Characterization of single-nucleotide polymorphisms in coding regions of human genes

Michele Cargill[1*], David Altshuler[1,2*], James Ireland[1], Pamela Sklar[1,3], Kristin Ardlie[1], Nila Patil[5], Charles R. Lane[1], Esther P. Lim[1], Nilesh Kalyanaraman[1], James Nemesh[1], Liuda Ziaugra[1], Lisa Friedland[1], Alex Rolfe[1], Janet Warrington[5], Robert Lipshutz[5], George Q. Daley[1,4] & Eric S. Lander[1,6]

*These authors contributed equally to this work.

A major goal in human genetics is to understand the role of common genetic variants in susceptibility to common diseases. This will require characterizing the nature of gene variation in human populations, assembling an extensive catalogue of single-nucleotide polymorphisms (SNPs) in candidate genes and performing association studies for particular diseases. At present, our knowledge of human gene variation remains rudimentary. Here we describe a systematic survey of SNPs in the coding regions of human genes. We identified SNPs in 106 genes relevant to cardiovascular disease, endocrinology and neuropsychiatry by screening an average of 114 independent alleles using 2 independent screening methods. To ensure high accuracy, all reported SNPs were confirmed by DNA sequencing. We identified 560 SNPs, including 392 coding-region SNPs (cSNPs) divided roughly equally between those causing synonymous and non-synonymous changes. We observed different rates of polymorphism among classes of sites within genes (non-coding, degenerate and non-degenerate) as well as between genes. The cSNPs most likely to influence disease, those that alter the amino acid sequence of the encoded protein, are found at a lower rate and with lower allele frequencies than silent substitutions. This likely reflects selection acting against deleterious alleles during human evolution. The lower allele frequency of missense cSNPs has implications for the compilation of a comprehensive catalogue, as well as for the subsequent application to disease association.

Introduction

The human population has relatively limited genetic diversity, reflecting its young age and historically small size[1]. Many rare genetic variants exist in the human population, but most of the heterozygosity in the population is attributable to common alleles (that is, those that are present at a frequency of >1% in the general population). The infrequent variants include the primary causes of rare, mendelian genetic diseases, with these alleles typically being recent in origin and highly penetrant. By contrast, some authors have recently hypothesized that the common variants may contribute significantly to genetic risk for common disease[2–4]. If this common disease-common variant (CD-CV) hypothesis is true, it permits a conceptually straightforward approach to identifying disease-causing mutations: build a comprehensive catalogue of the limited number of common gene mutations in the human population and test them directly for association to clinical phenotypes. Such an approach is only now becoming possible due to progress in the human genome project in identifying genes and in the technology for discovering and typing DNA sequence variants. One advantage of such an approach is the potentially greater statistical power of association compared with linkage-based designs[2]. Although the CD-CV hypothesis remains speculative as to its generality, important examples of associations with common alleles are known, including the *APOE*E4* allele in Alzheimer disease[5], the *F5* 1691G→A allele (also known as FV Leiden) in deep-venous thrombosis[6] and *CCR5*Δ32 in resistance to HIV infection[7]. The high frequency of these alleles contributes to their population impact: *APOE*E4*, with a threefold effect and allele frequency of 10–20%, accounts for approximately half of the population-attributable risk for Alzheimer disease[8].

Any such approach to human disease genetics will require an understanding of human gene variation. Although there is a rich literature concerning nucleotide variation in model systems, particularly in *Drosophila melanogaster*[9], sequence variation in human genes has been studied only in limited ways. Classical studies of protein variants offered the first view of human gene diversity[10], but provided an incomplete picture. A few recent studies have focused on individual genes (such as those encoding β-globin[11] and lipoprotein lipase[12]) in many individuals, and one study examined 49 genes by comparing two independent sequences deposited in public databases[13]. It has been difficult to compare variation among classes of sites within genes, among genes and between populations, owing to the small sample sizes and to differences in the populations studied. To define the nature of variation

[1]Whitehead Institute/MIT Center for Genome Research, One Kendall Square, Building 300, Cambridge, Massachusetts 02139, USA. Departments of [2]Endocrinology, [3]Psychiatry and [4]Hematology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [5]Affymetrix, Inc., Santa Clara, California 95051, USA. [6]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. Correspondence should be addressed to E.S.L. (e-mail: lander@genome.wi.mit.edu).

Table 1 • Summary of polymorphisms in 106 human genes

Gene	Coding bp screened	No. synonymous polymorphisms	No. non-synonymous polymorphisms	Non-coding bp screened	No. non-coding polymorphisms
AADC	1,229	0	2	311	0
ADORA2	332	0	1	75	0
AHC	1413	0	0	63	1
ANX3	929	2	4	725	6
APOD	570	1	3	383	1
AR	2,759	3	0	300	1
AT3	1,357	3	0	121	0
BDNF	744	0	1	212	0
CD36	1,209	1	1	252	0
CETP	1,397	4	4	299	0
CGA	349	1	0	235	0
CLanalog	1,461	3	2	12	0
CNTF	603	0	1	154	0
COMT	783	2	1	241	1
CRH	51	0	0	745	3
CYP11A	1,556	1	1	547	0
CYP11B1	1,410	7	7	496	9
CYP11B2	1,512	7	8	906	4
CYP17	1,395	3	0	36	0
CYP21	1,488	7	7	1,400	24
DBH	1,266	0	2	49	0
DRD1	1,341	1	0	81	0
DRD2	1,032	2	0	1,379	3
DRD3	719	0	1	145	0
DRD5	1,408	2	1	34	0
F10	1,369	3	2	416	1
F11	1,878	7	4	1,312	2
F13A1	2,199	3	6	948	4
F13B	1,952	4	6	2,339	4
F2	1,740	3	2	292	0
F2R	1,202	2	1	13	0
F3	875	0	1	92	0
F5	6,564	13	16	1,542	8
F7	1,262	4	2	1,209	2
F9	1,364	0	1	1,062	2
FGA	1,935	2	2	490	0
FGB	1,476	7	3	1,057	0
FGG	1,252	0	2	1,392	2
FSH	355	1	1	44	0
FSHR	1,683	1	3	0	0
GABRB1	1,425	5	0	804	2
GAP43	675	1	1	79	0
GH1	644	0	1	426	5
GHR	1,765	1	6	391	1
GNRHR	237	0	1	513	0
GP1BA	1,881	2	2	48	0
GP1BB	1,238	0	0	73	0
GP5	1,683	0	0	52	0
GP9	534	1	0	143	0
GRF	224	0	0	239	0
GRIN1	1,681	1	0	553	0
GRL	2,334	4	3	4,028	5
HCF2	1,500	3	3	64	1

in human genes, as well as provide a catalogue of gene polymorphisms for association studies, we performed an extensive survey of coding sequence diversity of many genes in many individuals.

Results

Screening of 106 genes for sequence variants

We assembled a collection of 106 genes whose protein products have roles in cardiovascular, endocrine and neurological systems (Table 1). The genes were chosen because of their potential relevance to common, clinically significant diseases, such as coronary artery disease, type II diabetes and schizophrenia. They encode proteins involved in coagulation, lipid metabolism, energy metabolism, neuroendocrine physiology, neurotransmission and central nervous system development. Variation in these genes was studied in a sample including Europeans, African-Americans, African Pygmies and Asians, with an average of 114 chromosomes screened for each gene. The sample size provides greater than 99% power to detect alleles of 5% or greater, and 65% power to detect alleles with a frequency of 1%. The most relevant variants are likely to be those in coding and regulatory regions of genes. As regulatory regions of most genes remain poorly characterized at present, we focused our efforts on characterizing variation in and around the coding regions.

Overall, we screened the sample of 114 chromosomes for SNPs in 196.2 kb, consisting of 135.8 kb of coding regions and 60.4 kb from adjacent non-coding regions (UTRs and introns). Sequences were amplified by PCR and screened by two independent methods. The first method involved hybridization of labelled PCR products to variant detector arrays (VDAs; that is, high-density DNA probe arrays containing oligonucleotides specific for the sequences under study): variant sequences typically give rise to altered hybridization patterns[14,15]. The second method involved subjecting PCR products to denaturing HPLC (DHPLC; ref. 16) at a critical temperature: heterozygous individ-

Table 1 • Summary of polymorphisms in 106 human genes (continued)

Gene	Coding bp screened	No. synonymous polymorphisms	No. non-synonymous polymorphisms	Non-coding bp screened	No. non-coding polymorphisms
HMGCR	1,724	0	1	12	1
HSD3B1	1,122	3	2	653	1
HSD3B2	1,122	1	1	723	2
HTR1A	1,272	1	0	1,189	1
HTR1D	1,134	1	1	46	0
HTR1DB	1,173	2	0	85	1
HTR1E	1,098	1	1	70	0
HTR1EL	1,101	1	0	46	0
HTR2A	1,398	2	3	1,709	9
HTR2C	1,245	0	1	138	0
HTR5A	1,062	2	0	34	0
HTR6	437	1	0	34	0
HTR7	1,279	0	0	138	0
IGF1	630	0	0	7,250	8
IGF2	546	0	0	610	1
ITGA2B	2,833	4	3	707	0
ITGB3	2,131	4	3	163	0
KLK2	297	0	1	279	2
LCAT	1,289	3	0	90	0
LDLR	2,101	7	3	38	0
LIPC	1,471	4	3	754	4
LPL	409	1	1	48	0
MAOA	1,032	1	0	69	0
MAOB	980	1	0	135	0
MPL	1,748	1	2	903	1
NGFB	726	1	1	1,186	5
NOS1	127	0	0	56	0
NT3	774	1	0	150	0
NTRK1	1,961	5	2	1,106	0
PACE	1,500	2	0	1,095	4
PAI1	1,171	1	2	911	1
PAI2	1,248	5	4	915	5
PC1	1,881	1	3	456	1
PCI	1,221	5	5	576	4
POMC	132	0	0	520	0
PRL	633	1	1	180	1
PROC	1,334	3	0	114	0
PROS1	1,868	1	0	557	0
PTAFR	1,029	0	2	13	0
PTH	348	1	0	230	2
PTHLH	634	0	0	2,342	13
SELP	2,096	5	8	14	0
SHBG	1,209	1	3	494	1
SLC6A1	1,388	2	0	547	2
SLC6A3	1,496	6	1	205	0
SLC6A4	1,623	1	2	824	1
TBXA2R	1,006	1	0	12	0
TBXAS1	1,605	1	6	1,411	1
TFPI	806	0	1	139	0
TH	965	1	1	104	0
THBD	1,728	0	0	26	0
THPO	1,049	0	0	632	2
VLDLR	2,391	3	1	850	2
All genes	13,5823	207	185	60,410	168

uals typically give rise to heteroduplex products with altered denaturation and migration properties.

Because both screening methods can generate false positives, we confirmed every reported SNP. Samples implicated by either method as containing a candidate SNP were thus subjected to fluorescent dideoxy sequencing, either to confirm the presence of the SNP (in the case of VDA) or to identify and confirm the presence of the SNP (in the case of DHPLC). Such confirmation proved essential for eliminating false positives.

We identified 560 confirmed SNPs in the 196.2 kb surveyed, consisting of 168 non-coding SNPs and 392 cSNPs. Most true SNPs were likely to have been identified, as determined by comparing these results with those from comprehensive, double-stranded sequencing of 10 genes in 20 people. Specifically, VDA and DHPLC identified 85% and 87%, respectively, of variants found by sequencing. Where VDAs and DHPLC overlapped, the combination of the two methods identified all polymorphisms. We screened approximately one-third of individuals with both methods, and one-third with either method alone. Thus, we estimate the overall sensitivity of polymorphism discovery to be in excess of 90%. The complete data are available (http://www.genome.wi.mit.edu/cvar_snps).

Nucleotide diversity of human genes

A SNP survey can be characterized in terms of either K, the observed number of variant sites, or π, the observed heterozygosity per base pair. Because K increases with the number of chromosomes (n) studied and the total sequence length L, it is preferable to use the normalized number of variant sites

$$\hat{\theta}=K/\sum_{i=1}^{n-1} i^{-1} L,$$

which corrects for sample size. Under the neutral theory of molecular evolution and infinite sites model, $\hat{\theta}$

Table 2 • Polymorphism rates for different classes of sites

Polymorphism type	bp screened	No. polys	Frequency (SNP/bp)	$\hat{\theta}$	Π	Adjusted for the frequency of sites* Frequency (SNP/bp)	$\hat{\theta}$	Π
Non-coding	60,410	168	1/354	5.30±1.33	5.19±2.47			
Coding	135,823	392	1/346	5.43±1.36	5.00±2.38			
synonymous		207	1/656	2.86±0.72	3.05±1.45	1/187	10.03±2.52	10.67±5.07
non-synonymous		185	1/734	2.56±0.64	1.96±0.93	1/523	3.59±0.90	2.75±1.31
conservative		119	1/1,141	1.65±0.41	1.43±0.68	1/390	4.81±1.21	4.17±1.98
non-conservative		66	1/2,058	0.91±0.23	0.53±0.25	1/763	2.46±0.62	1.43±0.68
fourfold degenerate sites	21,645	112	1/193	9.73±2.46	11.18±5.31			
twofold degenerate sites	34,294	125	1/274	6.85±1.72	6.27±2.98			
nondegenerate sites	79,659	155	1/514	3.66±0.92	2.93±1.39			
Total	196,233	560	1/348	5.39±1.36	5.05±2.40			

*The number of synonymous sites was calculated as the sum of fourfold degenerate sites and half the number of twofold degenerate sites; the number of non-synonymous sites is the sum of the non-degenerate sites and half the twofold degenerate sites. The number of conservative and non-conservative sites is estimated as the proportion of non-synonymous sites at which a nucleotide substitution would create a conservative or non-conservative substitution.

and π are both estimators of the population genetic parameter $\theta=4N\mu$, where N is the effective population size and μ is the mutation rate[17].

SNPs were found at a similar overall frequency in coding and non-coding regions. SNPs in coding regions occurred at a frequency of 1 per 346 bp, corresponding to $\hat{\theta}=5.43\times10^{-4}$ and $\pi=5.00\times10^{-4}$. SNPs were observed in non-coding DNA at a similar frequency of 1 per 354 bp. The normalized number of variant sites was $\hat{\theta}=5.30\times10^{-4}$, and the mean heterozygosity, $(\pi)=5.19\times10^{-4}$ (Table 2). For both classes of SNPs, the similar value for $\hat{\theta}$ and π is consistent with a population of constant size evolving according to neutral expectation. These results are consistent with most human genetic polymorphism (weighted by allele frequency) reflecting the period before the recent expansion in human population size.

Distribution of polymorphisms in genes

The 392 cSNPs were roughly equally divided between synonymous (207 cSNPs) and non-synonymous (185 cSNPs) changes. As approximately two-thirds of random coding mutations alter an amino acid, the fact that non-synonymous cSNPs comprise slightly less than one-half the cSNPs implies strong selection against amino-acid altering changes. To address this issue directly, we examined the nucleotide diversity at fourfold degenerate sites, twofold degenerate sites and non-degenerate sites. Changes at fourfold degenerate sites produce only synonymous changes, whereas those at non-degenerate sites are always non-synonymous. Nucleotide diversity ($\hat{\theta}$) was 9.73×10^{-4} at fourfold degenerate sites, 6.85×10^{-4} at twofold degenerate sites and 3.66×10^{-4} at non-degenerate sites (Table 2). Assuming that mutations occur at an equal rate at both classes of sites, non-synonymous variants survive to be detected in such a survey at only 38% the rate of synonymous changes.

The force of selection is also evident in comparing non-synonymous cSNPs causing a non-conservative amino acid alteration with those causing a conservative amino acid change. Non-conservative cSNPs represent only 36% of the non-synonymous cSNPs, whereas randomly distributed mutations would be expected to produce a higher proportion (52%) of non-conservative changes. This implies that non-conservative cSNPs survive to be detected in such a survey at only approximately one-half the rate of conservative, non-synonymous cSNPs.

Distribution of allele frequency

The various types of SNPs differ not only in the rate of their occurrence, but also in the frequency of their minor alleles. This can be seen in several ways. When SNPs were classified according to whether the frequency of the minor allele was high (≥15%), intermediate (5–15%) or low (≤5%), the non-synonymous cSNPs were enriched in low-frequency alleles compared with the rest of the collection (Fig. 1). The distribution of non-synonymous allele frequencies was significantly different than that of synonymous changes ($P=0.02$, Kolmogorov-Smirnov test). Indeed, more than half (59%) of non-synonymous cSNPs were found at a frequency below 5%, with this effect evident for both conservative and non-conservative substitutions.

The effect of selection can also be inferred by considering the average frequency of the minor allele: 7% for non-conservative cSNPs; 11% for conservative, non-synonymous cSNPs; 14% for synonymous cSNPs; and 13% for non-coding SNPs. In addition, the lower allele frequency of non-synonymous cSNPs is reflected in the fact that the heterozygosity π is lower than the normalized rate of variant sites $\hat{\theta}$ for this class of SNPs (Table 2). This divergence is in the direction predicted by the action of purifying selection, although it is not statistically significant[18] (Tajima's D statistic=–0.0037).

Distribution of polymorphisms among genes

We explored the distribution of SNPs among the 106 genes to detect differential effects of selection among genes. The number of cSNPs per gene ranged from 29 for *F5* to none for 13 genes, and the normalized rate, $\hat{\theta}$, similarly showed considerable variation (Fig. 2). The observed range in nucleotide diversity is similar in magnitude to that observed for *Drosophila*[9].

The most polymorphic gene studied was *CYP21* ($\hat{\theta}=24.7\times10^{-4}$), which encodes an enzyme involved in the biosynthesis of adrenal steroid hormones. Mutations in this gene are also the most frequent cause of the common monogenic disorder congenital adrenal hyperplasia (MIM 210910). There are many known factors that may contribute to the high rate of polymorphism at *CYP21*: (i) it is located in the HLA complex, thus selection for HLA diversity may maintain excess diversity at *CYP21*; (ii) *CYP21* has a nearby pseudogene, which may act as a reservoir of genetic diversity via gene conversion; and (iii) the role of *CYP21* in sex steroid metabolism may lead to balancing selection between the sexes. In addition, variation in local rates of mutation and recombination may effect diversity at each locus[19,20]. Disentangling such factors will require considerable work. A variety of population genetic tests are available for testing selection at individual loci[21]; such analyses are ongoing and will be reported elsewhere.

Ancestral allele of human cSNPs

The age of an allele has important implications for human genetic studies, because it bears on the extent of linkage disequi-

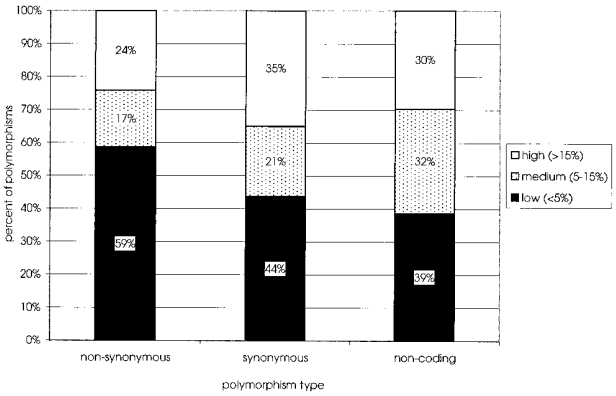

Fig. 1 Minor allele frequency by polymorphism type. The percentage of cSNPs having minor allele frequency classified as low (<5%), medium (5–15%) or high (>15%) frequency is shown for synonymous, non-synonymous and non-coding SNPs.

librium (retention of the ancestral haplotype on which the allele arose) that will be present. Linkage disequilibrium can be a powerful tool in mapping disease genes[22]. Linkage disequilibrium will typically be more extensive for recently arising alleles than for older alleles (because there has been less time for the haplotype to be pared down by recombination). Although the precise age of the SNPs cannot be directly assessed without extensive analysis of haplotypes using tightly linked markers, some information about the relative age of alleles can be inferred by determining which allele preceded human speciation and which arose thereafter. The ancestral human allele would be expected to show negligible levels of linkage disequilibrium.

An elegant prediction from theoretical population genetics states that, for any polymorphic locus under neutral evolution, the probability that a given allele represents the oldest allele in the population is equal to its frequency[23]. In practice, the oldest human allele will almost always be equivalent to that present in the chimpanzee, in view of the low mutation rate since the human-chimp divergence. Thus, to determine the ancestral allele, we sequenced each corresponding gene from the common chimpanzee (*Pan troglodytes*).

Both the ancestral allele and minor allele frequency were determined for 267 of the reported SNPs. (For 3 of 267 SNPs, the chimpanzee was homozygous for a third allele differing from both of the current human alleles. This is consistent with the overall 0.6% nucleotide sequence divergences seen between human and chimpanzee.) Among polymorphisms with a minor allele frequency below 10%, the average allele frequency was 3% and the proportion ancestral was 7% (11/158). Among polymorphisms with minor allele frequency exceeding 10%, the mean frequency was 28% and the proportion ancestral was 32% (35/109). These results agree with the theoretical prediction, providing the first reported test of this prediction in humans. It therefore follows that in a significant fraction of cases, the newer human allele has risen in frequency (due to drift or selection) to become the major allele in the current population. In such cases, the minor allele is actually older and the extent of linkage disequilibrium around this allele would be expected to be quite small.

Distribution of polymorphisms among populations

We examined the distribution of SNPs among the European, African-American, African and Asian samples. Although most SNPs were seen in multiple groups, there was a statistically significant excess of SNPs that were seen in only one of the sub-groups.

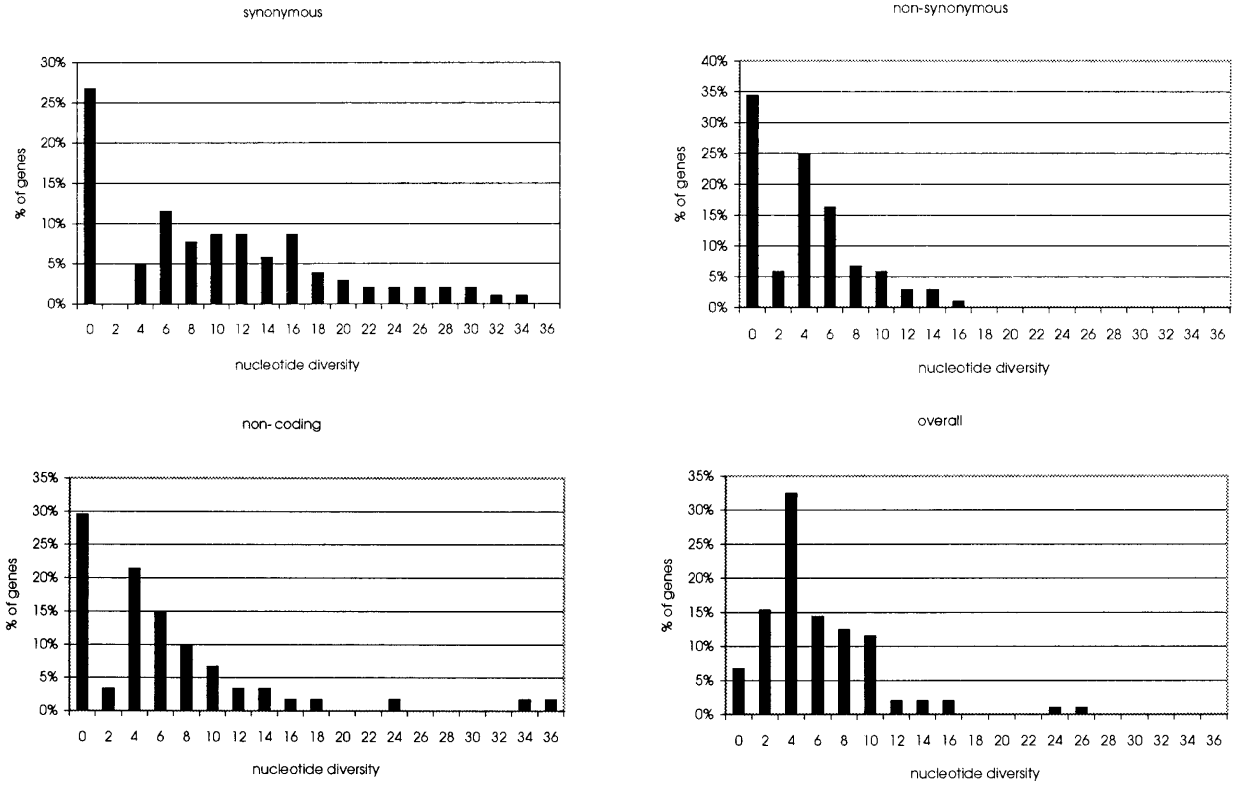

Fig. 2 Distribution of nucleotide diversity. Normalized frequency of variant sites, $\hat{\theta}$ was calculated for different classes of sites in each gene. Each graph shows the percentage of genes having $\hat{\theta}$ in the binned range.

Owing to the limited sample size, the fact that a SNP is found only in one sub-group does not necessarily imply that it is private to that group in the general population; it can, however, provide information about substructure. The greatest excess of SNPs observed in only one of the groups was found in the African-American and African samples; this finding is consistent with prior studies[24] and the accompanying paper of Halushka *et al.*[25]. Specifically, the number of SNPs observed more than once that were confined to the African-American, African Pygmy, European and Asian sub-groups were 17, 17, 12 and 9 (compared with expectations of 3.02, 1.34, 8.62 and 1.81). As a significant fraction of all polymorphisms were observed only in a single ethnic group, construction of a comprehensive SNP database may be more efficient and complete if screening includes a diverse set of DNA samples.

Discussion

The results of this survey provide a fundamental description of sequence variation in the coding regions of human genes. Our data indicate that the average gene contains only a handful (approximately four) of coding sequence SNPs with allele frequencies of at least a few per cent in the human population. Extrapolating these numbers over the human genome, we estimate that there exist only about 240,000–400,000 common cSNPs.

The extent of diversity can also be understood by considering the allelic diversity within a typical individual: the 2 copies of an average gene chosen at random will differ by roughly 1 base in 2 kb, corresponding to somewhat less than 1 heterozygous base in the coding region of a typical gene. Accounting for both the different rate and frequency of non-synonymous SNPs, only about 40% of these observed changes will alter the encoded amino acid. Thus, the typical person would be estimated to be heterozygous for about 24,000–40,000 non-synonymous (amino acid altering) substitutions.

The action of purifying selection during human evolution is evident from the comparatively lower rate of non-synonymous cSNPs, especially those that create non-conservative changes. Non-synonymous cSNPs not only occur less often, but also have lower minor allele frequencies: for this class, the one most likely to have the most dramatic effects on proteins, about 60% display a minor allele frequency below 5%. Recently, the deleterious mutation rate has been estimated by comparing the coding sequences of 46 genes between human and primate species; this analysis concluded that of amino acid altering mutations that have occurred since the chimp-human divergence, 38% have survived to be detected in existing populations[26]. Our estimate of the proportion of surviving non-synonymous mutations within current human populations (38%) is identical to that observed by comparing human with chimp DNA sequences. This indicates that the selective forces that limit survival of deleterious changes have, in sum, acted similarly in the development of current human populations when compared with the more distant divergence of human and chimp.

Of course, many polymorphisms that affect biological function will occur outside the coding regions of genes. Non-coding DNA contains sequences important for regulatory functions, and will therefore be constrained to some extent; the magnitude of this effect can be estimated by comparing the polymorphism rate in non-coding DNA with that observed at fourfold degenerate sites. The fourfold degenerate sites had the highest nucleotide diversity in our survey, and may most closely approximate the neutral rate of polymorphism. The diversity observed in the non-coding regions studied here was only half as great as at fourfold degenerate sites. Similarly high levels of constraint in non-coding DNA have been reported for *Drosophila*[9] and in a smaller human data set by Li and Sadler[13], who observed over three times the nucleotide diversity at fourfold degenerate sites ($\mathring{\theta}=11\times10^{-4}$) compared with that in both untranslated regions and non-degenerate sites ($\mathring{\theta}=3\times10^{-4}$).

It should be noted that the non-coding DNA studied here was in the immediate vicinity of genes. Such perigenic sequences may conceivably be subject to greater constraints than random non-coding DNA, because they contain sequences regulating gene expression and splicing. The expectation that perigenic sequences would show lower nucleotide diversity than random genomic DNA, however, is not supported by two recent surveys performed by our group (ref. 14 and D.A., C. Cowles and E.S.L., unpublished data). Specifically, these studies examined 2.3 and 1.8 Mb of random genomic DNA and estimated nucleotide diversity at 4.5×10^{-4} and 6.2×10^{-4}, respectively. These values more closely match that observed in this report for non-coding, perigenic sequences, and are well below that seen at fourfold degenerate sites.

Assuming that the diversity at fourfold degenerate sites most closely reflects the neutral polymorphism rate, and that the underlying mutation rate is similar across all classes of sites, our results indicate that non-coding changes survive at 55% the rate of neutral changes. These studies would indicate that non-coding DNA adjacent to coding regions is functionally constrained to a significant degree, perhaps comparable to that observed within the coding region. An alternative explanation is that fourfold degenerate sites display rates of polymorphism that exceed the neutral rate, perhaps due to mutagenic effects of transcription or other unforeseen mechanisms. Clearly, additional work is needed to define the neutral rate of human DNA diversity, and to understand the degree of constraint in different types of sequence.

We found individual genes to vary widely in their nucleotide diversity. Identification of genes that fall outside this distribution—such as *CYP21*—might identify specific loci that have undergone some manner of evolutionary selection. Many factors other than selection can influence diversity at each locus, however, such as local variation in mutation rate, gene conversion and recombination. Thus, it will be necessary to apply more sensitive tests to all the genes to search for signatures of selection. Many such tests have been described[21], and further analysis of these data is ongoing.

Comparisons of allele frequencies across different sub-populations can shed light on human population structure and history. Although the size of individual sub-populations was too small for such conclusions to be drawn, the observed number of private polymorphisms exceeded that expected by chance, and indicated sub-structure in our sample. It is worth noting that rare polymorphisms are most likely to be private to one ethnic group, and non-synonymous polymorphisms are more likely to be rare. Thus, to discover the polymorphisms of greatest interest for disease studies, it will be advantageous to explore a diverse sample.

Our overall results about human gene diversity agree with a companion study by Halushka *et al.*[25]. The authors surveyed SNPs in 75 genes in 160 chromosomes. The overall estimate of nucleotide diversity was higher than in our study (8.5×10^{-4} versus 5.4×10^{-4}), but the difference is entirely due to the fact that the sample studied by Halushka *et al.* was composed of equal numbers of Europeans and Africans. When diversity is calculated for each study based solely on the European sub-sample, the results are essentially identical (5.0×10^{-4} and 5.2×10^{-4} for Halushka *et al.* and our report, respectively). (The effect of the African sample on total diversity serves to underscore the existence of population substructure.) In addition, our findings regarding the relative frequency of coding versus non-coding SNPs and non-synonymous versus synonymous cSNPs agree well with those of Halushka *et al.*[25].

The relatively lower allele frequencies of non-synonymous cSNPs has important implications for the size of the study popu-

lation needed to produce a comprehensive catalogue of human variants for studies of disease association. It has been proposed, for example, that most human SNPs may be found by performing shotgun sequencing on a handful of individuals[27,28]. One proposal specifically suggests performing 10-fold shotgun coverage of 5 individuals (10 chromosomes); by Poisson resampling, such a design would sample an average of only 6.3 independent chromosomes. This should identify many SNPs, but the small sample size will fail to identify those cSNPs most likely to have pathophysiologic importance: those that alter the sequence of the encoded protein. Small sample sizes will instead bias in favour of synonymous, and typically neutral, polymorphisms.

A comprehensive collection of common (population frequency >1%), non-conservative cSNPs may require surveying as many as 50–100 chromosomes. Because coding sequence represents only about 3% of the genome, it may prove inefficient to obtain such deep coverage of cSNPs by shotgun sequencing of genomic DNA (refs 27,28). Instead, it may prove more efficient to perform shotgun sequencing on cDNA libraries from multiple individuals or for specific sets of genes by direct amplification of the target sequences from multiple individuals, as done here.

The allele frequencies of non-synonymous cSNPs also have implications for the sample size needed to perform disease association. Power calculations[2] show that detecting associations with alleles having frequency under 5% will typically require thousands of patients (except in rare cases in which the relative risk attributable to the allele is large). Thus, association studies involving most non-synonymous cSNPs will require larger patient populations that are currently available for most phenotypes, as well as higher-throughput genotyping technologies.

Genetic association studies can take two forms, which have been termed[3] 'direct' and 'indirect'. In direct association studies, one tests the hypothesis that a given SNP is a causative factor in a disease by examining whether its frequency is increased in affected individuals. In the alternative ('indirect') approach of linkage-disequilibrium (LD) studies, one uses multiple SNPs to search for evidence of an ancestral haplotype that is enriched in affected individuals. The latter approach does not require the discovery of each disease-causing allele, but does require an extremely dense SNP map and significant increases in genotyping capacity. In addition, its potential for success depends sensitively on untested population genetic assumptions regarding the number of founder chromosomes in the population. Thus, while the technical capacity to perform genome-wide LD studies is under development, it will be fruitful to examine in detail the characteristics of linkage disequilibrium around and among the SNPs described in this and related reports[29].

In the near term, direct tests of association with missense cSNPs may be the most practical as well as the most promising approach. SNPs that affect the amino acid sequence of a protein are likely to explain a significant fraction of disease variation. SNPs that affect gene regulation may be equally important in disease risk, but it is difficult to recognize such SNPs from among the much larger pool of non-coding SNPs given our limited knowledge of regulatory signals in DNA. Thus, until it becomes possible to genotype every SNP within a genomic locus or identify functional regulatory polymorphisms directly from primary sequence data, collecting and typing missense cSNPs is likely to provide the greatest efficiency and yield.

This study provides an initial catalogue of cSNPs in 106 genes relevant to common clinical conditions. Together with the work of Halushka *et al.*[25] describing variants in 75 genes, there is now a substantial database of cSNPs available for disease association studies. Moreover, ongoing studies by our group and others are likely to swell this collection to more than 1,000 genes in the near future. Assuming that similar advances occur in the technology for genotyping SNPs in large collections of patient samples, it should soon become possible to test directly the common disease-common variant hypothesis.

Methods

Collection of gene sequences. Gene sequences were obtained from the GenBank (http://www.ncbi.nlm.nih.gov/) and TIGR (http://www.tigr.org/) databases. When multiple sequence depositions were available, a consensus sequence was derived. Determination of coding sequence, UTRs and intronic regions was based on annotation in the databases, although we performed internal checks to ensure accurate determination of start and stop codons, ORFs and so on.

Samples used for polymorphism discovery. We obtained 51 cell lines (from 20 European individuals, 14 from Asian, 10 African Americans and 7 African Pygmies) from Coriell Cell Repository and prepared DNA according to standard protocols. In addition, 10 European samples were obtained as anonymous blood samples from C. Hennekens and J.M. Gaziano. The average number of individuals successfully screened for each gene was 57, with the precise number successfully screened varying among genes. Details about the DNAs used are available on request.

Amplification of samples. We designed PCR assays spanning each exon using Primer 3.0, release 0.7. For *CYP21*, which is known to have a highly homologous (98%) pseudogene, a primary amplification was performed using gene-specific primers as described[30]. For VDA samples, PCR was performed using DNA (25 ng), dNTPs (100 μM), $MgCl_2$ (1.5 mM) and AmpliTaq Gold (0.75 U; Perkin-Elmer) in 15 μl reaction volumes. Samples were amplified using an MJ Tetrad for 96 °C for 10 min; 35 cycles of 96 °C for 30 s, 59 °C for 2 min and 72 °C for 2 min; and 72 °C for 5 min. Assays destined to be hybridized to the same chip design were pooled together. Chip samples were prepared and hybridized as described[14,15], except that pools consisting of ~100 assays typically contained 5–6 μg of amplified material. In all, we amplified 854 assays (average size of 300 bp, covering 106 genes) from each individual and hybridized them to 12 distinct chip designs. We designed probe arrays to query only the coding sequence for some genes, but also included surrounding untranslated and intronic sequences for other genes. For DHPLC samples, sequences were amplified as above, except the final extension in the PCR protocol was followed by denaturation and slow reannealing to allow heteroduplex formation. We injected each individual's PCR product (6 μl) into Wave DNA Fragment Analysis System (Transgenomic). We successfully screened 592 of the VDA assays (covering the 89 genes attempted with this method) by DHPLC. We used only assays of >160 bp for DHPLC, because in our experience shorter assays performed unreliably for mutation detection. The DHPLC parameters (percentage of acetonitrile, column temperature) used for each fragment were automatically calculated using a novel predictive algorithm.

Data analysis. For VDAs, candidate SNPs were identified using a combination of three algorithms followed by visual inspection. The analysis software and guidelines have been described[14,15]. For each base position and strand queried there are four VDA features: one contains the expected base (the reference sequence) in the central position and the other three features contain central substitution bases (in the background of the reference sequence). The base-calling algorithm looked for positions at which hybridization to a substitution base gives a stronger signal than the reference base. The second algorithm (mutant fraction) examined the reference base and each of the substitution bases in turn and calculates the fraction of signal present in the non-reference base. The final algorithm (footprint detection) depends on a loss of signal at the reference positions surrounding a nucleotide substitution. These algorithms are combined to yield a confidence score of 'certain' or 'likely' for each candidate polymorphism. Two analysts independently scored the data, and candidate polymorphisms found by either observer were included in subsequent confirmation tests. For DHPLC, we transferred DHPLC trace files to a UNIX system and analysed them using the clustering program ASH v2.0 (J.I., unpublished data). A scoring algorithm was developed based on the similarity score produced by ASHv2.0 and contour of the elution profile. Full details of the predictive algorithm, ASHv2.0 and performance of the scoring algorithm will be reported elsewhere (D.A. *et al.*, manuscript in preparation).

DNA sequencing. We amplified sequences with PCR primers tailed with

standard M13 sequencing sites (–21 forward and –28 reverse) and performed conventional dye-primer sequencing on ABI 377 sequencers. For candidate SNPs discovered by VDAs, we chose one individual (a candidate homozygous variant, when available, or a candidate heterozygote) and performed sequencing on one strand to confirm by visual inspection the presence of the SNP at the indicated position. For amplicons found to be polymorphic by DHPLC, we selected two individuals representing each distinct elution pattern observed and sequenced them on both strands to discover the variant base or bases. Sequences were base-called by the Phred program, assembled by the Phrap program and polymorphism candidates identified by the PolyPhred program[31]. All results were visually inspected by at least two observers.

Determination of false-positive and false-negative rates. The overall false-positive rate for screening for SNPs using VDAs was 45%. The rate was much lower (~10%) for certain chip designs, synthesis protocols and for candidate polymorphisms scored as 'certain'. The false-positive rate among fragments displaying an altered elution pattern by DHPLC was similar (40%). The false-positive rates reflect the thresholds employed for declaring a candidate SNP, which were chosen to ensure high sensitivity. To directly determine the false negative rate of the screen, we sequenced ten genes (*CYP11B1*, *F10*, *GABRB1*, *GHR*, *HTR1A*, *HTR2A*, *IGF2*, *PTHLH*, *TBXA2R* and *THPO*) on both strands in 20 individuals and interpreted the traces using PolyPhred. Correcting for the specific individuals and regions interrogated by each method, VDA identified 34/40 SNPs found by ABI sequencing, and DHPLC discovered 46/53 SNPs identified by ABI.

Calculations of allele frequencies. Calculations involving allele frequency (including π) required accurate genotyping of samples. Polymorphisms identified by DHPLC alone were excluded from such calculations because we did not sequence all of the samples showing a variant DHPLC pattern and thus were not certain of allele frequency. The estimates of π were thus based on 420 of 560 polymorphisms. Although the VDAs were designed for polymorphism discovery rather than genotyping, the estimated allele frequencies for confirmed SNPs proved to be quite accurate. Specifically, genotyping assays (employing single-base extension assays) for 25 SNPs yielded allele frequencies that differed by an average of only 2% from those estimated on the basis of genotypes inferred from the VDA (J. Hirschhorn and S. Bolk, unpublished data).

Determination of conservative and non-conservative changes. Conservative and non-conservative amino acid substitutions were defined for this analysis according to the BLOSUM62 matrix, used in sequence comparison[32]. Conservative changes were those having a positive or neutral sign in the matrix, whereas non-conservative changes were those having a negative value. We calculated the proportion of non-synonymous SNPs expected to cause a non-conservative amino acid substitution on the basis of the actual codon usage in the 106 genes studied, the known frequencies of transitions and transversions and the definition of non-conservative changes employed in the BLOSUM62 matrix.

Chimpanzee sequencing. Each assay used in the human survey was amplified from a single chimpanzee and subjected to dye-primer sequencing on both strands. We obtained 136 kb of chimpanzee sequence, revealing an inter-species divergence of 0.6% in the regions studied. A single chimpanzee sample will accurately reveal the ancestral allele, except in cases where the site has mutated and fixed during the chimpanzee evolution or is polymorphic in the chimpanzee population and happened to be homozygous for the non-ancestral allele. These two cases are quite rare (probably <2%) and thus have been neglected for the purpose of estimating overall rates. We considered a human allele to be ancestral if it was present in the homozygous state in the chimpanzee sample.

Expected distribution of private polymorphisms. For the evaluation of private polymorphisms, a subset of the data was employed: polymorphisms identified by VDA or comprehensive ABI sequencing for which we had reliable individual genotyping data. We observed 253 such polymorphisms more than once. The probability that a SNP occurring k>1 times in an overall sample of n individuals would be found entirely within a given subset of m individuals is B(m,k)/B(n,k), where B(x,y) is the binomial coefficient x!/(x-y)!y!. In this fashion, we calculated the probability that each individual SNP would be confined to a particular ethnic subgroup within the sample and summed these probabilities to obtain the number of SNPs expected to be confined to the group within the sample.

Acknowledgements

We thank the High Throughput Screening Team from Affymetrix for processing some of the VDA hybridizations; and L. Linton and L. Kann for DNA sequencing support; S. Rozen for the initial VDA designs; and members of the Whitehead Institute Program in Functional Genomics for discussions. D.A. is a recipient of a postdoctoral fellowship for physicians from the Howard Hughes Medical Institute. P.S. is supported by a Young Investigator Award from NARSAD. This work was supported by a grant from Affymetrix, Bristol-Meyers-Squibb and Millennium Pharmaceuticals.

1. Ayala, F.J., Escalante, A., O'Huigin, C. & Klein, J. Molecular genetics of speciation and human origins. *Proc. Natl Acad. Sci. USA* **91**, 6787–6794 (1994).
2. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
3. Collins, F.S., Guyer, M.S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
4. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
5. Saunders, A.M. *et al.* Association of apolipoprotein E allele ε 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**, 1467–1472 (1993).
6. Bertina, R.M. *et al.* Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* **369**, 64–67 (1994).
7. Dean, M. *et al.* Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* **273**, 1856–1862 (1996).
8. Corder, E.H. *et al.* Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nature Genet.* **7**, 180–184 (1994).
9. Moriyama, E.N. & Powell, J.R. Intraspecific nuclear DNA variation in Drosophila. *Mol. Biol. Evol.* **13**, 261–277 (1996).
10. Harris, H. *The Principles of Biochemical Genetics* (North-Holland/Elsevier, Amsterdam, 1975).
11. Harding, R.M. *et al.* Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**, 772–789 (1997).
12. 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).
13. Li, W.-H. & Sadler, L.A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
14. Chee, M. *et al.* Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–614 (1996).
15. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
16. Underhill, P.A. *et al.* A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proc. Natl Acad. Sci. USA* **93**, 196–200 (1996).
17. Li, W.-H. *Molecular Evolution* (Sinauer Associates, Canada, 1997).
18. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
19. Begun, D.J. & Aquadro, C.F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1993).
20. Nachman, M.W., Bauer, V.L. Crowell, S.L. & Aquadro, C.F. DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133–1141 (1998).
21. Wayne, M.L. & Simonson, K.L. Statistical tests of neutrality in the age of weak selection. *Trends Ecol. Evol.* **13**, 236 (1998).
22. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
23. Watterson, G.A. & Guess, H.A. Is the most frequent allele the oldest? *Theor. Popul. Biol.* **11**, 141–160 (1977).
24. Zietkiewicz, E. *et al.* Nuclear DNA diversity in worldwide distributed human populations. *Gene* **205**, 161–171 (1997).
25. Halushka, M.K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes regulating blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
26. Eyre-Walker. A. & Keightley, P. High genomic deleterious mutations rates in hominids. *Nature* **397**, 344–347 (1999).
27. Weber, J.L. & Myers, E.W. Human whole-genome shotgun sequencing. *Genome Res.* **7**, 401–409 (1997).
28. Venter, J.C. *et al.* Shotgun sequencing of the human genome. *Science* **280**, 1540–1542 (1998).
29. Clark, A.G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
30. Day, D.J., Speiser, P.W., White, P.C. & Barany, F. Detection of steroid-21 hydroxylase alleles using gene specific PCR and a multiplex ligation detection reaction. *Genomics* **29** 152–162 (1995).
31. Nickerson D.A., Tobe, V.O. & Taylor, S.L. PolyPhred: automating the detection and genotyping of single nucleotide substitution using fluorescence-based resequencing. *Nucleic Acids Res.* **25**, 2745–2751 (1997).
32. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10915–10919 (1992)