Strategies to Assess Structural Variation in the Chicken Genome and its Associations with Biodiversity and Biological Performance

M. Soller,∗ S. Weigend,† M. N. Romanov,‡ J. C. M. Dekkers,§ and S. J. Lamont§

*Hebrew University of Jerusalem, 91904, Israel; †Institute for Animal Breeding Mariensee, Federal Agricultural Research Centre, 31535 Neustadt, Germany; ‡Arnold and Mabel Beckman Center for Conservation Research, Zoological Society of San Diego, Escondido, CA 92027; and §Department of Animal Science, Iowa State University, Ames 50011

ABSTRACT A primary goal in the assessment of structural variation in the avian genome is to understand the relationship of this variation with biodiversity and with biological performance. To develop such knowledge, certain essential tools are needed. One set of tools includes the laboratory techniques used to assess molecular genetic variation. The current time is a transitional one for this field, in that the recently sequenced chicken genome will add significantly to the portfolio of existing methods used to identify molecular markers. To most efficiently discover marker-trait associations, the experimental mapping populations must be appropriately designed and the relevant statistical analyses applied. This paper reviews methods for assessment of molecular markers in poultry and their use in the characterization of avian biodiversity and in studies to identify marker associations with biological traits, including important considerations of population structure and statistical analysis.

Key words: structural variation, genome, biodiversity, quantitative trait gene, candidate gene

2006 Poultry Science 85:2061–2078

TECHNIQUES FOR ASSESSMENT OF STRUCTURAL VARIATION IN THE CHICKEN GENOME TO EVALUATE GENETIC DIVERSITY

Genetic diversity within a given farm animal species refers to the variety of genetic variants evolved during domestication and is displayed by the existence of structural variation among genomes of individuals, families, strains, and populations. Poultry genetic resources require further identification and evaluation to set up efficient measures for their conservation and utilization. Recent advances in molecular technology have opened up new horizons for characterizing structural variation at the genome-wide level and for working out conservation measures. The link between biodiversity and modern genomics may facilitate a sustainable management of genetic diversity in farm animals and ensure its exploitation for human benefit. In this section, various aspects of the application of molecular markers to evaluate chicken biodiversity will be reviewed.

Molecular Markers in Chicken Biodiversity Studies

On the assumption that the more distant a breed or population is, the more likely it might carry unique genetic features, the assessment of genetic variation and distances by molecular tools may provide useful information for initial evaluation of chicken genetic resources (Weigend and Romanov, 2001). The evaluation of genetic diversity within and between different chicken populations, both native and commercial, has been undertaken by using several DNA marker systems (Table 1). Before contemporary molecular tools had been developed, researchers had to rely on techniques of indirect, quantitative evaluation of genetic diversity based on DNA and DNA duplex formation and RNA and DNA hybridization and using, for instance, repetitive DNA base sequences, rapidly labeled RNA, and 28S rRNA that were applicable to interspecific comparisons only. Over past 3 decades, the fundamental DNA technology developments—restriction enzymes coupled with Southern-blot hybridization, sequencing, and PCR—have contributed to a burst of applications in multiple research areas, including genetic variation and diversity in chickens (Weigend et al., 2004a).

RFLP. Originally, RFLP referred to analysis of band patterns derived from DNA cleavage using restriction endonuclease enzymes. Restriction fragment length polymorphism and related techniques are usually modifications of the Southern blot method when the whole genomic DNA or its fraction is cut with restriction enzymes, transferred to a membrane, and hybridized with radiolabeled or fluorescent probes. The latter can be cloned fragments of endogenous avian viruses, particular nuclear genes, MHC genes, EST, or mitochondrial DNA (mtDNA) genes.
Table 1. Molecular markers for assessment of genetic diversity

<table>
<thead>
<tr>
<th>Marker type</th>
<th>Polymorphism</th>
<th>Detection method</th>
<th>Sequence information</th>
<th>Locus specificity</th>
<th>Reproducibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFLP</td>
<td>Restriction fragment length polymorphism</td>
<td>Restriction enzyme cutting of DNA Gel electrophoresis Membrane blotting Hybridization Visualization</td>
<td>Not required Yes Good</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VNTR</td>
<td>Variable no. of tandem repeats A) Minisatellites (repeat unit 10 to 100 bp) B) Microsatellites (repeat unit 1 to 5 bp)</td>
<td>Multilocus fingerprints (similar to RFLP) Multilocus fingerprints (similar to RFLP)</td>
<td>Not required No Medium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAPD</td>
<td>Random amplified polymorphic DNA</td>
<td>Single-locus PCR, PAGE PCR, arbitrary primer (10 bp) Gel electrophoresis</td>
<td>Required Yes Very good</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFLP</td>
<td>Amplified fragment length polymorphism</td>
<td>Restriction enzyme cutting of DNA Ligation PCR (2 steps) PAGE</td>
<td>Not required No Good</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphisms</td>
<td>Allele-specific PCR Single-strand conformation polymorphism PCR-RFLP Hybridization to high-density DNA arrays Primer extension Mass spectrometry</td>
<td>Required Yes Very good</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DNA Fingerprinting.** A widespread derivative of the restriction enzyme-based method is DNA fingerprinting for detecting multiple anonymous loci across the whole genome that are either multilocus minisatellites or multilocus microsatellites, also known as variable number of tandem repeat loci. Minisatellites consist of tandem repeat units of a 10- to 100-base motif flanked by conserved restriction endonuclease sites. They can be revealed using a DNA probe containing multiple copies of the minisatellite core sequence, such as bacteriophage M13 DNA probe (Dawe et al., 1988), human hypervariable probes 33.6 (Dunnington et al., 1991) and α-globin 3’HVR (Meng et al., 1996), and cattle probe R18.1 (Dunnington et al., 1991). Multilocus microsatellites can be detected by using oligonucleotide probes that contain a shorter, simple core motif (e.g., CAC₅, GGAT₅, etc.) without flanking sequences, and are represented throughout the genome. They were frequently applied to chicken diversity studies in the last decade because no preliminary sequence information about these anonymous multilocus markers is necessary (unlike single-locus microsatellites, see below), and individual or pooled RFLP patterns can easily be compared with identity variation within and among populations studied. The technique is time consuming but might still be useful in species for which no or little sequence information is available.

**PCR-Based Techniques.** Amplification of noncoding or coding regions of a genome using PCR has revolutionized molecular genetics research and provided an impressive variety of new markers to tackle diversity problems:

1. **Random amplified polymorphic DNA markers.** The random amplified polymorphic DNA technique employs single short primers of random sequence, usually 10-mers, which produce multiband patterns similar to DNA fingerprints. No sequencing information is needed before genotyping. Use of these markers to study poultry genetic diversity was thought to be promising, and they were heavily exploited in the 1990s. However, because of poor PCR reproducibility and dominance mode of inheritance, they are no longer markers of choice.

2. **Amplified fragment length polymorphism markers.** As developed by Keygene (Keygene N.V., Wageningen, The Netherlands), the amplified fragment length polymorphism technique involves the restriction of genomic DNA, followed by ligation of adaptors complementary to the restriction sites and selective PCR amplification of a subset of the adapted restriction fragments. Although this type of markers is popular, especially among plant researchers, there are just a few examples of its application to examine genetic variation in chickens. Like random amplified polymorphic DNA markers, amplified fragment length polymorphism markers are characterized by a dominant nature, which is a main disadvantage of this technique.

3. **Microsatellites.** These types of single-locus markers are also known as short (or simple) tandem repeats, simple sequence repeats, or simple sequence-length polymorphisms and belong to a variable number of tandem repeat loci, the most extensively used class of highly polymorphic molecular markers. Unlike all the above techniques, prior sequence information of flanking regions is necessary to develop these markers. Major ad-
vantages of microsatellites are that they are detectable by PCR representing unique sequences in the genome that can be mapped and easily be exploited for many genetic applications. Also, they show extensive allelic differences in length, mainly based on variation in the number of repeats and partly on polymorphism of flanking regions.


**Sequencing Approach.** Direct sequencing became possible for wide application with the invention of PCR amplification. In chickens and their wild galliform relatives, this approach has been mostly used to explore variation in mtDNA genes as well as in nuclear genes. More recently, it has become popular in discovering single nucleotide polymorphism (SNP) markers (see below).

There are several other molecular techniques used for investigating inter- and intrapopulation genetic variation in chickens, for instance, PCR-RFLP. This is a combination of 2 major approaches when DNA regions of interest are amplified by PCR, and sequence polymorphisms of these fragments are subsequently detected by RFLP assays. Numerous examples of the implementation of various DNA markers in chicken biodiversity studies are listed elsewhere (http://www.msu.edu/~romanoff/biodiversity/studiesdb.htm).

**SNP as Molecular Markers in Poultry Biodiversity Studies**

A SNP is a minimal DNA variation that occurs as a replacement of a single nucleotide with 1 of the 3 other possible nucleotides. Single nucleotide polymorphisms are the most common class of genetic polymorphism, often outside of coding regions, and they make up a new generation of biallelic markers that become promising for use in biodiversity studies due to their abundance and applicability for high throughput analyses (Schmid et al., 2005). Estimates of SNP frequencies in poultry species range from 1:48 to 1:1632 bp (Table 2). Single nucleotide polymorphisms can be used in biodiversity studies as single loci or haplotypes, which are stable over the course of the evolution. Three different possible conditions of SNP in coding regions can lead to different phylogenetic outcomes (Weigend et al., 2004a):

1. **Synonymous SNP** should not be subject to selection, except cases when they are linked to genes controlling traits under selection. Thus, genetic relationships among

| Table 2. Frequencies and distribution of single nucleotide polymorphisms (SNP) in fowl. |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Population      | SNP frequency  | Source of DNA   | No. of individuals | Transition substitutions (%) | Non-synonymous substitutions (%) | Reference       |
| Chicken         | 1:77           | EST             | 10               |                           |                               | Smith et al. (2000)² |
| Guinea fowl     | 1:155          | EST             | 10               |                           |                               | Smith et al. (2000)² |
| Pigeon          | 1:55           | EST             | 10               |                           |                               | Smith et al. (2002)² |
| Japanese quail  | 1:165          | EST             | 10               |                           |                               | Smith et al. (2002)² |
| White Leghorn   | 1:185          | EST             | 10               |                           |                               | Twito et al. (2002) |
| Layers          | 1:140          | Introns and promoters | 200 from 20 populations |                           |                               | Wong et al. (2004)² |
| Broilers        | 1:145          | Introns and promoters | 1 from each of 4 populations |                           |                               | Schmid et al. (2005)² |

Modified from Schmid et al. (2005).

SNP within 4 different species.

SNP in White Leghorns; heterozygotes were identified as SNP.

SNP in 6 layer lines and 3 broiler lines; heterozygotes were identified as SNP.

SNP in 10 chicken populations within the AVIANDIV project.

SNP in 4 different species: Red Jungle Fowl, Cornish (broiler), White Leghorn (layer), and Silkie (Chinese native breed).
populations based on this type of SNP should reflect the evolutionary time more accurately than the other SNP types (Kimura, 1983).

2. Nonsynonymous SNP may result in mutated proteins, and, thus, their frequencies could be affected by selection and the accumulation of genetic differences may be accelerated.

3. Single nucleotide polymorphisms in the promoter region can cause mutations in binding sites of transcription factors and, thus, affect the transcription and, probably, the phenotype.

Single nucleotide polymorphisms can be discovered by direct sequence comparison of a certain genomic region among individuals, mining EST data or bacterial artificial chromosome libraries, or by in silico comparison. For SNP-based genotyping, the major techniques are oligonucleotide chips, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, and pyrosequencing (Schmid et al., 2005).

Recent completion of the chicken genome sequence draft and its ongoing improvement and annotation offer new, prodigious opportunities to poultry geneticists. We can now look deeply into the genomic architecture of biological processes, phenotypic traits, and biodiversity. Single nucleotide polymorphisms seem to be key markers in such investigations. With the help of the Gallus gallus genome sequence as a reference framework, a map of genetic variation was developed for 3 different strains of domestic chickens: broiler, layer, and Silkie (Wong et al., 2004). To make the map, researchers identified and analyzed over 3.6 million nonredundant, high-quality sequence variation sites, mostly SNP. The genetic variation data was deposited into the Chicken Variation Database (http://chicken.genomics.org.cn/index.jsp; Wang et al., 2005) and dbSNP (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp), from which the data are freely accessible to researchers worldwide. Because a circumstantial SNP map of the whole genome was created, we can screen genomes of various breeds and populations to determine the following: 1) SNP profile of a given breed, population, or both and SNP allele frequencies across populations, 2) the uniqueness of the breed, population, or both in terms of harboring rare SNP or haplotypes, or 3) the value of the breed, population, or both concerning economically important SNP.

One platform for massive SNP typing that is currently in usage for chicken is the Illumina BeadArray technology (Illumina Inc., San Diego, CA), which allows simultaneous analysis of multiple samples and SNP targets. Each assay associates a fluorescent label with an oligonucleotide sequence that is complementary to a particular SNP address on a bead (Oliphant et al., 2002). This association is made to have high allele and locus specificity. Two allele-specific oligonucleotides (ASO) and 1 locus-specific oligonucleotide (LSO) are designed for each SNP. Each ASO consists of one end that hybridizes to genomic DNA at the SNP locus, and the other incorporates a universal PCR primer (Fan et al., 2003). Initially, an allele-specific primer extension is carried out, followed by ligation of the extended ASO to their LSO to create PCR templates (Fan et al., 2003). Each SNP is assigned a different address sequence, which is identified by the LSO. This assay allows for single-base extension, to perform robust genotyping of samples (Oliphant et al., 2002).

These saturated SNP data will allow a new level of characterization, conservation, and utilization of the chicken gene pool that will aid in elucidating the biological role and function of numerous chicken SNP. Although the new, high-throughput SNP assays are promising, their expense will still make them inaccessible for many researchers and projects. Therefore, researchers are likely to continue to use, for years, the techniques that have generated existing data. Genomic resources available for chicken will also be indispensable in acquiring knowledge about genomes and biodiversity of other poultry species.

ASSESSMENT OF BIODIVERSITY BY MOLECULAR GENETICS

In recent years, numerous studies in various species, including farm animals, have demonstrated that the variability in DNA is a powerful source of information for examining diversity within and among individuals, families, and populations. Microsatellites are currently most commonly used, and successful application of these types of markers in biodiversity studies has been reported for all major livestock species (Weigend and Romanov, 2001; Delany, 2003). Using microsatellites, insight into the extent of diversity of a wide range of chicken breeds originating from various countries has been gained in the European research project AVIAN-DIV [EC contract no. BIO4-CT98-0342 (1998–2000); S. Weigend, (coordinator), M. A.M. Groenen, M. Tixier-Boichard, A. Vignal, J. Hillel, K. Wimmers, T. Burke, and A. Mäki-Tanila; Hillel et al., 2003] and follow-up studies (Weigend et al., 2004b). The set of chicken breeds studied consisted of widely distributed local populations from Europe, Asia, and Africa; purebred lines of commercial layers and broilers; 2 Red Jungle Fowl subspecies (Gallus gallus gallus and Gallus gallus spadiceus); and 1 inbred line. Overall, results suggest that Jungle Fowl populations and traditional unselected breeds are widely heterogeneous populations that may include a large portion of the genetic diversity. Within commercial chickens, broiler lines were slightly more polymorphic than layers. Among the layers, the white egger strains had the lowest polymorphism as compared with brown eggers. In recent years, there has been concern about reduced genetic variability in commercial white egg layers that originated from a sole breed, the Single Comb White Leghorn. Although findings of the AVIAN-DIV project support, to some extent, this concern, commercial lines still exhibit a considerable amount of variation at microsatellite loci. It should be noted, however, that microsatellites are predominantly in noncoding regions of the genome. Although we cannot rule out that a few of the loci used...
are linked to genes controlling traits under selection, they may collectively be considered as neutral to selection. Hence, differences in allele frequencies among populations are probably the result of genetic drift and founder effects, on the one hand, and crossbreeding, in many cases unrecorded, on the other hand. By studying polymorphisms in genes coding traits under selection, the biodiversity profile may be considerably different from that obtained from microsatellites.

In recent years, SNP have been discussed as a very promising class of molecular markers for biodiversity studies. It has been found that about 70% are common SNP, which segregate in many domestic breeds and originated before divergence of modern chicken breeds (Wong et al., 2004). The observation that there are similar degrees of differences among various domestic breeds and between those and the Red Jungle Fowl supports the results obtained with microsatellites that there is still a high level of genetic variation present in modern commercial chicken lines. Although chickens have undergone intensive selection, resulting in highly productive, specialized strains, it appears that breeding has not dramatically reduced genetic variation. However, when studying nonsynonymous SNP in functional loci, this view might change, and greater similarity may be found within and among commercial chicken lines (Ye et al., 1998, Delany, 2000). This is further supported by the observation that genetic differences among populations quantified with neutral molecular markers are often not highly correlated with differences in their performance and phenotypic traits.

In the AVIANDIV project, 145 SNP were revealed through sequencing of 6,952 bp from 15 genomic DNA fragments in noncoding regions (1 fragment out of the 15 was found to be monomorphic) in a subset of 10 populations (10 individuals in each population; Schmid et al., 2005). On average, 1 SNP per 50 bp were found in this study, which is much higher than the reported frequencies in chickens by the International Chicken Polymorphism Map Consortium (Wong et al., 2004). In the latter study, a SNP rate was found to be about 5 SNP per 1,000 bp by comparing sequences of different domestic breeds except for broiler-broiler and layer-layer combinations, which showed a rate of 4 SNP per 1,000 bp. The high frequency found in the AVIANDIV project presumably reflects the wide genetic spectrum of chicken breeds collected, suggesting a much greater actual number of SNP in the chicken genome. In addition, Twito et al. (2002) used 10 individuals from each of 20 diverse populations of the AVIANDIV project for genotyping at 25 coding SNP markers from 25 genes (1 SNP per gene). Frequencies of the rare SNP alleles in the synonymous sites, untranslated regions, and nonsynonymous sites were 0.29, 0.29, and 0.095, respectively.

In developing countries, village chickens may represent a diverse gene pool that could comprise unique genetic features. Due to their development in a given environment, they might be better adapted to survive under harsh conditions without a proper management program and a limited supply of resources. Until recently, limited efforts have been made to characterize the local poultry strains based on molecular markers (Sharma et al., 1998; van Marle-Köster and Nel, 2000; Wimmers et al., 2000). Over the last several years, the number of local breed genetic studies has increased each year, especially in Asia (http://www.msu.edu/~romanoff/biodiversity/studiesdb.htm). There is only little stratification into well-defined breeds, and the survival of local poultry breeds is threatened by several factors of economic and social needs; however, crossbreeding in an uncontrolled way is, in fact, one of the major causes that erode genetic diversity in the third world (Wollny, 2003). Similarly, in developed countries, the widely spread local middle-level food producers disappeared from drastically changed agricultural production systems, and part of them are nowadays maintained by fanciers without recording reliable pedigree information. Under such circumstances, the assumed relationship among breeds may give a misleading structure of populations. Individuals with no pedigree information might be of interest for conservation, but their breed or population affiliation is uncertain. Furthermore, the correct assignment of individual affiliation to a genetic group based on molecular information (i.e., average similarity) may be important for association studies between genotype at particular locus and a phenotype in a given environment.

The utility of individual microsatellite typing for clustering and assigning individuals to genetic groups was studied in a subset of 20 breeds within the AVIANDIV project (Rosenberg et al., 2001). Each of 600 individuals representing 20 distinct chicken breeds were genotyped at 27 microsatellite loci. The individual microsatellite data were used to deduce genetic clusters based on the algorithm implemented in the software STRUCTURE (Pritchard et al., 2000), which identifies clusters of related individuals from multilocus genotypes. Clustering analysis revealed that most individuals of the 20 breeds were correctly assigned to the original population. The clustering success rate was about 98%, using all 27 markers. When markers of highest heterozygosity or highest number of alleles were used, 8 to 10 marker loci were sufficient to achieve >95% clustering success. On the other hand, when 12 to 15 highly variable markers and only 15 to 20 of the 30 individuals per breed were included in the analysis, clustering success was at least 90%.

Based on the conclusion from the above mentioned study that the model-based algorithm in STRUCTURE has potential applications in defining the within-species population subdivision, Hillel et al. (2005) undertook a large-scale analysis, including 2,000 individuals from 65 populations representing different chicken types and various geographical regions, which were typed individually at 29 microsatellite loci. Preliminary results indicated that the 65 populations were clustered nicely into their geographic origin and cultivation history. There is obviously low admixture between genomes of nonselected populations originated from Asia vs. those from Europe. In addition, some mixed populations were identified that share...
genetic clusters with several other populations. Some of these populations, such as the Red Jungle Fowl, may have been the progenitor of the domesticated chicken, whereas others reflect genetic admix from several origins (Hillel et al., 2005).

As in other species, mtDNA sequences have been used to study relationships within and among poultry species. The mtDNA is a circular molecule that is 16,775 bp in size (Desjardins and Morais, 1990) and that has a maternal mode of inheritance. The displacement loop region of the mtDNA contains the elements that control the replication of the molecule, and it is highly polymorphic. These features make the displacement-loop region attractive for phylogenetic studies. One of the first attempts to look into the problem of the genealogical origin of the present domestic chickens at the molecular level was undertaken by Fumihito et al. (1994, 1996) by studying the displacement-loop region of various Gallus species, including Red Jungle Fowl (Gallus gallus) and diverse domestic breeds. Based on sequence differences, Fumihito et al. (1996) suggested that the origin of domestic chickens is mono- phyletic, a conclusion that supported Darwin’s hypothesis (Darwin, 1868). Darwin (1868) was the first to hypothesize that the Red Jungle Fowl was the direct ancestor of the chicken. In the study of Fumihito et al. (1996), a rather small number of wild chicken samples were studied, and they concluded that 1 continental subspecies of Red Jungle Fowl from Thailand is the maternal origin of all chickens. Sequence analysis of the displacement-loop region in some Chinese populations suggested that they originated from Thailand as well (Niu et al., 2002). However, recent work has challenged this view. Based on sequence comparison of whole mtDNA and 2 segments of the nuclear genome among several species of the genus Gallus, Nishibori et al. (2005) reported that interspecies hybridizations may have occurred between Gray Jungle Fowl (Gallus sonneratii), Red Jungle Fowl, and domestic chickens and between Gray Jungle Fowl and Ceylon Jungle Fowl (Gallus lafayettei). Liu et al. (2006) analyzed the mtDNA hyper-variable segment I for a large number of domestic chickens across Eurasia as well as wild Red Jungle Fowls from Southeast Asia and China. Results suggested that several ancestral chicken populations contributed to the maternal genetic makeup of the species, which supports the theory of multiple origins of domestic chickens in South and Southeast Asia. It would be interesting to compare results obtained from mtDNA with clustering based on nuclear polymorphisms and with a broader geographical sampling to obtain further insight into the origin of domestic chickens (Schmid et al., 2005).

Overall, improvement of our knowledge about the mechanisms underlying genetic diversity may assist in the reconstruction of domestication events, determination of relationships among populations, and assessment of genetic variation within populations. Molecular genetic tools will be of invaluable help to characterize and classify genetic resource populations and will provide information; these are all essential prerequisites for effective conservation and management of genetic diversity within avian species.

ASSOCIATION OF STRUCTURAL GENETIC VARIATION WITH BIOLOGICAL PERFORMANCE: EXPERIMENTAL APPROACHES

It is generally accepted that genetic variation in quantitative traits within populations of outcrossing species is due to the cumulative action of a number of genes, called quantitative trait genes (QTG) or QTL, all acting on the same trait (Falconer and Mackay, 1996). It is convenient to think of each QTG as presenting a “positive” allele, having an increasing effect on trait value, and a “negative” allele, having a decreasing effect on trait value. Allele effects of individual QTG are assumed to be small relative to the totality of genetic and environmental variation. Because of this genetic structure, it is not possible, using Mendelian principles alone, to infer a specific genotype of an individual for a quantitative trait. This enables individual QTG to be mapped to defined chromosomal locations. The map location of the QTG gives them a specific identity, enabling their number and individual properties to be determined.

Haplotypes

Gene mapping is made possible by the fact that genes located on the same chromosome are physically linked together. Consider an individual having genotype $A_1A_2B_1B_2$. If the 2 genes, A and B, are on the same chromosome, the specific coupling of the alleles on the individual’s homologous chromosomes into haplotypes can be specified (Weir, 1990); in the above individual, the haplotype arrangement might be $A_1B_2$ and $A_2B_1$ for its 2 homologous chromosomes. Another individual with the same genotype might have a different haplotype composition (e.g., $A_1B_1$ and $A_2B_2$).

Haplotypes and Preferential Association of Alleles: The Identical by Descent Principle

Because of their physical association on the same chromosome, alleles making up a haplotype tend to be transmitted as a group from parent to offspring. For this reason, offspring that received the same haplotype from their common parent are said to be identical by descent (IBD) for that haplotype. Thus, if the parent has haplotypes $A_1B_1$ and $A_2B_2$, the progeny population will primarily consist of 2 large IBD groups: those that received haplotype $A_1B_1$ and those that received $A_2B_2$. Considering these 2 progeny groups only, the marker alleles making
up the haplotypes derived from their common parent show complete “preferential” (as opposed to “independent”) association: Allele A1 is always found together with allele B1, and allele A2 is always found together with allele B2. Thus, in this case, knowledge of allele status at locus A conveys complete information about allele status at locus B. For reasons that will be given later, this situation of preferential association of alleles at linked loci is technically termed a state of linkage disequilibrium (LD; Weir, 1990). We will refer to this ability of identity by descent to produce progeny groups that are in LD as the IBD principle.

In addition to the 2 haplotype groups directly derived from the parental haplotypes, some of the progeny will receive novel haplotypes containing new combinations of the 2 parental haplotypes. This occurs because of a meiotic process termed “crossing over,” which recombines parental haplotypes into new “recombinant” haplotypes. In our example, these would be A1-B2 and A2-B1. The recombinant haplotypes between the 2 loci are produced in proportion r/2 each so that together they make up a proportion of the recombination rate of all meiotic products. Thus, an A1-B1/A2-B2 parent will form 4 meiotic products: 2 “parental types”: A1-B1 and A2-B2 and 2 recombinant types: A1-B2 and A2-B1. As a result of recombination, allele A1 is now associated with allele B2, as well as with allele B1; similarly, allele A2 is also now associated with both alleles B1 and B2. Clearly, then, the presence of the recombinant haplotypes reduces the degree of preferential association (or LD) present in the population. Nevertheless, as long as recombination rate for a pair of loci is <0.50 (which can only hold if the 2 loci are on the same chromosome), a certain degree of preferential association of alleles at these loci will be present in the progeny population as a whole. This preferential association tells us that the 2 loci are linked.

**Preferential Association and LD**

Because linkage between a pair of loci is declared from preferential association of alleles, a quantitative measure of degree of association of alleles in the population is needed. This is provided by the deviation of the observed association of alleles at the 2 loci from the situation expected under independence. Independence, in turn, is found when an allele at locus A is associated with a specific allele at locus B, according to what would be expected by chance from the simple frequency of the B locus allele. For example, if the frequency of allele B1 is 0.6, and that of allele B2 is 0.4, then, under independence, we expect allele A1 to be associated with allele B1 60% of the time and with allele B2 40% of the time. The same will hold for allele A2. In general, then, on the above definition of independence, it can easily be shown that \( p_{A_1B_1} = p_{A_1} \times p_{B_1} \), where \( p_{A_1B_1} \) is the expected frequency of haplotype \( A_1B_1 \); \( p_{A_1} \) is frequency of allele \( A_1 \); and \( p_{B_1} \) is frequency of allele \( B_1 \). Preferential association of 2 loci, A and B, is then said to be present when an allele at locus A is found associated with a specific allele at locus B more frequently than expected under independence. The degree of deviation is a measure of the degree of preferential association.

These concepts are illustrated in Figure 1, panel A, which shows the distribution of 200 haplotypes sampled from a population under independence. In the sample, there were 140 haplotypes carrying allele A1 and 60 carrying allele A2 (allele frequencies 0.7 and 0.3, respectively) and 120 haplotypes carrying allele B1 and 80 carrying allele B2 (allele frequencies 0.6 and 0.4, respectively). Examination of the body of the table shows that, of the 60 haplotypes carrying A2, 36 were A2-B2, and none were A2-B1. Of the 140 haplotypes carrying A1, 120 (86%) were A1-B1, whereas 20 (14%) were A1-B2; this was very different from the expected 60:40 under independence. Clearly, in this population, allele A2 is
preferentially associated with B2, whereas allele A1 is preferentially associated with B1.

As a result of recombination, independent association is the situation to which a population tends to return over the long run, even if it departs from this for some reason (i.e., independence of alleles at linked loci is the equilibrium situation. For this reason, the distribution of haplotypes at a pair of linked loci under independence is a state of linkage equilibrium (LE).

For a pair of diallelic loci, the basic quantitative measure of LD (D) is simply the absolute deviation of the observed frequency of any of the haplotypes from its expected frequency (for a pair of diallelic markers, this is necessarily the same for all 4 possible haplotypes), D = pA1 B1 − pA1 × pB1. For the example of panel B of Figure 1, and taking the A1B1 haplotype for calculating purposes, D = 0.60 − 0.60 × 0.70 = 0.18. Without going into detail, it can easily be shown that the maximum value that D can attain for a given set of allele frequencies (Dmax) is equal to A2 × B2, where, among all 4 alleles of the 2 loci, pA2 = the allele with the lowest frequency and pB2 = the allele with second highest frequency. Note that Dmax depends on allele frequencies and cannot be large if any of the alleles are present at low frequency. This limits the usefulness of the D statistic as a general measure of LD.

Lewontin (1964) suggested adjusting for the effect of allele frequency by standardizing D against its maximum value, defining D′ = D/Dmax, whereas Hill and Robertson (1968) suggested using the square of the correlation between Ai and Bj, r^2 = D^2/(pA_i × (1 − pA_i) × pB_j × (1 − pB_j)), where pA_i = frequency of allele A_i. This is equivalent to the regression of the allelic state at locus B on the allelic state at locus A and has been found to be the more useful measure when dealing with LD involving QTG. Both of these measures have been extended to polyallelic loci: D′, as Hedrick’s polyallelic extension of D′ (Hedrick, 1987), and r^2, as Yamazaki’s standardized χ^2, χ^2 (Yamazaki, 1977). The latter, similar to r^2, has been found to closely track the regression of the allelic state at locus B on the allelic state at locus A.

Recombination and Map Distance

Because crossing over occurs more or less at random along the chromosome, it follows that the further apart 2 linked genes are, the greater the likelihood that a point of crossing over will be found between them, generating recombinant haplotypes. Thus, the recombination rate between 2 loci can serve as an approximate measure of the distance between them. (Because of “double crossing over” a strict proportionality between r and distance holds true only for short distances). Because the chromosome is a linear structure, knowing the approximate distances among a series of markers enables them to be arranged in a linear order, with distances measured in units of “proportion of recombination.” For convenience, a proportion of 0.01 recombinants among all meiotic products involving a pair of linked genes is defined as 1.0 cM. A typical chromosome will have a total length of 100 cM, or 1 M. At the physical level, 1 cM in a mammalian species corresponds to about 1,000,000 bp and about 500,000 bp in chickens, with great variation among different genomic regions.

Gene Maps and Genetic Markers

Once sufficient Mendelian genes distributed along the chromosomes of a species are known, they can be organized by analysis of proportions of recombination into a complete ordered map covering the entire genome (Weir, 1990). Any newly uncovered Mendelian gene can then be readily “mapped” by evaluating the proportion of recombination between the new gene and the existing genes. When used simply as place markers to locate new genes, the mapped genes making up the map are termed “genetic markers.” The concept of genetic markers has been extended beyond the class of functional Mendelian genes, and additional classes of markers, particularly those at the DNA level, have come into wide use and are described in detail previously in this paper. Using such markers, comprehensive dense genome maps have been constructed for all of the major agricultural and experimental species and, of course, for humans as well.

QTG Mapping

Using these genomic maps, programs are now underway in all major agricultural species to map and identify the QTG underlying genetic variation in quantitative traits (Andersson, 2001; Andersson and Georges, 2004). All QTG mapping designs are based on construction or collection of mapping populations that contain large IBD haplotype groups descended from a single individual in the immediate or more removed past. These large IBD groups generate LD among alleles at linked loci, thereby providing information as to the groups of loci that are located on the same chromosome. In addition, the mapping populations allow recombination in proportion to the distance among the linked loci, thereby providing information as to the order and spacing of these loci.

LD and QTG Mapping. To map QTG, a mapping population is constructed, using the IBD principle, such that marker and QTG alleles are in LD. That is, if marker A and QTG B are linked, then the mapping population is constructed so that haplotypes carrying marker allele A1, for instance, might preferentially carry the positive QTG allele, B1. Haplotypes carrying marker allele A2 preferentially carry the negative QTG allele B2, as in the example of panel B of Figure 1. As a result of this preferential association, individuals that carry marker allele A1 and its associated positive allele B1 will have, on average, a higher phenotypic value than individuals that carry marker allele A2 and its associated negative allele, B2. That is, preferential association of marker and QTG alleles will result in a marker-associated effect on the quantitative trait. This is detected by genotyping individuals of the mapping population for the marker and comparing phenotypic means according to marker genotype. When
this is done, A_1A_1 individuals are expected to show a higher trait value than A_2A_2 individuals, whereas A_1A_2 individuals should be intermediate (Falconer and Mackay, 1996). This is general: LD between a marker locus and a QTG always generates a marker-associated effect whose magnitude will depend on the degree of LD (and allele effect at the QTG). In contrast, markers that are not in linkage with QTG will not show such an association of trait value with marker genotype. Consequently, in an appropriately designed mapping population, finding a significant difference in quantitative value among genotype groups for a specific marker is evidence for a QTG in linkage to the marker. Because the map location of the marker is known, the approximate location of the QTG is also known (i.e., the QTG has been mapped to the chromosome level).

More specific information as to the location of the QTG is provided by the second aspect of mapping population construction – recombination. As noted, recombination reduces the degree of preferential association (LD) between markers and linked QTG, and, hence, the marker-associated effect. However, the closer the QTG to the marker, the less the effect of recombination in breaking up marker-QTG LD and the larger the marker-associated effect. Consequently, if associated phenotypic effects are obtained for a series of markers that are spaced fairly evenly along a chromosome, the 2 markers flanking the QTG are expected to exhibit the largest associated effects, because they will be the closest to the QTG. Thus, consideration of the magnitude of the marker-associated effects along the chromosome immediately places the QTG within the interval flanked by the markers showing strongest effects and, by interpolation, can even be used to estimate the location of the QTG within that interval.

If the major IBD groups making up the mapping population differ with respect to numerous QTG distributed throughout the genome and also with respect to a large number of markers spanning the entire genome, a test for associated effects can be applied to query each marker in turn with respect to QTG linkage. In this way, a complete genome scan for all QTG differentiating the IBD groups can be implemented in a single mapping population.

### QTG MAPPING DESIGNS AND STRATEGIES

Experimental designs to generate and analyze QTG mapping populations fall into 3 classes, depending on how the IBD subgroups that generate LD are produced: 1) admixture designs, in which IBD subgroups are produced by crossing populations that differ in their equilibrium marker-QTG associations; 2) family designs, in which IBD subgroups are produced by marker-QTG associations present in the haplotypes of the founder parents of the family; and 3) population designs, in which population-wide IBD subgroups are produced by the operation of population-genetic factors such as mutation, selection, and drift. These designs are described in turn.

<table>
<thead>
<tr>
<th>Alleles</th>
<th>A_1</th>
<th>A_2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_1</td>
<td>68 (0.34)</td>
<td>32 (0.16)</td>
<td>100 (0.50)</td>
</tr>
<tr>
<td>B_2</td>
<td>32 (0.16)</td>
<td>68 (0.34)</td>
<td>100 (0.50)</td>
</tr>
<tr>
<td>Total</td>
<td>100 (0.50)</td>
<td>100 (0.50)</td>
<td>200 (1.00)</td>
</tr>
</tbody>
</table>

**Figure 2.** Allele and haplotype frequencies for an admixture population.

### Admixture Designs

In admixture designs, a population with large IBD subgroups (and, hence, extensive LD among linked loci) is generated by crossing breeds or lines. Breeds that differ extensively in many characteristics or that have been separated from a common ancestral breed for many generations are likely to differ in frequency for many QTG and markers. Because haplotype frequencies under LE are determined by allele frequencies, this means that haplotype frequencies will differ between the 2 breeds. This will automatically generate LD when the breeds are crossed. For example, assuming 2 biallelic loci, with allele frequencies, p_A_1 = p_B_1 = 0.8 in breed 1 and p_A_1 = p_B_1 = 0.2 in breed 2, the equilibrium frequency of the A_1B_1 haplotype will be 0.64 in breed 1 but only 0.04 in breed 2. When the 2 breeds are crossed, haplotype frequencies in the F_1 will be the average of those in the 2 parent breeds and will differ from those expected under independence from the pooled allele frequencies. Consequently, much LD will be present. Figure 2 shows observed and expected haplotype frequencies for the above example, from which it can readily be calculated that D = 0.09, D_max = 0.25, D’ = 0.36, and r^2 = 0.13.

Although the F_1 population is in strong LD, it cannot be used as a mapping population, because there has not been any opportunity for recombination to break up the LD in proportion to the distance separating pairs of loci. Consequently, all pairs of loci in the F_1, whether closely linked or even on different chromosomes, will show high LD. To allow recombination to come into play, it is necessary to produce a second generation out of the F_1 individuals. This can be a backcross (BC) to one of the parental breeds or an intercross among the F_1 individuals, forming an F_2 population. For illustrative purposes, we analyze the simplest case, a BC involving parental breeds that are fixed for alternative alleles at a pair of loci. Application to the F_2 design, and to the more general case in which the breeds differ but are not at fixation for alternative alleles, will be considered later.

**BC Design, Parental Breeds Fixed for Alternative Alleles at Marker and QTG.** Let locus A be a genetic
Figure 3. Population designs for quantitative trait gene (QTG) mapping by admixture.
parameters. However, as noted above, this limitation is easily removed by considering most markers on the same chromosome. When this is done, least squares regression (Haley and Knott, 1992; Haley et al., 1994) or maximum likelihood (Lander and Botstein, 1989) “interval mapping” methods can be used to obtain “best fit” estimates of $r$ and $d + h$, in this way mapping the QTG to a specific location and estimating its effect ($d + h$). Note, however, that $d$ and $h$ remain confounded in this design.

**F2 Design.** Based on genotype at the marker, the F2 progeny can be differentiated into 3 marker genotype classes ($A_1A_1$, $A_1A_2/A_2A_1$, and $A_2A_2$) with differing frequencies of QTG alleles and, therefore, different mean phenotypes (Figure 3). Using procedures similar to those described for BC progeny, the difference of phenotypic means among alternate homozygotes at the marker is expected to be $2(1 - 2r)d$ (Falconer and Mackay, 1996). Thus, in the absence of dominance, marker-associated effects in the F2 will be twice as large as those observed in BC progeny, with a corresponding increase in statistical power. The F2 individuals that are heterozygous at the marker will deviate from the mean of the marker homozygotes by $(1 - 2r)^2h$ (Falconer and Mackay, 1996). Combined with interval mapping, this provides unbiased estimates of $d$, $h$, and $r$.

**Crossovers Among Breeds that are not at Fixation for Alternative Alleles.** The previous section assumed that the breeds that created the BC or F2 were fixed for alternative QTG alleles. Generally, this will not be the case for a breed cross. Breed-cross designs can, however, detect QTG that are at different frequencies, $q_1$ and $q_2$, in the parental breeds. In this case, the expected marker-associated effect is reduced by a factor $q_1 - q_2$ relative to that expected under complete fixation for alternative alleles. If the markers are not at fixation for alternative alleles, marker-associated effects will be reduced by a further factor of $m_1 - m_2$, where $m_1$ and $m_2$ = frequencies of the associated marker allele in lines 1 and 2 (Soller et al., 1976; Alfonso and Haley, 1998). Effects of marker nonfixation in the 2 lines can be reduced, however, by tracing allele origin from parent to progeny within pedigreed families of the BC or F2 (Beckmann and Soller, 1988).

Breed-cross QTG mapping populations are primarily designed to detect QTG that differ in frequency among the parental breeds used to produce the cross. Although power for the F2 analysis is markedly reduced when $q_1$ and $q_2$ do not differ greatly, F2 or BC generations typically have a family structure resulting from the hierarchical mating structure of F1 parents (e.g., an F2 population can be composed of a large number of full-sib families within a small number of paternal half-sib families). In this situation, breed-cross designs also allow family-based analysis of QTG (which will be described in the following section), allowing QTG that segregate within the 2 parental breeds to be detected (de Koning et al., 2001). Combined breed-cross and within-family analyses based on least squares (Kim et al., 2005) or Bayesian methods (Perez-Enciso and Varona, 2000) can provide greater power to detect QTG than breed-cross analysis alone and some power to differentiate among QTG that are fixed or segregating within the parental breeds.

In pedigreed breed-cross populations, allele origin can be determined by Mendelian analysis, supplemented by probabilistic considerations when breeds 1 and 2 differ in marker allele frequencies (de Koning et al., 2000; Tsukula-Haavisto et al., 2004). Consequently, breed-cross designs allow detection of imprinted QTG, because, with imprinting, progeny that are heterozygous for the QTG will differ in phenotype, depending on whether the $B_1$ or $B_2$ allele came from the sire or the dam (Figure 3).

**Advanced Intercross and Full-Sib Intercross Lines.** In F2 and BC designs, chromosomes have only undergone 1 round of recombination since the parental generation. As we have seen, this results in a marker-associated effect proportional to $1 - 2r$ (e.g., 80% of the marker-associated effect is retained for markers located 10 cM from the QTG). This is a useful characteristic of F2 and BC populations, in that it allows QTG to be detected with a limited number of markers across the genome; even at a marker spacing of 20 cM, the maximum distance of a QTG from a marker is 10 cM, and the average distance is 5 cM. However, due to sampling variation of the marker-associated effect, it also means that it is possible for a marker located 10 to 15 cM from a QTG to show a larger marker-associated effect than a marker located at the QTG. Consequently, unless mapping population size is very large, confidence intervals of QTG map location typically extend over 20 cM or more (Darvasi and Soller, 1997; Weller and Soller, 2004). To overcome this limitation, Darvasi and Soller (1995) proposed the use of advanced intercross lines (AIL), in which the F2 generation is further intercrossed for several generations. Each round of recombination increases the proportion of recombinant haplotypes at the expense of the parental haplotypes, such that when the single generation recombination rate, $r$, is small, the proportion of recombinant haplotypes in generation F2, is approximately equal to $r(t + 2)/2$, and expected marker-associated effects are now proportional to $1 - r(t + 2)/2$ instead of $1 - 2r$, as in the F2 or BC generation (Darvasi and Soller, 1995). This reduces the marker-associated effect at given marker-QTG distance. Hence, for QTG to be detected with given power in an AIL, the marker has to be much closer to it than for an F2 of equivalent size, and the expected difference in marker-associated effects for close markers will be large relative to sampling variation. Thus, with sufficient effective population size ($N_e > 100$) during AIL development to reduce sampling effects, confidence intervals for QTG position can be reduced 5-fold in the F10 (Darvasi and Soller, 1995).

A full-sib intercross line is a modification of the AIL design produced by mating a single male and a single female from 2 breeds (or from the same population) that segregate at the same marker and QTG alleles, followed by repeated intercrossing within the resulting full-sib population (Song et al., 1999). The resulting population consists of 4 subgroups of IBD haplotypes (2 from each of the parents) and their recombinants. As in BC and F2 populations, the large IBD subgroups generate a marked
degree of marker-QTG LD and consequent marker-associated effects in the population, allowing mapping of QTG with power almost equivalent to an F2 design. If the full-sib intercross line is continued to advanced generations, the repeated recombination events, as in an AIL, strongly reduce marker-associated effects among all markers but those very tightly linked to QTG; this allows the QTG to be more precisely mapped.

**Family-Based Designs.** Mapping QTG within pure lines or breeds requires identification of QTG that segregate within the breeds, in contrast to the QTG that differ among breeds that are detected by the admixture designs described above. The simplest of these family designs is the half-sib design, in which a single sire is mated at random to many dams, and the resulting half-sib progeny are phenotyped and genotyped for markers across the genome. This can be expanded to analysis of multiple half-sib families and complex pedigrees, as shown below.

**Single-Sire Half-Sib Family Designs.** The progeny, derived from a sire heterozygous for a marker (i.e., having marker genotype $A_1A_2$) can be separated into 2 groups: those that received the $A_1$ vs. the $A_2$ marker allele from the sire, and the average phenotype of the 2 groups can be compared. If the marker is also heterozygous for a linked QTG (e.g., the sire has haplotype structure $A_1B_1/A_2B_2$), the analysis proceeds exactly as in a BC design (Figure 3), except that the mean difference between the 2 progeny groups is $(1 - 2r)\alpha$, where $\alpha$ = the allele substitution effect for the QTG (Falconer and Mackay, 1996).

**Multiple-Sire Family Designs.** The single-sire family design is limited by the size of the family, which can be produced by a single sire and its mates and by the fact that a single-sire family can map only those QTG for which the sire is heterozygous. For a population at Hardy-Weinberg equilibrium, this would apply at most to half of the QTG. Thus, to map all or most of the QTG segregating in the population, it is necessary to test a number of sire families. When considering multiple sires from the same population, however, marker alleles and QTG alleles across families are expected to be in LE. This means that although there is strong LD within each family, across the population as a whole, marker-QTG LD is not present. Consequently, it is not possible to detect QTG in a mapping population consisting of a number of sire half-sib families by simply comparing the average marker-associated effect across the population as a whole, as was done for a single sire. Instead, QTG mapping is based on a hierarchical ANOVA (markers nested within sire families; Weller, 2001). Such an analysis asks whether the marker explains some part of the variation within half-sib groups and is expressed as a significant marker component of variance in the ANOVA, rather than as a difference in marker genotype mean value. Such analyses can be expanded to full-sib families and to simultaneous analysis of multiple markers, allowing for interval mapping of QTG (Knott et al., 1996).

**Analysis of Complex Pedigrees in Outbred Populations.** When a population consists of individuals that have more complex family relationships than the full or half-sib family designs considered above, it is not possible to analyze the data in terms of simple contrasts of groups sharing the same degree of within-group IBD. Fortunately, a methodology for dealing with this situation was developed by animal breeders in the context of estimating breeding values of individuals. This methodology is based on the fact that, because of their sharing of genetic material, each relative of an individual provides some information as to the breeding value of that individual (Lynch and Walsh, 1998). The amount of information provided stands in proportion to the fraction of the genome that is IBD between the individual and its relative; this can be estimated from the pedigree. Thus, in this methodology, termed an “animal model,” all individuals in the population and their ancestors are organized into a large relationship matrix (termed the A matrix), showing the IBD relationship of each individual to every other individual in the present population and its ancestral populations, going back many generations. All individuals taken together now comprise a set of simultaneous equations that can be expressed in matrix form and solved for the breeding value of each individual in the population, using methods developed specifically to handle tre mendous matrices of this sort. Although not primarily intended for this purpose, using such models, it is possible to estimate the overall variance in breeding values, which is equivalent to the genetic variance of the trait in the population (Lynch and Walsh, 1998).

This latter procedure is readily adapted to estimate the genetic variance of the trait that can be explained by a QTG linked to a marker (Fernando and Grossman, 1989; Lynch and Walsh, 1998). To achieve this, in addition to generating the general IBD matrix, pedigree information is combined with marker genotype information to assign to each individual a marker-based IBD relationship at each marker with each of the other individuals in the population (including ancestors). The set of equations is now solved to provide an estimate of the marker-associated effect for each individual in the population. The variance of these marker-associated effects is directly comparable to the marker variance component of the multiple-sire half-sib family analyses. If it is significantly different from 0, it indicates that the marker is associated with genetic variation in the trait and, hence, is in linkage to a QTG.

**Strategies to Increase Power or Reduce Genotyping Costs**

Several strategies have been proposed and used to increase power or reduce the cost of genotyping for the designs described above. The main ones are listed below.

**Replicated Progenies.** The accuracy of phenotypic information used for QTG mapping is increased by using the average performance of their progeny as the phenotypic data for the genotyped individuals making up the mapping population. For family designs (e.g., dairy cattle), this results in the so-called granddaughter design (Weller et al., 1990), in which marker genotypes of sons...
of a grandsire are associated with the mean phenotype of each son’s progeny (granddaughters of the sire whose alleles are evaluated). In admixture designs, a similar strategy can be used by progeny-testing F2 or BC individuals (Soller and Beckmann, 1990).

Multitrait and Multilocus Analyses. Power can also be increased by simultaneously analyzing data on several traits (Korol et al., 1995) and by simultaneously fitting more than 1 QTG. Such models also provide some power to differentiate between the presence of a single QTG with pleiotropic effect on multiple traits vs. the presence of separate linked nonpleiotropic QTG (Knott et al., 1998).

Selective Genotyping with or without DNA Pooling. Most power to map QTG for a given trait is provided by individuals that have extreme high or low phenotypes. Consequently, genotyping only individuals that are in the top and bottom fraction of the population phenotype distribution can result in a substantial reduction in genotyping costs with limited reductions in power (Darvasi and Soller, 1992). Genotyping costs can be further reduced by creating pools of DNA from individuals that have high vs. low phenotypes and determining marker allele frequencies by a quantitative genotyping method rather than genotyping each individual (Darvasi and Soller, 1994). These strategies can be applied to both admixture and family designs, noting that with family designs, selection and pooling must be on a within-family basis to allow for differences in QTG genotype and linkage phases.

Multiple-Stage Genotyping Strategies. Costs of genotyping can also be reduced by first genotyping only a sample of individuals or only a sample of markers to exclude regions of the genome that are unlikely to contain QTG, followed by genotyping additional individuals, markers, or both (Motro and Soller, 1992).

POPULATION DESIGNS: POPULATION-WIDE LD MAPPING

Drift, mutation, selection, and inbreeding can generate marker-QTG LD within a closed random mating population. In this case, marker-QTG LD is detected operationally as a significant difference in means among marker genotypes (as in admixture and single-family designs) that is taken across the population as a whole or as a marker-associated component of variance (as in multiple-family and complex pedigree designs) that is also taken across the population as a whole. The special feature of designs based on population-wide LD is the expected limitation of appreciable LD generated by population forces to very short distances, as will be shown below. In principle, therefore, population designs can narrow QTG location to small confidence intervals, without requiring very large sample sizes. A further attractive feature is the possibility of applying LD mapping within livestock or poultry populations as they stand, without need for special crosses or pedigree information. In addition, marker-QTG associations based on population-wide LD can be very simply and powerfully utilized for marker-assisted selection (Dekkers, 2004). The corollary, however, is that for markers to show significant associated effects due to population-wide marker-QTG LD, the markers must be very close to the QTG. This can be achieved by saturating a chromosomal region with markers, using recently available high-density SNP maps, or by using SNP in genes that are hypothesized to contain QTG, the so-called candidate gene approach (Rothschild and Soller, 1997).

Factors Producing Population-Wide LD

Two population-genetic factors, genetic drift and mutation, act to produce LD among linked loci within a closed population. Their effectiveness is increased by selection and inbreeding. At the same time, recombination acts to break down LD as it is generated. Thus, the observed LD among linked loci depends on the balance between these opposing forces.

Genetic Drift. In a random mating population, chance sampling at various stages of the reproduction cycle cause haplotype frequencies in the progeny generation to differ from those of the parent generation. Thus, even if at some time in the past a population was at LE, these random changes can accumulate over the generations to produce large deviations from equilibrium. This process is termed “genetic drift.” Because drift is a sampling process, its magnitude is inversely proportional to Ne (Falconer and Mackay, 1996). Inbreeding and selection, by limiting the number of parents that participate in producing the next generation, decrease Ne and increase the effects of drift. In addition, at chromosomal regions under selection, an additional local “inbreeding” effect (a so-called “selective sweep”) occurs, further increasing the effectiveness of drift in generating LD in the region (Andersson and Georges, 2004).

Mutation. Consider a population that is segregating for alleles A1 and A2 at marker A but that is fixed for allele B2 at a linked QTG B. Only 2 marker-QTG haplotypes are present in the population, A1B2 and A2B2. Now, assume a 1-time novel mutation of QTG allele B2 to allele B1 taking place in an A1B2 haplotype, converting it to A1B1. As a result of drift or selection, the new haplotype A1B1 can increase in frequency while still avoiding recombination if A and B are very tightly linked. In this case, the reciprocal haplotype A2B1 will be very rare or even absent from the population. Thus, a situation of preferential association of marker allele A1 with QTG allele B1 is found in the population. Because marker allele A1 is found with both QTG alleles, however, preferential association is not complete, and LD will be <1.0. If selection for the new haplotype takes place, this will cause it to increase rapidly in frequency, with only rare recombination events generating the reciprocal haplotype. In this case, preferential association can be strong and LD can be high.

Recombination. Acting against LD generated by drift, mutation, and selection, we find recombination, which tends to break up LD as it is formed (Terwilliger et al., 1998). Thus, considering creation of LD by drift alone,
the amount of LD found between a pair of loci will be primarily determined by the balance of the strength of drift in generating LD, which is inversely proportional to effective population size, and the strength of recombination in breaking up LD, which is directly proportional to the distance between the loci. The approximate LD at equilibrium, using the $r^2$ measure of LD, is given by the expression (Sved, 1971):

$$r^2 = 1/(1 + 4N_e c),$$

where $c$ is measured in morgans. For example, if $N_e = 50$ and $c = 0.10, 0.05, 0.01,$ and $0.0025$ M, respectively, expected LD at drift or recombination equilibrium $= 0.04, 0.09, 0.33,$ and $0.67$ (for distances $<0.0025$ M, the Sved (1971) equation may not hold due to the importance of mutation-generated LD at these distances.). Thus, appreciable marker-QTG LD and, hence, appreciable population-wide marker-QTG associations are expected only for loci separated by distances of $0.01$ M (1 cM) or less. Consequently, searching for LD between markers and QTL requires closely spaced markers but could also provide very high mapping resolution.

**LD in Livestock and Poultry Populations**

Because effective population sizes of poultry and livestock populations are small, they are expected to be strongly affected by drift (Terwilliger et al., 1998). In addition, livestock and poultry populations have been subject to other forces affecting LD, such as selection and mutation, hybridization and admixture episodes, founder effects (when a breed is based on a limited number of selected individuals), bottlenecks (when a population goes through an episode when numbers are drastically reduced, markedly increasing drift), and family stratification (due to the presence of large IBD cohorts derived from influential animals). Thus, LD in these populations may extend over regions of appreciable size, greater than expected for drift alone, enabling LD mapping with much sparser marker density than required for most human populations (Terwilliger et al., 1998). To obtain an estimate of the actual marker density needed for effective QTL mapping, there is a great need for data-based estimates of the degree of marker-QTG LD in populations of interest. Because, with rare exceptions, QTL cannot be genotyped directly, marker-to-marker LD has been taken as a surrogate measure. Recently, a number of studies have reported extensive LD among microsatellite markers in dairy cattle (Farin et al., 2000), sheep (McRae et al., 2002), swine (Nsengimana et al., 2004), and poultry (Heifetz et al., 2005). These studies have indeed concluded that LD is much more extensive than in human populations, extending over regions of 5 cM or more. However, all these studies used microsatellite marker data and, except for Heifetz et al. (2005), used Hedrick’s (1987) multiallelic extension of $D'$ to assess the extent of LD. Zhao et al. (2005), however, showed that this measure greatly overestimates the magnitude of the LD of multiallelic markers with (presumed) biallelic QTL because of the possibility of low or 0 frequencies of some haplotypes, which strongly inflates $D'$. Using Yamazaki’s (1977) standardized $\chi^2$ statistic, $\chi^2$, which does not suffer from this limitation, LD in livestock and poultry populations, as estimated from actual data or by simulation for markers separated by a little as 1 cM, appears to follow expectations based on what would be generated by drift on the Sved (1971) equation (Sved, 1971; Heifetz et al., 2005; Zhao et al., 2005; Lipkin et al., 2006).

**Statistical Methods for LD Mapping**

Statistical methods used for LD mapping of QTL in livestock can roughly be classified into genotype- and haplotype-based methods, as reviewed by Dekkers et al. (2006). Genotype-based methods do not require knowledge of marker haplotypes and are therefore easier to implement. In the following sections, these methods will be further described. Although in the literature, LD mapping methods have primarily been developed for fine-mapping QTL within a previously identified QTL region (e.g., based on admixture of family-based mapping), what follows also applies to analysis of candidate gene regions and for implementation of a genome scan for QTL based on genome-wide high-density SNP data.

**Genotype-Based LD Mapping.** Linkage disequilibrium mapping, using genotype at single markers, is based on an association test between marker genotype and trait value by fitting marker genotype as an effect in the model of analysis. Consider first a diallelic marker in complete LD with the QTL. Because all 3 marker genotypes are present in the population, this is basically the same as an F2 mapping population, except that homozygous genotypes will not be present in equal numbers, somewhat reducing the power of the test. If the individuals used for analysis are related, the association analysis can be conducted by adding marker genotype as a fixed effect to an animal model analysis (Short et al., 1997; Israel and Weller, 1998). If multiple alleles are present at the marker, the effect of the various alleles vs. the most common allele can be included as fixed effects (Batra et al., 1989). Because additional df are used in this more complex model, power is reduced even further compared with an F2 design.

For QTL mapping in general, power at a given sample size will stand in proportion to the square of the marker-associated effect (Song et al., 1999). Using a measure of LD such as $r^2$ or $\chi^2$ that reflects the regression of QTL allele status on marker allele status, the expected magnitude of the marker-associated effect due to LD is equal to $LD(2\alpha)$, where $2\alpha_1$ = the quantitative difference among alternative genotypes at the QTL. Power will, therefore, be proportional to $LD^2$. Thus, the sample size required for LD mapping by association test stands in direct proportion to $N/\text{LD}^2$, where $N$ = the sample size required to distinguish allelic effects at the QTL itself (i.e., under complete LD). Thus, the most important factor for designing LD mapping experiments is the degree of LD expected
between marker and QTG, as a function of distance between them, as measured in morgans.

As shown above, at a marker-QTG spacing of 1 cM and \( N_e = 50 \), the expected value of LD on the Sved (1971) equation is 0.33. Thus, at this marker-QTG spacing, a mapping population size \( 9 \times t \) that required under complete LD would be needed for equivalent power to an \( F_2 \) design. However, it should be noted that an average marker-QTG spacing of 1 cM would be provided by a marker-to-marker spacing of 4 cM, which is sparse in the context of the spacing thought acceptable for LD mapping. Using a marker spacing of 1 cM, average distance of marker and QTG would be 0.25 cM, and expected LD is 0.67. This would provide very useful power in populations only 2.5-fold greater than would be required under complete LD. However, the extent to which the Sved (1971) approximation holds for such short distances has not been investigated experimentally or by simulation.

The single-marker methods can be easily extended to multiple markers by simultaneously fitting genotype at each marker within a window of multiple markers and sliding this window across the chromosome. Similar models have also been fitted using maximum likelihood for a (presumed) biallelic QTG, although approximations are needed to facilitate computation when fitting multiple markers (Farnir et al., 2002; Abdallah et al., 2004).

**Haplotype-Based LD Mapping.** As shown by simulation and experiments, LD among loci separated by the same distance can vary widely. The high variance of the sampling distribution of LD, when based on a single pair of loci, derives from the degree to which the marker genotype groups represent groups which are IBD with respect to the linked QTG. Clearly, when using a single marker only, this would be affected by many factors, and, hence, the effective LD will vary greatly for marker-QTG pairs separated by the same distance. Thus, in some cases, analysis of a population according to marker haplotypes may provide stronger differentiation into IBD groups with respect to the QTG and, hence, more powerful LD mapping.

Implementation of a multilocus haplotype-based approach to LD mapping, raises 2 practical problems: 1) how many markers to include in the haplotype, and 2) how to implement the association test, which now potentially involves many haplotypes, depending on the number and spacing of the markers included in the haplotype. Regarding the latter, various statistical models have been used, including using least squares (Long and Langley, 1999; Grapes et al., 2004), in which phenotype is regressed on the number of copies of each possible haplotype, Bayesian methods (Pérez-Enciso, 2003) to model the joint distribution for multiple-marker haplotypes, and IBD mixed linear models (Meuwissen and Goddard, 2000). For the latter, Meuwissen and Goddard (2000) investigated by simulation (using the gene-drop method), and later deterministically (Meuwissen and Goddard, 2001), the ability of shared marker haplotypes centered on a marker interval in a chromosomal region to determine the IBD relationship of individuals in a population with respect to a QTG in the haplotype interval as a function of founder population structure and the subsequent interaction of mutation, drift, and recombination over a large number (100) of generations. They found that, for a pair of individuals, the degree of sharing of marker haplotypes flanking a marker interval provided a useful guide to the likelihood that the individuals are IBD for a specific locus within the haplotype. The IBD probabilities were then used as a locus-specific relationship matrix to set up a variance-covariance matrix for the QTG. Similar to what was already described under analysis of complex pedigrees, this IBD matrix can be used to set up a system of equations for each individual in terms of its IBD with respect to all other individuals in the population and solve this for the associated effect at a given locus interval. By proceeding interval by interval along a chromosomal region, the marker interval with the highest interval-associated effect can be identified as the interval that contains a QTG.

In a sire half-sib family mapping design, the chromosomes of the progeny that derive from the sire can be analyzed for linkage mapping, using the standard single-marker interval mapping approach, whereas the chromosomes of the progeny that derive from the dams are more representative of the population as a whole and can be analyzed for LD mapping, using the above IBD haplotype-interval mapping approach. Thus, it is possible to implement a joint linkage and LD mapping analysis, utilizing both sources of information (Meuwissen et al., 2002).

**Genotype- and Haplotype-Based LD Mapping Compared.** When single-marker and haplotype analysis are compared by simulation for LD mapping across relatively small, densely marked regions, the best single marker in the group making up the haplotype does not perform appreciably less well than the haplotype as a whole when LD is generated by drift alone (Long and Langley, 1999; Fan and Xiong, 2002; Nielsen et al., 2004; Zhao et al., 2006). The lack of benefit from haplotype analysis may indicate that there is not a unique haplotype that is associated with the QTG allele, a situation that may be caused by the rather random pattern of LD that is created by drift (Dekkers et al., 2006). Alternatively, when the number of haplotypes segregating in a population is small, single markers will be in high LD with other markers making up the haplotype so that they convey most of the haplotype information. The high utility of single markers for LD mapping, if confirmed by experiment, has important implications, because association tests based on single markers can be implemented at very high density and very low cost through selective DNA pooling. Potentially, this could be very cost effective for LD mapping, particularly if combined with the new fractioned pool design developed by Korol and coworkers (Frenkel et al., 2005; Cohen et al., 2006) that permits application of many statistical procedures to pool analysis that were previously limited to individual genotyping. In analyses of actual data (Farnir et al., 2002; Meuwissen et al., 2002; Blott et al., 2003), haplotype analysis worked well in practice but was
not compared with single-marker LD analysis of the same data set. Thus, these critical comparisons remain to be performed on actual data.

**CONCLUSIONS**

As reviewed in this paper, laboratory techniques are currently available to provide detailed assessment of the molecular genetic variation in the structure of the avian genome. Especially for the chicken, for which the genome has been sequenced, molecular markers are abundant. The application of molecular data can enhance free-living populations by management decisions based on a more accurate understanding of their biodiversity and can make directed improvement in farmed populations by incorporating genomic marker information into breeding programs. Ultimately, the full utilization of marker-based breeding decisions will depend upon a thorough and accurate definition of the relationships of the molecular genetic markers and biological performance through appropriately designed experimental studies. Although there is much active research in this area, the complex molecular genetic architecture of most important traits in poultry remains to be elucidated.

**REFERENCES**


Lewontin, R. C. 1964. The interaction of selection and linkage.
Knott, S. A., L. Marklund, C. S. Haley, K. Andersson, W. Davies,
Kim, J. J., H. Zhao, H. Thomsen, M. F. Rothschild, and J. C.
Hillel, J., M. A. Groenen, M. Tixier-Boichard, A. B. Korol, L.
Hillel, J., Z. Granevitze, T. Twito, D. Ben-Avraham, S. Blum, U.
Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in
Helfer, C. S., J. E. Fulton, N. O’Sullivan, H. Zhao, J. C. Dekkers,
Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in
Hedrick, J. L., Z. Gravenitzte, T. Twito, D. Ben-Avraham, S. Blum,
Hillel, J., M. A. Groenen, M. Tixier-Boichard, A. B. Korol, L.
David, V. M. Kirschner, T. Burke, A. Barre-Dirie, R. P. Croojmans,


