Copyright © Munksgaard 2000
CLINICAL GENETICS
ISSN 0009-9163

Mini Review

Single nucleotide polymorphisms and the future of genetic epidemiology

Schork NJ, Fallin D, Lanchbury S. Single nucleotide polymorphisms and the future of genetic epidemiology.
Clin Genet 2000: 58: 250–264. © Munksgaard, 2000

In this review, we consider the motivation behind contemporary single nucleotide polymorphism (SNP) initiatives. Many of these initiatives are projected to involve large, population-based surveys. We therefore emphasize the utility of SNPs for genetic epidemiology studies. We start by offering an overview of genetic polymorphism and discuss the historical use of polymorphism in the identification of disease-predisposing genes via meiotic mapping. We next consider some of the unique aspects of SNPs, and their relative advantages and disadvantages in human population-based analyses. In this context, we describe and critique the following six different areas of application for SNP technologies:

- Gene discovery and mapping.
- Association-based candidate polymorphism testing.
- Diagnostics and risk profiling.
- Prediction of response to environmental stimuli, xenobiotics and diet.
- Homogeneity testing and epidemiological study design.
- Physiologic genomics.

We focus on key issues within each of these areas in an effort to point out potential problems that might plague the use of SNPs (or other forms of polymorphism) within them. However, we make no claim that our list of considerations are exhaustive. Rather, we believe that they may provide a starting point for further dialog about the ultimate utility of SNP technologies. In addition, although our emphasis is placed on applications of SNPs to the understanding of human phenotypes, we acknowledge that SNP maps and technologies applied to other species (e.g. the mouse genome, pathogen genomes, plant genomes, etc.) are also of tremendous interest.

NJ Schork[a,b,c,e], D Fallin[a] and JS Lanchbury[d]

[a] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA, [b] Program for Population Genetics and Department of Biostatistics, Harvard University of Public Health, Boston, MA, USA, [c] The Jackson Laboratory, Bar Harbor, ME, USA, [d] Molecular Immunogenetics, Department of Rheumatology, GKT School of Medicine, King's College, London, UK, [e] The GENSET Corporation of La Jolla, CA, USA and GENSET, SA of Paris, France

Key words: genetic epidemiology – polymorphism – SNPs

Corresponding author: Nicholas J. Schork, Ph.D., The Genset Corporation, 875 Prospect Street, Suite 206, La Jolla, CA 92037, USA. Tel: +858 551 3000; fax: +858 551 3001; e-mail: schorkn@genxy.comnjs2@po.cwru.edu

Received 14 June 2000, revised and accepted for publication 17 July 2000

The last decade or so has witnessed a veritable explosion in the design and development of molecular genetic technologies that can be used to understand the biological basis of complex traits and diseases such as hypertension and obesity. Although awesome in their innovation and construction, it is not at all clear how researchers should use these technologies to maximize scientific insight and practical application. An excellent case of this lack of clarity over a particular set of modern molecular genetic technologies concerns the identification, cataloguing, and mapping of human single nucleotide polymorphisms (SNPs): there are now at least four industrial group efforts (i.e. by the companies Genset, Incyte, Celera and CuraGen (1–3)), one large academic–industry consortium effort (4, 5), two government-sponsored efforts (one in the USA (6–8) and one in Japan (9)), and countless non-industrial scale academic programs attempting to identify thousands, if not hundreds of thousands, of SNPs for the purposes of advancing genetic analysis initiatives of some sort or another. In addition to these SNP-finding initiatives, there are numerous academic and industrial programs trying to develop technologies that will improve the efficiency, rapidity, and cost effectiveness of genotyping large numbers of individuals for identified SNPs (e.g. through microarray designs (10, 11)). Although few would question the scale of the committed resources and the scientific sincerity of the individuals behind SNP-related efforts, the question that inevitably arises in the wake of this intense interest in SNPs is: Why? The next few years will undoubtedly produce massive SNP collections and better technologies for SNP genotyping, but if these efforts are not to be wasted, greater emphasis needs to be placed on the application of these technologies and collections, rather than their mere creation.

The biology of polymorphism

The term 'polymorphism' is often used in rather vague and facile ways by geneticists. Technically, a polymorphic locus is one whose alleles or variants are such that the most common variant among them occurs with less than 99% frequency in the population at large (e.g. if the locus is biallelic, the rarer allele must occur with a frequency greater than 1% in the population). However, use of polymorphism in modern genetic initiatives ultimately emanated from the study of physiological and biochemical variation, such as that exhibited by protein isoforms and blood group antigens (for useful discussions see, e.g. (12–14)). This variation was thought to arise from actual DNA sequence variation, and when this was confirmed, a number of crucial questions inevitably arose. First, what kind of alterations exist in the genome that can be understood to impact phenotypic variation? Second, how are such variations maintained and what is their behavior in populations? Third, how can one 'link' relevant genetic alterations with phenotypic variations? And fourth, how are such alterations ultimately translated into overt biochemical and phenotypic variation? In theory, the answers to these questions seem straightforward to obtain by merely conducting studies examining the association of DNA variants with either the presence or absence of different phenotypes among individuals or among individuals from different populations. In practice, however, things have proven more difficult for many reasons. One simple reason for this difficulty is that the very definition of a DNA variant ranges from a single base pair to several hundred base pairs. Thus, it is not always straightforward to actually identify a specific genomic site that results in phenotypic variation.

Polymorphism arises as a result of mutation. The different types of polymorphism are typically referred to by the type of mutation that created them. The simplest type of polymorphism results from a single base mutation which substitutes one nucleotide for another. The polymorphism at the site harboring such changes has recently been termed a 'single nucleotide polymorphism (SNP)', although previously, in some instances, such variation was referred to by the particular methods used to detect it. For example, the first systematic studies of single base variants were pursued through the identification of restriction enzyme sites, where a single base pair change could result in the loss or gain of a restriction site. Digestion of a piece of DNA containing the relevant site with an appropriate restriction enzyme could then distinguish alleles or variants based on resulting fragment sizes via electrophoresis, and this type of polymorphism was thus referred to as 'restriction fragment length polymorphism (RFLP)' (15).

Other SNPs, which do not directly create or destroy a restriction site, have been identified, often by creating restriction sites via PCR primer design, by oligonucleotide probing, or by direct sequencing. Recent technological advances have greatly improved the ease and sophistication of such identification processes, as will be discussed in sections below. Although the frequency with which SNPs (of any kind) occur over the genome is certainly much greater than that of RFLPs alone, precise estimates are difficult to determine and often vary across different populations and genomic regions. Some studies have suggested that SNPs can be found, on average, every 0.3–1 kilobases (kb) within the genome, although most data used to address the question of SNP frequency were derived from studies of SNPs within specific genes and thus are likely biased. Thus, it is not clear whether or not current estimates can be extrapolated to the rest of the genome and to populations other than those studied (16, 17). Whatever the actual frequency of SNPs across the genome is, it is known to be greater than any other type of polymorphism (14, 18).

Other types of genetic polymorphism result from the insertion or deletion of a section of DNA. The most common type of such 'insertion/deletion' polymorphism is the existence of variable numbers of repeated base or nucleotide patterns in a genetic region (14). Repeated base patterns range in size from several hundreds of base pairs, known as 'variable number of tandem repeats' (VNTRs or 'minisatellites'), to the more common 'microsatellites' consisting of two, three or four nucleotides repeated some variable number of times. Microsatellites are often referred to as 'simple tandem repeats' (STRs). Repeat polymorphisms often result in many alleles or variants (e.g. several different repeat sizes) within the population and are thus considered 'highly polymorphic'. This can be extremely useful for population genetic studies since the probability that two individuals from, say, different populations (ethnic groups, diseased vs.

non-diseased populations, etc.), will have the same number of repeats can be quite low (19). The genome-wide frequency estimates for STRs are difficult to come by, though a range of figures of one STR every 3–10 kb seems reasonable (14, 20, 21).

Another type of insertion/deletion polymorphism involves the presence or absence of Alu segments at a genetic location. Alu segments are named according to the restriction enzyme used to detect them (e.g. AluI), and contain two sequences approximately 120–150 bases in length, separated by an A base-rich segment. Insertions of this type occur approximately every 3 kb on average (22). Large insertion/deletion polymorphism such as Alu insertions are easy to identify and genotype given the large differences in resulting amplified fragments.

The traditional uses of polymorphism in gene mapping

The study of variation at the DNA sequence level within the last 20 or so years has had an enormous impact on the belief that one could directly link specific variants with specific traits or diseases. This idea of connecting overt phenotypic diversity with DNA sequence diversity has been plagued by three very important issues, all of which have become more pronounced as a result of improved technologies for DNA sequencing. First, so much sequence variation has been identified across individuals that questions about the origin and maintenance of such variation in the population at large have been raised. The traditional belief that mutation drove the compilation or build-up of sequence variation and that most mutation was deleterious and subject to selection was thus challenged. In response, the 'neutral' theory of evolution was developed, most notably by Kimura and colleagues (23). This theory suggested that most sequence variation does not directly impact phenotypic variation and thus is not directly subjected to the forces of selection. Thus, neutral theory suggests that there is simply no guarantee that any identified polymorphism has an associated overt or clinically relevant phenotype.

Second, it is now generally acknowledged that most common traits and phenotypes are determined by a multitude of genetic and non-genetic (or 'environmental') factors (24). The multifactorial nature of most traits makes the identification of each individual factor influencing them difficult, if not impossible, simply because the effect of any one factor may be obscured or confounded by the effects of others (25–27).

Third, the fact that there is a great deal of sequence variation that is of no physiological or 'functional,' significance, and the fact that most traits and diseases of contemporary interest are multifactorial in nature, has led to a great deal of debate concerning methodological and study design issues for linking sequence variation with phenotypic variation (24, 28–31).

For the last 25 years, the most commonly used approach to identify genes that influence traits has been meiotic or 'linkage' mapping (24, 32, 33). Although linkage analysis comes in various guises (e.g. parametric or non-parametric, unipoint or multipoint, etc.), all linkage analysis methods basically work by assessing the transmission and co-segregation of alleles at landmark spots on the genome, known as marker loci, with putative disease alleles assumed to be carried by family members exhibiting the disease or trait of interest. Alleles at marker loci that are very close to the disease allele-bearing locus should be transmitted with the disease through pedigrees (see Schork and Chakravarti (33) for a non-technical overview of linkage analysis and related methods). Thus, the fundamental idea behind linkage analysis is that, in a few generations' time, chromosomal segments harboring disease alleles will bear alleles at closely neighboring loci that were on the original chromosome harboring the disease-predisposing allele. This would occur because sufficient time would not have elapsed for recombination, mutation, etc. to shuffle the alleles across different chromosomes. Thus, by tracing and examining the consistency of co-segregation of marker locus alleles with a disease from generation to generation in a family segregating the disease, one may be able to identify the approximate position of a locus harboring disease alleles relative to the positions of the marker loci studied. If evidence for such consistent co-segregation can be found within a large number of families (note: the actual co-segregating allele at the linked locus does not have to be the same in each family, especially if there are many alleles at the locus), then one might consider refining the linkage in the region of interest by using more markers in that region and looking for more compelling co-segregations (34). This gene detection process, i.e. finding a linked marker and then refining the linkage until an actual offending disease locus is found, is often referred to as 'positional cloning'. Note that microsatellites are suited for such studies since they typically have a large number of alleles (i.e. repeat lengths), making it easy to identify alleles uniquely co-segregating with a disease in different families.

To facilitate this process of linkage mapping, 'maps' of the genome with identified landmark sites whose alleles could be used in the co-segregation analysis are needed (see Lander and Weinberg (35) for an excellent historical account of the place of genetic map development and linkage analysis in genetics research). Botstein and colleagues proposed the use of a map of RFLPs for such purposes in 1980s (15, 36, 37), but the first 'complete' linkage map of the genome (i.e. where the landmark sites had been ordered on each chromosome and their rough or exact locations are known), only contained around 400 markers (i.e. had a density of, roughly, 1 marker every 7–10 megabases (38). Due to their increased level of allelic polymorphism, a growing emphasis on microsatellite markers in the early 1990s shifted map construction efforts away from RFLPs. In 1990, a large collaborative effort to develop a linkage map of the human genome was proposed at the Centre d'Etude du Polymorphisme Humain (CEPH) in Paris, France (39). The motivation for the CEPH initiative was to provide researchers with DNA from families that could be used to order markers. The identified and mapped markers were then placed in a repository which was accessible to researchers interested in conducting linkage and gene mapping studies for various traits and diseases. Currently, there are over 12000 markers in the CEPH database, the vast majority of them being microsatellites. The most reliable map of the genome constructed from the CEPH repository consists of over 7500 markers with an intermarker distance of roughly 500 kb (40). General use of an extremely dense map of markers, however, is not pursued often, both because of the expense in genotyping and because the size of the chromosomal segments likely to be kept intact over a few generations does not require the use of so many markers to detect co-segregation within families. Maps for initial linkage analysis studies generally exploit a marker every 10 cM (see, e.g. the study by Xu et al. (41)).

Unfortunately, linkage analysis and the use of maps designed for linkage analysis studies have not proven powerful enough to detect genes influencing many common multifactorial diseases. The reasons for this are simple: many of the analytical strategies used to assess the within-family co-segregation phenomena at the heart of linkage analyses are not very powerful for genes with small to moderate effects on a trait or disease, often requiring the collection of hundreds if not thousands of families for reliable results (29, 30). To combat this, emphasis has been placed on the use of different sets of polymorphisms and different analytical strategies for gene mapping efforts.

The emergence of SNPs in genetic analysis

SNPs have many advantages over other sorts of polymorphism in the genetic dissection of complex traits and diseases, and for population-based gene identification studies, generally. Below, we briefly describe some of these advantages that have contributed greatly to the emergence of SNPs as an alternative form of sequence variation for gene identification and mapping studies (6, 8). In addition, they have also sparked intense interest in a reevaluation of methodologies and study designs that a researcher might exploit to explain phenotypic variation on the basis of sequence variation (see, e.g. (29, 31)).

Abundance

The high frequency with which SNPs are found on the genome gives them definite utility for trait or disease gene discovery purposes. Thus, one can use SNPs as markers for very dense gene mapping studies in positional cloning efforts (42), or, more importantly, as candidate polymorphisms to be tested directly as the functional or causal mutations for a trait or disease.

Position

SNPs are found throughout the genome, e.g. in exons, introns, intergenic regions, in promoters or enhancers, etc. Hence, they are more likely to yield, upon collection, a functional or physiologically relevant allele than other sorts of polymorphism. What is of extreme interest in this regard is the nature of the effect that a simple base pair substitution can have on a trait or disease. Thus, a SNP in coding region may directly impact a relevant protein, an intronic SNP can influence splicing (43), a SNP in a promoter can influence gene expression (44), etc. The degree to which each kind of SNP influences phenotypic expression is likely to receive a great deal of attention as more and more SNPs are identified and studied.

Origins and haplotypic patterns

Because new SNP alleles arise as mutations at different loci and at different points in time, and because they occur with such great abundance over the genome, groups of neighboring SNPs may have alleles that show distinctive patterns of linkage disequilibrium (LD, i.e. LD is the phenomenon whereby the presence of one allele on a chromosome may suggest a high probability that a particular allele will be present at a neighboring site on the same chromosome) and as such may create a haplotypic diversity that can be exploited in both genetic linkage and direct association studies. Nickerson and colleagues appear to have been the first to emphasize this fact (45).

Ease of genotyping

Because of their simple structure as base changes, microarray and other technologies can be (and are being) developed to allow the rapid and efficient genotyping of hundreds (or thousands) of individuals for hundreds (or thousands) of SNPs (10, 11).

Allele frequency drift

Because they typically only have two alleles (with a maximum of four, obviously – one for each base (18)), SNPs will have allele frequencies that will 'drift' as a function of the dynamics of different populations. This can create allele frequency differences that can be exploited in many population-based studies, such as admixture mapping (46).

Less mutable

SNPs generally are less mutable than other forms of polymorphism (14, 18, 47). This increased 'stability' could afford geneticists a more reliable way of assessing LD relationships, locus associations, and co-segregation phenomena since associations would not be confounded by alleles having mutated to different forms in the course of the transmission of alleles from generation to generation.

Recombinational oddities

Since SNPs can occur very close to one another, study of the patterns of LD they show may reveal sites for recurrent mutation, gene conversion, or recombination 'hot-spots'. Such information may be very useful when assessing a genomic region for linkage or association with a particular trait or disease (48, 49).

Contemporary and future uses of SNPs

The examination of DNA sequence variation to identify genes that influence multifactorial diseases and traits will undoubtedly benefit from the SNP craze. However, just how this benefit will occur depends on how SNPs will be exploited in relevant study designs, and what traits and diseases will be of focus in these studies. In the sections that follow we consider areas of application that are meant to address these issues as well as some problems that might plague relevant applications.

Gene discovery and mapping

There are currently many ways of identifying genes that underlie a particular human disease, e.g. comparing gene expression patterns in diseased vs. non-diseased tissues; exploiting homology between human genes and those identified in model organisms with a similar disease; using *in silico* methods to uncover functionally relevant gene families from available sequence databases, etc. Many of these approaches are not necessarily focused on the study of inherited differences in sequence variation. An alternative approach focuses on inheritance of sequence variants underlying a given phenotype. As pointed out in the discussion on traditional uses of polymorphism, one of the most (if not 'the' most) widely used current methods for disease gene discovery that exploits inheritance and patterns of heredity is meiotic mapping (24) (33). As noted before, meiotic mapping involves tracing co-segregation and recombination phenomena between alleles at loci whose chromosomal locations are known (i.e. 'marker' loci) with putative or hypothetical alleles at disease-influencing loci. This is achieved by assessing co-segregation of a phenotype or disease and marker alleles among related individuals. If evidence for such co-segregation is found, one can infer the existence of a disease-influencing locus near the marker locus. Meiotic mapping, however, actually comes in two varieties. *Linkage mapping*, as discussed previously, exploits *within*-family associations between marker alleles and putative trait-influencing alleles arising from the 1–4 generations' worth of co-segregation and recombination phenomena traceable within such families (33). Because the focus of linkage mapping is on the small number of meiotic events observable within a family, it does not require a very dense map of markers to find initial evidence for possible co-segregation of a disease-influencing gene with marker locus alleles. However, refining the location of the disease locus may require a much denser map. *Linkage disequilibrium (LD)* mapping, unlike linkage mapping, exploits *across*-family associations (50). LD mapping requires a dense map of markers since the size of the genomic regions harboring alleles that co-segregate with a disease across different families may be very small, due to the large number of meiotic and recombination events in the genealogical links connecting the families. Since linkage and LD mapping exploit co-segregation of known alleles with disease-influencing alleles, access to a large number of SNPs can only increase their resolution and power (29, 42).

Fig. 1 describes some issues in linkage and LD mapping. The figure depicts the origin and transmission of a disease gene (shaded individuals have the disease gene which is assumed to be dominant and fully penetrant) and alleles at a marker locus residing near it. The disease allele is denoted 'D'. Marker alleles are denoted by numbers. Note that in the left-most ancestral family, the '1' marker allele was co-segregating with the trait. This co-segregation was not disrupted in the descendants of the left-most grandchild in that family (i.e. no recombination or mutation event occurred during the time those descendants lived that would reshuffle the disease allele away from the 1 allele). This was not the case for the descendants of the other two grandchildren, where a recombination and mutation event put the disease allele on a chromosome with either a '2' or '3' allele at the neighboring marker locus. Thus, for descendants of these individuals who lived after these recombination and mutation events occurred, their families would be segregating '2' or '3' marker locus alleles along with the disease allele, rather than a '1' allele. This, in effect, would create strong 'within-family' associations but weak 'across-family' associations. Additional factors such as immigrant families moving into a population that are segregating for the disease gene but have a different associated marker locus allele could create further across-family divergence. Linkage analysis exploits within-family associations by tallying up the number of times any allele is segregating with a disease in a family and LD mapping analysis exploits the commonality of an associated allele across families. For LD mapping to work in the hypothetical situation in Fig. 1, either analysis would have to be restricted to subsets of families (e.g. the left-most in the latest generations) or through the use of a different marker locus – preferably one that is closer to the disease locus so that recombination is not as likely to have occurred and washed away any across-family associations. LD mapping is also facilitated through haplotype analysis, as described in Fig. 2, which depicts the hypothetical 'signature pattern' of alleles surrounding an ancestral disease mutation transmitted to descendants. Finding SNPs close enough to each other to reveal such a signature can be difficult but is more powerful than focusing on a single locus.

There are a number of difficult issues plaguing meiotic mapping studies, however. First, statistical methods for conducting linkage studies are notoriously non-powerful for detecting all but genes with a relatively large effect on the trait or disease of interest (29, 30, 51–53). Since most diseases of contemporary interest, such as obesity and hypertension, are likely influenced by a number of genes and environmental factors, each contributing gene is likely to have a small effect on disease susceptibility and pathogenesis. Second, although more powerful, LD mapping requires across-family associations (i.e. LD) between alleles at marker loci

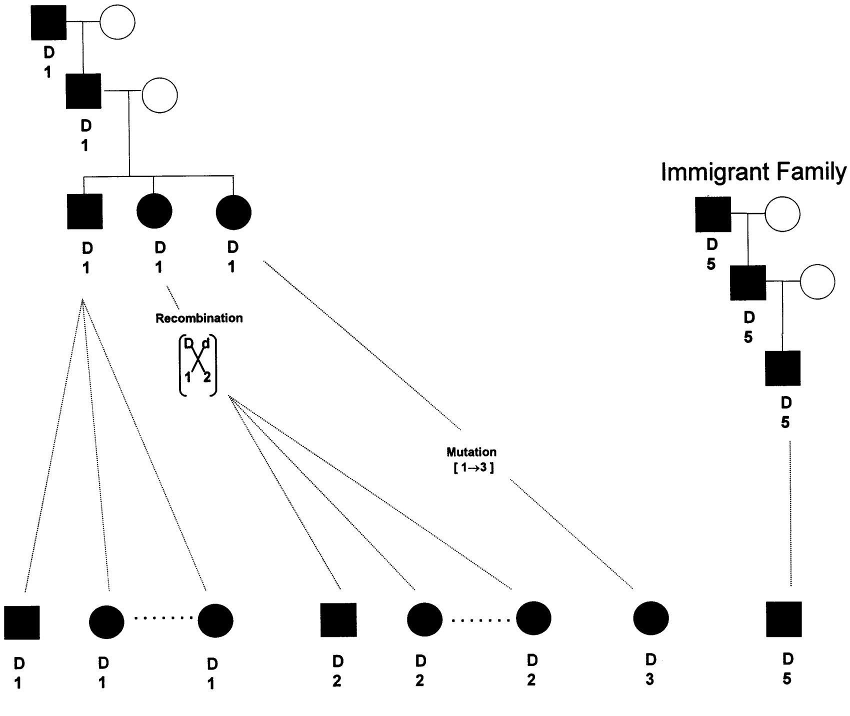

Fig. 1. Graphical depiction of hypothetical lines of descent involving a disease gene.

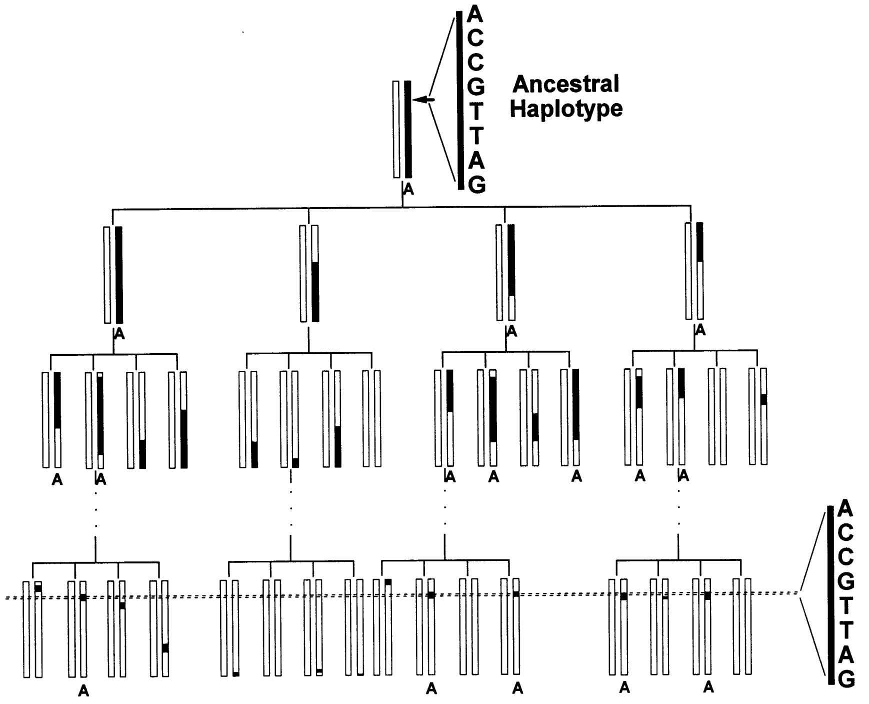

Fig. 2. Hypothetical conserved haplotype associated with a disease pre-disposing mutation.

and the sought-after trait influencing loci. Knowledge of how strong LD is and how far it extends away from a particular locus in a particular population is thus crucial for LD-based mapping studies and can help guide the researcher as to the number and density of markers needed for a study. Since LD mapping is currently in vogue, this point demands further emphasis since, unfortunately, a great number of factors influence LD strength. These factors run the gamut from population founder size and expansion rate to mutation and gene conversion rate. The interplay of all these factors creates tremendous variation in LD strength not only between populations of different origins (54), but also between different genomic regions (see Table 1) (55, 56). As a result, the LD strength exhibited among a set of alleles at different genomic loci in different populations is largely contingent on a historical fact that can really only be assessed empirically. With this in mind, the assessment of SNP utility in modern mapping efforts should combine traditional theory (57–59) about expected LD properties with complementary empirical data from the population and genomic region under study. Further, since SNPs occur with great frequency over the genome, a map composed of them may be of sufficient density to account for the possibility that some SNPs neighboring a disease locus may only exhibit weak LD with the disease locus alleles.

Also crucial for SNP-based mapping efforts are analytic techniques and study designs that can be used to assess evidence of co-segregation and LD phenomena. Developing appropriate methodologies is non-trivial, given the complexity of most traits and diseases of contemporary interest. Newer haplotyping methods (60), distance-based mapping measures (61, 62), combined linkage and LD analyses (63), and random effect models that can accommodate a wide variety of trait-influencing factors (64–66) show promise, especially if the use of SNPs is advocated.

Candidate polymorphism testing

Linkage and LD mapping ultimately assess and exploit co-segregation of marker and trait-influencing alleles within families or populations and, as such, assume that the markers used have an *indirect* association with the trait (i.e. the markers are merely acting as 'surrogates' for the actual disease alleles they neighbor and co-segregate with). As an alternative to assessing indirect association, one could test alleles at polymorphic sites for direct and *physiologically relevant* associations with a trait or disease. Since the identification of a polymorphic site whose alleles are causally related to a trait or disease is the ultimate goal of many genetic analysis studies, the more polymorphic loci one can type on a set of individuals, the more likely it will be that one of those loci harbors alleles that causally influence the trait in question. Large SNP collections can therefore provide the necessarily high volume of SNPs from which the few functional polymorphisms can be obtained. Table 2 offers some recent examples of such association studies.

Unfortunately, as SNP databases are still in their infancy (see the section that follows on SNP databases), tests of direct association are currently applicable only if one is initially drawn to a limited number or a particular set of polymorphisms based on ancillary information about them (e.g. based on knowledge of the putative function or pathological implication of the gene or genomic region within which the polymorphisms reside). This is due to the fact that most identified polymorphisms are known to be biologically inert (i.e. 'neutral') or have unknown functions. In addition, the mere testing of tremendous numbers of polymorphisms raises serious statistical issues, such as the determination of the false positive rate of the tests and the level of statistical significance to be adopted (29). Although one could combat these issues with better and more compelling statistical methods and study designs, better insight into the biology of the genes for determining candidate polymorphisms may be harder to come by.

Diagnostics and risk profiling

Once a SNP or SNP-based haplotype is identified which is associated (hopefully in a causal manner) with a trait or disease, one could potentially use the information to develop diagnostic or prognostic tools based on that SNP. The assessment of the utility and applicability of a SNP-based risk assessment tool for a complex disease will require well-designed, large-scale epidemiological studies. Consider again the fact that most diseases of contemporary interest are influenced by a number of genetic and non-genetic factors, each of which may make only a small contribution to disease risk in the population at large. If a gene (or series of genes) has been identified which is associated with a disease through the use of a special study design (e.g. affected sibpair linkage study, gene expression analysis, LD mapping in an isolated population, etc.) then it is an open question as to how large a role that gene has with respect to the disease burden in the population at large or with respect to the (clinical) populations for which the diagnostic will be used. Population genetic phenomena, such as disease heterogeneity, differential disease gene frequency, etc. will therefore undoubtedly impact the sensitivity and specificity of any diagnostic test based on an associated polymorphism(s). To obtain information about the putative population-based sensitivity and specificity of a DNA-based diagnostic for a complex or multifactorial disease, studies that assess the simultaneous influence of multiple relevant susceptibility factors on disease risk (e.g. diet, lifestyle, other gene effects, etc.) in the populations for which the diagnostic might be applied are needed. This is due to the fact that the influence of SNPs on the trait of interest may not be large enough for accurate prediction without ancillary information about an individual's risk. Large-scale population studies of this type are hardly unprecedented in traditional epidemiologic contexts, but their design will need to be refined to accommodate facts about genes and their role in disease susceptibility in populations. Thus, a greater adherence to population genetic and evolutionary principles will need to guide such studies (31).

Prediction of response to xenobiotics

An area of application of SNPs that is related to disease diagnostic, prognostic and risk assessment,

Table 1. Some recent (>1994) studies evaluating linkage disequilibrium or polymorphism content as a function of physical distance

Study/reference	Region or locus	# Loci/length	Comments
Jorde et al. (55)	APC	7/550 kb	Physical distance correlates with LD strength
Watkins et al. (56)	vWF	5/144 kb	Physical distance correlates with LD strength
Watkins et al. (56)	Multiple sites	5/144 kb	Stronger LD among loci closer to centromeres
Purandare et al. (81)	NF1	6/∼200 kb	CEPH and Japanese show stronger LD than an African cohort
Jeunemaitre et al. (82)	AGT	10/3.5 kb	Strong LD among CEPH, French, and Japanese
Ajioka et al. (83)	HH (MHC)	24/ 5 Mb	Strong LD within the very large region studied
Cox et al. (84)	IL-1	8/430 kb	Physical distance correlates with LD strength
Nickerson et al. (85)	LPL	88/9.7 kb	Sequence-based, exhaustive polymorphism search
Clark et al. (48)	LPL	88/9.7 kb	Evolutionary interpretation of extreme sequence diversity
Huttley et al. (86)	Genome-wide	–	Heterogeneity in LD by chromosome
Goodard et al. (87)	Multiple sites	–	LD strength variation between different US subpopulations
Collins et al. (88)	Multiple sites	–	LD extends to intermarker distances of up to 100 kb

Note: For references with results published prior to 1994, see Jorde et al. (55).

Table 2. Association studies between alleles at specific polymorphic loci and a complex disease/trait or drug response

Reference	Gene	Disease/trait	Comments
Dean et al. (89)	CCR5	HIV infection/AIDS	Physiological verification (90)
Corder et al. (91)	APOE	Alzheimer's	Replicated (92)
Cambien et al. (93)	ACE	Heart disease	Replicated (94)
Tiwari and Terasaki (95)	HLA	Many autoimmune diseases	Multiple studies
Carrington et al. (96)	HLA	AIDS	Heterozygosity of HLA loci
Jeunematrie et al. (97)	Angiotensinogen	Hypertension	Replicated (98)
Ebstein et al. (99)	Dopamine D4 receptor	Novelty seeking	Replicated (100)
Poirier et al. (101)	APOE	Response to tacrine therapy	Pharmacogenetic
Carmena et al. (102)	APOE	Response to lovastatin therapy	Pharmacogenetic
Flint et al. (103)	Alpha-thalassaemia	Malaria-associated anemia	Argument from selection
Drazen et al. (44)	15-LO	Asthma therapy	Pharmacogenetic analysis
Pianezza et al. (70)	CYP2A6	Cigarette consumption	Candidate gene
Montgomery et al. (104)	ACE	Physical performance	Replicated (105)
El-Omar et al. (106)	Interleukin-1	Gastric cancer	Candidate gene

is stratifying populations for the purposes of improving the effectiveness of interventions of one sort or another. Pharmacogenetic initiatives, in which the primary aim is to identify groups of diseased patients possessing a common genetic profile for which a particular compound is either ideally suited or likely to induce a side effect, are enjoying great interest (67, 68). The problems plaguing diagnostic studies also plague pharmacogenetic studies; however, if a gene that influences responsiveness to a particular compound has been identified via association analysis in a standard clinical trial sample or via molecular physiologic studies, then one will need to assess the frequency or penetrance of that gene in other populations and in light of other factors that may influence response to the compound. If such studies are not pursued, there is no guarantee that the polymorphism will adequately discriminate between populations likely to respond to a drug and those not likely to respond in typical, heterogeneous clinical populations. Although replication and follow-up efforts can be minimized through the pursuit of clever study designs, such as sequential clinical trial designs and multiarmed trials, they will still require considerable effort and resource investment.

The practice of identifying groups of individuals likely to benefit from (or adversely react to) some pharmacologic intervention can be extended to xenobiotic substances of all types. Consider toxic substances, cigarette smoke, exposure to ultraviolet light rays and diets (69). Response to each of these stimuli is likely to be under genetic control, for which prediction in the population at large would be of major public health and economic value (see the initiative sponsored by the National Institute of Environmental Health Sciences at http://www.niehs.nih.gov/envgenom/concept.htm). For example, the CYP2A6 polymorphism reflecting altered nicotine metabolism has been shown to correlate with cigarette consumption and therefore one would expect an indirect relationship between genotypes at this locus and tobacco-related disease such as lung carcinoma and emphysema (70). Of the possible stimuli, one might look to test associations with responsiveness to diet, and dietary interventions that may be of greatest market potential. A growing number of companies and scientists have begun to develop 'functional' foods and nutraceuticals whose ultimate efficacy will require verification and scientific validation (71). Thus, marrying genetic or biochemical profiles to 'optimal' diets, lifestyles, and interventions – though not unprecedented and in fact a motivating factor for all medical and public health practice – will likely expand and receive even greater attention in the not-so-distant future. Studies similar in orientation to traditional pharmacogenetic studies but with nutritional and dietary substances as outcomes can be initiated for which SNPs will play as large a role as in pharmacogenetics. Relevant genetic–epidemiologic study designs and statistical analysis tools will therefore likely be necessary for a wide range of xenobiotic response initiatives for which SNPs will have a predominant role.

Homogeneity testing and design of studies

It is quite commonplace in epidemiologic studies investigating the impact of a putative risk factor (e.g. smoking) on some outcome (e.g. lung cancer) to assess the homogeneity (or lack thereof) of the sample or population with respect to potential confounding variables (e.g. age, gender). The purpose of such heterogeneity testing is not only to protect against false inferences concerning the relationship of the set of primary endpoints and risk factors but also to assess the generalizability of the results. One particularly relevant source of heterogeneity in all human population studies is genetic heterogeneity. In the absence of information about the genetic profile of individuals at specific loci that influence, for example, response to a drug or predisposition to the multifactorial disease under study, one can assess homogeneity of the genetic backgrounds of the study participants using a panel of randomly distributed SNPs. By examining the commonality of the alleles at the loci among the study participants, one can look for clusters of genetically similar individuals that may reflect heterogeneity of genetic background and therefore possibly heterogeneity with respect to alleles that influence the outcome of interest. Obviously, the larger this panel and the easier it is to genotype individuals on this panel, the better. Since traditional phenotypic markers for genetic profiling, such as skin color, are ineffective in capturing genetic diversity (72–74), molecular profiling of the proposed type is a more effective and appropriate alternative.

As an example of an application of genetic homogeneity assessment, consider a large multicenter clinical trial in which participants are from all over the world or from different subpopulations within large urban centers associated with heavily admixed populations like the USA. Assume for the moment that the pharmacogenetic principle is valid, i.e. that individuals will respond to a particular compound according to their unique genetic and biochemical profile. Then it is arguable that by mixing individuals of different races, ethnic groups, or genetic subpopulations, a heterogeneity in responsiveness will arise (due to the mixture of different responsiveness genes in the sample) and should not be ignored. If evidence for genetic background heterogeneity is found, one can test the hypothesis that aspects of the heterogeneous response or outcome are due to this background genetic heterogeneity. Alternatively, one could verify that a rather homogeneous response to a compound, despite the background genetic heterogeneity of the subjects participating in the trial, is evidence that the compound in question could work ubiquitously or at least independently of genetic background.

Other study designs that more directly capitalize on the use of genetic background profiling, such as the genetic matching of subjects to test the effect of a particular allele on an outcome, will likely motivate additional work in this area. However, there are a number of methodological questions that need to be addressed for making homogeneity assessment and genetic study designs reliable and useful (e.g. how many random markers should be used? What statistical method should one use to assess the heterogeneity issue? etc.) Such questions will take on tremendous importance in the future (75, 76).

Physiological genomics

The identification of the function of a gene (i.e. its biological and physiological significance) has received considerable attention recently, especially as a motivating force and rallying cry for post-Human-Genome-Initiative scientific undertakings (77, 78). However, the current pools of 'functional genomic' study designs include highly specialized model organism-based experiments such as knockout, transgene and subsequent homology studies (78). These study designs can easily shed light on the physiologic role of a gene, but they are, in a sense, too contrived to do anything more than merely contribute partial insight into human pathogenic processes for which interventions could be designed effectively. Consider the fact that most common chronic human diseases are due to naturally occurring variation in genes (possibly at multiple sites or within multiple genes) as opposed to complete or partial gene deletions. Consider also that interventions for multifactorial diseases will not likely utilize direct insertion or removal of the responsible genes (at least in the foreseeable future). Therapeutic strategies will rely instead on the control of the expression of relevant genes or the manipulation of the sensitivity of gene products that are directly involved in the abnormal biochemical or physiological pathway mediating the disease. Thus, the gross interruption of a genome, as in a knockout experiment, may give erroneous impressions about not only the role of that gene in more natural settings but also about its potential as a therapeutic target. In addition, a host of redundancy, feedback and compensatory mechanisms present in physiologic systems has been shown to complicate the generalizability of current functional genomic study design results for this reason (79). Thus, knockout, transgene and related experiments are simply not sufficient to provide an adequate and comprehensive picture of human disease processes that could lead to the design of effective therapies and pharmacologically based prevention strategies. Furthermore, with the technology currently available it is not economically reasonable to envision selective investigation of the estimated 80000 genes that constitute the mammalian genome. Alternate methods are therefore needed to ascertain the function and pathogenic role of any newly discovered gene.

To address this concern, one could marry modern molecular phenotyping assays with polymorphism (i.e. SNP-based) analysis in order to determine the effect that naturally occurring genetic variation has on the network of interacting molecular and physiologic systems under normal and pathological human conditions. This, of course, is a non-trivial activity and will pose one of the greatest challenges facing modern medical scientists. However, one can think of designing studies investigating molecular physiologic phenotypes (e.g. circulating factors, differential gene expression in tissue samples, protein levels, etc.) in individuals with and without a disease-susceptible (i.e. SNP or mutation-based) genetic profile. Table 3 offers a list of studies investigating the impact of a polymorphism on a molecular or low-level physiological phenotype.

The goal of such studies would be to 'reverse engineer' relevant disease processes and outcomes through the comparison of individuals with and without these naturally occurring disease predisposing genetic profiles (80). It is only through detailed analysis of specific phenotypic abnormalities in humans with common disease susceptibility genetic profiles that definitive insight into what needs to be 'corrected' therapeutically can be obtained. This insight can only be achieved by first determining what those susceptibility profiles are, identifying individuals with those profiles and examining detailed phenotypes of these individuals. Although not perfect, as substitutes for the same reason as knockout studies, *in vitro* transfection studies may be amenable to high-throughput analysis of human polymorphism studies and serve as an initial foray into characterizing the functional significance of polymorphisms (see Table 3). Ultimately, the combination of SNP-based genetic technologies and resources (which focus on gene variations in the population) with genomic technologies (such as sequence structure analysis, expression profiling and protein level assays, etc., which assess molecular physiology and pathology) is the real future of medical and pharmaceutical research, rather than either one in isolation.

Resources

There are currently many publicly available resources for scientists interested in taking advantage of SNP technologies. The following describes the major websites offering collections of SNPs and information about those SNPs.

The genetic annotation initiative (http://cgap.nci.nih.gov/GAI/)

A National Institutes of Health (NIH)-operated site which contains information on candidate SNPs thought to be related to cancer and tumorigenesis generally.

dbSNP polymorphism repository (http://www.ncbi.nlm.nih.gov/SNP/)

A more comprehensive NIH-operated database containing information on SNPs with broad applicability in biomedical research.

HUGO mutation database initiative (http://ariel.ucs.unimelb.edu.au:80/ ~ cotton/mdi.htm)

Table 3. Polymorphism studies using molecular phenotypes derived from transfection and gene expression analysis

Reference	Gene	Cellular phenotype	Clinical phenotype
Janitz et al. (107)	HLA DRA	Transcription	Allergy/autoimmune
Hunault et al. (108)	Factor VII	Factor VII secretion	Coagulation
Walker et al. (109)	CDKN2A	Cell growth	Melanoma susceptibility
Crawley et al. (110)	Interleukin-10	Transcription	Juvenile idiopathic arthritis
Huizinga et al. (111)	Interleukin-10	Transcription	Erosive rheumatoic arthritis
Rood et al. (112)	Interleukin-10	Transcription	Systemic Lupus Erythamatosus
Fishman et al. (113)	Interleukin-6	Expression	Juvenile chronic arthritis
Zhao et al. (114)	Angiotensinogen	Transcription	Hypertension
Arinami et al. (115)	DRD2	Expression	Schizophrenia
Porzio et al. (116)	IRS-1	Expression & binding	Type II diabetes
McGraw et al. (117)	Beta 2 AR	Expression	Respiratory disease
Zenner et al. (118)	HDR4	Binding	Delusional disorder
Gill et al. (119)	CYP2C9	Hydroxylation	Drug responsiveness
Edenberg et al. (120)	ADH4	Expression	Alcoholism
Cravchik et al. (121)	DRD2	Binding	Schizophrenia
Bruss et al. (122)	5-HT1B receptor	Binding	Drug responsiveness
Wilson et al. (123)	TNF alpha	Transcription	Lupus

A database meant to provide systematic access to information about human mutations including SNPs. This site is maintained by the Human Genome Organization (HUGO).

Human SNP database (http://www-genome.wi.mit.edu/SNP/human/index.html)

Managed by the Whitehead Institute for Biomedical Research Genome Institute, this site contains information about SNPs resulting from the many Whitehead research projects on mapping and sequencing.

SNPs in the human-genome SNP database (http://www.ibc.wustl.edu/SNP)

This website provides access to SNPs that have been organized by chromosomes and cytogenetic location. The site is run by Washington University.

HGBase (http://hgbase.cgr.ki.se/)

HGBase is an attempt to summarize all known sequence variations in the human genome, to facilitate research into how genotypes affect common diseases, drug responses, and other complex phenotypes, and is run by the Karolinska Institute of Sweden.

The SNP consortium database (http://snp.cshl.org/db/snp/map)

A collection of SNPs and related information resulting from the collaborative effort of a number of large pharmaceutical and information processing companies.

GeneSNPs (http://www.genome.utah.edu/genesnps/)

Operated by the University of Utah, this site contains information about SNPs resulting from the US National Institute of Environmental Health's initiative to understand the relationship between genetic variation and response to environmental stimuli and xenobiotics.

Conclusion

The SNP craze is not likely to diminish soon. The amount of money invested in SNP technologies alone is enough to sustain enthusiasm for years to come. However, as with all large-scale, resource and investment intensive scientific research initiatives, qualifications, cautions, and appropriations must occur if success is to be had. There is no doubt that modern SNP initiatives depend critically on the development of novel assays, experimental apparati, database constructions, high-throughput devices of all sorts and analytic methods. But the existence of, and access to, such technologies is only half the battle for success. The other half is quite simply the appropriate and practical application of those technologies.

Acknowledgements

The authors would like to thank Hemant Tiwari for reading and commenting on the manuscript. DF is supported in part by NIH grants HL94-011 and HL54998-01 awarded to NJS.

References

1. Marshall E. Snipping away at genome patenting. Science 1997: 19: 1752–1753.
2. Marshall E. A second private genome project. Science 1998: 281: 1121.
3. Branca MA, Rubenstein K. Single Nucleotide Polymorphisms: Commercial and Scientific Prospects. Cambridge: Cambridge Healthtech, 1999: 91.
4. Masood E. As consortium plnas free SNP map of human genome. Nature 1999: 398: 545–546.
5. Hodgson J. Analysts, firms pour cold water on SNP Consortium. Nature Biotech 1999: 17: 526.
6. Collins FS, Geyer MS, Chakravarti A. Variations on a theme: cataloguing human DNA sequence variation. Science 1997: 278: 1580–1581.
7. Marshall E. Playing chicken' over gene markers. Science 1997: 278: 2046–2048.
8. Collins FS, Patrinos A, Jordan E et al. New goals for the US Human Genome Projects: 1998–2003. Science 1998: 282: 682–689.
9. Saegusa A. Japan bids to catch up on gene sequencing. Nature 1999: 399: 96–97.
10. Marshall A, Hodgson J. DNA chips: an array of possibilities. Nature Biotech 1998: 16: 27–31.
11. Ramsay G. DNA chips: state of the art. Nature Biotech 1998: 16: 40–44.
12. Cavalli-Sforza LL, Bodmer WF. The Genetics of Human Populations. San Francisco: W.H. Freeman and Company, 1971.
13. Brown TA. Genomes. New York: John Wiley, 1999.
14. Cooper DN, Krawczak M. Human Gene Mutation. New York: Academic Press, 1999.
15. Botstein D, White RL, Skolnick M et al. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 1980: 32 (3): 314–331.
16. Halushka MK, Fan JB, Bentley K et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 1999: 22 (3): 239–247.
17. Cargill M, Altshuler D, Ireland J et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes [published erratum appears in Nat Genet 1999 Nov: 23(3): 373]. Nat Genet 1999: 22: 231–238.
18. Brookes AJ. The essence of SNPs. Gene 1999: 8: 177–186.
19. Shriver MD, Smith MW, Jin L et al. Ethnic affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 1997: 60: 957–964.

20. Beckman JS, Weber JL. Survey of human and rat microsatellites. Genomics 1992: 12: 627–631.
21. Weber JL, Wong C. Mutation of human short tandem repeats. Hum Mol Genet 1993: 2 (8): 1123–1128.
22. Smit AF. The origin of interspersed repeats in the human genome. Curr Opin Genet Dev 1996: 6: 743–748.
23. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press, 1983.
24. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994: 265: 2037–2048.
25. Frankel WN, Schork NJ. Who's afraid of epistasis? Nat Genet 1996: 14: 371–373.
26. Schork NJ. Genetically complex cardiovascular traits: origins, problems, and potential solutions. Hypertension 1997: 29: 145–149.
27. Schork N. Genetics of complex disease. Am J Respir Crit Care Med 1997: 156: S103–S109.
28. Weeks DE, Lathrop GM. Polygenic disease: methods for mapping complex disease traits. Trend Genet 1995: 11: 513–519.
29. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996: 273: 1516–1517.
30. Risch N, Botstein D. A manic depressive history. Nat Genet 1996: 12: 351–353.
31. Schork NJ, Cardon LR, Xu X. The future of genetic epidemiology. Trend Genet 1998: 14: 266–271.
32. Ott J. Analysis of Human Genetic Linkage. Baltimore: Johns Hopkins University Press, 1991.
33. Schork N, Chakravarti A. A nonmathematical overview of modern gene mapping techniques applied to human diseases. In: Mockrin S, ed. Molecular Genetics and Gene Therapy of Cardiovascular Disease. New York: Marcel Dekker, Inc, 1996: 79–109.
34. Schuler GD, Boguski MS, Stewart EA et al. A gene map of the human genome. Science 1996: 274: 540–546.
35. Lander ES, Weinberg RA. Genomics: journey to the center of biology. Science 2000: 287: 1777–1782.
36. Lander ES, Botstein D. Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 1986: 51 (Pt 1): 49–62.
37. Lander ES, Botstein D. Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc Natl Acad Sci USA 1986: 83 (19): 7353–7357.
38. Donis-Keller H, Green P, Helms C et al. A genetic linkage map of the human genome. Cell 1987: 51 (2): 319–337.
39. Dausett J, Cann H, Cohen D et al. Centre d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. Genomics 1990: 6: 575–577.
40. Broman KW, Murray JC, Sheffield VC et al. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 1998: 63: 861–869.
41. Xu X, Rogus JJ, Terwedow HA et al. An extreme-sib-pair genome scan for genes regulating blood pressure. Am J Hum Genet 1999: 64 (6): 1694–1701.
42. Kruglyak L. The use of a genetic map of biallelic markers in linkage studies. Nat Genet 1997: 17: 21–24.
43. Krawezak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum Genet 1992: 90: 41–54.
44. Drazen JM, Yandava CN, Dube L et al. Pharmacogenetic association between ALOX5 promoter genotype and the response to anti-asthma treatment. Nat Genet 1999: 22: 168–170.
45. Nickerson DA, Whitehurst C, Boysen C et al. Identification of clusters of biallelic polymorphic sequence-tagged sites (pSTSs) that generate highly information and automatable markers for genetic linkage mapping. Genomics 1992: 12: 377–387.
46. McKeigue PM. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet 1998: 63: 241–251.
47. Stallings RL, Ford AF, Nelson D et al. Evolution and distribution of (GT)n repetitive sequences in mammalian genomes. Genomics 1991: 10: 807–815.
48. Clark AG, Weiss KM, Nickerson DA et al. Haplotype structure and population-genetics inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 1998: 63: 595–612.
49. Chakravarti A. It's raining SNPs, hallelujah? Nat Genet 1998: 19: 216–217.
50. Jorde LB. Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet 1995: 56: 11–14.
51. Kruglyak L, Lander ES. High-resolution genetic mapping of complex traits. Am J Hum Genet 1995: 56: 1212–1223.
52. Schork NJ, Xu X. Sibpairs versus pedigrees: what are the advantages? Diab Rev 1997: 5: 116–122.
53. Schork NJ, Theil B, StJean P. Linkage analysis, kinship, and the short-term evolution of chromosomes. J Exp Zoo 1997: 34: 101–115.
54. Tishkoff SA, Dietzsch E, Speed W et al. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 1996: 271: 1380–1387.
55. Jorde LB, Watkins WS, Carlson M et al. Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. Am J Hum Genet 1994: 54: 884–898.
56. Watkins WS, Zenger R, O'Brien E et al. Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willebrand Factor Region. Am J Hum Genet 1994: 55: 348–355.
57. Slatkin M. Linkage disequilibrium in growing and stable populations. Genet Soc Am 1994: 137: 331–336.
58. Thompson EA, Neel JV. Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. Am J Hum Genet 1997: 60: 197–204.
59. Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 1999: 22 (2): 139–144.
60. Fallin D, Schork NJ. The accuracy of haplotype frequency estimation involving biallelic markers and genotypic data. Am J Hum Genet 2000 (in press).
61. Terwilliger JD. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 1995: 56: 777–787.
62. Valdes AM, Thomson G. Detecting disease-predisposing variants: the haplotype method. Am J Hum Genet 1997: 60: 703–716.
63. MacLean CJ, Morton NE, Yee S. Combined analysis of genetic segregation and linkage under an oligogenic model. Comput Biomed Res 1984: 17 (5): 471–480.
64. Boerwinkle E, Charkraborty R, Sing CF. The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. Ann Hum Genet 1986: 50: 181–194.
65. George VT, Elston RC. Testing the association between polymorphic markers and quantitative traits in pedigrees. Genet Epidemiol 1987: 4: 193–201.

66. Amos CI, Zhu DK, Boerwinkle E. Assessing genetic linkage and association with robust components of variance approaches. Ann Hum Genet 1996: 60: 143–160.
67. Nebert DW, Weber WW. Pharmacogenetics. In: Pratt WB, Taylor P, eds. Principles of Drug Action. Churchill-Livingstone: New York, 1990.
68. Ball S, Borman N. Pharmacogenetics and drug metabolism. Nature Biotech 1997: 15: 925–926.
69. Schork NJ. Xenobiotics, dietary interventions, and genetically-mediated therapies. Curr Hypert Rep 2000: 2: 11–12.
70. Pianezza ML, Sellers EM, Tyndale RF. Nicotine metabolism defect reduces smoking [letter]. Nature 1998: 393 (6687): 750.
71. Farr DR. Functional foods. Cancer Lett 1997: 114: 59–63.
72. Lewontin RC. The genetic basis of evolutionary change. New York: Columbia University Press, 1974.
73. Barbajuni G, Magagni A, Minch E et al. An apportionment of human DNA diversity. Proc Nat Acad Sci 1997: 94: 4516–4519.
74. Templeton AR. Human races: a genetic and evolutionary perspective. Am Anthro 1999: 100: 632–650.
75. Schork NJ, Fallin D, Thiel B et al. The future of genetic case/control studies. In: Rao DC, Province MA, eds. Advances In Human Genetics. New York: Academic Press, 2000.
76. Schork NJ, Fallin D, Lanchbury JS. Clearing the air over association studies or wither association? 2000 (in preparation).
77. Lander ES. The new genomics: global views of biology. Science 1996: 274: 536–539.
78. Fields S. The future is function. Nat Genet 1997: 15: 325–327.
79. Schork NJ, Lanchbury JS. Integrated phenotyping, disease models, and pathophysiologic databases. Trends Biotech 2000 (submitted).
80. Anderson NG, Anderson LH, Hofmann JP. Research instrumentation for the 21st century: progress toward complete genomic maps and sequence data bases, and indexes of protein gene products. In: Beecher GB, ed. Research Instrumentation for the 21st Century. Dordrecht: Martinus Nijhoff, 1988.
81. Purandare SM, Cawthon R, Nelson LM et al. Genotyping of PCR-based polymorphisms and linkage disequilibrium analysis at the NF1 locus. Am J Hum Genet 1996: 59: 159–166.
82. Jeunemaitre X, Inoue I, Williams C et al. Haplotypes of angiotensinogen in essential hypertension. Am J Hum Genet 1997: 60: 1448–1460.
83. Ajioke RS, Jorde LB, Gruen JR et al. Haplotype analysis of hemochromatosis: evaluation of different linkage-disequilibrium approaches and evolution of disease chromosomes. Am J Hum Genet 1997: 60: 1439–1447.
84. Cox A, Camp NJ, Nicklin MJH et al. An analysis of linkage disequilibrium in the Interleukin-1 gene cluster, using a novel grouping method for multiallelic markers. Am J Hum Genet 1998: 62: 1180–1188.
85. Nickerson DA, Taylor SL, Weiss KM et al. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 1998: 19: 233–240.
86. Huttley GA, Smith MW, Carrington M et al. A scan for linkage disequilibrium across the human genome. Genetics 1999: 152: 1711–1722.
87. Goddard KA, Hopkins PJ, Hall JM et al. Linkage disequilibrium and allele frequency distributions for 114 single nucleotide polymorphisms in five populations. Am J Hum Genet 2000: 66: 216–234.
88. Collins A, Lonjou C, Morton NE. Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci 1999: 96: 15173–15177.
89. Dean M, Carrington M, Winkler C et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Science 1996: 273: 1862–1865.
90. Mummidi S, Ahuja SS, Gonzalez E et al. Genealogy of the CCR5 locus and chromokine system gene variants associated with altered rates of HIV-1 disease progression. Nat Med 1998: 4: 786–793.
91. Corder EH, Saunders AM, Strittmatter WJ et al. Gene dose of apolipoprotein E type 4 allele and risk of Alzheimer's disease in late onset families. Science 1993: 261: 921–923.
92. vanDuijn CM, deKniff P, Wehnert D et al. The apolipoprotein E epsilon 2 allele is associated with an increased risk of early-onset Alzheimer's disease and a reduced survival. Ann Neurol 1995: 37: 605–610.
93. Cambien F, Poirier O, Lecerf L et al. Deletion polymorphism in the gene for angiotensin-converting enzyme is a potent risk factor for myocardial infarction. Nature 1992: 359: 641–644.
94. Morris BJ, Zee RY, Schrader AP. Different frequencies of angiotensinogen-converting enzyme genotypes in older hypertensive individuals. J Clin Invest 1994: 94: 1085–1089.
95. Tiwari JL, Terasaki PI. HLA and Disease Associations. New York: Springer-Verlag, 1985.
96. Carrington M, Nelson GW, Martin MP et al. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. Science 1999: 283: 1748–1752.
97. Jeunemaitre X, Soubrier F, Kotelevtsev YV et al. Molecular basis of human hypertension: role of angiotensinogen. Cell 1992: 71: 169–180.
98. Hegele RA, Brunt H, Connelly PW. A polymorphism of the angiotensinogen gene associated with variation in blood pressure in a genetic isolate. Circulation 1994: 90: 2207–2212.
99. Ebstein RP, Novick O, Umansky R et al. Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of novelty seeking. Nat Genet 1996: 12: 78–80.
100. Benjamin J, Li L, Patterson C et al. Population and familial association between the D4 dopamine receptor gene and measures of novelty seeking. Nat Genet 1996: 12: 81–84.
101. Poirier J, Marie-Claude D, Quiron R et al. Apolipoprotein E4 allele as a predictor of cholinergic deficits and treatment outcome in Alzheimer disease. Proc Natl Acad Sci 1995: 92: 12260–12264.
102. Carmena R, Roederer G, Mailoux H et al. The response to lovastatin treatment in patients with heterozygous familial hypercholesteremia is modulated by apolipoprotein E polymorphism. Metabolism 1993: 42: 895–901.
103. Flint J, Hill AV, Bowden DK et al. High frequencies of alpha thalassemia are the result of natural selection by malaria. Nature 1986: 321: 744–750.
104. Montgomery HE, Marshall R, Hemingway H et al. Human gene for physical performance [letter]. Nature 1998: 393 (6682): 221–222.
105. Williams AG, Rayson MP, Jubb M et al. The ACE gene and muscle performance. Nature 2000: 403 (6770): 614.
106. El-Omar EM, Carrington M, Chow WH et al. Interleukin-1 polymorphisms associated with increased risk of gastric cancer [In Process Citation]. Nature 2000: 404 (6776): 398–402.

107. Janitz M, Reiners-Schramm L, Lauster R. Expression of the H2-Ea gene is modulated by a polymorphic transcriptional enhancer. Imm Genet 1998: 48: 266–272.

108. Hunault M, Arbini A, Lopaciuk S et al. The Arg353Gln polymorphism reduces the level of coagulation factor VII. *In vivo* and *in vitro* studies. Arterioscler Thromb Vasc Biol 1997: 17 (11): 2825–2829.

109. Walker G, Gabrielli B, Castellano M et al. Fuctional reassessment of P16 variants using a transfection-based assay. Int J Cancer 1999: 82 (2): 305–312.

110. Crawley E, Kay R, Sillibourne J et al. Polymorphic haplotypes of the interleukin-10 5′ flanking region determine variable interleukin-10 transcription and are associated with particular phenotypes of juvenile rheumatoid arthritis. Arthritis Rheum 1999: 42 (6): 1101–1108.

111. Huizinga TWJ, Keijsers V, Yanni G et al. Differences in IL-10 production are associated with joint damage. 2000 (submitted).

112. Rood MJ, Keijsers V, van der Linden MW et al. Neuropsychiatric systemic lupus erythematosus is associated with imbalance in interleukin 10 promoter haplotypes. Ann Rheum Dis 1999: 58: 85–89.

113. Fishman D, Faulds G, Jeffery R et al. The effect of novel polymorphisms in the interleukin-6 (IL-6) gene on IL-6 transcription and plasma IL-6 levels, and as association with systemic-onset juvenile chronic arthritis. J Clin Invest 1998: 102 (7): 1369–1376.

114. Zhao Y, Zhao J, Narayanan C et al. Role of C/A polymorphism at −20 on the expression of human angiotensinogen gene. Hyperten 1999: 33 (1): 108–115.

115. Arinami T, Gao M, Hamaguchi H et al. A functional polymorphism in the promoter region of the dopamine D2 receptor gene is associated with schizophrenia. Hum Mol Genet 1997: 6: 577–582.

116. Porzio O, Federici M, Hirbal M et al. The Gly972-Arg amino acid polymorphism in IRS-1 impairs insulin secretion in pancreatic beta cells. J Clin Invest 1999: 104: 357–364.

117. McGraw D, Forbes S, Kramer L et al. Polymorphisms of the 5′ leader cistron of the human beta2-adrenergic receptor regulate receptor expression. J Clin Invest 1998: 102: 1927–1932.

118. Zenner M, Nobile M, Henningsen R et al. Expression and characterization of a dopamine D4R variant associated with delusional disorder. FEBS Lett 1998: 422: 146–150.

119. Gill H, Tjia J, Kitteringham N et al. The effect of genetic polymorphisms in CYP2C9 on sulphamethoxazole N-hydroxylation. Pharmacogenet 1999: 9: 43–53.

120. Edenberg H, Jerome R, Li M. Polymorphism of the human alcohol dehydrogenase 4 (ADH4) promoter affects gene expression. Pharmacogenet 1999: 9: 25–30.

121. Cravchik A, Sibley D, Gejman P. Analysis of neuroleptic binding affinities and potencies for the different human D2 dopamine receptor missense variants. Pharmacogenet 1999: 9: 17–23.

122. Bruss M, Bonisch H, Buhlen M et al. Modified ligand binding to the naturally occurring Cys-124 variant of the human serotonin 5-HT 1B receptor. Pharmacogenet 1999: 9: 95–102.

123. Wilson AG, Gordon C, diGiovine FS et al. A genetic association between systemic lupus erythematosus and tumor necrosis factor alpha. Europ J Immun 1994: 24: 191–195.