The Application of the HapMap to Diabetic Nephropathy and Other Causes of Chronic Renal Failure

Sudha K. Iyengar, PhD, and Sharon G. Adler, MD†*

Summary: The human nuclear genome consists of approximately 3 billion nucleotides. Human beings are 99% similar in DNA sequence to each other, but natural genetic variation in approximately 1% of the DNA sequence is responsible for interindividual differences, including determining who will develop disease and who will remain healthy. The pace and timing of disease initiation also is regulated by exposure to individual-level environmental factors and other random causes. Therefore, an examination of the DNA sequences of individuals with and without diabetic nephropathy, or, more broadly, chronic renal failure, can predict which sequence differences vary with disease (or health). The technology is not yet economical enough to analyze large numbers of individuals down to each nucleotide, but standardized dense genotyping sets for interrogating 1 marker for every 5,000, 10,000, or 15,000 nucleotides now are affordable even in large samples. The swiftness with which disease-gene associations can be mined has improved radically as a result of the availability of discovery human genetic variation data from large-scale public and private initiatives, such as those provided by the International Haplotype Map Consortium and Perlegen Sciences, Inc. (Mountain View, CA). These projects have captured many of the common genetic variants (>1%) in the genome. This information has been buttressed with improvements in large-scale genotyping technologies and statistical methods for data analysis. In summary, the renal community is now poised for discovery of genes for chronic renal failure using these resources.

Semin Nephrol 27:223-236

Keywords: *Genome-wide association, admixture mapping, heterogeneity, environmental correlates*

To map and characterize all the genes in the human genome, an international alliance funded by governmental agencies worldwide was developed and initiated in 1988. The project, named the Human Genome Project (HGP), produced a draft sequence encompassing 90% of the human genome in February 2001, followed by the full sequence in April 2003,[1] using clone-based technology. Commencing soon after, these efforts were paralleled by Celera Genomics (Alameda, CA), a private venture that used shotgun sequencing in the race to complete the human genome.[2] To accommodate, parse, and analyze the data generated by the HGP, technologic platforms for molecular analysis and informatics tools were developed. These data currently reside in the public domain, and is accessible to all interested parties (http://genome.ucsc.edu/, http://www.ncbi.nlm.nih.gov/, http://www.ensembl.org) and has sparked other large-scale genomic endeavors, for example, sequencing of other species,[3-9] structural genome annotation and mining,[10-14] ENCyclopedia of DNA Elements (the ENCODE project),[15-17] and so forth.

*Departments of Epidemiology and Biostatistics, Ophthalmology and Genetics, Case Western Reserve University, Cleveland, OH

†Division of Nephrology and Hypertension, Los Angeles Biomedical Research Institute, Torrance, CA

Supported by research grants from the National Institute of Diabetes and Digestive and Kidney Diseases U01DK57292 and R01DK069844 (S.K.I. and S.G.A.), and from the National Center for Research Resources grant M01 RR00425.

Address reprint requests to Sudha K. Iyengar, PhD, Case Western Reserve University, Wolstein Research Building, 1315, 2103 Cornell Rd, Cleveland, OH 44106-7281. E-mail: ski@case.edu

0270-9295/07/$ - see front matter

 doi:10.1016/j.semnephrol.2007.01.003

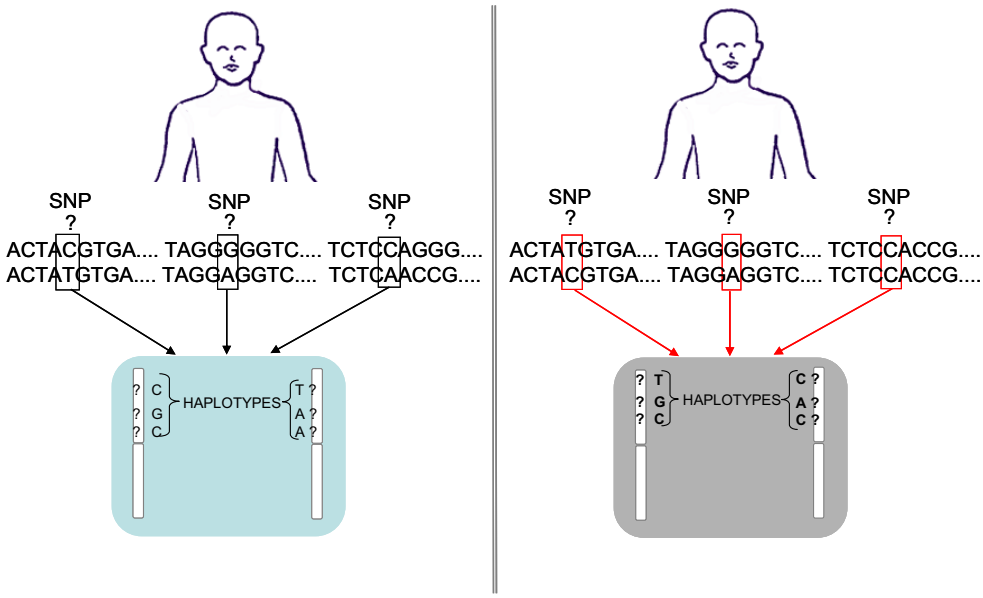

Figure 1. Two individuals with a stretch of DNA on 2 homologous chromosomes are shown. Both show variations at 3 SNPs marked by the grey and black boxes in the left and right panels, respectively. The haplotypes (linear arrangement of alleles at each SNP on the homologous chromosome) for each individual are shown below, omitting all other sequence information on the chromosome.

Although the completion of the HGP was a momentous event in genetics history, the project did not describe how individuals differed from each other at the level of the sequence. This information is crucial to comprehending why some individuals are susceptible to disease while others remain healthy. Thus, in an effort to characterize the extent of natural genetic variation in ethnic populations worldwide, a second large-scale undertaking, the International Haplotype Map (HapMap) Consortium project, was launched in 2003[18-21] (http://www.hapmap.org/). The premise of the HapMap project was to examine variations in 4 different ethnic groups worldwide and to determine where and how frequently sequence changes occurred in the genome.

HAPMAP AND THE COMMON DISEASE–COMMON VARIANT HYPOTHESIS

Even before HGP and HapMap, geneticists had recognized that variations in the DNA sequence are heterogeneous. The human nuclear genome consists of 2 copies each of 22 autosomes ordered sequentially in descending size from 1 to 22 (each matched pair is called a *set of homologous chromosomes*), and either 2 X chromosomes or 1 X and 1 Y chromosome. A parent donates one copy of each autosome and a sex chromosome to each child. The homologs furnished by separate parents usually differ from each other, and from any other individual in the population at approximately 0.1% of the nucleotide sites or 1 change per every 1,000 bases on average.[22-25] Thus, aligning the sequences of both chromosomes from head to tail or short arm (p-arm) to long arm (q-arm), the most elemental variation can be at a single base or nucleotide, such as a G nucleotide on one homolog and a T nucleotide on the other (Fig. 1). Although 4 bases (A, G, T, and C) can contribute to the variation at a particular locus, only 2 of the alleles or variant forms are present in most populations. By using the previous example of the variant G and T alleles, it is thus possible to resolve which individuals are GG, GT, or TT. Extending this concept, the genetic constitution of individuals within the population can be determined at a specific nucleotide by interrogating their DNA individually using molecular assays. This information enables us to record the genotype (pair of alleles) for an individual, and enumerate the frequency for each allele in a population. Other types of se-

Table 1. List of Polymorphisms in the Human Genome

Type of Variant	Frequency in the Genome	Detection Method	Large-Scale Genotyping Method Available
SNP (substitution)	Common in the genome	Multiple methods	Several platforms available;
(Insertion/deletion)			Sequenom, Inc., San Diego, CA; ParAllele Bioscience, Inc, Santa Clara, CA; Affymetrix Inc., Santa Clara, CA; Illumina Inc., San Diego, CA; Applied Biosystems, Foster City, CA
VNTR	Less frequent in the genome	Traditional gel-based methods only	Modest throughput methods available;
Microsatellites Other repeats			Applied Biosystems, agarose gels, sequencing gels
Inversions	Frequency in multiple populations unknown	Traditional methods and SNP-based methods	Fewer samples can be processed;
Translocations Insertion/deletions			Fluorescence in situ hybridization, array comparative genomic hybridization

Abbreviation: VNTR, variable number tandem repeats.

quence changes, such as duplications of segments of chromosomes, deletions of segments of chromosomes, inversions and translocations of chromosomal segments, and duplication of entire chromosomes as observed in some genetic disorders (eg, trisomy 21 in Down syndrome) also are possible.[26-28] For a more extensive description of the potential variants in the genome see Table 1. But the majority of the genome modifications involve changes in single nucleotides as described previously. These variant forms are called *single-nucleotide polymorphisms* (SNPs).

There are about 3 billion nucleotides in the human genome and an estimated 30 million putative variants. Studies surveying worldwide SNP variation contend that 10 to 13 million common SNPs exist, with 1 variant site present on average every 300 nucleotides in cosmopolitan genome comparisons.[29,30] These studies further maintain that alleles present in at least 1% of the chromosomes worldwide represent the majority (90%) of the variations in multiethnic populations representing about 3 million SNPs; other variants are rare, sometimes private to one or a few individuals. The former have been designated as *common variants* and their universality in individuals of diverse ancestry defines their utility in a specific genetic study design to locate disease susceptibility genes. The theory that encapsulates these ideas is generally referred to as the *common disease–com-*

mon variant (CDCV) hypothesis and is described in greater detail later.

The term *mutations* refers to a change in the DNA structure, but the use of this terminology implies that a cellular, perhaps systemic, effect is associated with these alterations that relates to the disease process. In general, the vast majority of SNPs are thought to be neutral and without consequence, but some are undiscovered disease-associated mutations, whether they are situated in exons or in noncoding regions of the genome. Our knowledge of disease pathogenesis is gleaned via insights from previously explored molecules in a variety of contexts. With large portions of the genome being uncharted with respect to function, it is difficult to predict based on prior knowledge which of the 10 million plus SNPs is coupled with a specific disease. Genetic linkage and association studies that contrast individuals with and without disease, or even comparing varying degrees of severity, can fill the gap between the vastly uncharacterized functional aspects of the genome and the role of particular variants in disease. Specifically, they can assist in identifying a relationship between a clinical entity (phenotype) and the location(s) of its molecular determinants.

The CDCV hypothesis asserts that for common diseases such as diabetes, cardiovascular disease, or diabetic nephropathy (DN), the disease-associated mutation(s) is a common SNP(s) that lies somewhere in the genome, and can be isolated by using other SNPs as signposts to lead investigators to its location, with little initial knowledge of the function. Distinguishing CDCV from other genetic study designs are very specific tenets and assumptions.[31] The central premise of the CDCV hypothesis is that variants that cause common diseases are reasonably frequent in the population, ranging from 1% to 10%.[32,33] Another cardinal feature of the hypothesis is that the original disease-causing mutations (the common variants) arose 100,000 years ago in a small group of founders, and that these variants have been propagated to the existing 6 billion descendents. The mutations would be carried on chromosomes from parent to offspring and onto subsequent descendents; other non–mutation-bearing chromosomes also exist.[34,35]

Chromosomes may or may not be inherited intact from parent to offspring; 2 forces, recombination and new mutations, scramble the information on the chromosome partially before passage to each new generation. The theory presumes that affected individuals (even from different families or populations) are in principle related by deep-rooted ancestry at the disease locus, and share a tiny residual piece of the old chromosomal segment surrounding the original mutation. This collective signature at a disease locus shared by the affected individuals may be identified by SNPs in close proximity to the disease allele acting as surrogates because they frequently have been transmitted with the disease allele.

In addition to finding cosmopolitan common variants, the responsibility of the HapMap was to investigate whether information regarding shared haplotypes could be inferred from contiguous SNPs. A consecutive arrangement of alleles at different SNPs on the same chromosome is considered a haplotype (Fig. 1), hence the name HapMap. The HapMap characterized the 13 million common SNPs (without knowledge of which were mutations) to determine how big these tracks of shared DNA segments were across 4 multi-ethnic populations. The conclusion of the HapMap was that DNA is inherited in blocks (haplotypes) that are shared by individuals of diverse ancestries. The length of the blocks and the number of such shared haplotypes varies among and between populations and is discussed further in section Comparing Ethnic Groups and Availability of Cosmopolitan Maps. In general the architects of the HapMap concluded that it should be possible to use this information to map common disease genes. In summary, disease-associated variants could be localized via an indirect method of interrogating surrogate SNPs residing on common cosmopolitan haplotypes identified through the HapMap.

As an example, if one considers a simple case-control study, individuals who are affected would share the affected haplotype, and those without disease would be depleted in that particular haplotype. This frequency difference could be captured in association studies that contrast case versus control individuals using SNP markers. There have been many debates in

the literature as to which SNPs carry the most information necessary to facilitate inference about the frequency difference of the causal allele,[36-44] the most common of which are tag SNPs. Tag SNPs are a selected subset of SNPs that are able to represent information from multiple SNPs in the general vicinity because there is sufficient redundancy in the information carried by the SNPs that genotyping all of them is unnecessary and wasteful. Some study designs for various types of association studies are described later.

There exists an opposing view to the CDCV hypothesis, called the *common disease–rare variant* (CDRV) hypothesis.[34,45] In this scenario, rare mutations and environmental factors are the biggest contributors to common disease. Rare mutations can either have a large effect, as seen in autosomal-dominant polycystic kidney disease, or may act in concert with many other genes such that each only contributes a little to the overall effect. If this theory holds true for common diseases, then mapping of disease genes through the use of surrogate SNPs may not be feasible, especially using unrelated individuals with and without disease because there will be very few common genetic features among them that can be captured. In this case, a previously successful strategy of using individuals of known relationships, such as in families, or an inbred population, or a founder population, will provide more information because the actual inheritance of entire chromosomes from recent ancestors can be traced.

A review of the recent literature supports both the CDCV and the CDRV hypotheses for many chronic diseases. Exemplifying the CDRV theory is the discovery of a gene in the Mexican American population for type 2 diabetes[46]: calpain-10. The initial discovery was subject to controversy but now has been replicated,[47] and its role in diabetes pathogenesis has been established. In contrast, 2 other genes, the peroxisome proliferative–activated receptor gamma gene[48] and the transcription factor 7–like 2 gene,[49] were identified following the CDCV paradigm. For the latter, replication of the association between the gene and type 2 diabetes swiftly followed.[49-58] These examples show that disease gene mapping for both rare and common variants is now tractable using the resources generated by large-scale projects, such as the HapMap, although replication still may prove to be difficult for rare variants. It will depend on the strength of the association signal, the effect size and penetrance of the gene, and the history of the population in whom the original discovery was made.

COMPARING ETHNIC GROUPS AND AVAILABILITY OF COSMOPOLITAN MAPS

The practical consideration behind these developments was to reduce the genotyping from 30 million SNPs in cosmopolitan populations to a manageable number that can be bundled in massively parallel assays. For this strategy to be cost effective, these SNPs must be portable across populations, have usable allele frequencies, and show similar haplotype structures across multiple ethnic groups. To determine the utility of this information, multiple independent surveys were conducted that differed in choice of SNPs, sample size, and the ethnic group(s) under examination.[31,59-64] The 2 largest surveys were by a private firm, Perlegen Sciences, Inc. (Mountain View, CA), with 1.6 million SNPs in 71 samples,[64] and by the publicly funded International HapMap Project, with 1 million SNPs in phase I[65] and 13 million SNPs in phase II[20] in 269 samples. Although immensely useful in generating information on millions of SNPs, a limitation of these surveys was the restricted number of individuals in the discovery sets.

Based on these surveys, genotyping technology has graduated from 10,000 SNPs to 100,000 plus SNPs multiplexed in single assays, with the promise of 1 million or more SNPs being available in comprehensive packages.[66-73] More modest platforms for replication and smaller-scale studies also are available.[74,75] The question of portability and suitability of the tag SNPs for multi-ethnic cohorts remains controversial.[76-80] Two recent studies by Conrad et al[80] and de Bakker et al[81] endeavored to provide a more comprehensive view of worldwide variation by examining worldwide populations and multi-ethnic cohorts, but limited the number of genes or regions examined. Conrad et al[80] examined 2,834 SNPs distributed in 36 genomic regions in 927 individuals, and de Bakker et al[81]

examined 1,679 SNPs in 25 genes in 1,000 DNA samples from the Multi-Ethnic Cohort study with 15 member populations. The consensus from all these studies was that the 4 representative populations tested in the HapMap, Caucasians, Chinese, Yoruba, and Japanese, possess the majority of the common haplotypes that will be useful for designing association studies, although significant variability exists between populations. Individuals of African descent had the greatest amount of variability and the least amount of representation in the HapMap. The general patterns of linkage disequilibrium (LD), haplotype block structure, and hotspots for recombination were similar across populations, with reduced LD and smaller blocks in populations of African descent. The consequence of low LD in populations of African descent is that a larger number of SNPs are necessary to obtain sufficient genome-wide coverage so as not to miss an association signal.

Because of the variability in allele frequencies of SNPs across populations, comparing data between populations on association results is a difficult task, and some have recommended that a common map that is less dependent on allele frequencies be developed using a common LD metric.[82-84] Such a map would provide a basis for a comparison of cases and controls across multi-ethnic cohorts. The strategy has not been used commonly, but is likely to come into wider use as predesigned and post hoc consortia with larger samples and a more extensive ethnic diversity come into existence. The motivation to aggregate samples from modest collections is their ability to leverage the superior numbers to find genes with smaller effect sizes. Transethnic comparisons can assist in localizing the disease variant in 2 ways. First, comparing the difference in LD between populations can hone the association signal by providing a better resolution of the LD contrast between cases and controls or between families of diverse ancestry. McKenzie et al[85] used this strategy at the angiotensin converting enzyme (ACE) locus, where the SNPs were in significant LD, and many individuals showed broad segments of DNA being shared in the British families. They were mapping circulating ACE levels to the structural locus for ACE on chromosome 17q23. By using transethnic comparisons from an African Caribbean population they narrowed the area of interest to a region near a specific variant G2350A. This type of analysis can be extended to other cohorts. The second use of transethnic comparisons is for deciding which SNPs are causal after sequencing experiments. If a common variant is the cause of the disease in question, then transethnic comparison of the sequence should show common profiles between cases from many ethnic groups. Variants that are not shared between these groups could be eliminated from consideration, or at least receive a lower priority for analysis. If a rare variant is the cause of disease, then the first rationale may be valid, but the direct sequence comparison from individuals of diverse ancestries may not be meaningful.

ASSOCIATION STRATEGIES FOR MAPPING GENES FOR DIABETIC NEPHROPATHY

There are multiple study designs that have been used for disease gene mapping that contrast information between cases and controls to formulate connections between a gene (or a region on a chromosome, namely, locus) and disease. Named *association mapping*, these include: (1) candidate gene analysis using individual genotyping, (2) whole-genome scanning using individual genotyping, (3) whole-genome or candidate gene association using DNA pooling, and (4) admixture mapping. The basis for all these methods is that genes for the disease (eg, DN) reside in the genome. The difference between these methods is in the genetic assumptions that are made when initiating a particular design. The first 3 association mapping methods make very similar assumptions. The most simple example is that of a molecule that has been studied through biochemical means and shown to be expressed in a relevant functional region in the kidney (eg, the podocyte for DN) and is nominated as a candidate for disease causation (a functional candidate gene). This is the candidate gene approach and is most appealing to scientists who can follow the biological rationale. However, the drawback of this method is that a limited number of candidate genes are characterized functionally at the biochemical level. A method that would canvas the

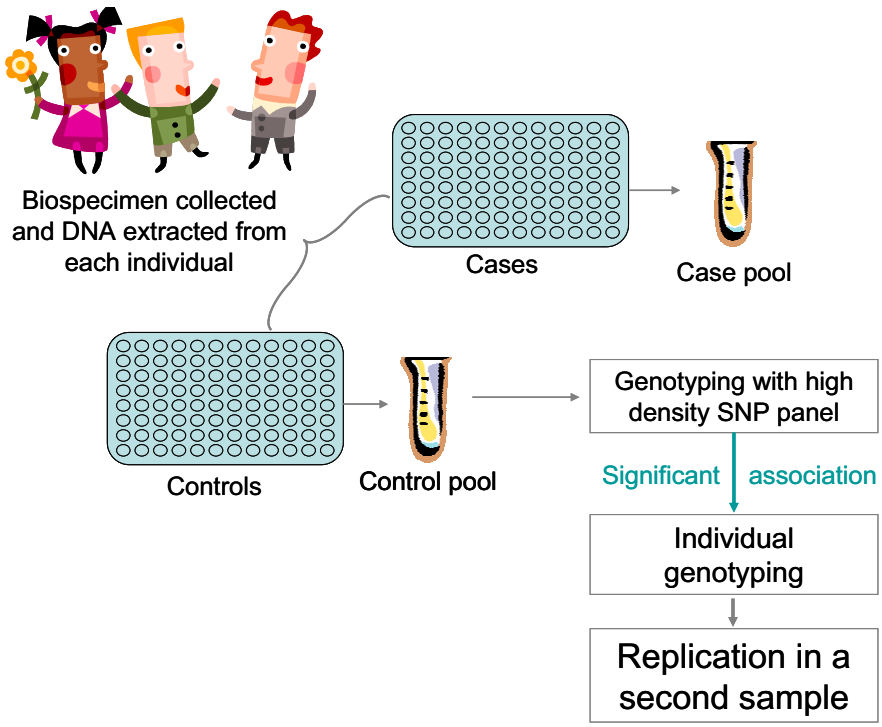

Figure 2. A study design incorporating DNA pooling. Cases and controls are collected from a population and the DNA is extracted individually. The DNA is arrayed on multisample formatted plates and then eventually pooled into a case pool and a control pool. The pools are subject to genotyping using a high-density SNP panel and the results were evaluated using statistical methods. The best signals are prioritized for follow-up evaluation using individual genotyping. The best results from this set are subject to replication in a second (and perhaps third) sample to confirm the validity of the results.

entire genome rather than a handful of genes is intuitively more interesting because the number of genes in the human genome is large (30,000 – 40,000 projected genes). As described previously, it is not feasible to sequence the entire genome in large cohorts, so a genome-wide association study is merely an extension of the candidate gene approach to both other known and unknown genes; in most cases non–gene-bearing regions also are covered, but pursuing the functional consequence of a polymorphism in a region that does not bear a gene is a more difficult task. As described earlier, this method would rely on shared ancestry around the disease-causing mutation or LD. The latter method does not require any knowledge of biological function, and may in fact lead to novel discoveries. The third method is very similar in concept but an additional cost-saving step is added. DNA from individuals is pooled into case-pools and control-pools (Fig. 2). Thus, it only is necessary to genotype the pools versus the actual individual samples. The last method, admixture mapping, can be applied only to admixed populations, such as African Americans and Mexican Americans, who have ancestry reflecting diverse geographic origins and a set of alleles that derive from distinct parental populations, such that the ancestry can be inferred (Fig. 3). Many of these designs are being used by the Family Investigation of Nephropathy and Diabetes (FIND) study to identify genes for DN. To simplify the explanation we only discuss how to assess differences between unrelated cases and controls, although case-only methods also exist. In 1999, before the publication of the complete human genome, the National Institute of Diabetes and Digestive and Kidney Diseases funded the FIND study to recruit cohorts of subjects to identify DN chromosomal risk loci. FIND comprised 3 study designs. It included a cohort of siblings who could be either concordant or discordant for the DN phenotype, and with whose DNA classic

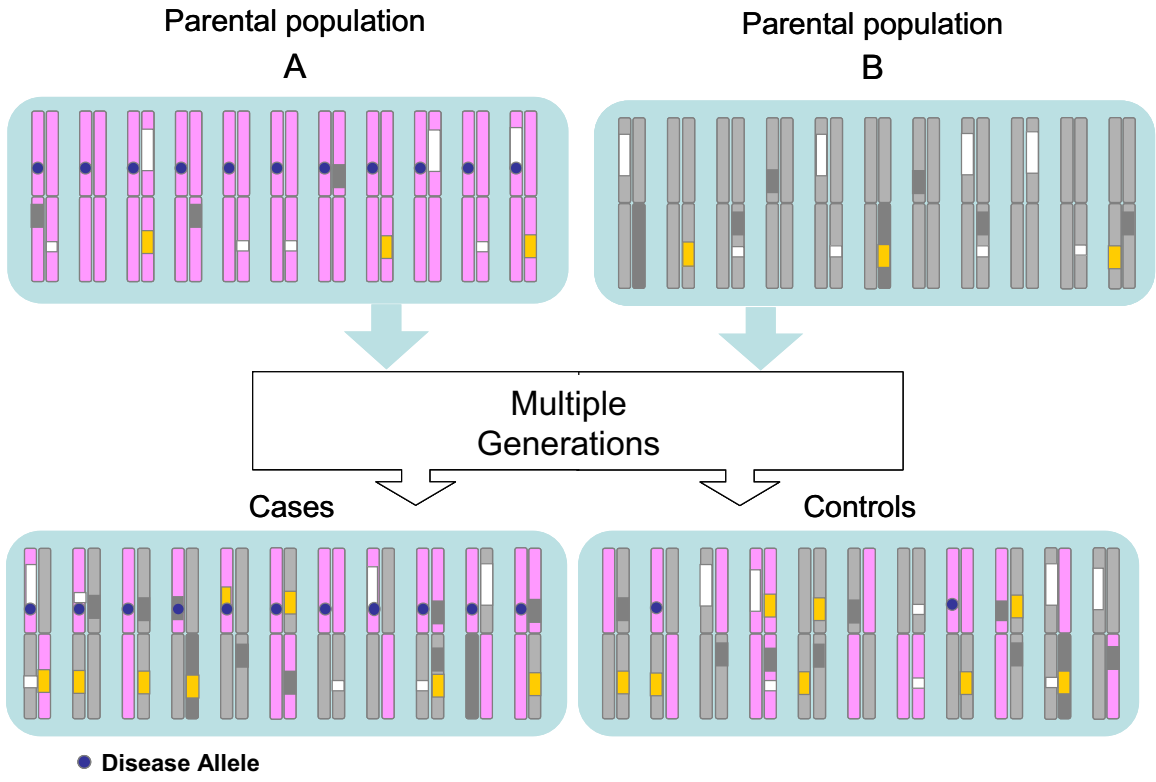

Figure 3. MALD is based on the premise that one of the parental populations has a higher degree of risk because the ancestral (parental) population had a risk allele at a particular locus. (A) Above in grey is a parental population A, which has a higher risk than parental population B. This information is depicted with a black dot on each parental chromosome (the site of the mutation). After multiple generations of intermixing the extant admixed population consists of cases that are enriched for the mutation (black dot), compared with the controls. This information tracks with the ancestry of these individuals, illustrated by the shared grey haplotype. Thus, finding the regions of common ancestry would enable us to locate the regions that are shared around mutations that arose in the parental population many generations ago.

linkage analysis studies were performed.[86] This study design also is described in the article entitled "Mining the genome for susceptibility to diabetic nephropathy: The role of large-scale studies and consortia," in this issue. In addition, FIND included 2 case-control (association) cohorts of Mexican Americans and African Americans in whom DN chromosomal risk loci were to be sought by the genetic strategy of mapping by admixture linkage disequilibrium (MALD, see later). An additional strategy, transmission disequilibrium testing, is another method for seeking disease risk genes. At the time of FIND's inception, it required the collection of DNA from an index patient with the disease and both parents. Given that FIND sought nephropathy genes in subjects with type 2 diabetes, the recruitment of both parents of these middle-aged to elderly probands was impractical, and this strategy was not used. In the intervening time, this strategy has become possible, owing to genetic analytic techniques capable of reconstructing the parent genome if enough information from siblings is available.[87]

The best genetic strategy to use to find susceptibility genes in a given disease is dependent in large part on whether the disease reflects contributions from one or a few genes (eg, monogenic or oligogenic), or is polygenic. To date, a single gene with a large effect or a few genes with moderate effects have not been identified consistently for DN. One exception is the identification of polymorphisms for the ACE gene, which in at least one meta-analysis has been estimated to contribute approximately 30% to 40% of the identifiable genetic risk,[88] although it is not always implicated in candidate gene studies, nor always implicated in the

full genome scans that have been completed. Another exception is the recent identification of carnosinase as a risk factor for DN.[89] Most common diseases are not caused by single gene mutations, but probably reflect the cumulative effects of many genes, several of which are likely to be found commonly in individuals in the general population in the absence of disease. These genes also may interact with environmental factors, further complicating the identification of the genetic factors. Candidate gene strategies may be useful even if the overall contributing effect of the gene is small.[90]

Finding susceptibility genes in complex diseases also is influenced by the sibling (or other relative) recurrence risk ratio (the risk of expressing the phenotype [eg, DN] in siblings divided by the prevalence of the general population); the genotypic risk ratio (the risk of expressing the phenotype if the gene is present in an individual divided by the risk of not expressing the phenotype if the gene is not present); the number of susceptibility genes; how the susceptibility genes interact; how much the disease is monogenic; and, depending on the experimental design, the penetrance. These data are not defined precisely for DN, despite many years of research. As a result, using multiple genetic strategies with complementary approaches that use genetic material from independent patient pools of large number may optimize the probability of locating DN genes.

Case-control association studies are particularly advantageous in oligogenic and polygenic models with low genotypic relative risk and moderate to high allele frequency in the population. Thus, in situations of low genotypic relative risk and high frequency of the gene in the population, hundreds of subjects may provide sufficient power to detect susceptibility genes using association studies compared with the potential need for thousands using sibling pairs (linkage). Case-control studies also may be superior to linkage approaches in defining the genetic effects of specific loci. However, association studies have specific shortcomings. The major potential disadvantages of association strategies are as follows: (1) LD may extend for only short genomic intervals; and (2) case-control methods may be confounded by unanticipated genetic differences between the proband and control subjects (population stratification).[91] Population stratification as a result of ethnic differences can result in spurious SNP associations. Recent progress in the ability to detect and correct for population stratification has provided a potential solution to this problem in case-control studies. By genetically defining the populations using molecular markers and appropriately selecting case and control samples to balance the origin of ancestry, the effect of stratification can be minimized, thereby avoiding the identification of false-positive associations.[91-93] *Nature Genetics*[94] published suggested guidelines for the performance of association studies to minimize the likelihood of false-positive results. Association studies should have "large samples sizes, small p values, report associations that make biologic sense and alleles that affect the gene product in a physiologically meaningful way. In addition, they should contain an initial study as well as an independent replication, the associations should be observed in both family-based and population-based studies, and the odds ratios and/or the attributable risk should be high."[94] Attributable risk is defined as that portion of the risk that can be eliminated if the risk factor (in this case the susceptibility allele) is removed. Research using SNP haplotype markers for genomic scanning also recently has been facilitated by the discovery of enough SNPs to provide a sufficiently dense genome-wide SNP marker map to cover the genome adequately. Costs for the performance of these full genome scans are becoming more practical, although pooling of samples for an initial analysis, followed by individual genotyping with a smaller number of promising markers as a follow-up analysis, is one way to cut costs even at this time. Initial genotyping requires on the order of 250,000 to 600,000 SNP markers, with the current trend nearing 1 million markers in 2007.

Although many methods have been used for SNP genotyping, including gel-based assays for single-strand conformation polymorphism, enzyme cleavage methods, mass spectrometry, allele-specific polymerase chain reaction, single nucleotide primer extensions, and oligonucleo-

tide ligation and pyrosequencing, most of these methods are not cost effective for large-scale, genome-wide analyses. Of these, only real-time polymerase chain reaction and mass spectrometry are suitable for inferring allele frequency differences using pooled DNA samples.[95,96] High-density oligonucleotide arrays offer the advantage of being able to process large numbers of SNP markers in parallel using automated methods across multiple samples simultaneously.[25] One method involves light-directed photolithography in conjunction with chemical coupling to direct the synthesis of a high density of oligonucleotides of specific DNA sequence in predetermined positions on a glass surface.[97,98] Hybridization of labeled DNA targets to these arrays then allows effective and accurate genotyping of SNP alleles.[25,99]

Choosing the SNP markers to cover the genome adequately is a challenge. In one study,[93] 250,000 SNPs were used that were in Hardy-Weinberg equilibrium for every population studied. The selected SNPs are distributed evenly across the genome, with a spacing of 1 SNP every 10 kb. Each selected SNP has a minor allele frequency of 10% or greater, and was chosen to maximize information concerning the haplotype structure in the vicinity of the SNP. This strategy was used to identify candidate regions in association with HDL cholesterol levels. The study design incorporated stratification analysis, pooled genotyping, confirmation of promising candidate loci by individual genotyping, and replication in an independent cohort.[93] By using a similar study design, we are conducting an ancillary study to the FIND study by using the same Mexican American cohort that was collected for gene mapping using the MALD technique. Although the 250,000 markers described previously are anticipated to provide excellent power to cover the genome and to identify chromosomal loci for DN risk, the application of this technique to the problem of DN is considerably more complex than what has been achieved previously for high-density lipoprotein (HDL) cholesterol association. This DN study involves an initial population of a similar order of magnitude to the earlier-described HDL study (eg, 475 subjects with and 475 subjects without DN), and a replicate population with the same number of subjects as the initial population also has been proposed.

There are some differences worth contrasting between studies. First, the HDL cholesterol work used SNP haplotyping to focus on 71 specific candidate gene regions, whereas the DN scans will use the technical and analytic strategies that cover the full genome. Second, the HDL cholesterol work used pooled samples from a total of 345 low HDL and 321 high HDL samples for the initial study and a considerably smaller replicate group.[93] Finally, the replicate population in the HDL cholesterol study was ethnically similar to the population of the initial study. In the DN study, the replicate population will be intentionally of a different ethnicity than the initial population. Although most of the published studies identifying chromosomal risk loci are confounded by ethnic differences that limit reproducibility, the SNP haplotype genomic scan envisioned in this case-control population, by using SNP markers that transcend ethnic differences, is intended to provide information regarding risk loci that are replicable across ethnicities, rather than those that are ethnicity specific. These techniques are only just being scaled up for application to full genomic scanning in large populations. The primary analysis will consist of logistic regressions with DN status (as a dichotomous trait) as the outcome, against SNP genotype plus terms for relevant covariates representing known risk factors for DN. Novel analytic techniques may be required to best interpret the data that emerge. A high-density genome-wide scan to identify SNPs that associate with the predisposition to DN is a first step toward discovering genes that play a role in the disease process. Additional efforts to understand genetic variations and disease processes then can focus on these loci to discover the causative mutations through disequilibrium analysis with the associated SNP, and on follow-up studies to ascertain the biological importance of these genes.

MALD

The prevalence and incidence of DN is well known to differ among ethnic groups, supporting the concept of a genetic contribution. However,

it also suggests that a strategy that uses special genetic markers that attempt to link disease risk to ancestral inheritance of particular polymorphisms may be a powerful tool to help to identify these loci. Thus, this strategy is used in what are termed *admixed* populations (eg, those populations in which the genome shows ancestry from 2 continents). For instance, the genome in Mexican Americans represents contributions from both Amerindian and European ancestors. The genome in African Americans represents contributions from both African and European ancestors. In any given individual from a so-called *admixed* ethnic group, the relative proportion of the ancestries differ at the genome level, such that an individual may be either more like one of the parental populations or some intermediate mix of the two.

Admixture mapping, or MALD, is a special type of case-control association study (Fig. 3). It traditionally has been used by gene mappers in areas other than human genetics (eg, intercrosses and back-crosses). Admixture mapping requires that the parentage (origin) of the genomes that contribute to the hybrid population be known and that the 2 (or more) parents be distinguishable at the genetic level at several, preferably many, loci. The easier it is to distinguish between the 2 parental populations, the greater the power of the method. This method assumes that one of the parental populations has a higher risk for disease (eg, DN) than the other, and that the locus (or loci) contributing to this excess risk tracks with the ancestry of the chromosomal segments in the hybrid population. Special markers for genomic scanning are required for this type of analysis, and have been developed for African Americans,[100,101] and for Mexican Americans,[102] although for the latter, many more markers have since been developed (M. Seldin, unpublished data). It has been argued that admixture mapping may offer more power to detect risk loci than linkage analysis for polygenic disorders, although requiring significantly fewer markers, providing an economic advantage.[103,104] The use of admixture mapping recently resulted in the identification of a chromosomal locus predisposing to prostate cancer in an African American cohort.[104] Within the FIND study, admixture mapping is currently in progress to identify DN risk loci in African American and Mexican subjects.

There are some caveats to admixture mapping, including location of genes/regions other than those bearing disease-risk alleles. If a region in the genome undergoes selection for another cause (eg, an infectious disease), then it is possible that unidentified differences in the parental population that are unrelated to disease per se could be selected for and lead to spurious results in admixture mapping. The increase in the density of markers through HapMap has increased the information being gained for admixture mapping and there are some methods to guard against identification of genes (loci) unrelated to disease, the primary among which would be replication in an independent sample.

CONCLUSIONS

The past decade witnessed the generation of maps of the human genome. Limited genome-wide analyses for DN and numerous candidate gene studies have been performed. Large banks of genomic samples from patients of varying ethnicities with and without diabetes and DN have now been assembled. The difficult task of the next decade will be to use the various genomic markers and analytic strategies to attempt to define the complex nature of the relationship between genes, gene expression, environmental impact, and the disease phenotype of DN.

Acknowledgment

The authors would like to thank Paula Wedig and Sarah Ialacci for technical assistance with the manuscript.

REFERENCES

1. The International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431:931-45.
2. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science. 2001;291:1304-51.
3. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;437:69-87.
4. Culotta E. Genomics. Chimp genome catalogs differences with humans. Science. 2005;309:1468-9.

5. Dennis C. Chimp genome: branching out. Nature. 2005;437:17-9.
6. Pennisi E. Evolution. Chimp genome draft online. Science. 2003;302:1876.
7. Xuan Z, Wang J, Zhang MQ. Computational comparison of two mouse draft genomes and the human golden path. Genome Biol. 2003;4:R1.
8. Carucci DJ, Gardner MJ, Tettelin H, et al. Sequencing the genome of Plasmodium falciparum. Curr Opin Infect Dis. 1998;11:531-4.
9. Hillier LW, Miller W, Birney E, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432:695-716.
10. Carson AR, Feuk L, Mohammed M, et al. Strategies for the detection of copy number and other structural variants in the human genome. Hum Genomics. 2006;2:403-14.
11. Derti A, Roth FP, Church GM, et al. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat Genet. 2006;38:1216-20.
12. Sharp AJ, Locke DP, McGrath SD, et al. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005;77:78-88.
13. She X, Jiang Z, Clark RA, et al. Shotgun sequence assembly and recent segmental duplications within the human genome. Nature. 2004;431:927-30.
14. Tuzun E, Sharp AJ, Bailey JA, et al. Fine-scale structural variation of the human genome. Nat Genet. 2005;37:727-32.
15. Guigo R, Flicek P, Abril JF, et al. EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 2006;7 Suppl 1:S2-31.
16. Yoo YK, Ke X, Hong S, et al. Fine-scale map of encyclopedia of DNA elements regions in the Korean population. Genetics. 2006;174:491-7.
17. Harrow J, Denoeud F, Frankish A, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006;7 Suppl 1:S4-9.
18. The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. Nat Rev Genet. 2004;5:467-75.
19. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437:1299-320.
20. Thorisson GA, Smith AV, Krishnan L, et al. The International HapMap Project Web site. Genome Res. 2005;15:1592-3.
21. The International HapMap Consortium. The International HapMap Project. Nature. 2003;426:789-96.
22. Cargill M, Altshuler D, Ireland J, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet. 1999;22:231-8.
23. Halushka MK, Fan JB, Bentley K, et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet. 1999;22:239-47.
24. Li WH, Sadler LA. Low nucleotide diversity in man. Genetics. 1991;129:513-23.
25. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science. 1998;280:1077-82.
26. Mills RE, Luttig CT, Larkins CE, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006;16:1182-90
27. Locke DP, Sharp AJ, McCarroll SA, et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet. 2006;79:275-90.
28. Hinds DA, Kloek AP, Jen M, et al. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet. 2006;38:82-5.
29. Kruglyak L, Nickerson DA. Variation is the spice of life. Nat Genet. 2001;27:234-6.
30. Reich DE, Gabriel SB, Altshuler D. Quality and completeness of SNP databases. Nat Genet. 2003;33:457-8.
31. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet. 2001;17:502-10.
32. Lander ES. The new genomics: global views of biology. Science. 1996;274:536-9.
33. Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. 1998;8:1229-31.
34. Wright AF, Hastie ND. Complex genetic diseases: controversy over the Croesus code. Genome Biol. 2001;2:comment2007.
35. Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. Nat Rev Genet. 2004;5:89-100.
36. Sham PC, Ao SI, Kwan JS, et al. Combining functional and linkage disequilibrium information in the selection of tag SNPs. Bioinformatics. 2006;129-131.
37. Willer CJ, Scott LJ, Bonnycastle LL, et al. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. Genet Epidemiol. 2006;30:180-90.
38. Wiltshire S, de Bakker PI, Daly MJ. The value of gene-based selection of tag SNPs in genome-wide association studies. Eur J Hum Genet. 2006;14:1209-14.
39. Moller S, Koczan D, Serrano-Fernandez P, et al. Selecting SNPs for association studies based on population frequencies: a novel interactive tool and its application to polygenic diseases. In Silico Biol. 2004;4:417-27.
40. Tantoso E, Yang Y, Li KB. How well do HapMap SNPs capture the untyped SNPs? BMC Genomics. 2006;7:238.
41. Xu H, Gregory SG, Hauser ER, et al. SNPselector: a web tool for selecting SNPs for genetic association studies. Bioinformatics. 2005;21:4181-6.
42. Butler JM, Bishop DT, Barrett JH. Strategies for selecting subsets of single-nucleotide polymorphisms

to genotype in association studies. BMC Genet. 2005;6 Suppl 1:S72.
43. Burkett KM, Ghadessi M, McNeney B, et al. A comparison of five methods for selecting tagging single-nucleotide polymorphisms. BMC Genet. 2005;6 Suppl 1:S71.
44. Hampe J, Schreiber S, Krawczak M. Entropy-based SNP selection for genetic association studies. Hum Genet. 2003;114:36-43.
45. Terwilliger JD, Weiss KM. Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol. 1998;9:578-94.
46. Horikawa Y, Oda N, Yu L, et al. Genetic variations in calpain-10 gene are not a major factor in the occurrence of type 2 diabetes in Japanese. J Clin Endocrinol Metab. 2003;88:244-7.
47. Weedon MN, Schwarz PE, Horikawa Y, et al. Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. Am J Hum Genet. 2003;73:1208-12.
48. Altshuler D, Hirschhorn JN, Klannemark M, et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet. 2000;26:76-80.
49. Grant SF, Thorleifsson G, Reynisdottir I, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet. 2006;38: 320-3.
50. Cauchi S, Meyre D, Dina C, et al. Transcription factor TCF7L2 genetic study in the French population: expression in human beta-cells and adipose tissue and strong association with type 2 diabetes. Diabetes. 2006;55:2903-8.
51. Damcott CM, Pollin TI, Reinhart LJ, et al. Polymorphisms in the transcription factor 7-like 2 (TCF7L2) gene are associated with type 2 diabetes in the Amish: replication and evidence for a role in both insulin secretion and insulin resistance. Diabetes. 2006;55:2654-9.
52. Florez JC, Jablonski KA, Bayley N, et al. TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program. N Engl J Med. 2006; 355:241-50.
53. Groves CJ, Zeggini E, Minton J, et al. Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. Diabetes. 2006;55:2640-4.
54. Humphries SE, Gable D, Cooper JA, et al. Common variants in the TCF7L2 gene and predisposition to type 2 diabetes in UK European Whites, Indian Asians and Afro-Caribbean men and women. J Mol Med. 2006;1-10.
55. Saxena R, Gianniny L, Burtt NP, et al. Common single nucleotide polymorphisms in TCF7L2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. Diabetes. 2006;55:2890-5.
56. Scott LJ, Bonnycastle LL, Willer CJ, et al. Association of transcription factor 7-like 2 (TCF7L2) variants with type 2 diabetes in a Finnish sample. Diabetes. 2006;55:2649-53.
57. van Vliet-Ostaptchouk JV, Shiri-Sverdlov R, Zhernakova A, et al. Association of variants of transcription factor 7-like 2 (TCF7L2) with susceptibility to type 2 diabetes in the Dutch Breda cohort. Diabetologia. 2006;59-62.
58. Zhang C, Qi L, Hunter DJ, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene and the risk of type 2 diabetes in large cohorts of U.S. women and men. Diabetes. 2006;55:2645-8.
59. Johnson GC, Esposito L, Barratt BJ, et al. Haplotype tagging for the identification of common disease genes. Nat Genet. 2001;29:233-7.
60. Patil N, Berno AJ, Hinds DA, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science. 2001;294: 1719-23.
61. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science. 2002;296:2225-9.
62. Carlson CS, Eberle MA, Kruglyak L, et al. Mapping complex disease loci in whole-genome association studies. Nature. 2004;429:446-52.
63. Ke X, Cardon LR. Efficient selective screening of haplotype tag SNPs. Bioinformatics. 2003;19:287-8.
64. Hinds DA, Stuve LL, Nilsen GB, et al. Whole-genome patterns of common DNA variation in three human populations. Science. 2005;307:1072-9.
65. Altshuler D, Clark AG. Genetics. Harvesting medical information from the human family tree. Science. 2005;307:1052-3.
66. Lamy P, Andersen CL, Wikman FP, et al. Genotyping and annotation of Affymetrix SNP arrays. Nucleic Acids Res. 2006;34:e100.
67. Nicolae DL, Wen X, Voight BF, et al. Coverage and characteristics of the Affymetrix GeneChip Human Mapping 100K SNP set. PLoS Genet. 2006;2:e67.
68. Fan JB, Gunderson KL, Bibikova M, et al. Illumina universal bead arrays. Methods Enzymol. 2006; 410:57-73.
69. Galver LM, Ng PC, Kuhn KL, Gunderson R, Shen R, Murray SS. A 300K phase I HapMap tag SNP panel. San Diego, CA: Illumina, Inc.; 2006.
70. Gunderson KL, Steemers FJ, Ren H, et al. Whole-genome genotyping. Methods Enzymol. 2006;410: 359-76.
71. Gunderson KL, Kuhn KM, Steemers FJ, et al. Whole-genome genotyping of haplotype tag single nucleotide polymorphisms. Pharmacogenomics. 2006;7: 641-8.
72. Peiffer DA, Le JM, Steemers FJ, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res. 2006;16:1136-48.
73. Peacock E, Whiteley P. Perlegen sciences, Inc. Pharmacogenomics. 2005;6:439-42.
74. Tobler AR, Short S, Andersen MR, et al. The SNPlex

genotyping system: a flexible and scalable platform for SNP genotyping. J Biomol Tech. 2005;16: 398-406.
75. Del V, Lazaruk KD, Rhodes MD, et al. Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. Mutat Res. 2005;573: 111-35.
76. Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat Genet. 2006;38: 659-62.
77. Sawyer SL, Mukherjee N, Pakstis AJ, et al. Linkage disequilibrium patterns vary substantially among populations. Eur J Hum Genet. 2005;13:677-86.
78. Bonnen PE, Pe'er I, Plenge RM, et al. Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. Nat Genet. 2006; 38:214-7.
79. Gonzalez-Neira A, Ke X, Lao O, et al. The portability of tagSNPs across populations: a worldwide survey. Genome Res. 2006;323-330.
80. Conrad DF, Jakobsson M, Coop G, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet. 2006; 38:1251-60.
81. de Bakker PI, Burtt NP, Graham RR, et al. Transferability of tag SNPs in genetic association studies in multiple populations. Nat Genet. 2006;38:1298-303.
82. Tapper W, Collins A, Gibson J, et al. A map of the human genome in linkage disequilibrium units. Proc Natl Acad Sci U S A. 2005;102:11835-9.
83. Maniatis N, Morton NE, Gibson J, et al. The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. Hum Mol Genet. 2005;14:145-53.
84. Zhang W, Collins A, Maniatis N, et al. Properties of linkage disequilibrium (LD) maps. Proc Natl Acad Sci U S A. 2002;99:17004-7.
85. McKenzie CA, Abecasis GR, Keavney B, et al. Transethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). Hum Mol Genet. 2001;10:1077-84.
86. Knowler WC, Coresh J, Elston RC, et al. The Family Investigation of Nephropathy and Diabetes (FIND): design and methods. J Diabetes Complications. 2005;19:1-9.
87. Burdick JT, Chen WM, Abecasis GR, et al. In silico method for inferring genotypes in pedigrees. Nat Genet. 2006;38:1002-4.
88. Fujisawa T, Ikegami H, Kawaguchi Y, et al. Meta-analysis of association of insertion/deletion polymorphism of angiotensin I-converting enzyme gene with diabetic nephropathy and retinopathy. Diabetologia. 1998;41:47-53.
89. Janssen B, Hohenadel D, Brinkkoetter P, et al. Carnosine as a protective factor in diabetic nephropathy: association with a leucine repeat of the carnosinase gene CNDP1. Diabetes. 2005;54:2320-7.
90. Krolewski AS. Genetics of diabetic nephropathy: evidence for major and minor gene effects. Kidney Int. 1999;55:1582-96.
91. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 1999;65:220-8.
92. Ardlie KG, Lunetta KL, Seielstad M. Testing for population subdivision and association in four case-control studies. Am J Hum Genet. 2002;71:304-11.
93. Hinds DA, Stokowski RP, Patil N, et al. Matching strategies for genetic association studies in structured populations. Am J Hum Genet. 2004;74:317-25.
94. The Nature Genetic Journal. Freely associating. Nat Genet. 1999;22:1-2.
95. Germer S, Holland MJ, Higuchi R: High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. Genome Res. 2000; 10:258-66.
96. Laken SJ, Jackson PE, Kinzler KW, et al. Genotyping by mass spectrometric analysis of short DNA fragments. Nat Biotechnol. 1998;16:1352-6.
97. Fodor SP, Read JL, Pirrung MC, et al. Light-directed, spatially addressable parallel chemical synthesis. Science. 1991;251:767-73.
98. Lipshutz RJ, Fodor SP, Gingeras TR, et al. High density synthetic oligonucleotide arrays. Nat Genet. 1999;21:20-4.
99. Mei R, Galipeau PC, Prass C, et al. Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. Genome Res. 2000;10: 1126-37.
100. Smith MW, Patterson N, Lautenberger JA, et al. A high-density admixture map for disease gene discovery in African Americans. Am J Hum Genet. 2004; 74:1001-13.
101. Smith MW, Lautenberger JA, Shin HD, et al. Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. Am J Hum Genet. 2001;69:1080-94.
102. Collins-Schramm HE, Chima B, Morii T, et al. Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. Hum Genet. 2004;114:263-71.
103. Smith MW, O'Brien SJ. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat Rev Genet. 2005;6:623-32.
104. Freedman ML, Haiman CA, Patterson N, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci U S A. 2006;103:14068-73.