Journal of Biochemistry and Molecular Biology, Vol. 37, No. 1, January 2004, pp. 11-27

JOURNAL OF Biochemistry and Molecular Biology

Review

Global Genetic Analysis

Elahe Elahi[†], Jochen Kumm[‡] and Mostafa Ronaghi[‡,§,*]

[†]Faculty of Science, Tehran University, Tehran, Iran

[‡]Stanford Genome Technology Center, Stanford University, Palo Alto, California 94304, USA

[§]Institute for Biophysics and Biochemistry, Tehran University, Tehran, Iran

Received 12 January 2004

The introduction of molecular markers in genetic analysis has revolutionized medicine. These molecular markers are genetic variations associated with a predisposition to common diseases and individual variations in drug responses. Identification and genotyping a vast number of genetic polymorphisms in large populations are increasingly important for disease gene identification, pharmacogenetics and population-based studies. Among variations being analyzed, single nucleotide polymorphisms seem to be most useful in large-scale genetic analysis. This review discusses approaches for genetic analysis, use of different markers, and emerging technologies for large-scale genetic analysis where millions of genotyping need to be performed.

Keywords: Disease gene identification, Genetic variation, Large-scale genotyping technology, Single nucleotide polymorphism

Introduction

"Connecting phenotype with genotype is the fundamental aim of genetics" (Botstein and Risch, 2003). This connection in model organisms has most often been made by traditional genetic mapping that relies on recombination analysis. Recombination analysis is based on the premise that the closer two locations are on the genome, the less likely a recombination event will occur between them during a meiosis. Therefore, a direct relationship between distance on the genome and recombination frequency is presumed. This approach until recently has been of very limited use in human studies.

The tide of investigations in humans changed when in 1980 it was proposed that DNA sequences which show variations among individuals could per se serve as reference points ("markers") with respect to which genes determining specific traits can be mapped (Botstein *et al.*, 1980). At the time the proposal was made, only a few sites having variable DNA sequences were known. Since then, the traits most often investigated have been diseases. About 1,200 genes causing human diseases or traits have been identified (Botstein and Risch, 2003). Pursuant to gene identification, analysis of protein products and disease causing mutations creates venues which lead to better understanding of the biochemical and physiological basis of the disease phenotype. Improved methods for diagnosis and treatment may also become available.

During the past two decades, four approaches have dominated genetic analysis of human disease traits: linkage analysis, allele sharing paradigms, association studies and model organism-based studies. Animal studies, in addition to having sometimes suggested candidate genes or loci for human traits, have been of utmost importance in confirming genotype-phenotype relations found in human studies and in analysis of functional pathways by which genotypes determine phenotypes (Kash *et al*., 1997). Nevertheless, this article will emphasize only the first three approaches and animal studies will not be further discussed.

The immediate objective in linkage analysis and allele sharing paradigms is to determine maximum co-inheritance of disease phenotype with a specific locus defined by a marker. In association studies, it is to determine maximum co-inheritance with specific alleles of markers whose positions are known. The two types of studies merge in certain situations. With good experimental design, it is expected that the co-inheritance is due to physical proximity between the trait locus and the marker locus, which should minimize separation due to recombination. Previous knowledge of the position of the marker locus then considerably confines the potential position of the trait locus.

The effectiveness of the approaches is strongly affected by the validity of the assumption that the disease causing

*To whom correspondence should be addressed.
Tel: 650-812-1971; Fax: 650-812-1975
E-mail: mostafa@stanford.edu

variation in the genome arose only once in the history of the pedigree or population under study. It is evident from what has been said that prerequisites for such gene hunting studies include "good" markers and "good" families or populations. Additionally, in order to reap maximum information from and make valid conclusions based on obtained data, fairly sophisticated mathematical and statistical analysis and algorithms are needed.

Markers for Genotyping

Any Mendelian character can in principle be used as a genetic marker. The most important property of a good marker is that it be sufficiently polymorphic such that a randomly chosen individual will be heterozygous for it. This feature is quantitatively assessed by the "mean heterozygosity" or "polymorphism information content" parameters assigned to each marker (Botstein *et al.*, 1980). It is an added advantage if the marker alleles are easily and inexpensively distinguishable. For any gene hunting analysis, many of these need to be spread out through the genome.

Until the 1970s, blood group variants, electrophoretic variants of serum proteins and variations in HLA types constituted the available markers for human genetic analysis. All these had severe limitations (Strachan and Read, 1999). Since the discovery of restriction fragment length polymorphisms (RFLPs) in the human genome, DNA sequences which are variable and generally not associated with known functional variations became used almost exclusively as genetic markers. More than anything else, the availability of DNA sequence markers as reference points in the genome has been responsible for the notable progress in human gene mapping during the past two decades.

RFLPs are restriction enzyme digests of the genome or PCR amplified portions of the genome which are variable in size due to sequence differences which destroy enzyme recognition sites in some individuals. Shortly after the existence of RFLPs in the human genome was recognized, several diseases displaying simple Mendelian inheritance with unknown molecular etiology were localized to limited chromosomal regions by demonstrating close linkage to RFLP loci. These include the autosomal dominant Huntington disease and polycystic kidney disease, and autosomal recessive cystic fibrosis and X-linked Duchenne muscular dystrophy (Gusella *et al.*, 1983; Reeder *et al.*, 1985; Koenig *et al.*, 1987; Kerem *et al.*, 1989). But RFLPs suffer from the major handicap of limited polymorphism; each has only two alleles which allows a maximum probability of heterozygosity equal to 0.5. More polymorphic marker such as variable number tandem repeats (VNTRs) whose notable polymorphism is due to variability in the number of tandem repeats of unit sequences each of less than 100 nucleotides, have setbacks associated with difficulties in detection and uneven distribution in the genome (Jeffreys *et al.*, 1985).

The remaining two classes of DNA sequence markers, microsatellites and single nucleotide polymorphisms (SNPs) are very important. Microsatellites have been the workhorse of human genetic analysis since the late 1980s (Weber and May, 1989). Their polymorphism is due to variations in the number of tandem repeats of short sequence units typically ranging from two to four nucleotides in size. Many have 5-10 alleles and heterozygosity levels of 0.75 or greater. About 10,000 microsatellites have been identified (http://www.cephb.fr/cephdb) and they are spread throughout the genome with normal spacing approaching 5-10 cM. Their alleles are easily and rapidly distinguished on the basis of variations in electrophoretic movement of fluorescent labeled PCR products amplified with primers complementary to conserved sequences flanking the repeats.

SNPs came into use a decade after microsatellites (Sachidanandam *et al.*, 2001; Venter *et al.*, 2001). Each represents a nucleotide variation at a single nucleotide site in the genome and most are bialleleic. A site is generally considered a SNP if the minor allele frequency reaches 1% in a population (Botstein and Risch, 2003). The major advantage of SNPs is their abundance, theoretically allowing detection of tighter linkages and associations. Millions of SNPs have already been identified, corresponding to a frequency of about 1/1,000 bp. Results of fairly comprehensive analysis of short regions of the genome predicted that their frequency may approach 1/200 bp. (Patil *et al.*, 2001). Although found throughout the genome, as expected, they are more abundant in non-coding sequences. In addition to frequency, SNPs have the benefit of being more stable and easily amenable to automation for assessment in large scale experiments (Wang *et al.*, 1998; Goddard *et al.*, 2000) and the price is expected to decrease as the scale of the experiment increases, as experienced with microsatellite genotyping in the last decade.

The first generation tools widely in use for SNP analysis include TaqMan probes, Pyrosequencing, single-base extension and mass spectrometry (Ronaghi *et al.*, 1998; De La Vega *et al.*, 2002; Jurinke *et al.*, 2002; Marnellos *et al.*, 2003). To reduce the cost, strategies using pooled DNA samples in the above-mentioned techniques have also been explored (Bansal *et al.*, 2002; Mohlke *et al.*, 2002; Yang *et al.*, 2003). However, the sheer magnitude of additional experimental data to be generated and analyzed accurately for both rare and common polymorphisms has ignited development of new generation of tools and resources for pharmacogenomic and genetic studies, including studies that are focused on candidate genes, candidate regions, and associations throughout the whole genome. The common attribute for these new tools are high level of multiplexing which will be discussed in this review.

The notable drawback of SNPs as markers, as with RFLPs, is their limited polymorphism. Most have a maximum theoretic heterozygosity frequency of 0.5 which is rarely achieved. To circumvent this limitation, the notion of using haplotype blocks has been given considerable attention (Daly

et al., 2001; Patil *et al.*, 2001; Gabriel *et al.*, 2002). Each block is defined by a set of alleles at neighboring SNPs rarely separated by recombination because of proximity. The extent of polymorphism in such haplotypes is expected to be considerable and partly depends on the number of SNP loci each encompasses. SNP analysis has been beneficial in the search for loci of several human traits (Kim *et al.*, 2003; Shin *et al.*, 2003) and it is likely that it will be the marker of choice for genetic analyses.

Knowledge of proximity of a disease locus to a marker locus provides guidance to the physical location of the disease locus only if the physical locus of the marker is known. As one of the preliminary goals of the Human Genome Project, high density framework maps of thousands of highly polymorphic microsatellite markers were developed (Collins *et al.*, 1996; Broman *et al.*, 1998). These maps have been indispensable for disease gene localization. Genetic distances between markers were established by recombination analysis in the CEPH collection of multi-progeny families (Dausset *et al.*, 1990). Distances between markers on the map were calculated using map functions which quantitatively describe a relation between map distance (x in cM) and probability of recombination (θ) (Broman and Weber, 1998; Ott, 1999). Morgans map function ($x = \theta$) is the simplest, but warranted only for very closely linked loci ($\theta < 0.01$) because it assumes complete interference Haldanes function [$x = -1/2 \ln(1 - 2\theta)$] is based on the assumption of zero interference. The degree of interference in different organisms and probably different regions of any one genome differs, being relatively strong for example in mice. Kosambis map function ($x = 1/4 \ln[(1 + 2\theta)/(1 - 2\theta)]$) which assumes an interference of value of 0.5 seems most appropriate for distances greater than 5 cM.

The chromosomal locations of markers have been determined using FISH or radiation hybrid mapping, thus relating genetic maps to physical maps. The exact physical position of many markers is now becoming known as the sequencing of the human genome approaches completion. SNP localization has been the product of many joint efforts; the search for SNPs continues (Sachidanandam *et al.*, 2001; Venter *et al.*, 2001; Stanssens *et al.*, 2004). It is interesting that correspondence between genetic and physical maps has shown that the frequency of recombination in male meiosis is significantly less than in female meiosis, in accordance with early observations of lower chiasma counts in male meiosis (Morton *et al.*, 1982; Collins *et al.*, 1996). Thus 1 male cM averages 1.05 Mb, whereas 1 female cM averages 0.70 Mb. The approximation of 1 cM = 1 Mb is commonly used. It is important with respect to gene mapping that recombination frequencies in humans also differ in different chromosome regions (http://research.marshfieldclinic.org/genetics), some regions being hotspots of recombination and others relatively free of recombination (Goddard *et al.*, 2000; Reich *et al.*, 2002; Clark *et al.*, 2003).

Linkage Analysis

In linkage analysis, the locus associated with a genetic disease is sought in one or more family pedigrees in which the disease phenotype segregates. Straightforward linkage analysis is parametric, meaning that a model is proposed to explain the inheritance pattern observed in a pedigree and the model is then tested. A feature of the model is mode of inheritance, the possibilities being dominant, co-dominant, recessive and X-linked. Additionally, penetrance and phenocopy effects are also designated. The greater the deviation from complete penetrance is, and the larger the number of phenocopies is, the less likely a linkage analysis is fruitful. Because disease loci will be sought with reference to co-inheritance with marker loci, disease and marker allele frequencies need to be provided. If more than one marker is being used as is generally the case, genetic distances between markers must also be known. As in all genetic analysis, accurate assessment of the phenotype is imperative.

The CFTR (cystic fibrosis transmembrane regulator) gene responsible for cystic fibrosis was the first disease gene identified by this protocol (Kerem *et al.*, 1989). A gene associated with lactose intolerance is a more recent example (Enattah *et al.*, 2002). Linkage analysis is most appropriate and has been most productive in finding loci associated with diseases showing clear Mendelian inheritance, as it is relatively easy to propose and test inheritance models for these. In fact, the vast majority of the genes that have been positionally cloned were originally mapped in families showing straightforward Mendelian inheritance (Botstein and Risch, 2003).

An assumption, very likely to be valid within any single pedigree is that all affected individuals have inherited the disease causing allele (dominant traits) or alleles (recessive traits) from, respectively, one or two common ancestors. The endeavor then reduces to testing whether the alleles of the marker are segregating among affected and unaffected individuals of the pedigree in accordance with the Mendelian pattern of inheritance proposed (two-point analysis). Alternatively, a marker is sought whose alleles behave in this manner better than the alleles of the many other markers tested. The locus of the best marker provides a first approximation to the locus of the gene. If only one gene is responsible for the disease phenotype, the same marker locus should be identified in analysis of different pedigrees, but the segregating alleles may be different.

Statistical considerations come into play at several stages of linkage analysis. A necessary feature of pedigrees used in such analyses is that they contain “informative meioses.” The greater the number of informative meioses in a pedigree is, the more useful that pedigree will be for the analysis. Even if two loci are on different chromosomes, an allele of one will co-segregate with an allele of the other with a probability of 0.5.

The greater the amount of co-segregation is observed, that is the less frequent "recombination" of alleles is among the progeny, the greater the probability of physical linkage within a distance of less than 50 cM is. The accuracy of determination of probability of co-segregation partly depends on the number of meiosis (number of individuals) investigated, therefore the value of large pedigrees.

Furthermore, a segregation event contributes to the assessment of co-segregation only if it is telling with regards to whether the gamete is recombinant or not. Herein lays the value of high polymorphism for genetic markers used in linkage analysis. A meiosis event is always informative if both parents are heterozygous for different alleles of the marker, informative with a probability of 0.5 based on the condition that both parents are heterozygous for the same alleles and never informative if the parent transmitting the disease allele is homozygous. Clearly, the more polymorphic a marker is, and also the more nearly equal the frequency of the different alleles is, the more likely that a meiosis will be informative for that marker is.

At the stage of data analysis, unequivocal classification of progeny as recombinant or non-recombinant is often not possible. Morton in 1955 proposed that the best statistic to assess possible linkage relates to the ratio of the probability of occurrence of observed data in the framework of the proposed model under the assumption of linkage, to the probability of the same under the assumption of no linkage (Morton, 1955). The ratio of these two is essentially the odds of linkage and it is generally presented as the logarithm of odds or the LOD score (Ott, 1999). The probability of linkage is assessed at various recombination fractions ($0 < \theta < 0.5$) and no linkage means a genetic distance of more than 50 cM or presence on different chromosomes. Therefore, the product of linkage analysis is a table or graph of LOD scores at various recombination fractions.

For the relation between each marker locus and disease locus, one is in search of the è which gives a maximum LOD score and when many markers are tested, one wants to identify the marker which results in a maximum LOD score. A maximum LOD score is derived for each pedigree and the LOD scores of different pedigrees in the same analysis are addable. The credibility of the linkage suggested by the maximum LOD score is evaluated in terms of how it compares to a threshold level. Mortons proposal remains the basis of most commonly used linkage analyses today.

The LOD score threshold level for acceptance is a murky issue. The most striking demonstrations of the importance of giving careful consideration to this criterion come from linkage analyses of psychiatric disorders, whose reported linkages have failed to be confirmed (Sherrington *et al.*, 1988). The traditional threshold for acceptance of linkage in human studies has been a LOD score of 3, corresponding to odds of 1,000 : 1 in favor of linkage. This value is not as stringent as it appears at first sight, as it derives from Bayesian calculations which take into account the inherent improbability (1/50) that two loci in the human genome chosen at random should be linked (Ott, 1999). The LOD 3 value in effect corresponds to the standard probability of 5% false positive under condition of no linkage.

The standard LOD score for rejection of linkage is −2. The LOD 3 value is acceptable for a single interrogation of linkage to one marker. For genome-wide linkage analysis of Mendelian traits (or analysis of limited regions of the genome defined by the results of other investigators) in which many marker loci are investigated, a threshold of 3.3 is more reasonable (Weeks *et al.*, 1990). As increasing numbers of models of inheritance are tested, the threshold needs to be more stringent yet. The guideline remains keeping the probability of a false positive result at the 5% level. Simulation approaches randomly generated is considered a reliable, though not necessarily facile approach for arriving at false positive rates (Royer-Pokora *et al.*, 1986; Terwillinger and Ott, 1992).

In general, a LOD score of less than 5 should be considered tentative (Strachan and Read, 1999). Confidence in a linkage can best be achieved if linkage is confirmed in completely independent studies using a new set of samples (Botstein and Risch, 2003). Once linkage is accepted, the interval on each side of the locus of maximum LOD score delineated by one unit (LOD-1) is considered the "confidence interval" of linkage and pursued in further studies (Ott, 1999).

If linkage analysis is to be used in gene hunting, it should usually be considered a first step (Botstein and Risch, 2003). The usual approach for disease/marker mapping starts with a framework map in which markers are spaced 5-30 cM apart. After detection of initial linkage, the same samples are re-analyzed using additional markers more densely localized in the region initially identified. But the very nature of linkage mapping limits the resolution to approximately 1-10 cM. The limitation is determined by the reality that once a marker is identified with respect to which no recombinants are identified ($\theta = 0$), there is no way to get closer to the disease locus. The reason that recombinants with respect to that marker can not be found is that Mendelian diseases are generally rare and a linkage study is unlikely to have more than a few hundred meiosis events available for investigation.

The 1-10 cM interval potentially identified is usually too large to make an immediate search for disease mutations a reasonable endeavor. Unless extremely lucky (as in the cystic fibrosis case), one must resort to other options for more tightly delineating the disease locus. Some of the genes identified in the 1980s with the help of linkage analysis also used other information. The search for genes of chronic granulomatous disease and X-linked muscular dystrophy made use of chromosomal aberrations (Royer-Pokora *et al.*, 1986; Koenig *et al.*, 1987). Insight into reasonable candidate genes was immensely helpful in finding genes for rare Mendelian forms of blood pressure diseases (Lifton *et al.*, 2001). More recently, a taste receptor gene within a chromosomal region linked to taste sensitivity to phenylthiocarbamide (PTC) was followed

up as a reasonable candidate gene and finally shown to be responsible for the quantitative trait (Kim *et al*., 2003).

LOD scores for the simplest pedigrees can be calculated manually, but the work very soon becomes extremely tedious. In addition to known genotypes at loci investigated, all possible genotypes of individuals with unknown genotypes must be taken into account weighed by the probability of the possible genotypes. The probabilities will depend upon observed genotypes of relatives and allele frequencies in the population, all in accordance with inheritance by Mendelian rules.

The calculations are routinely done on computers. The two computer programs most often used are LINKAGE and GENEHUNTER. LINKAGE is based on the Elston-Stewart algorithm and has the advantage of being able to handle very large pedigrees (Elston and Stewart, 1971; Lanthrop *et al*., 1984). It is limited by the number of loci it can evaluate because its computation time increases exponentially with the number of possible haplotypes, in turn dependent on the numbers of alleles and marker loci. Analysis on GENEHUNTER is based on the Lander-Green algorithm (Lander and Green, 1987; Kruglyak *et al*., 1996; Kong and Cox, 1997). This algorithm can cope with many loci, but its computation time increases exponentially with the size of the pedigree. The GENEHUNTER program is appropriate for whole genome scans of moderate sized pedigrees, for example not exceeding 16 individuals.

Homozygosity Mapping

Approximately 1400 recessive diseases are recorded in the MIM database, but it is likely that many more in fact exist (Botstein and Risch, 2003). Although the gene for some of the common recessive diseases has been identified, straightforward linkage analysis is generally not suitable for investigation of the rare recessive diseases simply because of the unlikelihood of finding nuclear families with multiple afflicted members. Homozygosity mapping is a gene hunting protocol which circumvents this problem (Lander and Botstein, 1987). Garrot long ago noted that a disproportionate number of alkaptonuria patients were progeny of consangious marriages. Homozygosity mapping relates to this observation and is based on the fact that, compared to a child of unrelated parents, a larger fraction of the genome of a child from a consangious marriage is likely to be homozygous.

To perform homozygosity mapping, the investigator seeks homozygous marker loci in multiple consangious families, each of which has at least one afflicted child. Finding such families is more likely than nuclear families with multiple afflicted progeny. On one hand the disease will be relatively more prevalent in these and on the other hand families with only one afflicted child will do. Suitable families are more easily found in countries where inbreeding is common. In fact, some very rare recessive diseases may be found only in such families (Botstein and Risch, 2003).

Each child of a family will have multiple regions of homozygosity. For example, 1/16 of the genome of a child of a first cousin marriage (corresponding to approximately 30 cM) and 1/64 of a second cousin marriage is expected to be homozygous simply because of shared ancestry. But if it is assumed that a specific gene is always the cause of the disease being investigated, one seeks a region of homozygosity shared by afflicted individuals in all the families. (The length of shared homozygosity among children of *n* first cousin marriages is expected to be approximately 30 cM/*n*.) If the region is unique, one expects the disease causing gene to lie within this region (Mani *et al*., 2002). Even if the disease is heterogeneous, meaning that any of several genes may be responsible for the disease, the same approach can be used after limiting the analysis to families of a limited population. Here, it is assumed that all afflicted individuals inherited the same disease causing allele from a distant common relative. In this case, the alleles of the homozygous marker loci near the disease gene should be the same in all the families.

A more quantitative look at homozygosity mapping for recessive traits reveals that a single afflicted child of a first cousin marriage will contribute a value of 1.8 (= log64) to the LOD score of a linkage analysis. A nuclear family of three afflicted children would make the same contribution. In both cases, six meiosis events are being followed. If there were two more afflicted siblings in the first cousin marriage, a LOD score of 3 would be reached. The above analysis assumes all meiosis events are informative. But even using markers with an average heterozygosity of = 0.5 spaced at 10 cM intervals, it is expected that a rare recessive disease causing locus can be identified using ten single-child first-cousin marriages. The more polymorphic and densely packed the markers used are, the smaller number of families is needed to be investigated in minimum. Furthermore, the rarer the disease causing allele is, the more likely it is that homozygosity is due to identity by descent rather than identity by state. Finally, LOD scores obtained by homozygosity mapping can be added directly to LOD scores obtained by conventional linkage analysis of nuclear families.

The loci for autosomal recessive hearing loss are extremely heterogeneous and the application of homozygosity mapping to search for responsible genes represent a classic example of the method (Guilford *et al*., 1994; Chaib *et al*., 1996). In the case of the rare recessive condition of benign recurrent intrahepatic cholestasis, common ancestry was inferred in four afflicted individuals from an isolated Dutch village (Houwen *et al*., 1994). Several of the genes responsible for fanconi anemias and Charcot-Marie-Thooth phenotype were mapped by homozygosity (LeGuern *et al*., 1996). The method has been widely used for gene hunting related to rare and/or heterogeneous traits since 1995.

Complex Diseases

In the perspective of world healthcare, genetic components contributing to common diseases such as diabetes, cardiac diseases and cancer are more important than those contributing to Mendelian diseases. The straightforward linkage analysis approach is generally not appropriate for finding genes involved in these common diseases because of its parametric nature, meaning that it demands a precise genetic model. Such precise models do not exist for the common diseases and they are for this reason included in the class of complex diseases as apposed to Mendelian diseases. Some of the not so common diseases with a genetic component such as multiple sclerosis, genetic forms of alcoholism and ataxia-telangiectasia are also complex (Stansbury *et al*., 1983). A fundamental difficulty in the analysis of at least some complex diseases relates to diagnosis. Because phenotypic changes are often subtle, accurate diagnosis can be a challenge. The difficulty is best appreciated in the case of psychiatric disorders (Van der Bree and Owen, 2003). At the level of genotypes, features of incomplete penetrance, genetic heterogeneity and polygenic inheritance can each be problematic.

Some of the approaches to genetic analysis of complex diseases still rely heavily on traditional linkage analysis. For example, an affected-only analysis avoids ambiguities in disease status of ostensibly unaffected family members. In effect, as the phenotype of all unaffected individuals is designated as unknown, the approach avoids the need to specify penetrance. Affected-only analysis was used to show linkage of late onset Alzheimers disease to chromosome 19, ultimately leading to identification of the APOE gene (Kehoe *et al*., 1999). This gene is important not only in the etiology of the disease in patients of multi-case families, but also in a significant proportion of sporadic cases of the late onset form.

It may also become possible to use linkage analysis if a more stringent definition of disease phenotype leads to a more nearly Mendelian inheritance pattern. In effect, those defined as affected become a more homogeneous group, helping to overcome impasses due to heterogeneous or polygenic basis of the disease phenotype. In the cases of familial breast cancer, a linkage analysis in which early age of onset was included in the phenotypic description led to the discovery of the BRCA1 gene first, and then the BRCA2 gene (Hall *et al*., 1990; Wooster *et al*., 1994). These genes are rarely mutated in the far more common sporadic cases of breast cancer. Similarly, when only early onset cases of Alzheimer's disease were considered, an autosomal dominant form of Mendelian inheritance was found in some families. Standard LOD score analysis with these families led to the discovery of three separate loci (see MIM 104300, 104311, 600579). More recently, selection of families on the basis of onset age led to localization of a novel melanoma susceptibility locus (Gillanders *et al*., 2003). Restricting linkage analysis to colon cancer families with extreme polyposis led to the discovery of the APC gene (Joslyn *et al*., 1991). Phenotypic restrictions based on severity of disease or the number of affected relatives have also been applied (Lander and Schork, 1994). Finally, there is a trend towards genetic analysis of measurable quantitative traits that are correlated with disease risk rather than the disease phenotype itself (Mackay, 2001; Blangero *et al*., 2003).

Allele sharing methods offer an alternative approach to the analysis of complex diseases. These methods are non-parametric in the sense that no genetic model is presented and no assumptions about the genetics of the disease are made. The approach is based on the detection of inheritance of a chromosomal region in a non-random fashion. Analysis of affected-sibling-pairs (ASP) is the simplest allele sharing method (Kruglyak and Lander, 1995; Sham and Zhao, 1998).

For this analysis, many affected sibling pairs are compared to determine the number of marker alleles or haplotypes of marker alleles both siblings of each pair have inherited from their parents. Random inheritance predicts that siblings will share 0, 1 or 2 parental haplotypes with a frequency of 1/4, 1/2 and 1/4, respectively. If the inheritance of maternal and paternal haplotypes is independently assessed, sharing 0 or 1 haplotype is expected with a 1:1 frequency. In the analysis, one seeks to find deviations from the expected frequencies. For a dominant Mendelian trait, all affected siblings are expected to share at least one parental haplotype and for recessive Mendelian traits, they should share two parental haplotypes. In the case of complex diseases, chromosomal regions are sought whose inheritance deviates to some degree from the expected 1 : 2 : 1 or 1 : 1 ratios. The statistical analysis for detecting significant deviation uses a $\div^2$ test. An extension of ASP analysis is affected-pedigree-member (APM) analysis in extended families. Because identity by descent is being presumed, multipoint analysis with highly polymorphic markers is the best.

A well known instance of using ASP analysis as a mapping strategy led to the identification of a VNTR locus upstream of the insulin gene as a susceptibility locus for type 1, or insulin-dependent diabetes (Davies *et al*., 1994; Merriman *et al*., 1997). Extensive sibling-pair analysis has been of limited use in the case of multiple sclerosis, probably because no single locus confers high susceptibility in the populations tested (Sawcer *et al*., 1997). The angiotensinogen gene was shown by AMP analysis to be linked to essential hypertension in multiplex families (Jeunemaitre *et al*., 1992). This sort of analysis also confirmed linkage of late onset Alzheimers disease to chromosome 19 (Pericak-Vance *et al*., 1991).

Association studies constitute a non-parametric genetic analysis approach based on population genetics rather than pedigree analysis (Xiong and Guo, 1997). Its basis is that once a mutation has occurred, it tends to spread though the population along with whatever alleles were present on the ancestral chromosome at nearby loci (Lander and Botstein, 1986). The loci whose alleles or haplotypes are investigated in association studies are marker loci (McPeek and Strahs,

1999). Like homozygosity mapping, identity by descent is presumed, but from a more distant ancestor. As the population being studied becomes more strictly confined, for example to a particular ethnic group, the distinction between population and family becomes murky and association and linkage studies converge. Association studies are useful for moving in closer to a disease causing locus after linkage studies have identified a larger inclusive chromosomal region (Ivinson *et al*., 1989; Merriman *et al*., 1997; Varon *et al*., 1998). Furthermore, they are particularly suitable for complex diseases because, as compared to linkage analysis, low penetrance mutations likely to be relevant in these diseases are more likely to be identified (Botstein and Risch, 2003).

In association studies, linkage disequilibrium between alleles of markers and the disease phenotype is sought. Markers whose allele frequencies in disease affected individuals of the population are different from the frequencies of the whole population, may be located in or near disease relevant genes. For example, the HLA-DR4 and HLA-DR3 and/or -DR4 alleles are found in a much higher percent of, respectively, rheumatoid arthritis and type 1 diabetes patients in England than in the general English population (Braun, 1979). In some studies, simple deviations from Hardy-Weinberg equilibrium at markers in linkage disequilibrium with the disease phenotype were used for disease gene identification (Klein *et al*., 1999).

Disease related loci discovered in linkage analysis do not necessarily reveal association in population studies. This is because different pedigrees may have inherited different mutations (and therefore different alleles) of the locus. This is more likely for dominant or X-linked diseases where a rapid turnover of disease causing alleles may occur. But if the disease causing alleles in a population are inherited from a single or few common ancestors, then the same locus may be identified with linkage and association studies.

As the effect of the locus becomes increasingly subtle with respect to the disease phenotype, then its identification will be more probable in association studies. Of course, the number of generations during which the population under study has expanded from the common ancestors will affect the facility of discovery of associated marker alleles. If the recombination fraction between marker and disease loci is è, then a fraction $(1-\theta)^n$ of the disease causing mutations will retain the association after *n* generations. Under the assumption of random breeding, this suggests that alleles of a marker and a disease locus, if separated by less than 3 cM should remain together in about 25% of the chromosomes after approximately 500 years of 20-25 generations (Strachan and Read, 1999).

Based on population simulations, it has been suggested that linkage disequilibrium in the general human population is unlikely to be larger than 3 kb (Kruglyak, 1999). Furthermore, based on empirical data of two genome-wide polymorphism discovery studies, a tendency for genomic regions extending 10-100 kb to be poor or rich in sequence variations has been shown and mostly attributed to variations in recombination rates and less to variations in mutation rates (Reich *et al*., 2002; Wall and Pritchard, 2003). As already mentioned, disease associated loci are more likely to be homogeneous in populations which have rapidly expanded from a relatively small number of founder individuals. Not only is the number of disease related alleles expected to be fewer in these, but the chromosomal length of linkage disequilibrium will be larger (Kruglyak, 1999; Kidd *et al*., 2000). The populations of Finland and Iceland are prime examples of relatively large populations with this feature. The genes for diastrophic dysplasia (DTD) and torsion dystonia were found by analysis, respectively, of the Finnish and Ashkenazi Jewish populations (Hastbacka *et al*., 1992; Ozelius *et al*., 1992). Smaller young populations throughout the world offer a similar advantage.

Care must be taken in experimental design and interpretation of data to avoid mistaken extrapolation from observed associations. The two most likely problematic sources are population stratification and statistical considerations. If the population under study is heterogeneous, then it is possible that a disease is prevalent in a particular subset of that population. If the frequency of an allele of a marker is fortuitously also high in that population, association will be detected independent of linkage disequilibrium. To avoid this, care must be taken in selection of the control group and it is best to work with a homogeneous population (Lander and Schork, 1994).

Ultimately, after having identified an associated genomic interval in an association study, a transmission disequilibrium test (TDT) within the interval can clarify potential problems associated with population admixture (Schaid, 1998). For this test, a new set of patients is selected whose parents are heterozygous for the marker allele found to be associated with the disease phenotype. If linkage disequilibrium is the basis of the association, then that allele should be transmitted to the affected offspring more often than random independent of whether or not population stratification exists.

Statistical analysis of data from association studies is fairly straightforward. Analysis becomes more complicated when haplotype determinations are made with the hope of establishing critical cross-over points during the history of the population (McPeek and Strahs, 1999; Botstein and Risch, 2003). Linkage disequilibrium mapping programs are available (Lu *et al*., 2003). Statistical artifacts are generally due to use of insufficiently stringent criteria for assessing significant association. Because many markers each with several alleles are generally tested, this multiplicity should be taken into account in assessing significance. A very useful and much used guideline for assessing the significance of disequilibrium based on the probability of random detection of that degree of disequilibrium has been suggested (Lander and Kruglyak, 1995).

Reference to complications of linkage disequilibrium analysis due to population histories and variations in recombination rates in the human genome have already been made. The complications imply that for the study of some

diseases and other traits such as drug response, thousands of markers and thousands of chromosomes will need to be analyzed. The magnitude of such analysis is bringing increasing focus on the use of single nucleotide polymorphisms as markers and rapid and relatively inexpensive scoring protocols probably relying on high density DNA chips.

Genetic Markers Across the Genome

Recombination in the genome occurs preferentially at specific locations called "hot spots." As a consequence, LD regions occur as discrete low recombination segments interspersed between hot spots. Such high LD segments have been dubbed "haplotype blocks." Haplotype block structure has been shown on chromosome 5 for immunogloblins (Daly *et al*., 2001), on chromosome X for HLA (Taillon-Miller *et al*., 2000), and chromosomes 21 (Zhang *et al*., 2002) and 22 (Gabriel *et al*., 2002). Johnson *et al.* (2001) argued that a small number of SNPs can capture ~90% of the haplotype variation in a population. They and others (Patil *et al*., 2001; Zhang *et al*., 2002; Zhang and Jin, 2003) have independently argued that between 300,000 and 1,000,000 carefully selected SNPs are needed for genotyping an entire genome in a disease association study. However, in our recent analysis based on the existing models of genomes, 100,000 accurately genotyped SNPs in and around genes provide sufficient power to detect many disease causing variants. In order to maximize effectiveness of those SNPs, they should be selected based on allele frequency, spacing, and haplotype structure.

Much has been made of the potential benefits of candidate gene based studies in the context of disease over map-based approaches (Schwartz *et al*., 2003). It is true that fewer genetic markers are required to establish LD associated with known genes. However, we have seen that several multiplexing technologies are emerging as viable solutions. For these technologies the cost difference between candidate gene approaches and whole genome studies is relatively small. However, Carlson *et al*. (2003) show through re-sequencing of 50 genes that the SNPs available show significant differences in their ability to detect and estimate haplotype structure in different populations. While they estimate that using the entire set of then ~3 million SNPs in dbSNP will uncover 80% of genotypic variance in Europeans, this drops to 50% for African Americans. This is consistent with Gabriel *et al.* (2002) observation of smaller haplotype blocks containing a greater number of haplotypes in African populations. Carlson *et al*. (2003) and Zhang and Jin (2003) argued that more SNPs and further studies are needed for whole-genome association studies in non-Caucasian populations.

The International HapMap Project (www.hapmap.org), involving collaborators from 7 countries, aims to develop such a SNP set from 300 samples in four populations (Caucasian, African, Japanese and Chinese ancestry). SNP discovery efforts associated with the HapMap project have doubled the number of SNPs in dbSNP to ~6 million. While the initial HapMap will provide resolution at ~5 kb, variation in the extent of LD in the genome (Zhang *et al*., 2002; Phillips *et al*., 2003) is extensive and will ultimately make it necessary to saturate parts of the genome at one SNP/kb.

The HapMap project as well as other studies also provides an opportunity for data analysis. High-density SNP map pose challenges for discovering and defining haplotype blocks (Goldstein, 2001; Carlson *et al*., 2003; Zhang *et al*., 2003) and recombination hot spots (Stumpf and Goldstein, 2003). Likewise, the selection of SNPs that accurately identify haplotype structure across many populations will have to be addressed. Statistical analysis of high-density SNP maps poses challenges, since multiple comparisons across large regions potentially reduce statistical power (Aslaug *et al*., 2003). Gene expression studies using arrays have already provided a strong impetus to develop methodology for multiple testing, and sophisticated computational methods based on probabilistic models are capable of discovering LD structure across the genome (Zhang *et al*., 2002; Schwartz *et al*., 2003; Zhang and Jin, 2003). In addition, although genotype quality for many new technologies exceeds 99%, genotyping errors will be present in a set of 1,000,000 or more SNPs. Novel data analysis methods that detect and correct for such errors are needed.

The Scale of Genotyping

As new technologies provide the capability to cost-efficiently type many markers across the genome, genetic markers begin to make good on the promises of mapping complex trait loci, uncovering evolutionary processes in the genome and revealing much about the history of human populations. What remains costly and difficult is to obtain sufficient samples from populations to take advantage of the power of new technologies. As the mean study size in genotyping projects approaches the thousands, the cost associated with acquiring and processing these increases. Additionally storage and tracking of samples and associated information alone requires significant investment in infrastructure and informatics.

This directs towards an era of collaborative studies involving multiple institutions and often newly formed interactions between academia and industry. Nationally funded agencies, such as the National Cancer Institute in the US, are establishing themselves as significant entities in the genotyping space. Both Singapore and Iceland are examples of nationwide efforts to develop the technologies, infrastructure and clinical resources required for genomic medicine and the use of genetic or biomarkers. Funding in excess of $4 billion has been committed to technology and clinical research in Singapore, and the National Science and Technology Board oversees several institutes that provide a national infrastructure. Iceland-based deCode combines extensive pedigree information with clinical records and technology. Although it is a commercial venture it

Table 1. This table shows significant varaibility in drug metabolism exist for several drugs. Heritability of this variable response is very high

Drug	Dose	Measurement	Response	Heritability
Antipyrine	18 mg/kg po once	plasma half-life	5.1-1.6 h	0.99
Phenylbutazone	6 mg/kg po once	plasma half-life	1.2-7.3 d	0.99
Aspirin	18 mg/kg/da 3	plasma level	11.9-36.4 mg/dl	0.98
Dicumarol	4 mg/kg po once	plasma half-life	7.0-74.0 h	0.98
Halothane	3.4 mg iv once	excretion/24 h	2.4-11.4%	0.63
Ethanol	1.2 ml/kg once	absorption/h	0.40-2.24 mg/ml	0.57
Diphenylhydantoin	100 mg iv once	serum half life	7.7-25.5 h	0.85
Lithium	600 mg/da 7	plasma level	0.16-0.38 mg/l	0.86
Amobarbital	125 mg iv once	clearance	16.0-67.2 ml/min	0.83

is strongly tied to the national healthcare system in Iceland. Through collaborations with pharmaceuticals, deCodes (www.decode.com) genotyping efforts have led to promising discoveries for neurological disorders (stefansson *et al*., 2002), asthma (hakonarson *et al.*, 2002), diabtes (styrkarsdottir *et al*., 2003), and osteoarthritis (stefansson *et al*., 2003).

Drug Metabolizing Enzymes

The study of drug metabolizing enzymes illustrates both the power and the pitfalls of population level studies. Individuals ability to process pharmaceutical drugs shows remarkable variability, and this response shows high heritability (Table 1). For a drug like Dicumarol a 10-fold difference in response may occur and serious side effects may occur for those patients who are sensitive. In a study that examines populations from Africa, Asia, Europe and the Middle East, Wilson and Sorant (2000) presented data on allelic variants where haplotypic variants using multiple loci provide better prognostics for drug metabolism than apparent ethnic or geographic association (Rosenberg *et al*., 2002; Goldstein *et al*., 2003). This surprising result illustrates the applicability of genetic marker study to understand, classify and predict clinical phenotypes.

Emerging Approaches for Large-scale Genetic Analysis

There are two major driving forces behind the development of new high throughput technologies. These are cost and low amount of DNA sample to be analyzed. To fulfill these criteria, either single molecule analysis should be enabled or multiplexing in genotyping, amplification and detection to be developed. No single molecule approach is currently in commercial use. However, several multiplex approaches have been described and a few of them commercialized. The challenge with highly multiplex approach is development of multiplexing of each step involved in genetic analysis to enable proper workflow process while having minimized cross interactions. Three approaches are emerging for large-scale genetic analysis having the potential to reduce the cost, sample size, and analysis time by several orders of magnitude. The common attribute for these techniques are high level of multiplexing (above 1,000 multiplexed reactions in a single tube).

For SNP detection, MIP assay from ParAllele BioScience (Hardenbol *et al*., 2003), GoldenGate assay from Illumina (Oliphant *et al*., 2002), and WGSA (whole genome sampling analysis) from Affymetrix (Kennedy *et al*., 2003) have been developed and commercialized. For comprehensive genetic analysis based on SNP, more markers will be needed. Currently six million SNPs have been discovered, but it is predicted that approximately 15 million SNPs need to be generated. Two new multiplexed SNP discovery techniques will be reviewed, *i.e. in vivo* mismatch repair detection (MRD) from ParAllele BioScience (Faham *et al*., 2001) and re-sequencing by hybridization from Perlegen (Patil *et al*., 2001).

Molecular Inversion Probe (MIP) for Genotyping

This technology was originally developed at Stanford University and has been commercialized by ParAllele BioScience (Hardenbol *et al*., 2003). The underlying principle for this technique is the use of unique sequence tag (molecular barcode) in target-specific oligonucleotides called molecular inversion probes (MIP). These probes carries two target-specific regions each about 20 nucleotides, one unique sequence tag about 20 nucleotides, and two sequences each of 20 nucleotides serving as general site for two PCR primers. In addition, two endonuclease sites have been engineered in the backbone of these probes. Although, these probes have a length of about 110 nucleotides, not more than about 60 nucleotides are unique. In MIP assay for genotyping more than 12,000 of these probes (each specific for a polymorphism and carrying a unique tag) can be added to the genomic DNA prior to denaturation step (Fig. 1). Upon hybridization to the

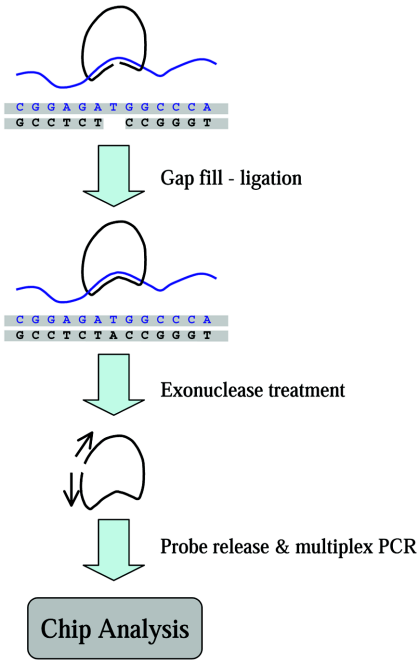

Fig. 1. Steps involved in molecular inversion probes assay for genotyping. A mixture of genomic DNA, more than 20,000 probes, and thermostable ligase and polymerase is heat denatured and brought to annealing temperature. Two sequences located at each termini of the probe hybridize to their respective complementary sites on the genome, thus forming a circular conformation with a single nucleotide gap between the termini of the probe. Subsequently, gap filling and ligation are performed using thermostable enzymes. Exonucleases are then added to digest linear probes in reactions where the added nucleotide was not complementary to the gap and excess linear probe in reactions where circular molecules were formed. The reactions are then heated to inactivate the exonucleases. To release probes from genomic DNA, uracil-N-glycosylase is added to depurinate the uracil residues in the probes. PCR reagents are then added including a primer pair that is common to all probes. The reactions are then thermocycled so that only probes that have been circularized in the allele-specific gap-filling reaction are amplified and analyzed on a universal chip.

target DNA, the ends are brought together, creating a double helix with a nick that can be sealed by a DNA ligase. The oligonucleotide is thereby circularized and due to the helical nature of double-stranded DNA, it encircles the target DNA strand (Fig. 1).

One of the most important characteristics of a MIP probe reaction is the circularization of the probe molecule after a correct identification of a specific sequence. The solution containing hybridized probes to the target is then divided into four different tubes each containing a single nucleotide species (dATP, dCTP, dGTP or dTTP). Extension of the probe takes place when the nucleotide is complementary to the DNA template. Only the extended templates are subsequently ligated and can be amplified. In the amplification the molecular barcodes are amplified and can be analyzed on a universal barcode chip. One of the PCR primers is labeled allowing detection of the hybridized DNA on the chip.

MIP assay provides very accurate SNP genotyping since it uses two hybridization events on a single probe, target-specific polymerization and target-specific ligation. Because both target-specific hybridization sites are located on a single probe, hybridization of one end increases the rate of hybridization of the other end providing very accurate hybridization. Therefore, probes can be used in attomole level, enabling multiplexing tens of thousand probes in a single tube.

In MIP assay, polymerization is used to fill the gap in the presence of a single nucleotide species to enable nucleotide specific ligation. This strategy provides a number of advantages: 1) one single probe can be used to genotype different variants of the polymorphism, reducing cost and intramolecular and intermolecular interactions, 2) higher accuracy can be obtained as fidelity of polymerase in the gap filling step is added to the system, and 3) any existing infrastructure for microarray facility can be used for this assay, diminishing the cost involved for equipment investment. Moreover, less quality control and tracking of probe will be required because there will be no complicated allele-specific reagents. Both two-color and four-color detection systems have been used. Using MIP assay it is likely to envision that all informative SNPs (~100,000 SNPs) in human be analyzed in a single tube on a single chip reducing the cost for analysis dramatically.

GoldenGate Assay for Genotyping

The GoldenGate technique is an OLA-based platform employing three oligonucleotides, one carrying the sequence tag and the other two are allele specific (Fig. 2). All three oligonucleotides contain target-specific regions and universal PCR primer sites; the locus-specific assay oligonucleotides also contains a unique address sequence that targets the complementary oligonucleotide probe for a particular bead type on the array (Oliphant *et al.*, 2002). These three assay oligonucleotide sets, for up to 1,152 different SNP loci, are pooled and hybridized to a small sample of genomic DNA. Hybridized allele-specific probes whose 3 ends are complementary to the dimorphic base at a SNP position are extended by a DNA polymerase, while allele-specific probes whose 3' ends mismatch with the base at the SNP position will not be extended efficiently.

Since all nucleotides are used for extension, strong strand-displacement activity of DNA polymerase limits the choice of DNA polymerases available because extended allele-specific assay oligonucleotides should terminate adjacent to the 5 phosphate of the locus-specific probe which are ligated to the locus-specific oligonucleotide. These full-length copies provide the amplification template for PCR using Cy3- and

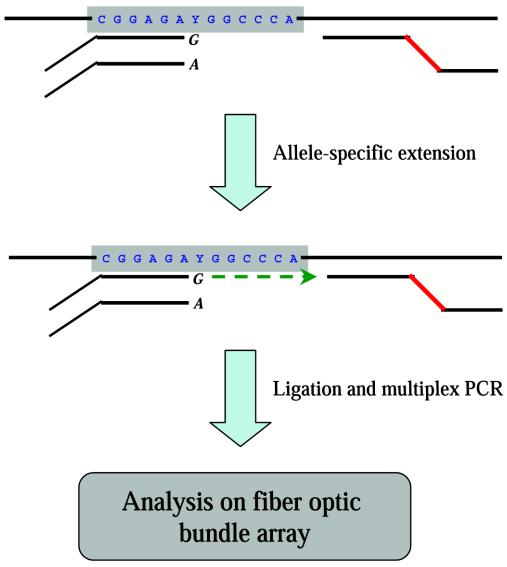

Fig. 2. GoldenGate technique for genotyping. Two allele-specific oligonucleotides and one oligonucleotide carrying sequence tag (red part of the oligonucleotide on the right side) are hybridized to the target. Subsequent to the enzymatic reaction involved in genotyping multiplex PCR is performed. PCR amplicons are then analyzed on a particular bead type on the array. Y represents T or C on the template.

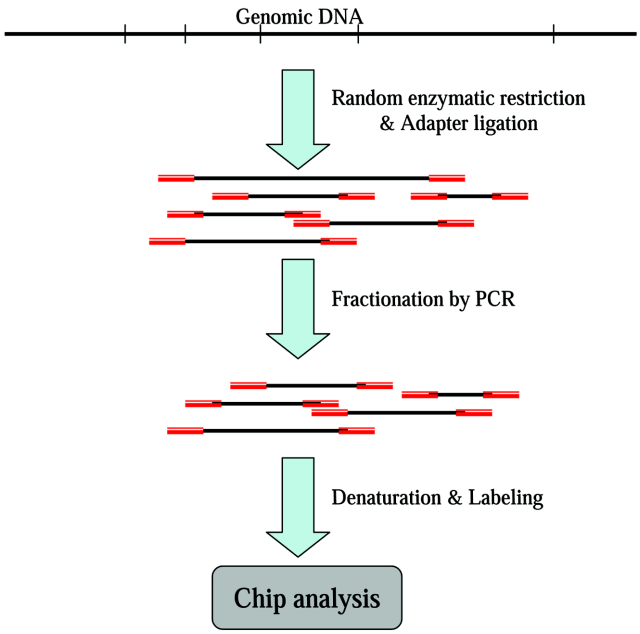

Fig. 3. Process involved in whole genome sampling analysis. Genomic DNA is digested with a restriction enzyme resulting in different fragment sizes. A universal adapter is ligated to restricted fragments and the fragments are subjected to multiplex PCR amplification. Subsequently, the amplified target is fragmented, labeled and hybridized to DNA microarray.

Cy5-labeled primers (for the allele-specific assay oligonucleotides) and a biotin-labeled primer. After thermocycling, the PCR products are bound to streptavidin-coated paramagnetic particles and the dye-labeled strands are isolated. The fluorescently labeled, single-stranded products are then hybridized to Illuminas pre-decoded random array. This assay allows multiplexing of up to 1152 SNPs in a single reaction tube. GoldenGate assay relies on Sentrix array and Sherlock scanner commercialized by Illumina.

Whole Genome Sampling Analysis (WGSA) for Genotyping

Recently, a hybridization-based technique was introduced for SNP genotyping (Kennedy *et al*., 2003). In this technique genomic DNA is digested by restriction endonuclease, namely *Eco*RI, *Bgl*II and *Xba*I to digest the genome. Followed by adapter ligation, multiplex PCR is performed to selectively amplify fragments between 400-800 nucleotides (Fig. 3). Amplification represents approximately 4×10^7 base pairs of genomic DNA (Kennedy *et al*., 2003). Successful genotyping of 14,548 SNPs has been reported. The number of SNPs to be analyzed on this platform can potentially be increased since the chip manufacturing technique is photolithographic and smaller feature size can be achieved, however, the technology demonstrates limited flexibility in the choice of SNPs to be genotyped and has a conversion rate approaching 15%. Because this product is not easily customizable, its applicability is limited primarily to genome wide mapping application.

In vivo Mismatch Repair Detection (MRD) for Variation Scanning

Recently, a new technique called *in vivo* mismatch repair detection, or MRD (Faham *et al*., 2001) was developed to enrich for variant alleles (Fig. 4). MRD detects variants utilizing the mismatch repair system of *Escherichia coli*. A specific strain is engineered to sort a mixture of transformed fragments into two pools: those carrying a variation and those that do not. MRD has been described before as a method for multiplex variation scanning (Faham *et al*., 2001). This technique has proven as an excellent enrichment method for dideoxy terminator sequencing to discover rare and common variant alleles in normal and disease populations. The cost involved in deep sequencing has been reduced by 70 folds. In addition, rare mutations and SNPs can be identified as each bacterial colony represents only one allele thereby homozygous sequencing trace is obtained.

One of the main advantages in using MRD for finding genes associated with disease is the ability of enriching the genes containing rare mutations, rare SNPs, common mutations and common SNPs as they will appear in individual colonies on MRD variant plate. This is unique feature of this technique, which is not possible with standard sequencing, as

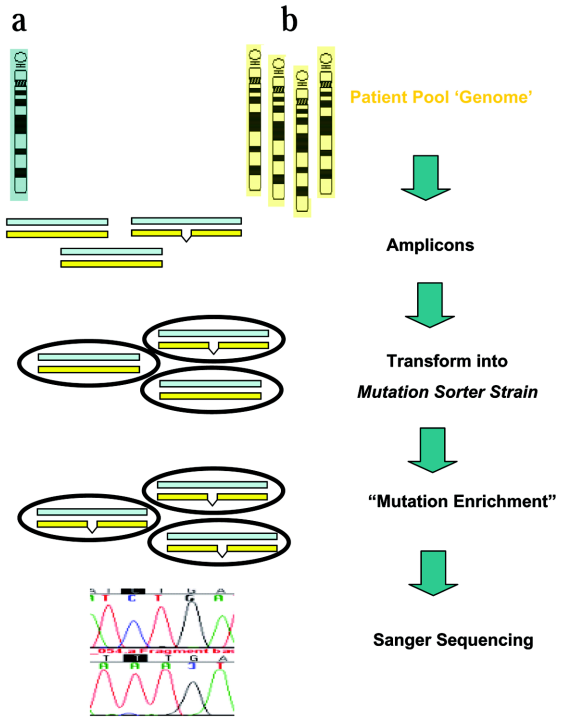

Fig. 4. Process involved in MRD assay for variation scanning. A reference library is made based on normal DNA. Samples to be analyzed for variation scanning are then amplified in pool and hybridized to the reference library prior to transformation. Bacterial strain selects heteroduplex fragments. Sequencing determines the nature of the variations.

these mutations and SNPs are often missed in single pass DNA sequencing without enriching the DNA fragment with MRD. This technique can be used to discover new genes containing rare and common mutations and polymorphisms correlated to different diseases.

Re-sequencing by Hybridization

Sequencing-by-hybridization was proposed by a few groups more than 15 years ago. However, it became in use for re-sequencing just a couple of years ago by Perlegen. Perlegen uses glass wafers on which high density arrays of DNA probes (short segments of DNA) have been placed. Each of these wafers holds approximately 60 million DNA probes that can be used to recognize longer sample DNA sequences (for example, from patients). The recognition of sample DNA by the set of DNA probes on the glass wafer takes place through the mechanism of DNA hybridization. When a DNA sample hybridizes with an array of DNA probes, the sample will bind to those probes that are complementary to target DNA sequence. By evaluating to which probes the sample DNA hybridizes more strongly, a genotype can be determined.

The flow of process is demonstrated in Fig. 5. Long range PCR is used to amplify fragments in a range of 3-12 kb from genomic DNA using target-specific primers. Subsequently, the product are purified and fragmented by a DNase which is later inactivated by heat. Thereafter, labeling is performed prior to hybridization of the fragments onto the wafer. Recently, 8 wafers were used to determine haplotype structure of chromosome 21 (Patil *et al*., 2001). Sample preparation steps are not highly multiplexed and the wafers are very expensive in this protocol. To reduce the cost involved in this protocol, hundreds of samples are typically genotyped in a single pool. Inaccuracies involved in pooling approach limit the sensitivity and specificity, making the technology best suited for finding very strong genetic effects with limited family groups available to provide linkage data, more specifically, adverse drug response studies.

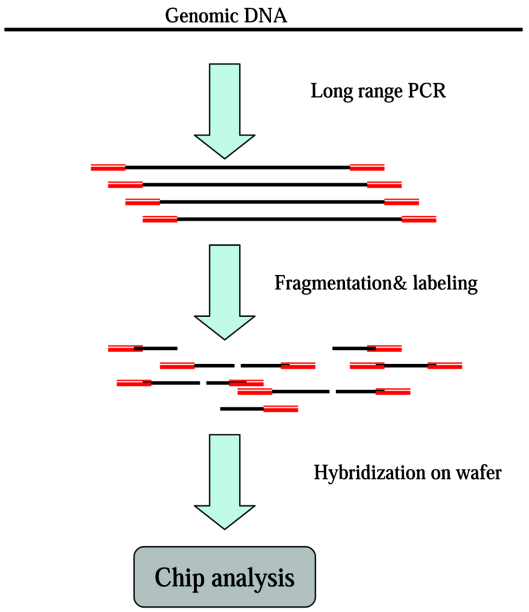

Fig. 5. Process involved in re-sequencing by hybridization. Long range PCR is used to amplify fragments in a range of 3-12 kb from genomic DNA using target-specific primers. Subsequently, the product are purified and fragmented by a DNase which is later inactivated by heat. Thereafter, labeling is performed prior to hybridization of the fragments onto the wafer.

Concluding Remarks

Technologies that provide a genome-wide view offer an unprecedented opportunity to scrutinize the genetic analysis for gene discovery, population studies and predictive medicine. As novel technologies make it possible to accurately type genetic markers at ever decreasing costs, these markers are of increasing importance as a tool in genetics and genomics. Genotyping studies have already shown significant successes in identifying genetic loci associated with diseases such as cystic fibrosis, Alzheimer's and diabetes to name a few. Secondly, large-scale studies of microsatellite markers are

providing a novel and detailed view of the structure and history of human populations (Rosenberg *et al*., 2002). Allelic variants associated with drug metabolizing enzymes also shed light on functional differences in drug response and, in some cases, prove to have greater predictive power than ethnic or environmental factors (Rosenberg *et al*., 2002). High-density genotyping using a growing repertoire of almost 10 million publicly known SNPs coupled with nation-wide clinical efforts promises to uncover associations between genetic factors in complex disease and aid in our understanding of human population history. This is greatly supported by the exploration of LD and haplotype patterns in the human genome. The HapMap project (www.hapmap.org) in particular is designed to provide a reference and resource for the distribution and properties of genetic markers. Success of these ventures depends in large part on a successful marriage of well-collected samples, technology, infrastructure and data analysis.

Large scale genotyping holds the promise of improving our understanding of genomics in the context of disease and evolution, but it brings significant challenges with it. Legal, ethical and regulatory guidelines must be established for genotyping studies that range across many populations and national boundaries. Geneticists need to learn from epidemiology and the methodology for clinical trials which address problems also crucial for genetic association studies in order to address the concern about reproducibility of findings of genetic association studies in complex diseases.

The cost of genotyping must decrease further, as larger cohorts of patients are typed. Although recent developments promise this as well as improved reliability, sensitivity, accuracy and resolution for genotyping, data handling and analysis techniques suitable for large-scale high-density studies have yet to be developed.

Guidelines and standards for experimental design and statistical analysis (Goldstein *et al*., 2003) should be established. This will significantly facilitate the difficult process of understanding and statistically testing large-scale genetic analysis. The importance of large-scale genetic analysis will depend on whether these issues can be addressed successfully. Recent and ongoing works suggest that we will be able to deliver the technological and analytical tools to utilize large-scale genetics. Building on the success and insight derived from single gene diseases, discovery of new factors in complex diseases and improvements in diagnosis and prognosis of medical conditions may depend on such a large-scale genome-wide view of genetic analysis.

Acknowledgments The second and third authors are supported by an NIH grant (2POI HG00205). We would like to thank Dr. Thomas Willis for useful discussions.

References

Aslaug, J., Thorlacius, T., Fossdal, R., Jonasdottir, A., Benediktsson, K., Benedikz, J., Jonsson, H. H., Sainz, J., Einarsdottir, H., Sigurdardottir, S., *et al*. (2003) A whole genome association study in Icelandic multiple sclerosis patients with 4804 markers. *J. Neuroimmunol.* **143**, 84-87.

Bansal, A., van den Boom, D., Kammerer, S., Honisch, C., Adam, G. and Cantor, C. R. (2002) Association testing by DNA pooling: an effective initial screen. *Proc. Natl. Acad. Sci. USA* **99**, 16871-16874.

Blangero, J., Williams, J. T. and Almasy, L. (2003) Novel family-based approaches to genetic risk in thrombosis. *J. Thromb. Haemost.* **1**, 1391-1397.

Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes fro Mendelian disease, future approaches for complex disease. *Nat. Genet.* **33**, 228-237.

Botstein, D., White, R. L., Skolnick, M. H. and Davis, R.W. (1980) Construction of genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314-331.

Braun, W. E. (1979) *HLA and Disease*, CRC, Florida, USA.

Broman, K. W., Murray, J. C., Scheffield, V. C. and White R. L. (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861-869.

Broman, K. W. and Weber, J. L. (1998) Characterization of human crossover interference. *Am. J. Hum. Genet.* **63** (suppl.) A1632.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Smith, J. D., Kruglyak, L. and Nickerson, D. A. (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole genome association studies in humans. *Nat. Genet.* **33**, 518-521.

Chaib, H., Place, C., Salem, N., Chardenoux, S., Vincent, C., Weissenbach, J., El-Zir, E., Loiselet, J. and Petit, C. (1996) A gene responsible for a sensorineural nonsyndromic recessive deafness maps to chromosome 2p22-23. *Hum. Mol. Genet*. **5**, 155-158.

Clark, A. G., Nielson, R., Signorovitch, J., Matise, T. C., Glanowski, S., Heil, J., Winn-Deen, E. S., Holden, A. L. and Lai, E. (2003) Linkage disequilibrium and influence of ancestral recombination in 538 single nucleotide polymorphism clusters across the human genome. *Am. J. Hum. Genet.* **73**, 285-300.

Collins, A., Frezal, J., Teague, J. and Morton, N.E. (1996) A metric map of humans: 23,500 loci in 850 bands. *Proc. Natl. Acad. Sci. USA* **93**, 14771-14775.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. and Lander, E. S. (2001) High resolution haplotype structure in the human genome. *Nat. Genet.* **29**, 217-222.

Dausset, J., Cann, H., Cohen, D., Lanthrop, M., Lalouel, J. M. and White, R. (1990) Centre d'Etude du polymorphisme humaine (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575-577.

Davies, J. L., Kawaguchi, Y., Bennett, S. T., Copeman, J. B., Cordell, H. J., Pritchard, L. E., Reed, P. W., Gough, S. C., Jenkins, S. C., and Palmer, S. M. (1994) A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371**, 130-136.

Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., *et al*. (2003) A first generation linkage disequilibrium map of

human chromosome 22. *Nature* **418**, 544-548.

De La Vega, F. M., Dailey, D., Ziegle, J., Williams, J., Madden, D. and Gilbert, D. A. (2002) New generation pharmacogenomic tools: a SNP linkage disequilibrium map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies. *BioTechniques* Suppl, 48-54.

Elston, R. C. and Stewart, J. (1971) A general model for the analysis of pedigree data. *Hum. Hered.* **21**, 523-542.

Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L. and Jarvela, I. (2002) Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233-237.

Faham, M., Baharloo, S., Tomitaka, S., DeYoung, J. and Freimer, N. B. (2001) Mismatch repair detection (MRD): high-throughput scanning for DNA variations. *Hum. Mol. Genet.* **10**, 1657-1664.

Fearnhead, P. and Donelly, P. (2001) Estimating recombination rates from population genetic data. *Genetics* **159**, 1299-1318.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., *et al*. (2002) The structure of haplotype blocks in the human genome. *Science* **296**, 2225-2229.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L. Y., Huang, W., Liu, B., Shen, Y., *et al*. (2004) The international HapMap project. *Nature* 426, 789-796.

Gillanders, E., Hank Juo, S. H., Holland, E. A., Jones, M., Nancarrow, D., Freas-Lutz, D., Sood, R., Park, N., Faruque, M., Markey, C., *et al*. (2003) Localization of a novel melanoma susceptibility locus to 1p22. *Am. J. Hum.Genet.* **73**, 301-313.

Goddard, K. A. B., Hopkins, P. J., Hall, J. M., Witte J. S. (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66**, 216-234.

Goldstein, D. B. (2001) Islands of linkage disequilibrium. *Nat. Genet.* **29**, 109-111.

Goldstein, D. B., Kourosh, R. A., Mike, E. W. and Nicholas, W. W. (2003) Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.* **19**, 615-622.

Guilford, P., Ben Arab, S., Blanchard, S., Levilliers, J., Weissenbach, J., Belkahia, A. and Petit, C. (1994) A non-syndromic form of neurosensory, recessive deafness maps to pericentromic region of chromosome 13q. *Nat. Genet.* **6**, 24-28.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, K., Wallace, M. C., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E. and Martin, J. B. (1983) A polymorphic DNA marker genetically linked to Huntingtons disease. *Nature* **306**, 234-238.

Hakonarson, H., Bjornsdottir, U. S., Halapi, E., Palsson, S., Adalsteinsdottir, E., Gislason, D., Finnbogason, G., Gislason, T., Kristjansson, K., Arnason, T., Birkisson, I., Frigge, M. L., Kong, A., Gulcher, J. R. and Stefansson, K. (2002) A major susceptibility gene for asthma maps to chromosome 14q24. *Am. J. Hum. Genet.* **71**, 483-491.

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B. and King, M. C. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684-1689.

Hardenbol, P., Jain, M., Namsaraev, E. A. Karlin-Neumann, G. A. Fakhrai-Rad, H. Ronaghi, M., Willis, T. and Davis, R.W. (2003) Highly multiplexed genotyping with molecular inversion probes. *Nat. Biotech.* **21**, 673-678.

Hastbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A. and Lander, E. (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat. Genet.* **2**, 204-211.

Houwen, R. H. J., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L. A. and Freimer, N. B. (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat. Genet.* **8**, 380-386.

Ivinson, A. J., Read, A. P., Harris, R., Super, M., Schwarz, M., Claton Smith, J. and Elles, R. (1989) Testing for cystic fibrosis using allelic association. *J. Med. Genet.* **26**, 426-430.

Jeffreys, A. J., Kauppi, L. and Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29**, 217-222.

Jeffreys, A. J., Wilson, V. and Thein, S. L. (1985) Hypervariable minisatellite regions in human DNA. *Nature* **314**, 67-73.

Jeunemaitre, X., Soubrier, F., Kotelevtsev, Y. V., Lifton, R. P., Williams, C. S., Charru, A., Hunt, S. C., Hopkins, P. N., Williams, R. R. and Lalouel, J. M. (1992) Molecular basis of human hypertension: role of angiotensinogen. *Cell* **71**, 169-180.

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., *et al*. (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233-237.

Joslyn, G., Carlson, M., Thliveris, A., Albertsen, H., Gelbert, L., Samowitz, W., Groden, J., Stevens, J., Spirio, L. and Robertson, M. (1991) Identification of deletion mutations and three new genes at the familial polyposis locus. *Cell* **66**, 601-613.

Jurinke, C., Van Den Boom, D., Cantor, C. R. and Köster, H. (2002) The use of MassARRAY technology for high throughput genotyping. *Adv. Biochem. Eng. Biotechnol.* **77**, 57-74.

Kash, S. F., Johnson, R. S., Tecott, L. H., Noebels, J. L., Mayfield, R. D., Hanahan, D. and Baekkeskov. S. (1997) Epilepsy in mice deficient in the 65-kDa isoform of glutamic acid decarboxylase. *Proc. Natl. Acad. Sci. USA* 14060-14065.

Kehoe, P., Wavrant-de Vrieze, F., Cook, R., Wu, W. S., Holmans, P., Fenton, I., Spurlock, G., Norton, N., Williams, H., Williams, N., *et al*. (1999) A full genome scan for late-onset Alzheimer disease. *Hum. Mol. Genet.* **8**, 237-246.

Kennedy, G. C., Matsuzaki, H., Dong, S., Liu, W. M., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., *et al*. (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol.* **21**, 1233-1237.

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M. and Tsui, L. C. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Scienc*e **245**, 1073-1080.

Kidd, J. R., Pakstis, A. J., Zhao, H., Lu, R. B., Okonofua, F. E., Odunsi, A., Grigorenko, E., Tamir, B. B., Friedlaender, J., Schulz, L. O., Parnas, J. and Kidd, K. K. (2000) Haplotype and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am. J. Hum. Genet.* **66**, 1882-1899.

Kim, U.-K., Jorgenson, E., Coon, H., Leppert, M., Risch, N. and Drayna, D. (2003) Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. *Science* **299**, 1221.

Klein, C., Vieregge, P., Hagenah, J., Sieberer, M., Doyle, E., Jacobs, H., Gasser, T., Breakefield, X. O., Risch, N. J. and Ozelius, L. J. (1999) Search for the PARK3 founder haplotype in a large cohort of patients with Parkinsons disease from northern Germany. *Ann. Hum. Genet.* **63**, 285-291.

Koenig, M., Hoffman, E. P., Bertelson, C. J., Monaco, A. P., Feener, C. and Kunkel, L. M. (1987) Complete cloning of the Duchenne muscular dystrophy (DMD): cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* **50**, 509-517.

Kong, A. and Cox, N. J. (1997) Allele-sharing models: LOD scores and accurate linkage tests. *Am. J. Hum. Genet.* **61**, 1179-1188.

Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22, 139-144.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. and Lander, E. S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, 1347-1363.

Kruglyak, L. and Lander, E. S. (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.* **57**, 439-454.

Kruglyak, L. and Nickerson, D.A. (2001) Variety is the spice of life. *Nat. Genet.* **27**, 234-236.

Lander, E. S. and Botstein, D. (1986) Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. *Cold Spring Harb. Symp. Quant. Biol.* 51, 49-62.

Lander, E. S. and Botstein, D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567-1570.

Lander, E. S. and Green, P. (1987) Construction of multilocus genetic maps in humans. *Proc. Natl. Acad. Sci. USA* **84**, 2363-2367.

Lander, E. S. and Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**, 241-247.

Lander, E. S. and Schork, N. J. (1994) Genetic dissection of complex traits. *Science* **265**, 2037-2048.

Lanthrop, G. M., Lalouel, J. M., Julier, C. and Ott, J. (1984) Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. USA* **81**, 3443-3446.

LeGuern, E., Guilbot, A., Kessali, M., Ravise, N., Tassin, J., Maisonobe, T., Grid, D. and Brice, A. (1996) Homozygosity mapping of an autosomal recessive form of demyelinating Charcot-Marie-Tooth disease to chromosome 5q23-q33. *Hum. Mol. Genet.* **5**, 1685-1688.

Lifton, R. P., Gharavi, A. G. and Geller, D. S. (2001) Molecular mechanisms of human hypertension. *Cell* **104**, 545-556.

Lu, X., Niu, T. and Liu, J. S. (2003) Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms. *Genome Res.* **13**, 2112-2117.

Mackay, T. F. C. (2001) The genetic architecture of quantitative traits. *Ann. Rev. Genet.* **35**, 303-339.

Mani, A., Meraji, S. M., Houshyar, R., Radhakrishnan, J., Mani, A., Ahangar, M., Rezaie, T. M., Taghavinejad, M. A., Broumand, B., Zhao, H., Nelson-Williams, C. and Lifton, R. P. (2002) Finding genetic contributions to sporadic disease: a recessive locus at 12q24 commonly contributes to patent ductus arteriosus. *Proc. Natl. Acad. Sci. USA* **99**, 15054-15059.

Marnellos, G. (2003) High-throughput SNP analysis for genetic association studies. *Curr. Opin. Drug Discov. Devel.* **6**, 317-321.

McPeek, M. S. and Strahs, A. (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing with application to fine-scale mapping. *Am. J. Hum. Genet.* **65**, 858-875.

Merriman, T., Twells, R., Merriman, M., Eaves, I., Cox, R., Cucca, F., McKinney, P., Shield, J., Baum, D., Bosi, E., *et al*. (1997) Evidence by allelic association methods for a type 1 diabetes polygene (IddM6) on chromosome 18q21. *Hum. Mol. Genet.* **6**, 1003-1010.

Mohlke, K. L., Erdos, M. R., Scott, L. J., Fingerlin, T. E., Jackson, A. U., Silander, K., Hollstein, P., Boehnke, M. and Collins, F. (2002) High-throughput screening for evidence of using mass spectrometry genotyping on DNA pools. *Proc. Natl. Acad. Sci. USA* **99**, 16928-16933.

Morton, N. E., Lindstein, J., Iselius, L. and Yee, S. (1982) Data and theory for a revised chiasma map of man. *Hum. Genet.* **62**, 266-270.

Morton, N. E. (1955) Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277-318.

Oliphant, A., Barker, D. L., Stuelpnagel, J. R. and Chee, M. S. (2002) BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *BioTchniques* Suppl. 56-61.

Ott, J. (1999) *Analysis of Human Genetic Linkage*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, USA.

Ozelius, L. J., Kramer, P. L., de Leon, D., Risch, N., Bressman, S. B., Schuback, D. E., Brin, M. F., Kwiatkowski, D. J., Burke, R. E. and Gusella, J. F. (1992) Strong allelic association between the torsion dystonia gene (DYT1) and loci on chromosome 9q34 in Ashkenazi Jews. *Am. J. Hum. Genet.* **50**, 619-628.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719-1723.

Pericak-Vance, M. A., Bebout, J. L., Gaskell, P. C. Jr., Yamaoka, L. H., Hung, W. Y., Alberts, M. J., Walker, A. P., Bartlett, R. J., Haynes, C. A. and Welsh, K. A. (1991) Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am. J. Hum. Genet.* **48**, 1034-1050.

Phillips, M. S., Lawrence, R., Sachidanandam, R., Morris, A. P., Balding, D. J., Donaldson, M. A., Studebaker, J. F., Ankener, W. M., Alfisi, S. V. and Kuo, F. S. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.* **33**, 382-387.

Reeders, S. T., Breuning, M. H., Davies, K. E., Nicolls, R. D., Jarman, A. P., Higgs, D. R., Pearson, P. L. and Weatherall, D. J. (1985) A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. *Nature* **317**, 542-544.

Reich, D. E., Schaffner, S. F., Daly, M. J., McVean, G., Mullikin, J. C., Higgins, J. M., Richter, D. J., Lander, E. S. and Altshuler, D. (2002) Human genome sequence variation and

the influence of gene history, mutation and recombination. *Nat. Genet.* **32**, 135-142.

Ronaghi, M., Uhlén, M. and Nyrén, P. (1998) A sequencing method based on real-time pyrophosphate. *Science* **281**, 363-365.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. and Feldman, M. W. (2002) Genetic structure of human populations. *Science* **298**, 2381-2385.

Royer-Pokora, B., Kunkel, L. M., Monaco, A. P., Goff, S. C., Newburger, P. E., Baehner, R. L., Cole, F. S., Curnutte, J. T. and Orkin, S. H. (1986) Cloning of the gene for an inherited human disorder - chronic graulmatous disease - on the basis of its chromosomal location. *Nature* **322**, 32-38.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., *et al*. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933.

Sawcer, S., Goodfellow, P. N. and Compston, A. (1997) The genetic analysis of multiple sclerosis. *Trends Genet*. **13**, 234-239.

Schaid, D. J. (1998) Transmission disequilibrium, family controls and great expectations. *Am. J. Hum. Genet*. **63**, 935-941.

Schwartz, R., Halldorsson, B. V., Bafna, V., Clark, A. G. and Istrail, S. (2003) Robustness of inference of haplotype block structure. *J. Comput. Biol*. **10**, 13-19.

Sham, S. and Zhao, J. (1998) Linkage analysis using affected sib pairs; in *Guide to Human Genome Computing*, 2nd ed., Bishop, M. J. (ed.) Academic Press, London.

Sherrington, R., Brynjolfsson, J., Petursson, H., Potter, M., Dudleston, K., Barraclough, B., Wasmuth, J., Dobbs, M. and Gurling, H. (1988) Localization of a susceptibility locus for schizophrenia on chromosome 5. *Nature* **336**, 164-167.

Shin, H. D., Park, B. L., Kim, L. H., Jung, J. H., Wang, H. J., Kim, Y. J., Park, H. S., Hong, S. J., Choi, B. W., Kim, D. J. and Park, C. S. (2003) Association of tumor necrosis factor (TNF) polymorphisms with asthma and serum total IgE. *Hum. Mol. Genet.* 10.1093/hmg/ddh036.

Stansbury, J. B., Wyngaarden, D. S., Fredrickson, J. L., Goldstein, J. L. and Brown, M. S. (1983) *The Metabolic Basis of Inherited Disease*. McGraw Hill, New York, USA.

Stanssens, P., Zabeau, M., Meersseman, G., Remes, G., Gansemans, Y., Storm, N., Hartmer, R., Honisch, C., Rodi, C. P., Bocker S. and Van Den Boom, D. (2004) High-throughput MALDI-TOF discovery of genomic sequence polymorphisms. *Genome Res.* **14**, 134-141.

Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., Brynjolfsson, J., Gunnarsdottir, S., Ivarsson, O., Chou, T. T., Hjaltason, O., Birgisdottir, B., Jonsson, H., Gudnadottir, V. G., Gudmundsdottir, E., Bjornsson, A., Ingvarsson, B., Ingason, A., Sigfusson, S., Hardardottir, H., Harvey, R. P., Lai, D., Zhou, M., Brunner, D., Mutel, V., Gonzalo, A., Lemke, G., Sainz, J., Johannesson, G., Andresson, T., Gudbjartsson, D., Manolescu, A., Frigge, M. L., Gurney, M. E., Kong, A., Gulcher, J. R., Petursson, H. and Stefansson, K. (2002) Neuregulin 1 and Suseptibility to schizophrenia. *Am. J. Hum. Genet.* **71**, 877-892.

Stefansson, S. E., Jonsson, H., Ingvarsson, T., Manolescu, I., Jonsson, H. H., Olafsdottir, G., Palsdottir, E., Stefansdottir, G., Sveinbjornsdottir, G., Frigge, M. L., Kong, A., Gulcher, J. R. and Stefansson, K. (2003) Genomewide scan for hand osteoarthritis: a novel mutation in matrilin-3. *Am. J. Hum. Genet.* **72**, 1448-1459.

Strachan, T. and Read, A. P. (1999) *Human Molecular Genetics*, 2nd ed., John Wiley & Sons, New York, USA.

Stumpf, M. P. and Goldstein, D. B. (2003) Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr. Biol*. **13**, 1-8.

Styrkarsdottir, U., Cazier, J. B., Kong, A., Rolfsson, O., Larsen, H., Bjarnadottir, E., Johannsdottir, V. D., Sigurdardottir, M. S., Bagger, Y., Christiansen, C., Reynisdottir, I., Grant, S. F., Jonasson, K., Frigge, M. L., Gulcher, J. R., Sigurdsson, G. and Stefansson, K. (2003) Linkage of Osteoporosis to Chromosome 20p12 and Association to BMP2. **PLoS Biol 1**, E69.

Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P. and Kwok, P. Y. (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.* **25**, 324-328.

Terwillinger, J. D. and Ott, J. (1992) A multisample bootstrap approach to the estimation of maximized-over-models lod score distributions. *Cytogenet. Cell Genet.* **59**, 142-144.

van den Bree, M. B. and Owen, M. J. (2003) The future of psychiatric genetics. *Ann. Med*. **35**, 122-134.

Varon, R., Vissinga, C., Platzer, M., Cerosaletti, K. M., Chrzanowska, K. H., Saar, K., Beckmann, G., Seemanova, E., Cooper, P. R., Nowak, N. J., *et .al*. (1998) Nibrin, a novel DNA double-stranded break repair protein, is mutated in Nijmegen Breakage Syndrome. *Cell* **93**, 467-476.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al*. (2001) The sequence of the human genome. *Science* **291**, 1304-1351.

Wall, J. D. and Pritchard, J. K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **4**, 587-597.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., *et al*. (1998) Large scale identification, mapping and genotyping of single nucleotide polymorphisms in the human genome. *Science* **280**, 1077-1082.

Weber, J. L. and May, P. E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet*. **44**, 388-396.

Weeks, D. E., Lehner, T., Squires-Wheeler, E., Kaufmann, C. and Ott, J. (1990) Localization of a susceptibility locus for schizophrenia on chromosome 5. *Genet. Epidemiol*. **7**, 237-243.

Wilson, A. F. and Sorant, A. J. (2000) Equivalence of single- and multi-locus markers: power to detect linkage with composite markers derived from biallelic loci. *Am. J. Hum. Genet.* **66**, 1610-1615.

Wooster, R., Neuhausen, S. L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T. and Averill, D. (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265,** 2088-2090.

Xiong, M. and Guo, S.-W. (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *Am. J. Hum. Genet.* **60**, 1513-1531.

Yang, Y., Zhang, J., Hon, J., Matsuda, F., Xu, p., Lanthrop, M.

and Ott, J. (2003) Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc. Natl. Acad. Sci. USA* **100**, 7225-7230.

Zhang, K., Akey, J. M., Wang, N., Xiong, M., Chakraborty, R. and Jin, L. (2003) Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. *Hum. Genet.* **113**, 51-59.

Zhang, K., Calabrese, P., Nordborg, M. and Sun, F. (2002) Haplotype block structure and its applications to association studies: power and study designs. *Am. J. Hum. Genet.* **71**, 1386-1394.

Zhang, K., Deng, M., Chen, T., Waterman, M. S. and Sun, F. (2002) A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci. USA* **99**, 7335-7339.

Zhang, K. and Jin, L. (2003) HaploBlockFinder: haplotype block analyses. *Bioinformatics* **19**, 1300-1301.

Zhang, K., Sun, F., Waterman, M. S. and Chen, T. (2003) Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *Am. J. Hum. Genet.* **73**, 63-73.

Zhang, W., Collins, A., Maniatis, N., Tapper, W. and Morton, N. E. (2002) Properties of linkage disequilibrium (LD) maps. *Proc. Natl. Acad. Sci. USA* **99**, 17004-17007.