

Resources for Genetic Variation Studies

David Serre and Thomas J. Hudson

McGill University and Genome Quebec Innovation Center, Montreal, Quebec H3A 1A4, Canada; email: david.serre@mail.mcgill.ca, tom.hudson@mcgill.ca

Annu. Rev. Genomics Hum. Genet. 2006.
7:443–57

First published online as a Review in
Advance on June 7, 2006

The *Annual Review of Genomics and Human
Genetics* is online at
genom.annualreviews.org

This article's doi:
10.1146/annurev.genom.7.080505.115806

Copyright © 2006 by Annual Reviews.
All rights reserved

1527-8204/06/0922-0443\$20.00

Key Words

single nucleotide polymorphisms, human diversity, haplotypes, linkage disequilibrium, complex diseases, natural selection, allele-specific expression

Abstract

The rapid growth of genome-wide diversity databases, as well as ongoing large-scale resequencing projects targeting genes and other functional components of our genome, provide valuable resources of natural variation at the DNA sequence level. In this review, we briefly summarize the wealth of data on DNA polymorphisms in humans, the distribution of this diversity in the genome as well as among individuals, and the consequence of recombination on its organization. These data provide a set of powerful tools that can be used to better understand inherited phenotypic variation in humans. We discuss the implications for the design of studies investigating correlations between genotypes and phenotypes, both at the fundamental level of genome function and regulation, and for the mapping of disease genes.

SNP: single nucleotide polymorphism

INTRODUCTION

The completion of the initial draft of the human genome sequence (38) provided the first in-depth description of the different components of the human genome and considerable information about its gene content and gene positions. The genome sequence has been used extensively as a starting point for many subsequent studies focusing on the function of genes or the functional elements of particular loci linked to disease. Recently, human genetics shifted to large-scale studies of human genome variation. The large-scale polymorphism data that are made available can now be used to efficiently study genetic differences among humans. This will dramatically increase our understanding of how primary sequence can influence inherited phenotypes. Many new methods allow the investigation of gene expression patterns or protein variability, but since the primary cause of inherited phenotypes is determined by the genome sequence, the understanding of these phenotypes requires the identification of the DNA polymorphisms responsible for them.

In this review, we first summarize recent progress in our understanding of the distribution of natural variation in humans. We then provide a succinct description of different resources on genetic diversity, and finally discuss some of their potential uses for functional studies. We do not aim to provide a detailed description of the organization of the genetic diversity in humans (that can be found elsewhere; see, e.g., Reference 3), but prefer to concentrate on how we can utilize recently generated information on natural variation at the DNA sequence level to better study and understand inherited phenotypic differences. Here we consider DNA polymorphisms resulting from single nucleotide differences (i.e., SNPs) and do not cover other types of DNA polymorphisms such as nucleotide insertions/deletions, copy number variation, chromosomal rearrangement, or structural variation of the genome, which are described elsewhere (6, 66, 71).

DISTRIBUTION OF NATURAL VARIATION IN HUMANS

Where Do SNPs Come From?

Single nucleotide polymorphisms (SNPs) most often result from base substitutions occurring through a nonrepaired error that occurs during DNA replication (other causes of DNA polymorphisms include gene conversion and duplication; see, e.g., Reference 21). The mutation rate at most positions of the human genome is relatively low (on the order of 10^{-8} substitution per base pair per generation) compared with the most recent common ancestor of any two individuals (on the order of 10,000 generations). Therefore, the vast majority of allelic differences between individuals are inherited rather than *de novo* mutations. (This is often represented by the infinite-site model in evolutionary genetics, in which each position can only mutate once in the genealogy of the sample). Two individuals sharing the same allele at one position are most likely identical by descent (for this specific portion of the DNA) rather than through two independent mutations (i.e., homoplasy). One notable exception to this model concerns CpGs, a guanine immediately following a cytosine on a DNA molecule (see below).

The types of mutations as well as their distribution across the genome are not random. Transitions (i.e., the substitution of a pyrimidine by a pyrimidine, or of a purine by a purine) are more likely to occur than transversions (i.e., the substitution of a purine by a pyrimidine, or inversely). This bias is due to differential repair mechanisms (68) and to the importance of CpGs. CpGs harbor a mutation rate 3–10 times higher than the average mutation rate: Cytosines in CpGs are often methylated and more likely to mutate into thymines (67).

After a mutant allele is introduced in the gene pool, its fate is determined by the interaction of two evolutionary forces: random genetic drift and natural selection. Drift affects the distribution of a SNP in the population

by random sampling of different alleles at each generation (only a small fraction of all possible gametes are transmitted to the next generation). If only drift acts on a particular region of the genome, a SNP frequency varies randomly from generation to generation until one allele eventually reaches fixation (either 100% or 0%), and the time before fixation occurs is mainly determined by the population size. In other words, if only genetic drift acts, polymorphisms are only transient products of random fluctuations. However, natural selection affects the probability that a particular variant is passed to the next generation. It can either increase the probability and speed of fixation of a newly arisen allele if the mutant allele confers a fitness advantage (i.e., selective sweep or positive selection), remove new deleterious variants from the gene pool (i.e., negative selection), or maintain several alleles in the gene pool over extended periods of time (i.e., balancing selection).

Distribution of Diversity in Humans

Whereas mutations are generated essentially randomly in DNA molecules and random genetic drift affects chromosomes as a whole, selection, in recombining genomes such as humans, acts differently on distinct regions of the genome. Consequently, the level of polymorphism varies greatly between different portions of the genome. For example, regions of noncoding DNA (introns and intergenic regions) typically harbor a much higher diversity than coding DNA (25), regulatory elements (such as promoters or splicing sites), or conserved nongenic elements (2, 17). Similarly, within exons, synonymous and nonsynonymous positions differ in their diversity, as do, at a lower scale, twofold from fourfold degenerated sites (i.e., nucleotides that can be changed into, respectively, one or any other nucleotide and still code for the same amino acid) (25). Diversity also varies along the genome over several Megabases, even after correcting for gene or Guanine-Cytosine

(GC) content (27). This large-scale variation probably represents differences in recombination rates that affect the frequency of SNPs (either directly through a putative mutagenic effect or indirectly via background selection) (27).

If one considers the distribution of genetic diversity, not across regions of the genome, but across individuals, the frequencies of DNA polymorphisms also differ greatly. For one locus this is often summarized by the frequency spectrum that displays the frequency of all SNPs in a population. In a random mating population of constant size and without selection acting at the locus considered, most SNPs are expected to be present at low frequency and only found in one or few individuals, whereas very few SNPs will be common (**Figure 1**). Deviations from this pattern may reflect violation of either the neutrality of the loci or of the demographic assumptions. For example, population growth and positive selection increase the proportion of rare alleles (i.e., alleles with low frequency), whereas balancing selection and population substructure increase the proportion of intermediate alleles (**Figure 1**). Analyzing the frequency spectrum is one of the most common tools in population genetics, and many tests of neutrality rely on it directly or indirectly (see, e.g., Reference 48 for a comprehensive review of selection and its footprints in DNA sequence). Most of the recent large-scale projects that generated genome-wide polymorphism data used genotyping of selected SNPs (rather than more expensive and less high-throughput resequencing), which introduces a skew in the frequency spectrum, especially if the project focuses on intermediate frequency polymorphisms, as in the HapMap project (**Figure 1**). Frequency spectra obtained from such data can be adjusted to take into account the ascertainment bias (see, e.g., 49, 50), but a reliable reconstruction of the evolutionary history of a particular locus currently still requires resequencing data that produce unbiased diversity estimates and allow more rigorous testing of neutrality.

Frequency spectrum: the distribution of polymorphisms according to their frequencies

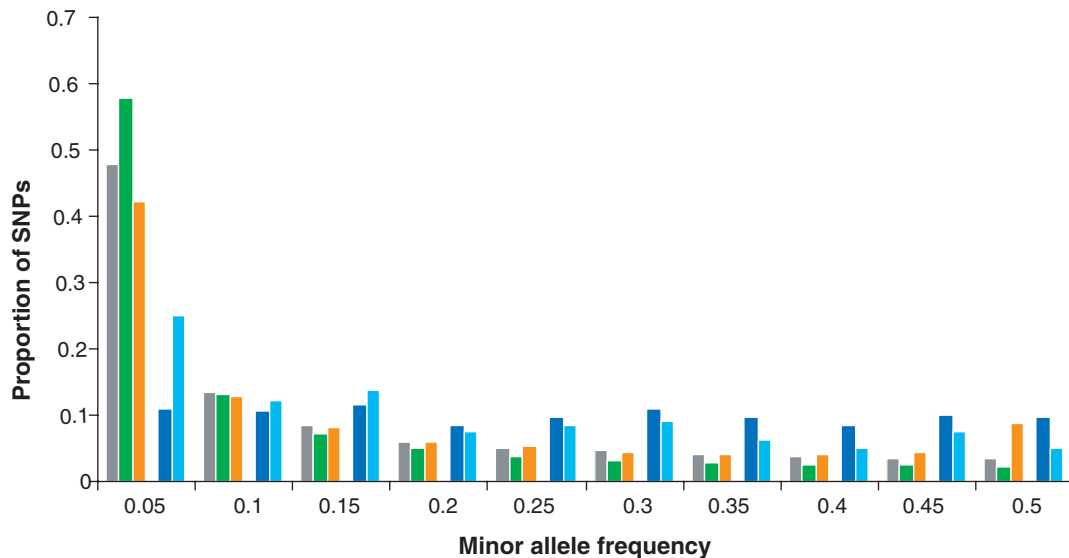


Figure 1

Frequency spectra for the standard neutral model, a scenario of population growth, and population substructure as it is observed in the HapMap and ENCODE data sets. This figure shows the distribution of polymorphic sites according to their minor allele frequency (on the x axis). The expected distribution under the standard neutral model (i.e., a pan-mictic population of constant size in which genetic variation is not affected by natural selection) is displayed in gray. The distributions of single nucleotide polymorphisms under a model of population growth and population substructure are displayed, respectively, in green and orange. The observed distributions in the HapMap and ENCODE data sets are shown in deep blue and light blue (for the individuals of European ancestry), respectively.

Analyses of the genetic diversity among humans reveal very little differentiation among populations. Grouping individuals according to their geographical origin is feasible but requires large data sets of highly informative genetic markers (see, e.g., 62). Differences between individuals from very distant populations only account for roughly 10% of the total variability at most loci, whereas differences between individuals within a particular population represent 90% (e.g., 3, 4, 19, 60, but see also 18). Additionally, the geographic distribution of the diversity seems to best be explained by large gradients of allele frequency rather than by well-defined and separated clades corresponding to continental or “racial” entities (11, 57, 60, 65, but see also 61). In this perspective, the data produced by Perlegen (29) and the International HapMap Consortium (2) should not be considered a

description of the entire worldwide human diversity. Nevertheless, these projects will be particularly useful in the investigation of other populations because they offer a detailed description of the allele frequencies at the extremes of some of the worldwide genetic diversity distributions (see discussion below).

Recombination, Haplotypes, and Haplotype Blocks

One of the most amazing developments in our understanding of the organization of genetic diversity concerns the phenomenon of recombination and its consequences on the allelic correlations between neighboring SNPs. Here we use the terminology “recombination” to refer to crossovers and the exchange of chromatid fragments during meiosis. (For the resolution of Holliday junctions resulting

in gene conversion and an analysis of the importance of this mechanism in the human genome, see Reference 51.) Without recombination, two polymorphisms can only yield three combinations of alleles on the same DNA strand (or haplotypes). Because each position can be hit only once by a mutation, one should end up with three different possible combinations if no haplotype is lost by genetic drift or eliminated by selection (**Figure 2d**). This nonrandom association of alleles (on the figure the blue allele is always associated with the red) is referred to as linkage disequilibrium (LD). If a crossover occurs between the two polymorphic positions, it may generate a fourth haplotype and break down LD between the two markers (**Figure 2e**). Interestingly, recombination does not occur evenly but happens much more frequently in a small fraction of the genome. These hot spots of recombination (typically 1–2 kb long) are frequent across the genome (roughly 1 every 50 kb) and account for more than 80% of all recombinations (15, 42, 46). Hot spots of recombination appear to be highly dynamic and at least some of them differ in intensity and/or location between humans and chimpanzees as well as possibly between human populations (15, 58, 59, 76). This may influence the design of association studies and the choice of markers for studying a given population, but because recombination rates seem conserved among human populations over a larger scale (>1 Mb), variations in recombination pattern among humans are unlikely to affect linkage studies (34, 64).

A popular and simple way to describe patterns of genetic diversity is to partition chromosomes into “blocks” of SNPs in high LD with each other (23). This usually allows summarizing polymorphism diversity into three to five common haplotypes that account for most of the sample variability. However, one should keep in mind that the definition and particularly the limits of these entities are highly arbitrary, as the resulting blocks

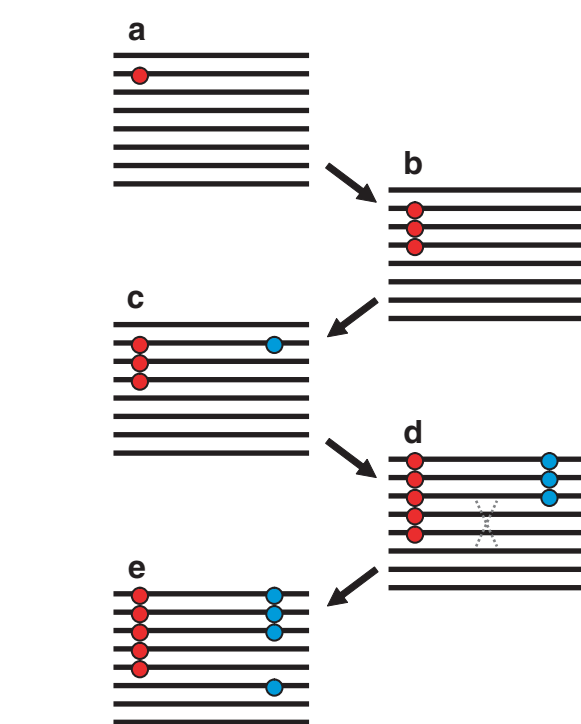


Figure 2

Effect of genetic drift and recombination on haplotypes. (a) A polymorphism is introduced by a mutation in a population of identical chromosomes. (b) The haplotype carrying the red allele increases in frequency in the population by genetic drift. (c) A second mutation, occurring on a haplotype carrying the red allele, introduces a new polymorphism. (d) The haplotype carrying both derived alleles increases in frequency in the population by genetic drift. At this point all chromosomes carrying the blue allele at the second polymorphic site also carry the red allele at the first site [resulting in high linkage disequilibrium (LD)]. (e) A crossover involving two different haplotypes (red-blue and ancestral) occurs between the two single nucleotide polymorphisms, leading to a reassortment of the alleles: The blue allele no longer occurs exclusively on haplotypes carrying a red allele (reducing the LD).

depend both on the SNP density and on the number of individuals studied. These “blocks” should not be a priori considered as stretches of DNA sequence where no recombination occurs, separated from each other by recombination hot spots (even if this will be true in some instances). Note that the LD patterns observed in a sample result both from recombination events and the population history of this sample (56, 74, 75).

Haplotype: a combination of alleles that are located close together

Linkage disequilibrium (LD): the nonrandom association of alleles at two or more loci

AVAILABLE RESOURCES FOR STUDYING HUMAN GENETIC DIVERSITY

The Single Nucleotide Polymorphism Database

The single nucleotide polymorphism database (dbSNP) serves as a central depository of SNPs in the public domain. It provides a description of the SNP and its flanking regions and links to multiple National Center for Biotechnology Information (NCBI) Internet pages. The latest release (Build 125) contains more than 27 million SNPs, more than 4 million of which lie within genes. More than 6 million SNPs have been validated and are likely relatively common in the population. Frequency descriptions are available for ~600,000 SNPs, but this figure will dramatically increase with the incorporation of the HapMap phase II data in dbSNP.

The HapMap and Perlegen Projects

Both the International HapMap Consortium (2) and Perlegen Sciences (29) recently produced and released genome-wide genotyping data in samples from several human populations. In both cases, the goal was to cover the entire genome with common DNA polymorphisms that are evenly spaced. The density obtained is of roughly 1 SNP every 2 kb for the Perlegen data set and of 1 every 3 kb for the first phase of the HapMap project (and the latter density will increase rapidly up to 1 SNP every kilobase when phase II is completed). Both data sets are freely available on the Internet.

The International HapMap Consortium (2) genotyped individuals from four populations: Centre d'Etude du Polymorphisme Humain (CEPH) (Utah residents with Northern or Western European ancestry), Yoruba from Ibadan (Nigeria), Han Chinese from Beijing (China), and Japanese from Tokyo (Japan). The sampling scheme varies for different pop-

ulations: The Yoruba and CEPH individuals consist of 30 unrelated trios while the Han Chinese and Japanese are represented by 45 and 44 unrelated individuals, respectively. The SNPs were chosen in order to obtain a high proportion of intermediate frequency polymorphisms (i.e., with a minor allele frequency higher than 5%). This makes this data set especially valuable for choosing markers for medical studies as they will likely maximize the information for each individual, but it complicates their use in population genetic analyses because the allele frequency spectra are biased (see above).

Perlegen Sciences analyzed three populations (CEPH, African Americans, and Han Chinese from the Los Angeles area) represented by 23 or 24 individuals each. The SNPs mostly came from a large resequencing effort using 48 individuals (for ~70% of the SNPs) and partly from dbSNP (29). The frequency spectrum is also biased toward intermediate frequency, but not as strongly as for the HapMap data set (8, 47).

Cell lines of the individuals genotyped in both projects are available at the Coriell Institute for Medical Research.

The Human Genome Diversity Project

The goal of the Human Genome Diversity project (10) is to provide researchers with a defined panel of DNAs (from cell lines) from more than 1000 individuals from 51 populations. These individuals have so far been genotyped for ~800 autosomal microsatellite markers that are evenly distributed in the genome (60, 62). Subsequent genotypes will soon be released in a public data base.

The ENCODE Project

The ENCODE project provides in-depth analyses of functional elements for 44 regions of 0.5–2 Mb distributed across many

chromosomes that represent 1% of the human genome (38). Some regions were chosen manually based on data already available [e.g., Cystic Fibrosis Transmembrane conductance regulator (CFTR) or β -globin loci], whereas others were picked randomly to ensure a global representation of the human genome diversity (i.e., representative of the diversity in gene content and nonexonic conservation). All these regions are analyzed by a large variety of methods (e.g., ChIP-chip experiments, expression microarrays, comparative genomics, or in silico predictions) in order to identify and characterize all functional elements and to understand the gene expression regulatory mechanisms.

From these 44 regions, 10 500-kb regions have been included in the HapMap project and resequenced entirely in 48 individuals (16 Yoruba, 8 Japanese, 8 Han Chinese, and 16 CEPH), and all identified SNPs were then genotyped in all HapMap individuals. These regions are densely covered with polymorphisms (with an average of 6 SNPs/kb) and are useful to develop and test new analysis methods or to estimate the effect of the SNP density on a particular type of analysis.

Gene-Based Resequencing Projects

Detailed descriptions of human diversity on a gene-by-gene basis are provided by ongoing projects such as SeattleSNPs (1) and the National Institute of Environmental Health Sciences (NIEHS) SNPs Program at the University of Washington (14). Both projects are based on resequencing exons, untranscribed regions, and (parts of) introns of genes involved in inflammation (SeattleSNPs) or environmental responses (NIEHS SNPs). Sequencing data are available on the Internet.

Similarly, the Sanger's Exon Resequencing project aims to resequence exons of all known genes in a panel of 48 individuals of European origin. Regulatory regions (promoters and enhancers) may be included later.

FROM NATURAL VARIATION TO PHENOTYPIC DIFFERENCES

Our knowledge of "natural" genetic diversity in humans and its organization has greatly progressed, notably with the results of genome-wide projects. The next step for human geneticists will be to use this tremendous amount of information on natural variation to unravel the still mostly mysterious connections between DNA and phenotypic differences. The consequences of this recent progress include:

- First and most obvious, the large amount of information concerning human diversity in individuals from different populations can be directly used to better design association studies for a particular phenotype.
- Second, the cell lines and/or DNA used for the Perlegen, HapMap, and Human Genome Diversity projects are available and allow researchers to study individuals that have been genotyped at numerous loci. Similarly, the ENCODE regions offer the opportunity to compare, in a single region, the descriptions of several potential regulatory mechanisms of gene expression as well as their associations with particular genotypes.
- Finally, these projects promote the development of new technologies that are rapid, low cost, and allow genome-wide genotyping from much less DNA.

Below we describe some of the new avenues to link genotypes to phenotypes, both in a fundamental or disease-related manner. We do not claim to deliver a "handbook of functional studies," but rather to introduce briefly some of the most interesting directions of research.

Variation Affecting Gene Expression

Given the relatively low genetic diversity that affects the amino acid composition of proteins among humans or between humans and our closest living relative the chimpanzee,

Allele expression difference: at some loci, the two alleles of a gene in a heterozygote individual are not expressed equally

it has been suggested that variation affecting gene expression may play a large role in the observed phenotypic variation (see, e.g., 35). Recent studies emphasize the complexity and variability of human gene expression patterns and their multiple levels of regulation (transcription, splicing, translation, or degradation). The HapMap and Perlegen projects genotyped several million DNA polymorphisms using cell lines that are available to the research community. We believe that these cell lines represent one of the most interesting materials to better understand how DNA polymorphisms influence mRNA diversity in humans. Cell lines are obviously simplified biological models (compared to actual complex tissues) but may turn out to be especially useful for understanding fundamental regulatory mechanisms of gene expression.

For example, large-scale gene expression measurements performed on these cell lines allow the correlation of total or allele-specific mRNA levels with particular DNA polymorphisms. Such experiments have been done in mouse (7, 12, 32), yeast (5), and to some extent in humans (45), but using much smaller polymorphism data sets. The high densities of current genome-wide polymorphism data sets have the potential to assist in the detection of small regions harboring functional regulatory variants for particular genes. This avenue is especially promising for the identification of *cis*-regulatory polymorphisms (13, 69) but, given the large amount of data, it may even highlight some master regulators simultaneously affecting many genes in *trans* (once the multiple testing issues involved in such experiments are appropriately and convincingly addressed). Additionally, the use of trios in the International HapMap project and the availability of cell lines from additional (nongenotyped) members of the same families provide opportunities to tease apart inherited variation (i.e., genetically determined) from environmental/stochastic variation, or gene expression variation due to individual

differences in epigenetic patterns such as imprinting (63), X inactivation (9), or random monoallelic expression (54).

Recent studies show that gene expression not only differs between individuals but also that, in some instances, the two alleles are differentially expressed within the cells of a single individual (24, 53, 54, 77). This phenomenon can be used to identify over- or underexpressing haplotypes and to define candidate regulatory elements of a given gene. The large number of coding SNPs provided by the Perlegen and HapMap projects allow testing particular heterozygote individual cell lines for allelic expression differences for almost all genes of the human genome. This approach may not only help identify regulatory elements of genes differentially expressed among humans but may also have some direct medical implications. Several loci have been implicated in disease etiology by reproduced linkage/association analyses without the identification of any coding mutation or splice variants. It is possible that, in such instances, gene expression-level differences may lead to the disease (see, e.g., 26, 31). Thus, identifying regulatory SNPs or, at least regulatory haplotypes, may greatly aid in the understanding of mechanisms leading to disease.

Finally, one important mechanism regulating mRNA diversity among individuals is the potential for a single mRNA molecule to be spliced into several variants and thus generate different proteins (33, 38). So far, alternative splicing has mainly been studied between different tissues from the same individuals, and it remains unclear whether, or to what extent, the distribution of the spliced forms differs among healthy individuals. One way to address this question is to use expression arrays covering several exons (or possible exon junctions) of human transcripts to obtain an overview of genome-wide exon expression (33, 39). Correlation between such data and genome-wide genotypes may allow identification of polymorphic splice regulators.

Such information would provide a new tool to investigate genetic disease as several recent studies shed light on the influence of splicing regulation (or misregulation) in causing disease (see, e.g., 20).

Human Evolution and Natural Selection

Using the HapMap and Perlegen data sets, several groups have identified regions with high levels of population differentiation, low levels of diversity, or unusually long stretches of DNA sequence in high or complete LD (2, 8, 50, 72). All these loci are promising candidates to further study the influence of natural selection on the human genome. Validating these regions, as well as integrating information from other species [such as the chimpanzee genome sequence that was recently released (43)], will shed light on where and how natural selection acted during human evolution to shape current human genome diversity. Additional studies are necessary to validate these potentially interesting loci as it is still unclear whether these observations are actual genomic differences or artifacts of the ascertainment in the choice of the SNPs. One cannot reject the possibility that some of these loci may be chance events, given the large amount of data analyzed. Resequencing the regions of interest in a large worldwide sample (e.g., using the CEPH-HGDP panel) would help solve some of these uncertainties and could contribute to the understanding of what makes or made humans human. These results can also have important medical implications by identifying alleles involved in common diseases. Under the assumption that at least some common diseases are due to common variants, one would expect that the disease alleles may, in particular environmental circumstances, confer a benefit to the carrier that would balance the deleterious effect of the disease. Such loci would evolve under balancing selection and thus could be identified by searching the entire genome for loci with

a local excess of diversity and highly differentiated haplotypes (70). Alternatively, some common diseases may result from a change in environment during the last phases of human evolution such that the derived protective alleles (formerly deleterious) have not yet reached fixation in humans (16). These alleles should harbor the footprints of a recent (in fact, ongoing) selective sweep [i.e., low diversity, excess of rare alleles, extended regions of high LD (48)] and could also be identified by a genome-wide search for selected loci performed in healthy individuals.

Medical Genetics: Toward Genome-Wide Studies

The most important contribution from these large resources of human genetic diversity will likely concern the medical field, with, hopefully, the identification of genes (and their variants) involved in genetic disorders and a better understanding of the molecular mechanisms leading to disease.

Both Perlegen and the International HapMap data sets can now be used to identify risk or protective alleles involving a particular disease. The HapMap project favored this application by preferentially selecting intermediate allele frequency polymorphisms that are more informative for linkage or association studies (compared to rare alleles). Additionally, these data lead to a very precise description of the LD patterns in the human genome. Using a set of markers evenly distributed on the physical map (e.g., one marker every ten kb) may be inefficient in testing a significant fraction of the genome with high recombination rates, given that markers in these regions are poorly correlated and thus may have not been “tagged” by the markers used. Incorporating information about the recombination patterns while choosing the markers allows one to use the LD information to aptly select SNPs [tag-SNPs (tSNPs)] for efficient coverage of the genome. This will lead to a higher marker density in regions of high

recombination rate and fewer markers (and less redundancy) in regions of low recombination rates. Thus, one can anticipate that these data will be successful to both narrow down previously identified candidate regions using a denser map of markers (37, 44) and to identify new candidate loci through whole-genome association scans (36). Completing these two large-scale projects promoted rapid developments in genotyping technologies. It is now possible to perform rapid genome-wide scans using hundreds of thousands of SNPs in many individuals. These new technologies will dramatically increase the power of association and linkage studies and allow the detection of alleles with lower individual contribution to the phenotype. The results of the first studies using these technologies for genome scan in case-control settings are very promising (28, 36, 37, 44), and the path of discoveries will likely increase rapidly. However, some issues will potentially become more problematic with the increase in sample size and markers and need to be appropriately addressed by each study [see, e.g., correction for multiple testing (30, 52) and epistasis analysis (41), population stratification (22, 40, 55), or cryptic relatedness (73)].

Finally, and perhaps more importantly, both Perlegen and HapMap projects enable one to efficiently investigate diseases in individuals of non-European ancestry. So far, most genetic studies have focused on North Americans or Europeans, which results in an important bias in the genomic markers that are available (including their allele frequency description). For many genes, resequencing efforts concentrated on populations of European ancestry and thus most SNPs reported were biased toward high allele frequencies in European individuals. The availability of allele frequency descriptions for populations of Southeast Asian and African ancestries will remove this bias and help the design of adequate (or better) sets of markers for each cohort tested. Even if the pop-

ulation tested in a particular medical study is geographically distant from the populations used in these projects, combining allele frequency and LD information from two or three of the populations will likely perform much better than a European-b(i)ased set of markers. One can therefore expect that the number of non-European association studies will rapidly increase and perhaps identify population-specific risk factors. This can also stimulate research on genetic diseases affecting mostly non-European populations that have been understudied so far (with a few exceptions, most notably on malaria resistance).

CONCLUSION AND FUTURE DIRECTIONS

The information and resources resulting from the recent completion of genome-wide diversity projects will surely catalyze the discovery of common genetic variants affecting disease risk, and increase our understanding of gene expression regulation. This is particularly promising for medical genetics, given the availability of numerous genetic markers and adequate technologies to analyze thousands of individuals. We note that the current mapping approaches that use common SNP markers are powerful in finding common disease alleles. Ongoing technological developments in DNA sequencing will eventually complement (or replace) genotyping approaches and allow the identification of rare disease variants. With microarray features shrinking, the development of ultra-high-throughput sequencing technologies, or the launching of projects such as the Medical Sequencing Program at the National Human Genome Research Institute, identification of disease variants by genome resequencing methods will likely appear in the lab very soon, less than 10 years after the completion of the first human genome sequence.

UNRESOLVED ISSUES

1. What is the proportion of common variants versus rare variants responsible for common genetic diseases?
2. What proportion of functional polymorphisms involved in complex diseases affects protein structure, gene expression, or splicing?
3. Can we develop robust statistical methods to reliably identify *trans*-regulator elements using association between genotypes and gene expression data?
4. What is the importance of insertions/deletions/rearrangements relative to SNPs in the etiology of common genetic diseases?

LITERATURE CITED

1. Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, et al. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2:e286
2. **Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. 2005. A haplotype map of the human genome. *Nature* 437:1299–320**
3. Barbujani G, Goldstein DB. 2004. Africans and Asians abroad: genetic diversity in Europe. *Annu. Rev. Genomics Hum. Genet.* 5:119–50
4. Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. 1997. An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci. USA* 94:4516–19
5. Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752–55
6. Buckley PG, Mantripragada KK, Piotrowski A, Diaz de Stahl T, Dumanski JP. 2005. Copy-number polymorphisms: mining the tip of an iceberg. *Trends Genet.* 21:315–17
7. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, et al. 2005. Uncovering regulatory pathways that affect hematopoietic stem cell function using “genetical genomics.” *Nat. Genet.* 37:225–32
8. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, et al. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15:1553–65
9. Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434:400–4
10. Cavalli-Sforza LL. 2005. The Human Genome Diversity Project: past, present and future. *Nat. Rev. Genet.* 6:333–40
11. Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The History and Geography of Human Genes*. Princeton, NJ: Princeton Univ. Press
12. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, et al. 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* 37:233–42
13. **Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437:1365–69**
14. Crawford DC, Akey DT, Nickerson DA. 2005. The patterns of natural variation in human genes. *Annu. Rev. Genomics Hum. Genet.* 6:287–312

2. One of two seminal papers (with Ref. 29) providing a detailed genome-wide description of the diversity patterns in humans.

13. Follow-up of a previous study associating genotypes to gene expression differences (restricted to *cis*-regulation).

16. Discusses possible links between natural selection and genetic disease etiology.

29. One of two seminal papers (with Ref. 2) providing a detailed genome-wide description of the diversity patterns in humans.

15. Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, et al. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.* 36:700–6
16. **Di Rienzo A, Hudson RR. 2005. An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet.* 21:596–601**
17. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, et al. 2006. Conserved non-coding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.* 38:223–27
18. Edwards AW. 2003. Human genetic diversity: Lewontin's fallacy. *Bioessays* 25:798–801
19. Excoffier L, Hamilton G. 2003. Comment on "Genetic structure of human populations." *Science* 300:1877; author reply
20. Faustino NA, Cooper TA. 2003. Pre-mRNA splicing and human disease. *Genes Dev.* 17:419–37
21. Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. 2004. Complex SNP-related sequence variation in segmental genome duplications. *Nat. Genet.* 36:861–66
22. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, et al. 2004. Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* 36:388–93
23. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. 2002. The structure of haplotype blocks in the human genome. *Science* 296:2225–29
24. Ge B, Gurd S, Gaudin T, Dore C, Lepage P, et al. 2005. Survey of allelic expression using EST mining. *Genome Res.* 15:1584–91
25. Graur D, Li W.-H. 2000. *Fundamentals of Molecular Evolution*. Sunderland, Massachusetts: Sinauer Assoc.
26. Gretarsdottir S, Thorleifsson G, Reynisdottir ST, Manolescu A, Jonsdottir S, et al. 2003. The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.* 35:131–38
27. Hellmann I, Prufer K, Ji H, Zody MC, Paabo S, Ptak SE. 2005. Why do human diversity levels vary at a megabase scale? *Genome Res.* 15:1222–31
28. Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, et al. 2006. A common genetic variant is associated with adult and childhood obesity. *Science* 312:279–83
29. **Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, et al. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* 307:1072–79**
30. Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6:95–108
31. Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, et al. 2000. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat. Genet.* 26:163–75
32. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, et al. 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* 37:243–53
33. Johnson JM, Castle J, Garrett-Engle P, Kan Z, Loerch PM, et al. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302:2141–44
34. Jorgenson E, Tang H, Gadde M, Province M, Leppert M, et al. 2005. Ethnicity and human genetic linkage maps. *Am. J. Hum. Genet.* 76:276–90
35. King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–16

36. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308:385–89
37. Laitinen T, Polvi A, Rydman P, Vendelin J, Pulkkinen V, et al. 2004. Characterization of a common susceptibility locus for asthma-related traits. *Science* 304:300–4
38. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
39. Le K, Mitsouras K, Roy M, Wang Q, Xu Q, et al. 2004. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res.* 32:e180
40. Marchini J, Cardon LR, Phillips MS, Donnelly P. 2004. The effects of human population structure on large genetic association studies. *Nat. Genet.* 36:512–17
41. Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* 37:413–17
42. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–84
43. Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87
44. Mira MT, Alcais A, Nguyen VT, Moraes MO, Di Flumeri C, et al. 2004. Susceptibility to leprosy is associated with PARK2 and PACRG. *Nature* 427:636–40
45. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, et al. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature* 430:743–47
46. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310:321–24
47. Nielsen R. 2005. Human genomics: disclosure of variation. *Nature* 434:288–89
48. Nielsen R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet.* 39:197–218
49. Nielsen R, Hubisz MJ, Clark AG. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168:2373–82
50. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15:1566–75
51. Padhukasahasram B, Marjoram P, Nordborg M. 2004. Estimating the rate of gene conversion on human chromosome 21. *Am. J. Hum. Genet.* 75:386–97
52. Palmer LJ, Cardon LR. 2005. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* 366:1223–34
53. Pastinen T, Hudson TJ. 2004. Cis-acting regulatory variation in the human genome. *Science* 306:647–50
54. Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, et al. 2004. A survey of genetic and epigenetic variation affecting human gene expression. *Physiol. Genomics* 16:184–93
55. Pritchard JK, Donnelly P. 2001. Case-control studies of association in structured or admixed populations. *Theor. Popul. Biol.* 60:227–37
56. Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69:1–14
57. Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* 15:R159–60
58. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, et al. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat. Genet.* 37:429–34
59. Ptak SE, Roeder AD, Stephens M, Gilad Y, Paabo S, Przeworski M. 2004. Absence of the TAP2 human recombination hotspot in chimpanzees. *PLoS Biol.* 2:e155

36. A success story in using whole-genome association to identify alleles involved in complex diseases.

40. & 41. Two theoretical papers addressing the influences of population stratification and the investigation of epistatic effects on large-scale association studies.

46. Describes the recombination patterns and the importance of recombination hot spots across the entire human genome.

53. Describes the role of polymorphisms on cis-acting regulation and their implications for disease mapping.

56. Reviews linkage disequilibrium in humans and discusses the importance of demography on the patterns observed and the different methods used to assess LD.

60. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* 102:15942-47
61. Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* 1:e70
62. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. 2002. Genetic structure of human populations. *Science* 298:2381-85
63. Sakatani T, Wei M, Katoh M, Okita C, Wada D, et al. 2001. Epigenetic heterogeneity at imprinted loci in normal populations. *Biochem. Biophys. Res. Commun.* 283:1124-30
64. Serre D, Nadon R, Hudson TJ. 2005. Large-scale recombination rate patterns are conserved among human populations. *Genome Res.* 15:1547-52
65. Serre D, Paabo S. 2004. Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* 14:1679-85
66. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al. 2005. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* 77:78-88
67. Shen JC, Rideout WM 3rd, Jones PA. 1994. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.* 22:972-76
68. Strachan T, Read AP. 1999. *Human Molecular Genetics*. New York: Wiley-Liss
69. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. 2005. Genome-wide associations of gene expression variation in humans. *PLoS Genet.* 1:e78
70. Tian D, Araki H, Stahl E, Bergelson J, Kreitman M. 2002. Signature of balancing selection in Arabidopsis. *Proc. Natl. Acad. Sci. USA* 99:11525-30
71. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. 2005. Fine-scale structural variation of the human genome. *Nat. Genet.* 37:727-32
72. Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72
73. Voight BF, Pritchard JK. 2005. Confounding from cryptic relatedness in case-Control association studies. *PLoS Genet.* 1:e32
74. Wall JD, Pritchard JK. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 4:587-97
75. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* 71:1227-34
76. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, et al. 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* 308:107-11
77. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. 2002. Allelic variation in human gene expression. *Science* 297:1143

RELATED RESOURCES

DbSNP: <http://www.ncbi.nlm.nih.gov/SNP/>

Perlegen Science Data Set: <http://genome.perlegen.com/>

International HapMap Project: <http://www.hapmap.org/>

CEPH-Human Genome Diversity project: <http://www.cephb.fr/HGDP-CEPH-Panel/>

Coriell Institute for Medical Research: <http://locus.umdj.edu/ccr/>
Sanger Exon Resequencing Project: <http://www.sanger.ac.uk/genetics/exon/>
ENCODE Project at UCSC Browser: <http://genome.ucsc.edu/ENCODE/>
NIEHS SNPs Project: <http://egp.gs.washington.edu/>
SeattleSNPs: <http://pga.mbt.washington.edu/welcome.html>



Contents

A 60-Year Tale of Spots, Maps, and Genes <i>Victor A. McKusick</i>	1
Transcriptional Regulatory Elements in the Human Genome <i>Glenn A. Maston, Sara K. Evans, and Michael R. Green</i>	29
Predicting the Effects of Amino Acid Substitutions on Protein Function <i>Pauline C. Ng and Steven Henikoff</i>	61
Genome-Wide Analysis of Protein-DNA Interactions <i>Tae Hoon Kim and Bing Ren</i>	81
Protein Misfolding and Human Disease <i>Niels Gregersen, Peter Bross, Søren Vang, and Jane H. Christensen</i>	103
The Ciliopathies: An Emerging Class of Human Genetic Disorders <i>Jose L. Badano, Norimasa Mitsuma, Phil L. Beales, and Nicholas Katsanis</i>	125
The Evolutionary Dynamics of Human Endogenous Retroviral Families <i>Norbert Bannert and Reinhard Kurth</i>	149
Genetic Disorders of Adipose Tissue Development, Differentiation, and Death <i>Anil K. Agarwal and Abhimanyu Garg</i>	175
Preimplantation Genetic Diagnosis: An Overview of Socio-Ethical and Legal Considerations <i>Bartha M. Knoppers, Sylvie Bordet, and Rosario M. Isasi</i>	201
Pharmacogenetics and Pharmacogenomics: Development, Science, and Translation <i>Richard M. Weinsilboum and Liewei Wang</i>	223
Mouse Chromosome Engineering for Modeling Human Disease <i>Louise van der Weyden and Allan Bradley</i>	247

The Killer Immunoglobulin-Like Receptor Gene Cluster: Tuning the Genome for Defense <i>Arman A. Bashirova, Maureen P. Martin, Daniel W. McVicar, and Mary Carrington</i>	277
Structural and Functional Dynamics of Human Centromeric Chromatin <i>Mary G. Schueler and Beth A. Sullivan</i>	301
Prediction of Genomic Functional Elements <i>Steven J.M. Jones</i>	315
Of Flies and Man: <i>Drosophila</i> as a Model for Human Complex Traits <i>Trudy F.C. Mackay and Robert R.H. Anbolt</i>	339
The Laminopathies: The Functional Architecture of the Nucleus and Its Contribution to Disease <i>Brian Burke and Colin L. Stewart</i>	369
Structural Variation of the Human Genome <i>Andrew J. Sharp, Ze Cheng, and Evan E. Eichler</i>	407
Resources for Genetic Variation Studies <i>David Serre and Thomas J. Hudson</i>	443

Indexes

Subject Index	459
Cumulative Index of Contributing Authors, Volumes 1–7	477
Cumulative Index of Chapter Titles, Volumes 1–7	480

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* chapters may be found at <http://genom.annualreviews.org/>