983) proposed using RFLP for detection and mapping of QTL. The first genome-wide scan for QTL using RFLP was performed on tomatoes by Paterson *et al.* (1988). In animal species, however, RFLP markers were homozygous in most individuals and therefore have not been as useful for QTL mapping.

A major breakthrough came at the end of the decade with the discovery of DNA microsatellites. Mullis *et al.* (1986) proposed the "polymerase chain reaction" (PCR) to specifically amplify any particular short DNA sequence. Using the PCR, large enough quantities of DNA could be generated so that standard analytical methods could be applied to detect polymorphisms consisting of only a single nucleotide. Since the 1960s, it has been known that the DNA of higher organisms contains extensive repetitive sequences. In 1989 three laboratories independently found that short sequences of repetitive DNA were highly polymorphic with respect to the number of repeats of the repeat unit (Litt and Luty, 1989; Tautz, 1989; Weber and May, 1989). The most common of these repeat sequences was poly(TG), which was found to be very prevalent in all higher species. These sequences were denoted "simple sequence repeats" (SSR) or "DNA microsatellites."

Microsatellites were prevalent throughout all genomes of interest. Nearly all poly(TG) sites were polymorphic, even within commercial animal populations. These markers, unlike most morphological markers, were by definition "codominant." That is, the heterozygote genotype could be distinguished from either homozygote. Furthermore, microsatellites were nearly always polyallelic. That is, more than two alleles were present in the population. Thus, most individuals were heterozygous. Relatively dense genetic maps based on microsatellites were generated nearly in all agricultural species (e.g., Ihara *et al.,* 2004), and these markers were also used to detect and map segregating QTL. The weaknesses of microsatellites are twofold: First their distribution throughout the genome is not sufficiently dense for determination of causative polymorphisms responsible for observed QTL. (The causative polymorphisms will be denoted "quantitative trait nucleotides" (QTN).) Second, due to the repeat structure of microsatellites, PCR amplification was generally not exact, and "stutter bands" with varying numbers of the repeat unit were generated. Various rules were developed to estimate the actual genotype from

the observed PCR product, but the analysis could not be fully automated. A technician still had to review each individual genotype, and error rates in genotype determination were in the range of 1–5%.

DNA-Level Markers Since 1995: SNPs and CNV

Since 2000 "SNPs" (reviewed by Brookes (1999)) have supplanted microsatellites as the marker of choice for genetic analysis. An SNP is generally defined as a DNA base pair location at which the frequency of the most common base pair is lower than 99%. Unlike microsatellites, which usually have multiple alleles, SNPs are generally biallelic, but are much more prevalent throughout the genome, with an estimated frequency of one SNP per 300–500 base pairs. In human populations differences in the base pair sequence of any two randomly chosen individuals occur at a frequency of approximately one per 1000 kb (Brookes, 1999). Thus, SNPs can be found in genomic regions that are microsatellite poor. SNPs are apparently more stable than microsatellites, with lower frequencies of mutation. Beginning in 2005, methods were developed for automated scoring of first thousands and then hundreds of thousands of microsatellites per individual. Genotyping error rates are in the range of 0.05–0.01% with "BeadChip" technology (Weller *et al.,* 2010). A detailed description of the technologies developed for high-throughput SNP analysis is beyond the scope of the current text. For details, see Matukumalli *et al.* (2009).

QTL Detection Prior to Genomic Selection

Generally both natural and commercial populations are at linkage equilibrium for the vast majority of the genome. The exception is genomic sites that are closely linked on the same chromosome. Unlike genetic linkage within families that extends over tens of centimeters, population-wide linkage disequilibrium (LD) extends in animals over less than 1 cM (Sargolzaei *et al.,* 2008; Qanbari *et al.,* 2010). Therefore, unless a segregating genetic marker is closely linked to a QTL segregating in the population with an effect on some trait of interest, no effect will be associated with the marker genotypes. Thus naturally occurring LD could not be exploited prior to the advent of high-density genome scans. To detect the effect of a single QTL in outbred populations prior to high-density genome scans, it was necessary to generate LD.

In an analysis of inbred lines we are confronted with the opposite problem. That is, a significant effect associated with a genetic marker may be due to many genes throughout the genome and not necessarily to genes linked to the genetic markers. In crosses between inbred lines it was necessary to devise an experimental design that isolates the effects of the chromosomal segments linked to the segregating genetic markers.

Experimental designs can be divided into designs that are appropriate for crosses between inbred lines and those designs that can be used for segregating populations. Most early analyses performed to detect QTL have been based on planned crosses, although studies on humans, large farm animals, and trees have used existing populations. For humans, most species of domestic animals, and fruit trees, it is impractical to produce the inbred lines. Instead, experimental designs were based on the analysis of families within existing populations. Three basic types of analyses have been proposed—the "sib-pair" analysis for analysis of many small full-sib families, the "full-sib" design for analysis of large full-sib families, and the "half-sib" or "daughter design" analysis for large half-sib families.

MAS Prior to Genomic Selection

Prior to genomic selection, two MAS breeding programs were initiated in dairy cattle based on microsatellites in German and French Holsteins (Bennewitz *et al.,* 2004b; Boichard *et al.,* 2006). Both programs computed marker-assisted genetic evaluations (MA-BLUP) based on the algorithm of Fernando and Grossman (1989).

In the German program, markers on three chromosomes were used. The evaluations were distributed to Holstein breeders who used these evaluations for selection of bull dams and preselection of sires for progeny testing. The algorithm only included equations for bulls and bull dams, and the dependent variable was the bull's daughter yield deviation (VanRaden and Wiggans, 1991; derivation and use of daughter yield deviations will be discussed in detail in Chapter 6). Linkage equilibrium throughout the population was assumed. To close the gap between the grandsire families analyzed in the German granddaughter design and the bulls in use in 2004, 3600 bulls were genotyped in 2002. Until 2008, about 800 bulls were evaluated each year. Only bulls and bull dams were genotyped, since tissue samples were already collected for paternity testing. Thus additional costs due to MAS were low, and even a very modest genetic gain could be economically justified. This scheme was similar to the "top-down" scheme of Mackinnon and Georges (1998) in that evaluation of the sons was used to determine which grandsires were heterozygous for the QTL and their linkage phase. This information was then used to select grandsons based on which haplotype was passed from their sires. It differed from the scheme of Mackinnon and Georges (1998) in that the grandsons were preselected for progeny test based on MA-BLUP evaluations (Fernando and Grossman, 1989), which include general pedigree information in addition to genotypes.

The French MAS program included elements of both the "top-down" and "bottom-up" MAS designs. Similar to the German program, genetic evaluations including marker information were computed by a variant of MA-BLUP, and only genotyped animals and nongenotyped connecting ancestors were included in the algorithm. Genotyped females were characterized by their average performance based on precorrected records (with the appropriate weight), whereas males were characterized by twice the "yield deviations" of their nongenotyped daughters (yield deviations will also be explained in Chapter 6). Twelve chromosomal segments, ranging in length from 5 to 30 cM, were analyzed. Regions with putative QTL affecting milk production or composition were assumed to be located on bovine chromosomes 3, 6, 7, 14, 19, 20, and 26; segments affecting mastitis resistance on chromosomes 10, 15, and 21; and chromosomal segments affecting fertility on chromosomes 1, 7, and 21. Each region was found to affect one to four traits, and on an average three regions with segregating QTL were found for each trait. Each region was monitored by two to four evenly spaced microsatellites, and each animal included in the MAS program was genotyped for at least 43 markers. Sires and dams of candidates for selection, all male AI ancestors, up to 60 AI uncles of candidates, and sampling daughters of bull sires and their dams are genotyped. The number of genotyped animals was 8000 in 2001 and 50,000 in 2006.

Guillaume *et al.* (2008) estimated by simulation the efficiency of the French program. Breeding values and new records were simulated based on the existing population structure and knowledge of the variances and allelic frequencies of the QTL under MAS. Reliabilities of genetic values of animals less than 1 year old obtained with and without marker information were compared. Mean gains of reliability ranged from 0.015 to 0.094 and from 0.038 to 0.114 in 2004 and 2006, respectively. The larger number of animals genotyped and the use of a new set of genetic markers can explain the improvement of MAS reliability from 2004 to 2006. This improvement was also observed by the analysis of information content for young candidates. The gain of MAS reliability with respect to classical selection was larger for sons of sires with genotyped daughters with records.

Summary

By 2005 dense genetic maps based on DNA-level genetic markers were developed for nearly all economically important animal species. Numerous studies demonstrated that QTL affecting traits of economic importance could be detected via linkage to genetic markers. Theory was developed for MAS based on selection of a relatively small number of chromosomal segments, and several MAS breeding programs for dairy cattle were implemented in two countries. The "rules of the game" were to change dramatically in 2006 with the development of high-throughput SNP chips, which will be discussed in detail in the next chapter.