Population genetics—making sense out of sequence

Aravinda Chakravarti

Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA. e-mail: axc39@po.cwru.edu

The complete human genome nucleotide sequence and technologies for assessing sequence variation on a genome-scale will prompt comprehensive studies of comparative genomic diversity in human populations across the globe. These studies, besides rejuvenating population genetics and our interest in how genetic variation is created and maintained, will provide the intellectual basis for understanding the genetic basis for complex diseases and traits.

Dissecting the human genome

Biomedical scientists, and indeed all of humanity, will be handed the entire human genome sequence in the very early dawn of the next century[1,2]. There will be many reasons to celebrate this event, but the principal reason will be our new-found ability to read the human sequence in its entirety at its most fundamental nucleotide level. It is a truism that the genome sequence will alter genetics fundamentally, perhaps unrecognizably so. Nowhere will the scientific impact be greater than upon human genetics. There are four reasons for this optimism: first, the breadth and depth of phenotypic characterization of the human remains unparalleled; second, knowledge of human history, geography and ecology is unmatched; third, our knowledge of cellular circuitry encoded by the genome is exploding; and fourth, the ability to identify all of the natural molecular variation in the human sequence is imminent.

Sequence variation is the currency of genetics; the central aim of all genetics is to correlate specific molecular variation with phenotypic changes. The human genome sequence is not one sequence but rather many variations on a common theme, each of which alters the inherent molecular circuitry, and thus, consequent phenotypes, in a specific manner. Currently, geneticists use positional cloning to identify the molecular changes underlying a mendelian phenotype[3]. However, with the human genome sequence in hand we can create a catalogue of all common variants, which will allow us to detect association between these variants and any hereditary (transmissible) phenotype[4–6].

The population genetics perspective

Extant human genetic variation is natural, created by mutation and vetted over history by biological, demographic and historical processes. Notably, the sequence variation that has survived in the human species is nonrandom; it has been shaped both by chance and natural selection, and by the demic organization and migrations of our ancestors[7]. A persistent feature of this evolution is extinction—only a tiny fraction of the variation ever created in the human lineage now remains; it is this fraction that currently impacts on human phenotypes. Thus the corollary: current patterns and distribution of human phenotypes, including disease, are the legacy of our genetic past. Just as history is critical to an understanding of current human social and political relationships, understanding the evolutionary context of human variation is indispensable to our ability to explain current phenotypic variation. Population genetics is the scientific enquiry of this central problem in biology.

The rediscovery of mendelism in 1900 triggered the transition—from controversy to orthodoxy—of Darwin and Wallace's theory of evolution by natural selection. Darwin and Wallace enunciated two major principles: they convincingly demonstrated that evolution and adaptation had occurred and they posited a plausible biological cause (natural selection), although the latter was not widely accepted at the time. Population genetics was sired by an attempt to entwine mendelism, darwinism and biometry in an attempt to determine how the gene could be used to explain the creation, maintenance and distribution of phenotypes in populations[8]. Mendelian segregation is a quantitative biological law and so its consequences in a population of individuals, closely related or not, are also quantitative. This is why population genetics, which seeks to quantify these genetic effects in time and space, has such a highly developed mathematical theory[9–11]. It is also observational and experimental[12,13]. The human genome sequence will emphasize the latter aspect—not only for verifying existing models of genetic change but, more importantly, to test alternative theories and generate totally new ideas of genomic change and evolution.

Technical challenges

Common types of sequence variation in the human include single nucleotide polymorphisms (SNPs), insertions and deletions of a few nucleotides, and variation in the repeat number of a motif (mini- and micro-satellites). The central challenge over the next five years will be to devise efficient and cost-effective methods and technologies for identifying and scoring all types of genetic variation in the human genome. Although several methods for identifying such genomic variants currently exist, none identify all types of variation, most do not specify the nucleotide change and most cannot contend with the entire genome. An ideal technology should assay all types of sequence change, use small amounts of biological material and do so at the nucleotide level with very high accuracy (see page 42 of this issue (ref. 14)). The development of microarray technologies are an exciting advance in this regard, as they are massively parallel and can, in principle, survey an entire genome. (See pages 5 (ref. 15), 10 (ref. 16) and 20 (ref. 17) of this issue for detailed descriptions).

The most common type of sequence variation is the SNP ; those occurring in coding sequences have been dubbed cSNPs (ref. 6). Recent studies have revealed SNPs to be very abundant in the

Table 1 • Features of nucleotide diversity in human genes

Region	Nucleotides compared	Nucleotide diversity (π)
Noncoding		
5′ UTR	3,624	0.0003±0.0003
3′ UTR	19,769	0.0004±0.0001
Coding		
Nondegenerate	34,869	0.0003±0.0001
twofold: nonsynmomymous	10,787	0.0001±0.0001
twofold: synonymous	10,787	0.0005±0.0002
fourfold: degenerate	8,537	0.0011±0.0004

human genome, occuring at a density of approximately 1 per kilobase (kb) of DNA when two alleles are compared. Preliminary studies suggest three classes of microarrays for assessing SNP variation: (i) arrays for 'resequencing' a sample of the human genomic sequence; (ii) arrays that contain all known SNPs/cSNPs in a contiguous genomic segment; and (iii) arrays that contain a sampling of SNPs/cSNPs mapped across the entire genome.

Protein polymorphism

Classical studies of human variation have explored the manifestation of genetic expression—either as antigenic or charge differences in soluble proteins. Variation has been assessed in two ways—either by the proportion of polymorphic proteins or the average gene diversity (heterozygosity expected under random mating). The first comprehensive human study, carried out by Harry Harris, showed that about 30% of human proteins are polymorphic and the average human gene is heterozygous no more than 10% of the time[18]. These data suggest that, at the protein level, two alleles differ at no more than 0.1 substitutions and an estimated 25% of these result in charge differences[19].

Gene diversity depends on the mutation rate of genes, the size and demographic history of the population in which these mutations occur, the time over which such diversity accumulates and biological factors such as selection. On an absolute scale, human gene diversity at 10% is low but typical of that observed in most vertebrates, which are less diverse than invertebrates[19]. There are, however, large differences in diversity between human proteins[20]. Underscoring this variation are: (i) the quaternary structure of proteins; (ii) the molecular weight of protein subunits; and (iii) species population size. It follows that diversity decreases with an increase in the number of protein subunits, due to a higher degree of functional constraint. Diversity directly correlates with subunit molecular weight, as a larger macromolecule has more 'material' to support mutations. Finally, the degree of diversity that can be maintained in a population is directly proportional to population size[20]. The lower gene diversity in humans is the result of the young age of our species and a small founding population[21].

Variation in soluble proteins may provide an incomplete view of genomic diversity. King and Wilson, on comparing variation between specific proteins in humans and chimpanzees, discovered a low diversity within each species and a high similarity between them[22]. This, of course, is in stark contrast to the phenotypic differences between humans and chimpanzees, leading the investigators to speculate that regulatory mutations may be the prime cause for biological differences and that protein diversity might not be the driving force effecting significant phenotypes. Genomic studies should confirm or refute this hypothesis.

Nucleotide sequence polymorphism

An unbiased, comprehensive study of human variation is only possible at the nucleotide level. In the 1980s, restriction fragment length polymorphisms (RFLPs) in human genes revealed a great deal of genetic diversity, but most studies did not systematically survey the genome or even part of it[23,24]. Nucleotide sequence provides a much clearer picture. For example, Li and Sadler obtained the genomic and cDNA sequences of 49 human genes to assess nucleotide diversity (π) or the average diversity (heterozygosity) per nucleotide[25] (Table 1). They found that humans have low sequence diversity, less than 1 variant per kb, and that it varies with functional constraint: noncoding changes (π=0.00039) are more frequent than coding changes (π=0.00026); and coding changes are quite heterogeneous, as changes at synonymous sites (π=0.0005) or at degenerate sites (π=0.0011) are more prevalent than at nonsynonymous sites (π=0.0001) or nondegenerate sites (π=0.0003). A typical human cDNA is about 3 kb, with approximately 1 kb of synonymous sites and 2 kb of nonsynonymous sites, and expected gene diversities of 33% and 25% at synonymous and nonsynonymous sites, respectively. At the protein level, synonymous changes go undetected and only a fraction (25% of the total or 38% of nonsynonymous sites) of the remainder would lead to charge differences, giving an expected protein heterozygosity rate of 38%×25%, or 9.4%—similar to the value of 10% observed by Harris[18].

Variation in nucleotide diversity is not a sole property of genes, whose functional constraints we understand, but of the genome as a whole. Nucleotide sequencing and the first microarray analyses have started to produce considerable data on variation in genes and in long, contiguous DNA segments. There are three types of information available (Table 2). First, a number of genes, such as those encoding the chemokine receptor 5 (*CCR5*; ref. 26), glial cell line-derived neurotropic factor receptor[27] (*GFRA1*) and lipoprotein lipase[28] (*LPL*), have been contiguously sequenced, with 1–10 kb in 140 or more alleles determined. These studies confirm the reduced variation in coding versus noncoding DNA and the gene-to-gene variation in coding region diversity. The high degree of noncoding sequence changes in *LPL* (π = 0.0021) is largely due to variation at known human repeat elements and probably indicative of preferential mutations at such sites[29]. Second, Lander and colleagues have examined over 2 megabases (Mb) of DNA in 14–20 alleles, revealing a nucleotide diversity rate (π=0.0004) typical of the 3′ untranslated regions (UTR) comprising over 70% of the DNA they screened[30], and confirming previous estimates (Table 1). Third, a couple of groups have characterized long, contiguous regions (of about 17–28 kb), containing genes and flanking sequence[31,32]. They found that overall rates of variation (π=0.0005 and 0.0008) are higher in genes than that of non-coding DNA immediately in the vicinity of genes (π=0.0004); one segment shows a diversity rate of as much as 0.005!

Comparative human genomic diversity

The interesting aspect of these results is the remarkable 50-fold difference in nucleotide diversity (π = 0.0001–0.005). Whatever its causes, identifying and classifying this variation will remain an important task. 'Resequencing' microarrays, which elucidate

Table 2 • A sampling of human genomic nucleotide diversity

Gene/Locus	Region	No. segregating sites	Length surveyed (bp)	No. alleles sequenced	Nucleotide diversity (π)
Xq28 SL6CA8	genomic	13	28,432	2	0.0005
22q11 IG35B9	genomic	17	22,593	2	0.0008
22q11 IG50D10	genomic	83	16,643	2	0.0050
STSs/3′ESTs	genomic/coding/noncoding	3,027	2,260,195	14–20	0.0004±0.0002
CCR5	coding	9	1,168	500	0.0011±0.0004
GFRA1	coding	5	2,564	180	0.0003±0.0002
GFRA1	noncoding	4	836	180	0.0008±0.0005
LPL	coding	7	998	142	0.0005±0.0005
LPL	noncoding	81	8,736	142	0.0021±0.0010

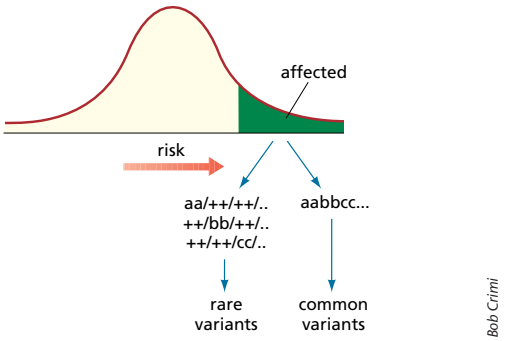

Fig. 1 Competing theories for complex disease inheritance (a, b, c and so on are susceptibility/protective alleles at multiple, independent loci). For a fixed disease incidence, individuals who are clinically affected can either have mutations at only one of many possible disease loci (in which case the mutant alleles are rare in the population) or harbour mutations at multiple loci simultaneously (in which case the mutant alleles are common in the population). These hypotheses are the extremes of many other possible intermediate scenarios.

nucleotide sequence in any specific genomic region in multiple individuals, will have a major role in the discovery of genome variation, particularly if we are able to scan large swathes of contiguous sequence.

I envision that comparative studies of genomic variation will become increasingly important and provide a different and unique view of the human genome. For these, we will need accurate estimates of sequence diversity, which will require analyses of larger DNA segments (of over 100 kb) and greater sample sizes. A diversity map of the human genome[29] will prompt the correlation of diversity with general features of the genome, such as gene content (UTR, exon versus intron, synonymous versus nonsynonymous), gene density, repeat element content, proximity to centromeres and telomeres, local recombination frequency, chromatin structure, banding pattern and any other imaginable feature. If evolution is the arbiter of how much variation survives, then comparative diversity studies may provide important clues to its nature.

Comparative diversity studies may also provide the answer to another question. We know that different genes show remarkably different rates of variation, but which general features do they associate with? We need to survey diversity in genes classified by function: function could be broadly assessed by expression type (for example, ubiquitous versus tissue-specific), tissue of expression (for example, brain versus immune system), timing of expression (for example, early development versus neonatal) or 'known' function (for example, ligands versus receptors). There is good reason to believe that these gene classes may show wide differences in variability, as they are likely to be subject to different selective forces. *De novo* resequencing and evaluating the frequencies of known SNPs/cSNPs in gene classes will contribute to the answer of this most important genomic property.

Maintenance of genomic diversity

How is diversity maintained? This is a most important question for population geneticists. A human genome diversity map, correlated with known sequence features, will generate widespread interest in this problem. There is strong evidence that deleterious mutations are rapidly eliminated from any population; these, such as mutations leading to most mendelian diseases, are present in all human groups at very low frequencies as a consequence of new mutations —the so-called mutation-selection balance. Thus, the existence of polymorphism raises, *per se*, the possibility of some active mechanism for their maintenance. Natural selection, acting through diverse mechanisms[9–11], is an attractive explanation for the maintenance of variability but its magnitude is still poorly known. Theoretical studies, carried out by Kimura in the mid-1950s, represented the apogee of this school of thought[33], but Kimura himself soon planted a bomb under the very edifice he created by declaring that much of evolution, at the protein level, is selectively neutral and determined by genetic drift[34] (chance). This 'neutralist' theory concedes that the vast majority of new changes are deleterious and that selection has a large role against deleterious mutations[35] ('purifying' selection). The controversy that arose around Kimura's proposals (the 'neutralist-selectionist' debate) centred on the nature of selection on the variation 'left behind'.

Molecular data are still largely on the side of Kimura's second thesis[20,35]; evidence of positive selection is generally rare except at some genes such as *HLA* (ref. 36). It seems likely that current methods for detecting seletion are not optimal and that future comparisons of within-species and between-species diversity will provide greater insight[37]; thus, for understanding patterns of variation in humans we will also need to study genomic evolution of the closely related great apes. Chip-based resequencing of human and great ape or primate genomes can be quite accurate, but further development of the technology is required to allow large-scale comparisons[38,39].

Geographic distribution of genomic diversity

Perhaps the most popular exploitation of extant genetic variation has been in the study of human origins and migration[7,40]. Anthropologists and geneticists have long been interested in reconstructing human evolutionary history using quantitative traits, polymorphic blood groups, serum proteins and enzymes; now, information on molecular variation, including nucleotide sequence data, is quickly becoming the norm. As human diversity is low, a great deal of data are necessary for discriminating between the different human origin stories with confidence. Although the current data are sparse, they permit some generalizations.

It is clear that sequence variation is ubiquitous throughout humanity, occurs continuously throughout the geographic range and occurs in most human populations; what differs are its frequencies. Comparative studies of within-group versus between-group genetic diversity have established that more than 90% of genetic variation is within human populations; thus, only a small fraction of the diversity is unique to groups[41,42]. This is why commonly used terms such as 'race' or 'deme' may not have scientifically useful definitions in humans. African populations, however, show greater diversity than other groups, reflecting their antiquity relative to populations in Asia, Europe and the Americas[43].

The new molecular data, on nuclear loci, mitochondrial DNA and Y-linked sequences, have given rise to an intriguing hypothesis.

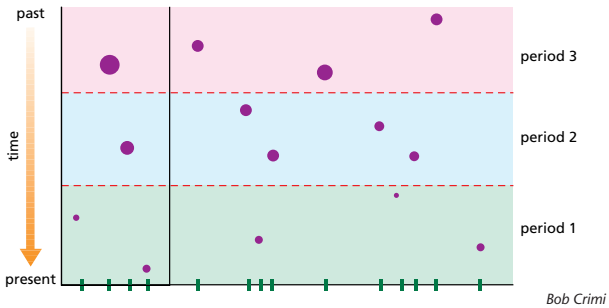

Fig. 2 The evolutionary origin of polymorphic sites in a genomic segment. All polymorphic sites (SNPs) within a region are not 'equal', in that they have arisen at different times in evolution (here arbitrarily designated and grouped as periods) and have different frequencies and other population genetic characteristics. Disease mutations that are older than a specific SNP (or a collection of SNPs) will have a smaller, if any, association with that SNP. Different collections of SNPs may be informative for mutations of different ages. The area of each circle denotes frequency.

Quantitative analysis of extant genetic variation increasingly support a single origin ('out-of-Africa') rather than a multi-regional scenario of human evolution. These data suggest that a severe population 'bottleneck' occurred approximately 100,000 years ago, during which our ancestors numbered only about 10,000 breeding adults[21]. Humans are thus a young species, explaining both the small gene diversity and the ubiquitous distribution of genetic variation around the world. A closer examination of the genome is unlikely to shake this view but may provide a more detailed history for us. The diversity at a very large number of genetic loci will increase the accuracy of estimates of ancestral size and bottleneck time, for African, Asian, European and American populations. Furthermore, by using loci that evolve at different rates, we can obtain a detailed demographic history over different time periods of human evolution, thereby clarifying the history of and relationships between the different continents.

The single-origin hypothesis and the global explosion of the human population over the last few centuries go some way to explaining why most rare alleles appear to have geographic specificity. These are probably mutations which have recently appeared and have not had sufficient time to diffuse across the global population. This is largely true for the many rare recessive mutations we observe in specific populations, such as the ΔF508 mutation that gives rise to cystic fibrosis[44]. It may also be true for the not-so-rare mutations leading to a complex phenotype, such as the Δ32 change in the CCR5 chemokine receptor that confers resistance to AIDS (ref. 45). To transform these speculations into concrete evidence, we need to date mutations. Knowledge of genomic diversity surrounding a specific sequence change will allow us to estimate the age of any mutation or variant and thus relate a sequence change to its frequency and history. DNA chips that allow the scoring of common SNPs and cSNPs in the region of a variant in multiple populations will be indispensable for this analysis.

The new data may lead to a revision, or rather a more accurate description, of our concept of population. Current studies of the geographic variation in genes are based on samples taken from populations defined by a host of social and cultural factors. Human populations sometimes have clear social boundaries, but whether they always represent a legitimate genetic unit is questionable[46]. The borders of any population can be quite diffuse, expanding and contracting over time. Thus, perhaps a broader survey of human genetic variation is in order. I would like to resurrect an idea initially proposed (to my knowledge) by the late Alan Wilson in the 1980s. Alan suggested that humans be sampled across the globe on the basis of geographic location and population density; in addition to donating a blood sample for DNA extraction, volunteers would also be asked about their population affiliation, spoken language, birth places of parents and other demographic and anthropological variables[47]. Such information, together with studies of genomic diversity, could not only answer major questions regarding genomic evolution and human evolution at the species level, but allow a critical evaluation of the utility and usefulness of population-based sampling. DNA chips that contain a sampling of the genomic variation across the genome, such as the recently constructed third-generation human SNP map[30], will have a critical role in these studies.

Genetic variation underlying complex phenotypes

One of the most significant genetic questions we shall be able to answer in the near future is: "what is the nature of genetic variation that underlies human phenotypes?". For the more than 100 identified genes associated with mendelian disease, the answer seems clear: mutational diversity at each locus is high; each mutation is rare, having occurred in recent human history (no older than 2,000 years); and each mutation is necessary and sufficient to cause the phenotype of interest[48]. However, most human phenotypes, including diseases, are 'complex' (in more ways than one!); mendelian patterns do not apply. It is suspected that the mutations that lead to a complex phenotype occur at multiple genes. The genetic model typically presented is one in which 'affected' individuals are those that lie above some biological threshold of risk (Fig. 1). If multiple genes are involved, then the central question is whether mutations at any one gene are necessary and sufficient to lead to the phenotype. The lack of mendelian segregation of a complex phenotype in most families argues against the sufficiency of a mutation at any one gene: either there is a strong environmental effect and/or multiple genes are involved. A mutation may be necessary if strong epistacy were to prevail. Consequently, the nature of genetic variation for complex traits determined by many loci is likely to be common alleles (polymorphisms) at these loci. And, if these alleles are common then they are probably very old, with ages of 10,000 years or more.

We need to answer some central questions regarding the nature of genetic variation of complex diseases. Are they at single or multiple genes? Is the mutational diversity high or low? Are the relevant alleles rare or common? Are they young or old? What is the nature of selection for or against them? Are individuals affected because they harbour too many susceptibility alleles or because they have too few protective alleles? Almost all of the contemplated studies of nucleotide sequence will assist in ferreting out the answers. These questions are not necessarily new: the bitter debate between the mendelists and biometricians earlier this century also centred on the nature of genetic variation and the role of selection in mendelian versus complex phenotypes[8].

We shall make little headway towards these goals unless we identify genes for complex diseases. The path ahead, however, is ill lit. The standard paradigm of positional cloning has been unsuccessful for complex traits. The multiplicity of genes underlying a complex phenotype can allow genetic mapping but frustrates refinement of the region, as recombinants cannot be unequivocally distinguished from the risk contributions of an unlinked locus[49]. Moreover, genetic mapping may not be an efficient way to identify genes that make only minor contributions to risk[5]. There are two closely related solutions to this dilemma, both enabled by the development of array technologies[4–6]. First, we can create a catalogue of common coding-sequence variants in human genes and test these directly for association with a phenotype. Second, we can use a genome-wide high-resolution map of known polymorphisms to scan the genome for marker-disease associations. The current excitement with SNPs, cSNPs and genome-wide SNP maps exists because it will allow us to travel these new paths.

Both of these strategies are based on assumptions that must be tested rigorously. First, constructing a catalogue of common cSNPs is feasible but assumes that common variants in coding sequences are the basis for many complex diseases. This is undoubtedly true in some cases, but what is the frequency distribution of these 'common' alleles? cSNPS of varying frequency need to be identified to test this view. Second, why limit the analysis to coding sequences? As the data obtained by King and Wilson suggest, it may well be that the majority of relevant mutations reside in regulatory regions[22]. Thus, we should identify variants in at least the proximal and distal regulatory sequences, bearing in mind that our poor understanding of 'regulatory' elements dictates the need for a more global approach. An approach in which marker-disease associations are sought takes into account our current ignorance. It will require the construction of a high-resolution map of genetic variants; SNPs are the natural candidates for this map because they are abundant, have a smaller mutation rate than microsatellites and can be genotyped *en masse* using microarray technology.

A map-based association search for multiple loci, each contributing to the total phenotype in a small yet measurable way, is also feasible but makes assumptions about the nature of genetic variation underlying complex traits. The success of this approach requires that at any 'culprit' locus, one variant allele dominates all other variant alleles in frequency. If so, then this allele can be indirectly recognized by its historical association with other SNPs in its neighbourhood—by haplotype analysis. This approach will be familiar to many, given its marked success in identifying mutations that give rise to mendelian phenotypes and relevance to gene tracking in isolated populations[50]. However, the association between a functional variant and neighbourhood SNPs is quite dependent on the history of that functional variant; its age and frequency determine the physical distance over which such associations persist and can be detected. The association of variants with one another (linkage disequilibrium) is a well-known feature of the human genome but its characteristics are still poorly understood[23,24,51]. The recent history of the human population, with its dramatic expansions in population size over the last few centuries, is a unique scenario in which linkage disequilibrium across much of the genome is expected.

Some have expressed doubt as to whether the 'linkage disequilibrium' approach to gene mapping of complex traits is a viable one[28,52]. Association patterns in a genomic segment are frequently complex, with no strict relationship between physical distance and the degree of association; this is only exacerbated with high allelic diversity at the culprit locus[28]. I suspect that this view arises from our less than optimal understanding of genomic variation patterns. SNPs which occur in a genomic segment have a specific spatial pattern, but each SNP also has a unique history, or moment of origin (Fig. 2). Thus, the distance between two polymorphisms is not the only arbiter of linkage disequilibrium between them; their times of origin, on average reflecting their relative frequencies, are also important. Thus, association between SNPs needs to be gauged quite differently than is currently the case, to account for both distance and time.

A simple but powerful approach would be to ask how much of the genome is shared between two randomly chosen chromosomes harbouring a specific variant. Younger SNPs have, on average, a lower frequencey than older ones, and occur on a background of haplotypes with different constellations of 'older' SNPs. A corollary to this is that two chromosomes harbouring a specific SNP will show greater identity over its length when the SNP is young. These lengths of identity are the critical parameters that define both the effects of recombination and history. A functional variant will show greater identity for SNPs in its neighborhood than those further away, enabling disease-gene mapping using a SNP map. Microarrays will have a major role, not only in the creation of this map, but also in mapping the components of complex phenotypes.

1. Collins, F.S. *et al.* New goals for the U.S. human genome project. *Science* **282**, 682–689 (1998).
2. Venter, J.C. *et al.* Shotgun sequencing of the human genome. *Science* **280**, 1540–1542 (1998).
3. Collins, F.S. Positional cloning moves from perditional to traditional *Nature Genet.* **4**, 347–350 (1995).
4. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
5. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
6. Collins, F.S., Guyer, M.S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
7. Cavalli-Sforza, L.L, Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton University Press, Princeton, 1994).
8. Provine, W.B. *The Origins of Theoretical Population Genetics* (The University of Chicago Press, Chicago, 1971).
9. Fisher, R.A. *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford, 1930).
10. Haldane, J.B.S. *The Causes of Evolution* (Longmans and Green, London, 1932).
11. Wright, S. *Evolution and the Genetics of Populations, Volume 2: The Theory of Gene Frequencies* (The University of Chicago Press, Chicago, 1997).
12. Wright, S. *Evolution and the Genetics of Populations, Volume 3: Experimental Results and Evolutionary Deductions* (The University of Chicago Press, Chicago, 1977).
13. Wright, S. *Evolution and the Genetics of Populations, Volume 3: Variability Within and Among Natural Populations* (The University of Chicago Press, Chicago, 1977).
14. Hacia, J. Resequencing and mutational analysis using oligonucleotide microarrays. *Nature Genet.* **21**, 42–47 (1999).
15. Southern, E. Mir, K. & Shchepinov, M. Molecular interactions on microarrays. *Nature Genet.* **21**, 5–9 (1999).
16. Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J. Expression profiling using cDNA microarrays. *Nature Genet.* **21**, 10–14 (1999).
17. Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**, 20–24 (1999).
18. Harris, H. Enzyme polymorphisms in man. *Proc. Royal Soc. London B* **164**, 298–310 (1966).
19. Nei, M. *Molecular Population Genetics and Evolution* (North-Holland Publishing Company, Amsterdam, 1975).
20. Nei, M. *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).
21. Harpending, H.C. *et al.* Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. USA* **95**, 1961–1967 (1998).
22. King, M.C. & Wilson, A.C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
23. Jeffreys, A.J. DNA sequence variants in the G gamma-, A gamma-, delta- and beta-globin genes of man. *Cell* **18**, 1–10 (1979).
24. Chakravarti, A. *et al.* Nonuniform recombination within the human beta-globin gene cluster. *Amer. J. Human Genet.* **36**, 1239–1258 (1984).
25. Li, W.H. & Sadler, L.A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
26. Ansari-Lari, M.A. *et al.* The extent of genetic variation in the CCR5 gene. *Nature Genet.* **16**, 221–222 (1997).
27. Angrist, M. *et al*. Human GFRA1: cloning, mapping, genomic structure, and evaluation as a candidate gene for Hirschsprung disease susceptibility. *Genomics* **48**, 354–362 (1998).
28. Nickerson, D.A. *et al.* DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
29. Chakravarti, A. It's raining SNPs, hallelujah? *Nature Genet.* **19**, 216–217 (1998).
30. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
31. Kawasaki, K. *et al.* One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome Res.* **7**, 250–261 (1997).
32. Eichler, E.E. *et al.* Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Human Mol. Genet.* **5**, 899–912 (1996).
33. Kimura, M. Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symp. Quant. Biol.* **20**, 33–53 (1995).
34. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
35. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
36. Hughes, A.L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class 1 loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
37. Hudson, R.R. Levels of DNA polymorphism and divergence yield important insights into evolutionary processes. *Proc. Natl Acad. Sci. USA* **90**, 7425–7426 (1993).
38. Hacia, J.G. *et al.* Detection of heterozygous mutations in *BRCA1* using high-density oligonucleotide arrays and two-colour fluorescence analysis. *Nature Genet.* **14**, 441–447 (1996).
39. Halushka, M. *et al.* Analysis of gene variation within and between humans and primates using high-density olligonucleotide arrays. in: *Genome Mapping, Sequencing and Biology* (Cold Spring Harbor, New York, 1998).
40. Seielstad, M.T., Minch, E. & Cavalli-Sforza, L.L. Genetic evidence for a higher female migration rate in humans. *Nature Genet.* **20**, 278–280 (1998).
41. Lewontin, R.C. The apportionment of human diversity. *Evol. Biol.* **6**, 381–398 (1972).
42. Nei, M. & Roychoudhury, A.K. Gene differences between Caucasian, Negro, and Japanese populations. *Science* **177**, 434–436 (1972).
43. Vigilant, L. *et al.* African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).
44. Kerem, B. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).
45. Smith, M.W. *et al.* Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. *Science* **277**, 959–965 (1997).
46. Juengst, E.T. Groups as gatekeepers to genomic research: Conceptually confusing, morally hazardous, and practically useless. *J. Kennedy Inst. Ethics* **8**, 183–200 (1998).
47. Committee on Human Genome Diversity, National Research council. *Evaluating Human Genetic Diversity* (National Academy Press, Washington, 1997).
48. McKusick, V.A. *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders* (12th edition)(Johns Hopkins University Press, Baltimore, 1998).
49. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
50. de la Chapelle, A. & Wright, F.A. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proc. Natl Acad. Sci. USA* **95**, 16–23 (1998).
51. Jorde, L.B. Linkage disequilibrium as a gene-mapping tool. *Amer. J. Human Genet.* **56**, 11–14 (1995).
52. Clark, A.G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Amer. J. Human Genet.* **63**, 595–612 (1998).