Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mutation Research 526 (2003) 53–61
Fundamental and Molecular Mechanisms of Mutagenesis
www.elsevier.com/locate/molmut
Community address: www.elsevier.com/locate/mutres

SNP and haplotype variation in the human genome

Benjamin A. Salisbury*, Manish Pungliya, Julie Y. Choi, Ruhong Jiang, Xiao Jenny Sun, J. Claiborne Stephens

Genaissance Pharmaceuticals, Five Science Park, New Haven, CT 065511, USA

Received 30 September 2002; received in revised form 4 November 2002; accepted 28 January 2003

Abstract

We have surveyed and summarized several aspects of DNA variability among humans. The variation described is the result of mutation followed by a combination of drift, migration and selection bringing the frequencies high enough to be observed. This paper describes what we have learned about how DNA variability differs among genes and populations. We sequenced functional regions of a set of 3950 genes. DNA was sampled from 82 unrelated humans: 20 African-Americans, 20 East Asians, 21 Caucasians, 18 Hispanic-Latinos and 3 Native Americans. Different aspects of variability showed a great deal of concordance. In particular, we studied patterns of single nucleotide polymorphism (SNP) allele and haplotype sharing among the four, large sample populations. We also examined how linkage disequilibrium (LD) between SNPs relates to physical distance in the different populations. It is clear from our findings that while many variants are common to all populations, many others have a more restricted distribution. Research that attempts to find genetic variants that explain phenotypic variants must be careful in their choice of study population.

Keywords: Single nucleotide polymorphism; Haplotype; Linkage disequilibrium

1. Introduction

With the advent of the first draft of the human genome sequence [1,2], research into the genetic causes of phenotypic differences among humans has been brought to a new scale. To plan such research, it is vital to understand the characteristics of the genetic variation that exists. Genaissance Pharmaceuticals has undertaken a broad and deep survey of that variation. Our interests are directed primarily towards explaining differential drug response among patients, yet the patterns we have uncovered may lend insight into other areas of study. Stephens et al. [3] detailed results from surveying 313 genes. Schneider et al. [4] extended that analysis to over 2000 genes. In this paper, we update these analyses based on the study of 3950 genes and their single nucleotide polymorphism (SNPs) and haplotypes.

Given the enormity of the human genome, one must usually place filters on which areas to investigate. One logical starting point is to restrict analysis to known and putative genes. We have taken that approach. Beyond eliminating extragenic regions, we have also focused in on the likely functional regions of each gene. Specifically, we approach SNP and haplotype discovery by sequencing exons, 100 bp of introns on each side of each exon, and up to 1 kb upstream and 100 bp downstream of each gene. In this way, functionally important variants in coding, intron–exon splice junction, and

* Corresponding author. Tel.: +1-203-786-3678; fax: +1-203-562-9377.
E-mail address: b.salisbury@genaissance.com (B.A. Salisbury).

0027-5107/03/$ – see front matter
doi:10.1016/S0027-5107(03)00014-9

proximal promoter regions are likely to be discovered [5].

As the results reported below show, it is crucial in studying genetic variation to sample subjects of diverse ethnogeography. Chromosomes sampled from different populations typically have meaningful, and unpredictable differences. With this understanding in mind, we sampled from 82 unrelated humans with the following ethnogeographic self-identifications: 21 Caucasians (CA), 20 African-Americans (AF), 20 Asians (AS), 18 Hispanic-Latinos (HL), and 3 Native Americans (NA). Most of these subjects described their parents and grandparents as having the same background. In addition to these individuals, we also included a three-generation European-American (CEPH) family (four grandparents, two parents, four offspring) and a two-generation African-American family (two parents, five offspring)—the eldest generation of these families contributed to the unrelated totals above. All sequencing, genotyping, SNP, and haplotype results will refer to this sample of individuals unless otherwise stated.

The genes in this study represent all manner of biochemical and physiological function. Because of the focus of Genaissance on discovering genetic predictors of drug response, the gene list has that overall flavor. The genes include drug targets and proteins involved in disease, metabolism, absorption, excretion, and transport. Additional genes have been processed because of functional, animal model, or evolutionary considerations. Furthermore, knowledge of the exon–intron structure of such genes was an important criterion in prioritizing the genes we sequenced for SNP and haplotype discovery. Despite the method of prioritization, these genes, which represent roughly 10% of the currently estimated number of human genes, seem a good set for studying patterns of human genetic variation.

2. SNPs

Single nucleotide polymorphisms are an atomic form of genetic variation [6]. They can be discovered through shotgun sequencing, or, as we have, through resequencing of targeted genomic regions. After discovery, SNPs may be assembled into haplotypes as discussed in Section 3 to examine allelic variation at a larger scale, such as the level of the whole gene. SNPs can be measured easily through many “genotyping” laboratory technologies.

SNPs are of interest for a variety of reasons. First, a SNP, particularly when found in a functional gene region, may itself encode differences in protein form and expression, which in turn lead to disease and other, often subtler, phenotypic differences. Second, SNPs may mark or track the presence of other, perhaps less easily detected and processed, genetic differences that cause phenotypes of interest. Third, they are useful in studying mutation rates and evolutionary history, as we do below in Section 4.

An additional useful trait of SNPs is their ubiquity. There are high frequency SNPs in all human populations and they are plentiful in the genome. Current estimates are that SNPs with minor allele frequencies of at least 1% are found at the rate of 1 in every 200–300 bp in the genome [3,7]. Extrapolating to the entire genome, this suggests that perhaps as many as 15 million SNPs are ultimately available to study, or even more when many populations and their specific SNPs are included. Already, a few million are described in public databases.

In the 82 subjects and 3950 genes we have examined, we have independently discovered 64,255 SNPs. Thirty-nine percent of these SNPs had minor allele frequencies below 1% because the minor allele was seen only once in the 164 unrelated chromosomes sampled. Although some might dismiss these variants as “sub-polymorphic,” this view is unreasonable because within the population in which the allele was observed, the frequency was approximately 2.5% (1 in 40). Several lines of evidence detailed elsewhere [4] indicate that these SNPs are generally real, rather than process artifacts. Fig. 1 shows a histogram of minor allele frequencies in the 164 chromosomes. These rare alleles are especially relevant when studying phenotypes that are similarly rare. When measured in a disease population, their frequencies may be greatly elevated.

SNPs with minor allele frequencies of 5% or greater make up 29% of the total. These SNPs may be more useful than rarer variants for understanding the occurrence of relatively common phenotypes (e.g. [8,9]. They are also valuable in pursuing linkage mapping across the genome [10]. The newly funded NIH Haplotype Map project [11] aims to use only SNPs with these relatively high allele frequencies to establish

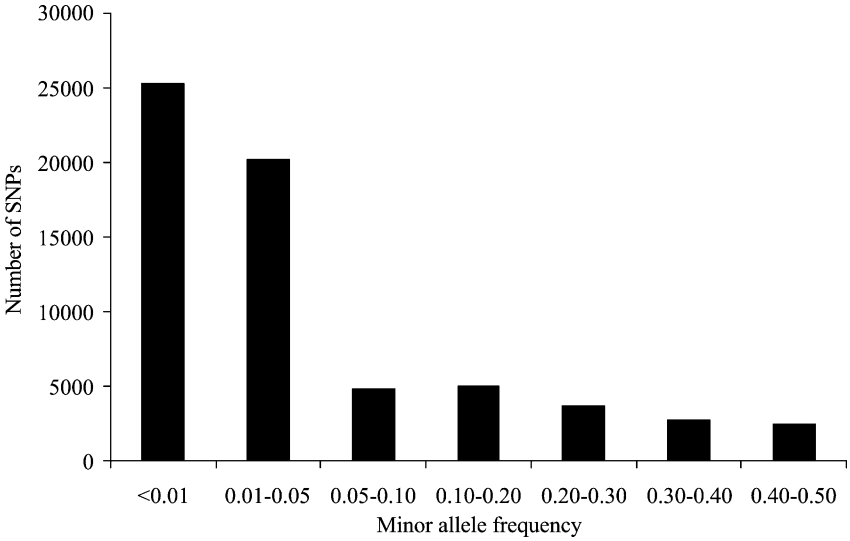

Fig. 1. Distribution of SNP minor allele frequencies for 64,255 SNPs discovered in 3950 genes in 82 unrelated individuals from diverse ethnicity.

"haplotype blocks" useful for linkage analysis with diseases and other phenotypic variation.

Fig. 2 depicts the distribution and sharing of SNPs among the four large populations. At the base of each column are the 13,680 SNPs that were observed as polymorphic in all four samples. The top layer indicates that half of the SNPs observed in only one group were seen in the African-American sample. The Asian

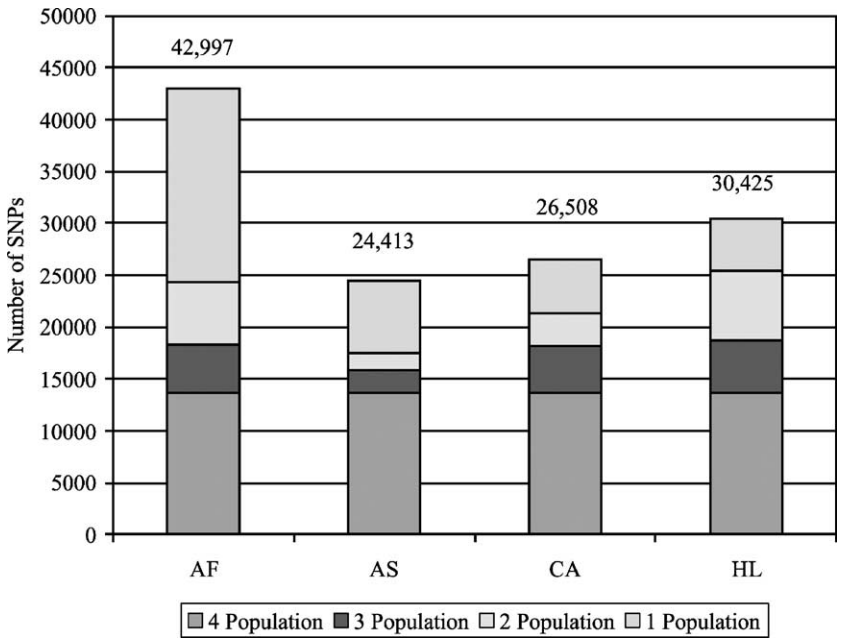

Fig. 2. Population distribution of 64,255 SNPs discovered in 3950 genes. Degree of population sharing is indicated.

Table 1
Pairwise population sharing of SNPs. Eight thousand seven hundred and seventy SNPs that were variable in two, but only two, of the population samples are allocated as to which two populations shared them

	Asian	Caucasian	Hispanic-Latino
African-American	701	1069	4304
Asian		296	550
Caucasian			1850

sample was second by that measure, although overall the least diverse. From Table 1, it is clear that many polymorphisms are shared among the non-Asian population samples.

As mentioned above, linkage disequilibrium (LD) between SNPs is an important topic. Fig. 3 addresses how linkage disequilibrium corresponds to distance between SNPs. It also shows how population structure matters. We measured $|D'|$ between each pair of SNPs in a gene for the 2597 genes where distances were reliable. $|D'|$ ranges from 0 to 1. Lower values indicate a stronger effect of recombination and recurrent mutation. Only SNPs with minor allele frequencies of at least 0.1 in the population were examined. Two trends are obvious: (1) LD decreases with distance; and (2) LD extends farther in non-African than in African populations. These trends are well documented by others [4,11–14]. The causes are understood respectively to be (1) that recombination breaks up LD, and greater distances provide more opportunity for recombination, and (2) the Eurasian population underwent a more recent population bottleneck/founder event which decreased diversity and elevated LD.

3. Haplotypes

A haplotype is simply the set of polymorphism alleles that co-occur on a chromosome. We have estimated haplotypes statistically for the SNPs in each gene using the program *HAP*™ Builder [15].

The patterns of haplotype diversity and sharing in and among populations closely parallel those described for SNPs. Furthermore, the number of haplotypes for a gene is strongly correlated with the number of SNPs. Fig. 4 depicts this relationship. In the absence of recombination, gene conversion, and recurrent mutation, the number of haplotypes could never be more than the number of SNPs plus one. Although the average in these genes is close to 1:1, regardless of the number of SNPs, many genes

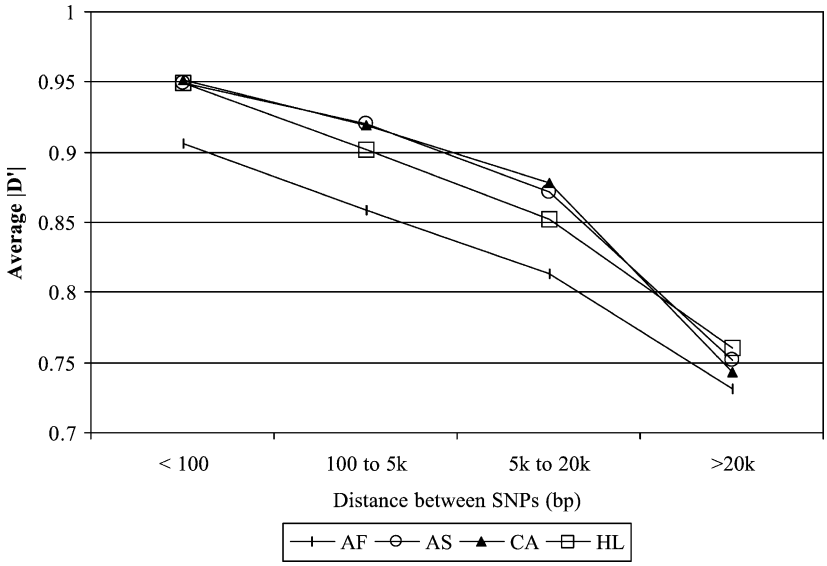

Fig. 3. Average linkage disequilibrium, $|D'|$, vs. distance between SNPs for 2597 genes in which accurate distances were available. For each population, only SNPs with minor allele frequency greater than 0.1 were used.

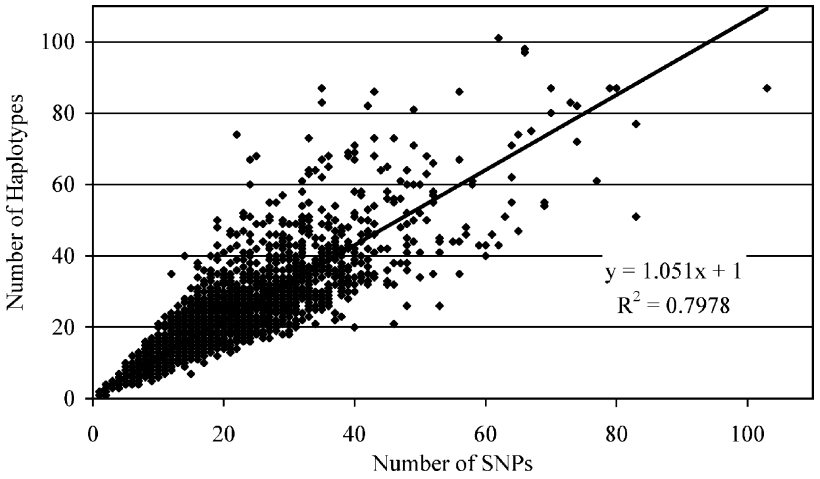

Fig. 4. Correlation of the number of haplotypes to the number of SNPs for 3931 genes.

deviate strongly from this ratio. When there are fewer haplotypes than SNPs, one can infer that some of the SNPs distinguish haplotypes that would be distinct even without the SNP; in other words, there must be redundancy in the information contained in the polymorphisms. One would also conclude that there is strong linkage disequilibrium within the gene.

When there is an excess of haplotypes, the most likely explanation is recombination, although recurrent mutation and gene conversion may have occurred as well. Recombination is an important process because it brings alleles onto the same chromosome where they previously did not co-occur. It could thus lead to new protein forms and different expression and protein combinations.

To get a better sense of the prevalence of recombination, we estimated haplotypes for the genes described by Stephens et al. [3] using only those SNPs where the minor allele was observed at least twice among unrelated individuals. Stated differently, we excluded "singleton" SNPs because they cannot provide evidence of recombination. The result was that the average number of SNPs was around 8 and the average number of haplotypes was 10.3. As expected, the ratio is higher than when singletons were included. However, the excess haplotypes ($10.3 > (8+1)$) is still relatively small, meaning that, on average, recombination has been a moderate force in the production of diversity in these genes. Point mutations, which create new SNPs and new haplotypes, seem to be relatively more productive in the generation of haplotype diversity.

One interest that many researchers may have regarding genes is their diversity. One message to take from Fig. 4 is that simply counting the number of SNPs or even haplotypes does not measure diversity very well. For instance, a gene might have many haplotypes while just one of those haplotypes describes 80% of chromosomes. Table 2 quantifies genetic diversity of genes and populations using a few common measures. All four measures indicate the same trend; in fact the four measures all have linear correlation with $r^2 = 0.99$ or higher. What these measures indicate is that the African-American population has much more diversity than the others, and among the others, Asian has the least diversity followed by Caucasian and Hispanic-Latino.

Another common interest is the degree to which populations carry different forms of the genes. The classical measure for this is the F_{ST} statistic, which contrasts observed heterozygosity in a set of subjects with the heterozygosity one would expect if the subjects all belonged to one population. When gene flow is low between populations and different genetic variants become prevalent in the different populations F_{ST} will be high. While most genes have low to modest F_{ST} values, we note that a substantial number of genes have high levels of population differentiation (Fig. 5).

Table 2
Haplotype diversity in 3931 genes

Population	Average heterozygosity	Percent of genes with expected heterozygosity >0.50	Percent of genes with maximum haplotype frequency <0.50	Mean Shannon–Wiener
African-American	0.64	75.7	52.2	2.23
Asian	0.49	54.2	29.9	1.48
Caucasian	0.52	58.8	35.2	1.59
Hispanic-Latino	0.54	63.4	37.6	1.72

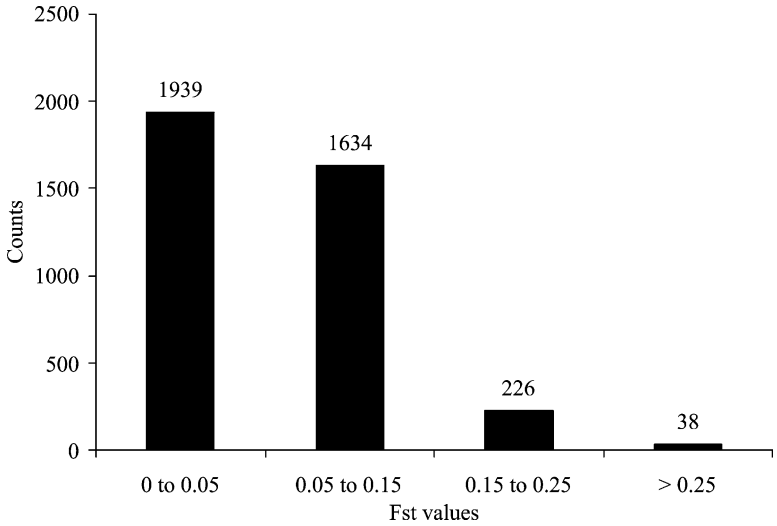

Fig. 5. Distribution of haplotype F_{ST} values for 3837 autosomal genes across four large sample populations.

4. Mutation and selection

Beyond looking at overall diversity characteristics of these polymorphisms, we can also learn about mutation and selection by studying differences between classes of change. Fig. 6 shows the counts of each of the 12 classes of base changes where the change is specified as common allele/rare allele. Two obvious trends are observed. First, transitions greatly outnumber transversions. Second, the G and C alleles tend to be the major alleles and A and T the minor ones, by a ratio as great as 2:1 for a pair of bases.

The two findings above both seem to reflect mutational biases. Transitions are known to occur more often at the mutational level [16]. CpG dinucleotides are also known to be prone to point mutation [17–20].

An intriguing finding is that complementary changes are at roughly equal frequency. For instance, a G/A as seen on one strand is the same as a C/T on the other, and such changes are very close to evenly occurring. The implication is that the changes are largely unaffected by which strand is sense and which is anti-sense in these genes.

Selection does appear to have a role, however. When we measure the rate of SNP discovery (SNPs per kb of sequence) in different gene regions, we find that SNPs are less abundant in coding regions than in introns and other regions (Fig. 7). Furthermore, when we examine the effect of coding changes, we find that more conservative changes are more common than radical changes (Table 3) based on Grantham's physico-chemical difference measure of alternative amino acids [21,22]. Nearly half of coding changes are silent. Only 0.7% are nonsense substitutions. For comparison, Table 3 also depicts the frequency of these classes of SNPs in a survey of mammalian pseudogenes [21,22]; these

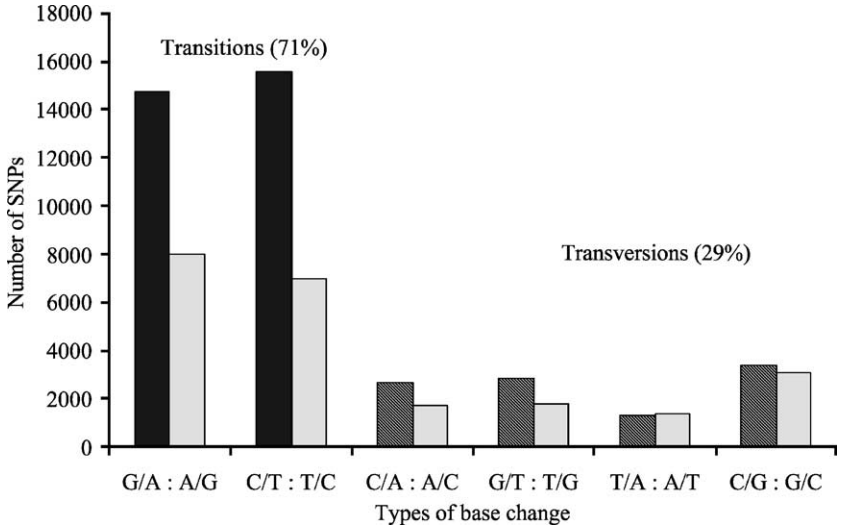

Fig. 6. Frequencies of SNP categories (63,609 SNPs in 3927 genes).

frequencies should reflect the net result of only mutation and drift. These patterns suggest that natural selection generally acts to limit change in coding regions. Those changes that selection does allow are likely to be silent or to replace one amino acid with a similar one.

Tajima's D is a useful measure based on the difference between θ and π [16,23] two measures of nucleotide diversity. Under the "standard neutral model," θ would equal π. Fig. 8 shows the overwhelmingly negative distribution of Tajima's D in our gene sample. Negative values equate to an excess of rare variants. This departure from the standard neutral model expectation implies that some of the assumptions of the model do not hold. Two violations that would cause the negative values to predominate are an expanding population, which is of course the case for the human

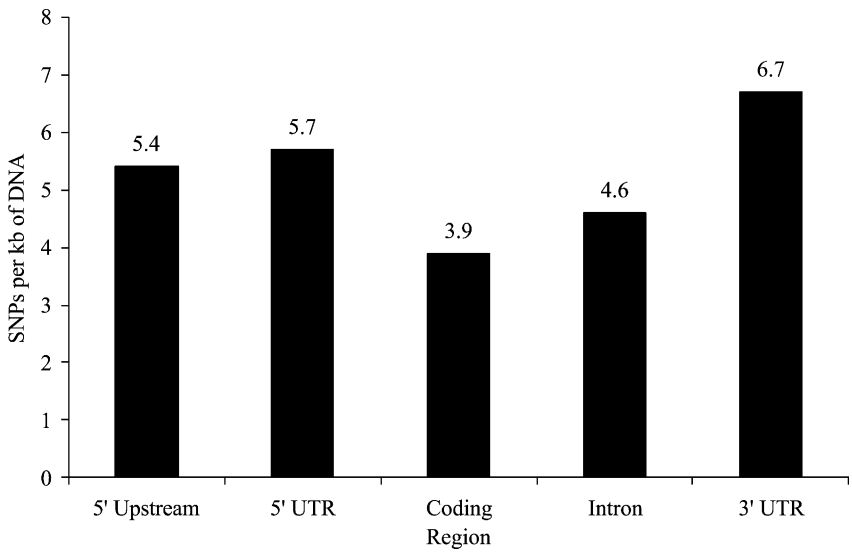

Fig. 7. SNP distribution per kb of functionally defined genomic region (3393 genes, 54,829 SNPs).

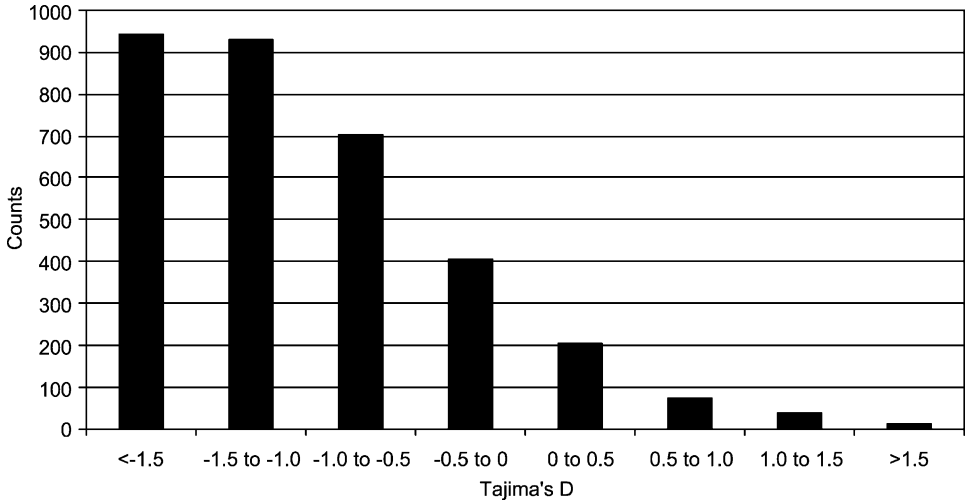

Fig. 8. Tajima's D for 3310 autosomal genes for the full population of 82 unrelated individuals.

Table 3
Inferred functional consequence of coding region SNPs (18,597 SNPs in 3688 genes)

Type of change	This study (%)	Mammalian pseudogenes (%)
Silent	49.0	27.9
Conservative	19.5	20.1
Moderately conservative	20.7	28.2
Moderately radical	6.8	12.9
Radical	3.3	6.6
Nonsense	0.7	4.4

Mammalian pseudogene data from [22].

species over the course of modern history, or purifying selection, i.e. a large fraction of mutations are deleterious, which is also conventionally believed. Currently, our tests do not determine the relative contribution of these factors.

5. Conclusions

A tremendous amount of variability exists in the human genome, even within the functional regions of genes. Researchers who intend to correlate specific phenotypes (e.g. disease susceptibility or variable drug response) to genomic variation must be aware of how this variation is organized and distributed among genes, gene regions, and populations. We have now surveyed these patterns in at least 10% of all known human genes. We can generalize that while recombination plays a role in generating diversity for many genes, only rarely does it greatly increase (e.g. double) the number of haplotypes observed. This conclusion is consistent with the generally high level of linkage disequilibrium observed between SNPs within 20–30 kb of each other, a typical physical span for human genes. Our survey also highlights the fact that many genes, SNPs, and haplotypes have patterns of variation that differ considerably among populations. This finding suggests the need for rigor in the ascertainment of population origin when conducting association studies. It will be important to extend our observations beyond the populations studied here. Finally, it is clear that the pattern of human genetic variability bears the imprint of demographic events and past, and presumably ongoing, episodes of natural selection. Our current repertoire of analytical tools for discerning among these factors needs to be augmented. As our knowledge of genomic patterns of variation grows more sophisticated, we will also need more sophisticated algorithms for conducting and interpreting phenotypic association analyses.

References

[1] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.

[2] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, The sequence of the human genome, Science 291 (2001) 1304–1351.
[3] J.C. Stephens, J.A. Schneider, D.A. Tanguay, J. Choi, T. Acharya, S.E. Stanley, Haplotype variation and linkage disequilibrium in 313 human genes, Science 293 (2001) 489–493.
[4] J.A. Schneider, M. Pungliya, J.Y. Choi, R. Jiang, X.J. Sun, B.A. Stephens, J.C. Stephens, DNA variability of human genes, Mech. Ageing Dev. 124 (2003) 17–25.
[5] H.K. Tabor, N.J. Risch, R.M. Myers, Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations, Nat. Rev. Genet. 3 (2002) 391–397.
[6] F.S. Collins, M.S. Guyer, A. Charkravarti, Variations on a theme: cataloging human DNA sequence variation, Science 278 (1997) 1580–1581.
[7] L. Kruglyak, D.A. Nickerson, Variation is the spice of life, Nat. Genet. 27 (2001) 234–236.
[8] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, Characterization of single-nucleotide polymorphisms in coding regions of human genes, Nat. Genet. 22 (1999) 231–238.
[9] M.K. Halushka, J.B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis, Nat. Genet. 22 (1999) 239–247.
[10] R.C. Lewontin, The detection of linkage disequilibrium in molecular sequence data, Genetics 140 (1995) 377–388.
[11] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, The structure of haplotype blocks in the human genome, Science 296 (2002) 2225–2229.
[12] S.A. Tishkoff, E. Dietzsch, W. Speed, A.J. Pakstis, J.R. Kidd, K. Cheung, Global patterns of linkage disequilibrium at the CD4 locus and modern human origins, Science 271 (1996) 1380–1387.
[13] Z. Zhao, L. Jin, Y.X. Fu, M. Ramsay, T. Jenkins, E. Leskinen, World-wide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 11354–11358.
[14] D.E. Reich, M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, Linkage disequilibrium in the human genome, Nature 411 (2001) 199–204.
[15] J.C. Stephens, A. Windemuth, Method and system for determining haplotypes from a collection of polymorphisms, World Intellectual Property Organization, Geneva, Switzerland, 2001.
[16] W.-H. Li, Molecular Evolution, Sinauer Associates Inc., 1997.
[17] W.M. Rideout III, G.A. Coetzee, A.F. Olumi, P.A. Jones, 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes, Science 249 (1990) 1288–1290.
[18] A.N. Magewu, P.A. Jones, Ubiquitous and tenacious methylation of the CpG site in codon 248 of the p53 gene may explain its frequent appearance as a mutational hot spot in human cancer, Mol. Cell Biol. 14 (1994) 4225–4232.
[19] A.R. Templeton, A.G. Clark, K.M. Weiss, D.A. Nickerson, E. Boerwinkle, C.F. Sing, Recombinational and mutational hotspots within the human lipoprotein lipase gene, Am. J. Hum. Genet. 66 (2000) 69–83.
[20] S.M. Fullerton, A.G. Clark, K.M. Weiss, D.A. Nickerson, S.L. Taylor, J.H. Stengard, Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism, Am. J. Hum. Genet. 67 (2000) 881–900 (in process citation).
[21] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (1974) 862–864.
[22] W.H. Li, C.I. Wu, C.C. Luo, Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications, J. Mol. Evol. 21 (1984) 58–71.
[23] F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, Genetics 123 (1989) 585–595.