

Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis

Marc K. Halushka¹, Jian-Bing Fan², Kimberly Bentley¹, Linda Hsie², Naiping Shen², Alan Weder³, Richard Cooper⁴, Robert Lipshutz² & Aravinda Chakravarti¹

Sequence variation in human genes is largely confined to single-nucleotide polymorphisms (SNPs) and is valuable in tests of association with common diseases and pharmacogenetic traits. We performed a systematic and comprehensive survey of molecular variation to assess the nature, pattern and frequency of SNPs in 75 candidate human genes for blood-pressure homeostasis and hypertension. We assayed 28 Mb (190 kb in 148 alleles) of genomic sequence, comprising the 5' and 3' untranslated regions (UTRs), introns and coding sequence of these genes, for sequence differences in individuals of African and Northern European descent using high-density variant detection arrays (VDAs). We identified 874 candidate human SNPs, of which 22% were confirmed by DNA sequencing to reveal a discordancy rate of 21% for VDA detection. The SNPs detected have an average minor allele frequency of 11%, and 387 are within the coding sequence (cSNPs). Of all cSNPs, 54% lead to a predicted change in the protein sequence, implying a high level of human protein diversity. These protein-altering SNPs are 38% of the total number of such SNPs expected, are more likely to be population-specific and are rarer in the human population, directly demonstrating the effects of natural selection on human genes. Overall, the degree of nucleotide polymorphism across these human genes, and orthologous great ape sequences, is highly variable and is correlated with the effects of functional conservation on gene sequences.

Introduction

Studies of nucleotide polymorphism have three distinct advantages: (i) detection of all types of sequence change (single nucleotide substitutions, insertion/deletions and copy number variation in nucleotide repeat motifs), of which SNPs are the most common; (ii) variant detection in coding (cSNPs) and non-coding DNA, including those in putative regulatory regions; and (iii) accomplishment of both detection and genotyping using a single method. Early attempts to survey human genetic variation at the DNA level were non-systematic and used either restriction fragment length polymorphism (RFLP) analysis of single genes in small samples, which sampled only a fraction of the extant variation¹⁻⁴, or the chance availability of two allelic sequences of the same gene⁵. SNP discovery in large samples has been reported, but these analyses are still restricted to single genes^{6,7}. To enable large-scale surveys of human genomic variation, rapid, efficient and high-throughput methods based on hybridization to high-density DNA probe arrays, or microarrays, have come to the forefront^{8,9}. Recently, Lander and colleagues¹⁰ have systematically searched for sequence variants in approximately 2.3 Mb of anonymous and 3' UTR DNA using VDAs of immobilized oligonucleotides^{9,11} and report over 3,200 candidate SNPs.

We report here the results of systematic and comprehensive studies of multiple human genes using high-density VDAs

(refs 9,11) in a large and diverse human sample. In an analogous study, Cargill *et al.*¹² have performed a similar analysis on a different set of genes and DNA samples to obtain similar results. Our aims were to detect, quantify and establish the pattern of sequence variation of each gene examined in the entire coding DNA and the immediately adjacent non-coding DNA, including both 5' and 3' UTRs and a portion of each intron. SNPs are used as markers in human genetic studies ranging from comparative population variation^{8,13} to disease linkage studies^{14,15}. SNPs also have potential as direct functional polymorphic variants involved in common and genetically complex human diseases^{8,13-17} and pharmacogenetic traits¹⁸. We have sought to identify a collection of potentially functional SNPs. For genetically complex diseases, we and others have outlined the 'common disease-common variant' hypothesis, which states that common disease susceptibility or resistance variants are expected to be few at each locus, relatively common in the human population and enriched in the coding and regulatory sequence of genes^{8,13,16}. Thus, polymorphism screening can lead to identification of an informative set of common gene-based SNPs for later use in complex disease gene identification. Gene-based SNPs can, of course, also serve as surrogates for unrecognized neighbouring functional SNPs that may be identified by disease-marker linkage disequilibrium studies^{8,13,14,16}.

¹Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA. ²Affymetrix, Inc., 3380 Central Expressway, Santa Clara, California 95051, USA. ³Department of Medicine, University of Michigan, Ann Arbor, Michigan 48109, USA. ⁴Department of Epidemiology, Loyola University Medical Center, Maywood, Illinois 60153, USA. Correspondence should be addressed to A.C. (e-mail: axc39@po.cwru.edu).

A model complex phenotype amenable to SNP identification and association analysis in the near term is essential hypertension. Essential hypertension, a leading cause of human morbidity and mortality, has a heritability of 35–45%, and environmental risk factors also affect an individual's hypertensive state^{19–21}. The identity of specific genes involved in mediating hypertension is still unclear, although physiologic, pharmacologic and genetic studies in animal models and human populations have identified a number of genes and pathways that regulate blood-pressure homeostasis^{22–25}. We studied 75 genes that encode hormones, growth factors and other ligands, receptors, adaptors, channels, transporters, cytoskeletal proteins, and enzymes known or suspected to regulate blood pressure and that may harbour variants in hypertensive patient populations. To maximize the chances of SNP identification, we screened independent DNA samples from 40 Africans and 34 Americans of Northern European descent, to include both a range of human genetic diversity and a range of hypertension phenotype diversity. Human polymorphism is known to be greater in African populations compared with European, Asian or American populations²⁶. Furthermore, there are significant differences in the prevalence and phenotype of hypertension between Africans (or African-Americans) and Northern Europeans^{19–21} (or European-Americans). Moreover, the individuals we sampled were ascertained from the top and bottom 2.5th percentile of a normalized blood-pressure distribution. Consequently, these individuals are most likely to differ with respect to genetic changes (SNPs) that contribute to blood-pressure differences.

Results

Large-scale systematic surveys of gene variation using VDAs

We scanned approximately 190 kb of genomic sequence comprising 75 candidate genes in 74 humans (148 alleles), or approximately 28 Mb of DNA (Table 1). The sequence scanned consisted of approximately 87 kb of coding, 25 kb of intron and 77 kb of 5' and 3' UTR sequences. We screened genomic sequence using VDAs. The 190 kb of sequence required 9 distinct chip designs. To detect polymorphic nucleotides, we analysed chip hybridization patterns using several algorithms and the computer program Ulysses. For each candidate SNP, we classified each sample as being homozygous for the nucleotide in the reference sequence, homozygous for an alternative nucleotide, or heterozygous. We provide here results from two parts of the polymorphism survey: first, a pilot study of 11 genes (chip designs 1 and 2) in which all SNPs were confirmed to estimate error rates, and subsequently a larger survey of 64 genes (chip designs 3–9) in which a 14% sample of SNPs was confirmed. This experimental design allowed us to obtain an unbiased estimate of the performance and characteristics of VDA screening for SNPs as well as the resultant error rates.

We identified 874 candidate SNPs in 74 samples over 190 kb of DNA, or a frequency of 1 SNP per 217 bp. Identified SNPs are deemed 'candidates', as technical artefacts can lead to the appearance of both false positives and the failure to detect all SNPs (false negatives). The false-positive rate in SNP detection may depend on VDA design (density and the specific sequences synthesized), DNA samples and specific experimental protocols used, and the analytical procedures used for identifying the position and variant site of each SNP. In this study, we kept all aspects of SNP detection constant, varying only chip design. Consequently, we estimated the 'error' rates in SNP detection based on both low-density (192,136 different 25-mer oligonucleotides/VDA) and high-density (304,512 different 25-mer oligonucleotides/VDA) SNP design. We identified 27 'certain' and 104 'likely' SNPs on low-density chip

designs 1 and 2 (Table 1), of which 27 and 76, respectively, were also identified by sequencing (that is, discordancy rates of 0% and 27%, respectively, and 21% overall). For high-density chip designs 3–9, we randomly sampled 57 SNPs across each of the classes (Table 2) to avoid biasing the results towards a specific category. These included 11 'certain' and 46 'likely' SNPs, of which 11 and 34 were also identified by sequencing (that is, discordancy rates of 0% and 26%, respectively, and 21% overall). Thus, we find no significant differences in the detection and accuracy of SNPs based on the density of sequence on the VDA. The observed discordancy rate is a combination of the VDA and sequencing false-positive rates. The sequencing false-positive rate based on single-pass sequencing may be as high as approximately 13% (ref. 10) and has not been independently estimated for these data; thus, VDA analysis has a false-positive rate of 11–21% for all sites depending on the sequencing error rate. We have achieved confirmation of an additional 45 SNPs—27 in the 'certain' and 18 in the 'likely' categories—while developing individual genotyping assays using oligonucleotide arrays. Consequently, we have confirmed 195 SNPs, or 22%, of the total set; the remainder consists of 130 'certain' and 549 'likely' SNPs, of which 130 and 401 are expected to be true SNPs based on the above error rates. Thus, 726 of 874 (83%) are true positives. We also assessed data quality by testing mendelian inheritance for unconfirmed SNPs in one CEPH reference family. We observed 118 SNP (non-reference) alleles in the offspring, of which 115 (97%) were also found in one or both parents, implying a lower error rate than the above figures. To estimate how many variants were missed by VDA screening, we sequenced 4 kb in multiple individuals to find 12 SNPs, 11 of which were also identified by VDAs in the same samples; the false negative rate is thus approximately 8%. The additional SNP identified was located 4 bp from an identified variant, which made its detection by hybridization difficult.

Characteristics of the detected SNPs classified by location in coding or non-coding DNA and length of DNA screened are provided (Table 2). Of 874 candidate SNPs, 387 (44%) were in coding sequence (cSNPs), 150 (17%) in introns and 337 (39%) in 5' and 3' UTRs. Of cSNPs, 178 and 209 SNPs led to synonymous (silent) and non-synonymous (replacement) substitutions in the translated protein, respectively, showing that 54% of all cSNPs lead to replacement of an amino acid residue and probably impact protein function. These frequencies are not significantly different between confirmed SNPs on chip designs 1 and 2 versus all SNPs on chip designs 3–9 when the sequence length scanned is taken into account. There is also a largely non-random nature of mutations leading to polymorphisms (Table 2). Although a random mutation process predicts that two-thirds of all changes are transversions, the data show that 64% of all SNPs arise as transitions. The mutation rate between all nucleotides is not equal, but shows a bias toward a greater frequency of transitions than transversions and a greater proclivity to mutate at CpG dinucleotides due to deamination²⁷. The transition frequency showed significant variation ($\chi^2=48.8$, 4 d.f., $P=6.6\times 10^{-10}$) across the 5 categories of 5' UTR, 3' UTR, intron, silent sites and replacement sites. CpG dinucleotides, which are preferred sites of mutation, explain 31% of all SNPs we identified. CpG frequency also showed significant, albeit less, variation ($\chi^2=11.3$, 4 d.f., $P=0.025$) across the five categories. The major source of this variation is nucleotide sequence composition, particularly the higher GC content in coding sequences. Future experiments might exploit this nonrandom occurrence of SNPs to enhance rates of detection.

Table 1 • Candidate genes for blood-pressure homeostasis and hypertension screened for human SNPs

Gene	Sequence screened (bp)					Total	No. of candidate SNPs
	5' UTR	Coding	Intron	3' UTR			
<i>ADD1</i>	387	2,341	517	1,555	4,800	14	
<i>ADD2</i>	153	2,238	393	0	2,784	8	
<i>ADM</i>	1,246	558	108	1,198	3,110	12	
<i>ADORA2A</i>	269	1,239	36	863	2,407	4	
<i>ADRB3*</i>	215	1,227	36	654	2,132	7	
<i>AGT*</i>	705	1,458	101	609	2,873	17	
<i>AGTR1*</i>	610	1,080	144	516	2,350	8	
<i>AGTR2*</i>	181	1,092	72	362	1,707	5	
<i>ALDR1</i>	108	951	324	396	1,779	10	
<i>APOA1</i>	252	804	762	348	2,166	13	
<i>APOA2</i>	557	303	857	259	1,976	7	
<i>APOA4</i>	773	1,191	1,134	98	3,196	22	
<i>APOC1</i>	403	252	1,280	5	1,940	10	
<i>APOC2</i>	116	255	463	240	1,074	8	
<i>APOC3</i>	316	300	1,972	488	3,076	19	
<i>APOC4</i>	220	384	2,644	301	3,549	26	
<i>AVP</i>	155	495	78	186	914	1	
<i>AVPR2</i>	84	1,116	72	577	1,849	9	
<i>BDKRB2*</i>	238	1,095	72	739	2,144	5	
<i>BRS3</i>	562	1,200	72	7	1,841	6	
<i>CALCA</i>	1,646	882	108	1,065	3,701	8	
<i>CLCNKB</i>	0	1,692	0	8	1,700	10	
<i>CYH</i>	205	744	144	173	1,266	7	
<i>CYP11B1</i>	403	1,512	288	805	3,008	29	
<i>CYP11B2</i>	189	1,512	288	402	2,391	24	
<i>DBH</i>	149	822	294	0	1,265	9	
<i>DCP1*</i>	46	2,114	108	95	2,363	15	
<i>DRD1</i>	158	1,341	0	14	1,513	8	
<i>EDN1</i>	86	639	127	173	1,025	2	
<i>EDNRA</i>	0	855	75	1,639	2,569	12	
<i>EDNRB</i>	2	1,305	0	3	1,310	8	
<i>GALNR</i>	735	1,042	62	551	2,390	12	
<i>GCG</i>	18	540	659	735	1,952	10	
<i>GCGR</i>	69	1,146	0	278	1,493	5	
<i>GH1</i>	132	654	143	39	968	1	
<i>GH2</i>	137	654	144	27	962	1	
<i>GIPR</i>	353	1,355	466	0	2,174	3	
<i>GNB3</i>	1,070	1,023	360	512	2,965	8	
<i>GYS1</i>	130	2,125	515	0	2,770	19	
<i>HP</i>	448	1,221	216	746	2,631	13	
<i>HSD11B1*</i>	112	879	162	192	1,345	0	
<i>HSD11B2</i>	1,130	1,218	144	587	3,079	3	
<i>IAPP</i>	156	268	350	1,041	1,815	8	
<i>ICAM1</i>	680	1,599	170	1,057	3,506	12	
<i>ICAM2</i>	277	812	15	186	1,290	3	
<i>INS</i>	244	333	36	36	649	4	
<i>KCNJ11</i>	664	1,173	0	1,345	3,182	19	
<i>KLK</i>	618	776	141	178	1,713	6	
<i>LRP8</i>	510	2,028	0	1,125	3,663	13	
<i>MLR*</i>	480	2,867	18	2,566	5,931	19	
<i>NPPA*</i>	228	375	72	292	967	6	
<i>NPPB</i>	258	405	260	42	965	4	
<i>NPPC</i>	1,555	336	36	1,414	3,341	9	
<i>NPY</i>	17	294	72	62	445	6	
<i>NPY1R</i>	7	1,363	72	830	2,272	2	
<i>PI4</i>	89	1,288	108	194	1,679	6	
<i>PLA2G1B</i>	141	444	358	13	956	7	
<i>PNMT</i>	127	849	72	104	1,152	7	
<i>PTGER3</i>	249	1,504	3,313	2,023	7,089	37	
<i>PTGIS</i>	576	1,461	3	4,048	6,088	54	
<i>PTGS2</i>	471	1,815	324	633	3,243	10	
<i>REN*</i>	60	1,224	324	214	1,822	5	
<i>SAH</i>	375	1,737	468	139	2,719	9	
<i>SCNN1G</i>	470	1,950	571	1,075	4,066	24	
<i>SELE</i>	351	1,833	414	1,562	4,160	17	
<i>SLC2A2</i>	586	1,575	360	203	2,724	16	
<i>SLC2A4</i>	1,812	1,518	360	1,471	5,161	29	
<i>SLC2A5</i>	197	123	36	0	356	1	
<i>SLC4A1</i>	472	2,736	684	2,090	5,982	29	
<i>SLC6A2</i>	18	1,854	468	49	2,389	13	
<i>SLC8A1*</i>	146	2,996	396	3,120	6,658	19	
<i>TBXA2R</i>	895	1,024	42	1,054	3,015	20	
<i>TBXAS1</i>	314	1,607	432	289	2,642	16	
<i>TRH</i>	130	729	12	738	1,609	13	
<i>TRHR</i>	609	1,197	53	1,939	3,798	14	
Totals	28,550	86,947	25,480	48,577	189,554	874	

*Indicates all human SNPs confirmed and sequence assayed for variation in great apes.

Table 2 • Total numbers and nucleotide diversity of human SNPs classified by gene sequence class

	Chip design no.		
	1,2	3–9	All
Number of genes	11	64	75
Total bp screened	30,292	159,261	189,553
Coding	16,407	70,539	86,946
synonymous	3,992	17,164	21,156
non-synonymous	12,415	53,375	65,789
Non-coding	13,885	88,722	102,607
5' UTR	3,021	26,577	29,598
intron	1,505	23,999	25,504
3' UTR	9,359	39,218	48,577
SNPs detected	105 ^a	769 ^b	874
Coding	41	346	387
synonymous	20	158	178
non-synonymous	21	188	209
Non-coding	64	423	487
5' UTR	12	96	108
intron	7	143	150
3' UTR	45	184	229
Nucleotide diversity, θ ($\times 10^4$)	6.2 \pm 1.5	8.6 \pm 2.0	8.3 \pm 1.9
Coding	4.5 \pm 1.2	8.8 \pm 2.1	8.0 \pm 1.9
synonymous	9.0 \pm 2.9	16.5 \pm 4.0	15.1 \pm 3.6
non-synonymous	3.0 \pm 1.0	6.3 \pm 1.5	5.7 \pm 1.4
Non-coding	8.2 \pm 2.2	8.5 \pm 2.0	8.5 \pm 2.0
5' UTR	7.2 \pm 2.7	6.6 \pm 1.6	6.8 \pm 1.7
intron	8.5 \pm 3.8	10.9 \pm 2.7	10.5 \pm 2.6
3' UTR	8.8 \pm 2.4	8.5 \pm 2.1	8.4 \pm 2.0
% transitions among SNPs	69	64	64
% SNPs occurring in CpG	39	30	31

^a100% of SNPs identified were confirmed by gel-based sequencing. ^bRandom sample of 14% of SNPs identified were confirmed by gel-based sequencing.

Nucleotide diversity shows variation across genes and functional class

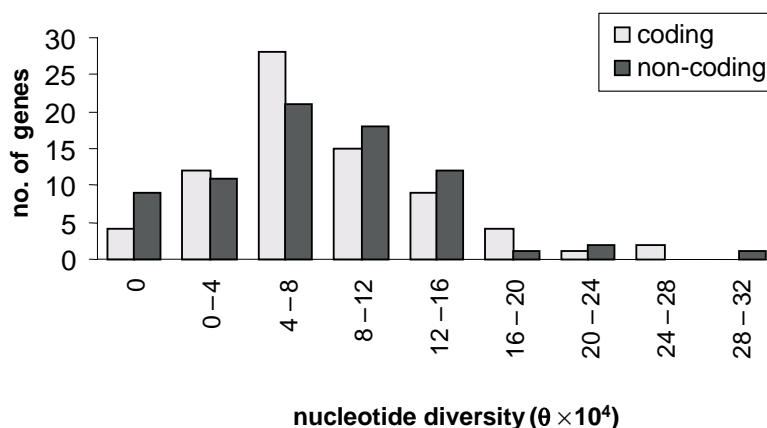
For comparative studies of sequence diversity, we estimated the heterozygosity per nucleotide site by either the empirical heterozygosity per site (π), termed nucleotide diversity, or the mutation parameter (θ), based on standard population genetic principles. These measures correct for both sample size and the length of the region surveyed^{28,29} and are nearly identical; in this study we report θ . We estimated the nucleotide diversities to be almost identical for coding (0.0008 \pm 0.0002) and non-coding (0.0009 \pm 0.0002) regions (Table 2), but these values fail to uncover the heterogeneity in diversity by functional class. Thus, in coding sequences, silent SNPs (0.0015 \pm 0.0004) show 2.5-fold more diversity than replacement SNPs (0.0006 \pm 0.0001), a statistically significant difference ($\chi^2=5.9$, 1 d.f., $P=0.015$) reflecting functional constraint and selection against changes in the protein sequence.

There is heterogeneity in polymorphism rates in the non-coding class as well because introns (0.0011 \pm 0.0003) are about 50% more variable than either the 5' (0.0007 \pm 0.0002) or 3' UTR (0.0008 \pm 0.0002), these differences not being statistically significant ($P>0.24$). The greater diversity in the 3' than 5' UTR and the relative patterns of non-coding sequence diversity can also be correlated with functional conservation of sequence^{28,29}. Estimates of θ in this study were generally twofold higher than those observed in other gene-based⁵ or genomic studies¹⁰. A number of factors may contribute to this difference: the use of genetically diverse samples, screening a larger number of genes and adjacent regions, and screening larger samples. It is unlikely that the increase is attributable solely to the use of candidate rather than confirmed SNPs, as we have estimated that more than 83% of the SNPs are true. Moreover, comparisons of diversity between confirmed SNPs on chip designs 1 and 2 and the remainder on chip designs 3–9 (Table 2) show that θ values of non-coding SNPs are

nearly identical; the pattern of diversity of cSNPs is the same although the numerical values are larger for chip designs 3–9 than for 1 and 2 (Table 2). We largely attribute this to the specific selections of genes we studied on each chip.

Gene-to-gene differences in SNP diversity are the most important of all factors that contribute to such variation. We identified 12 SNPs per gene, on average, but this number ranged from 0 (*HSD11B1*) to 54 (*PTGIS*), with 36 genes (48%) harbouring 10 or more SNPs (Table 1). Some of this variability is due to differences in the sequence length screened for each gene, but natural selection and frequency of meiotic recombination, both of which vary in local genomic segments, may also regulate levels of variation^{30,31}. Under the neutral allele infinite site models, the number of SNPs detected is proportional to the product of the sequence length screened and the natural logarithm of sample size. Consequently, we performed regression analysis of the natural logarithm of the number of coding and non-coding SNPs identified for each gene against the natural logarithm of sequence length screened; sample size being constant across genes was not a contributory factor. We ignored data from regions with no SNPs and studied 70 coding and 65 non-coding segments. Both regression analyses were highly significant and demonstrated that this length dependence explained 29% of the variation for cSNPs ($R^2=0.29$, $F(1,69)=27.99$, $P=1.3\times 10^{-6}$), but 57% for non-coding sequences ($R^2=0.57$, $F(1,64)=85.45$, $P=2.1\times 10^{-13}$). These analyses emphasize that, although sequence length is one parameter determining SNP detection, an additional factor must be intrinsic gene-specific diversity, particularly for coding regions. Consequently, we estimated the nucleotide diversity (θ) separately for coding and non-coding sequences of each gene (Fig. 1). There is 15-fold variation in θ across genes (coefficient of variation: 67% coding; 74% non-coding), with coding segments being less diverse than non-coding sequences. Thus, diversity in human genes is a gene-specific characteristic. Although mutation is responsible for creating

Fig. 1 Distribution of nucleotide diversity in human genes in coding and non-coding segments. The coding sequence comprises both silent and replacement cSNPs; the non-coding sequence comprises 5' UTR, 3' UTR and intronic sequences.



SNPs, their maintenance probably depends on natural selection on coding sequences, which in turn is regulated by its precise functional role⁸ as well as meiotic recombination^{30,31}.

SNP allele frequency patterns correlate with allele age and function

For human genetic applications, allele frequencies at SNPs are an important characteristic defining utility¹⁷. The allele frequency distribution of all 874 SNPs by the frequency of the minor (<50%) allele is shown (Fig. 2). The allele frequency spectrum shows that most genic SNPs are relatively uncommon, with 60% having a frequency less than 10%; the distribution fits the expectations of the infinite sites neutral allele model using Tajima's test²⁹ (Fig. 2). As many sources of variation determine SNP frequency, we next studied frequency spectra for each functional class. The 874 candidate SNPs have an average minor allele frequency of 11% and average heterozygosity of 17%, but there is considerable variation because the standard deviation of this frequency is 12%. This variation is explained by the correlation of SNP allele frequency with its functional role. First, silent and replacement SNPs have significantly different ($\chi^2=12.3$, 1 d.f., $P=0.0005$) average allele frequencies of $11\pm 0.9\%$ and $7\pm 0.7\%$, respectively (the expected heterozygosity for silent and replacement SNPs is 20% and 13%, respectively). Second, the frequency of detected SNPs with respect to their population identity also revealed variation. Analysis of the number of SNPs

identified across all 75 genes in individuals of African (y) versus Northern European (x) descent ($y=2.02+1.02x$; $R^2=0.62$; $F(1,72)=118.99$; $P 6.6\times 10^{-17}$) indicated that the African sample produced on average two additional SNPs per gene, reflecting the greater diversity expected in this population²⁶. In addition, estimates of nucleotide diversity were greater in African than European-American samples (Table 3) for all functional classes, although the relative values across categories were identical between the two populations. The allele frequency distribution of SNPs restricted to individuals of African (414 SNPs) or Northern European (256 SNPs) descent or shared (204 SNPs) by both is also shown (Fig. 2). Thus, Africans have both greater diversity and a greater number of unique SNPs, reflecting the antiquity of the population. Furthermore, population-specific SNPs have an inverse-J shaped distribution, with most sites (80%) having frequency less than 10%. Among this latter class, the smaller the minor allele frequency the greater the population-specific nature of a SNP. The minor allele frequency of population-specific SNPs is, on average, $7\pm 0.4\%$ (heterozygosity, 13%) compared with shared SNPs, which have a uniform distribution across the allele frequency range with frequency of $23\pm 1\%$ (heterozygosity, 35%); this difference is statistically significant ($P<10^{-6}$ by a two-tailed *t* test). It is notable that non-synonymous (replacement) SNPs, which have lower frequency and are more likely to be population-specific, are also the ones most likely to have a direct impact on protein function.

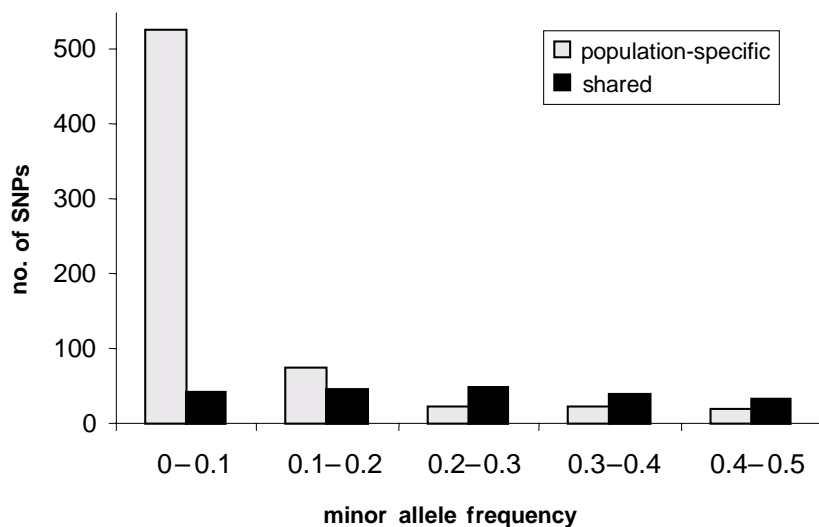


Fig. 2 Distribution of minor allele frequency of SNPs classified by their occurrence among individuals of either African or European descent (population specific) or their presence in both (shared). The frequency spectrum shown used the minor allele frequency (<50%) at each variant site. These data are explained by the neutral infinite sites model, as the observed (expected) values of 569 (533), 118 (124), 72 (83), 61 (69) and 50 (61), respectively, did not differ from theoretical predictions ($\chi^2=11.6$, 4 d.f., $P>0.02$).

(Table 2); $\theta=0.0006\pm 0.0004$ in chimpanzee). The finding of four shared polymorphisms is statistically significant. Of 12,987 bp examined, the frequency of detection of human and chimpanzee polymorphism in our samples is 53/12,987 and 18/12,987, respectively. If all polymorphisms occurred after the separation of the human and chimpanzee lineages then the expected number of sites showing variation in both species is 0.07. The identification of 4 sites compared with the expected number and variance of 0.07, which have identical SNPs in both species, is highly statistically significant ($P < 10^{-6}$). It is likely that these four SNPs, three of which are transitions, are either shared sites that predate the divergence of human and chimpanzee evolution (>5 mya) or exist at a mutation hot spot.

Discussion

The variation survey we describe is instructive for evaluating the performance of VDAs and the rates of false positives and negatives for SNP discovery. The data, as obtained and presented for chip designs 3–9, allow us to evaluate the inherent advantages and disadvantages of a VDA survey. SNP discovery in these early stages is akin to EST surveys, in which individual data items are prone to error, but an entire collection contains valuable genetic information. There is no doubt that we will eventually require confirmation of each SNP, but the strategies needed for such confirmation may well depend on the nature, functional class and frequency of SNPs. Our study has, however, shown by comparative analysis of confirmed (chip designs 1 and 2) and unconfirmed (chip designs 3–9) SNPs that consistent and reliable estimates of genomic diversity and its patterns can be obtained. This study also emphasizes the need for statistical error analysis of SNP data and the scoring of each SNP with respect to quality and probability of being correct. These data are the basis for new experiments and tests of hypothesis regarding human patterns of diversity and the effects of natural selection.

The gene-based SNP survey reported here shows substantial degrees of human and primate population diversity in coding and non-coding sequences. In comparison with some surveys^{5,10}, we have identified greater diversity, with average nucleotide diversity of approximately 0.0008, or 1 difference per 1,200 bp when two human genomes are compared. This estimate is based on the number and frequencies of SNPs identified in two populations and a large sample size. The sample sizes we used permitted a better chance of identifying the functionally relevant rare and low-frequency SNPs; we thus established the pattern of molecular diversity more accurately. A corollary is that some contemplated SNP surveys using small sample sizes³⁴ will miss identification of numerous functionally relevant SNPs. In addition, we used a larger sample of genes to obtain a more reliable estimate of genomic variation.

There may be some concern that estimates of SNP diversity in this study are biased and non-representative because we specifically sampled equal numbers of individuals from the two ends of the blood pressure (BP) distribution. Our data suggest otherwise. A SNP may be detected in the low BP group (denoted '+-'), the high BP ('-+') group, or both ('++'). The numbers of SNPs in the (+-, -+, ++) classes are (31,23,124) (41,45,123) and (63,71,353) for synonymous, non-synonymous and non-coding SNPs, respectively, demonstrating no statistically significant differences between SNPs detected in the low or high BP groups. Nevertheless, it is still possible that estimates of nucleotide diversity for some functional classes are larger in this study based on the sampling we used. This may be further compounded by the use of genes that have a high likelihood of being candidates for blood-pressure homeostasis. Our observations do not support such views, as comparison of this study with the results of Cargill *et al.*

*et al.*¹² demonstrate. Although these two studies have focused on different genes and used different DNA samples, and only confirmed SNPs were used in Cargill *et al.*¹², our diversity estimates are highly similar across all categories. The largest overlap in these two studies comprises samples of Northern European descent. For these samples, the nucleotide diversity estimates from our data are 0.00045, 0.00090, 0.00031 and 0.00054 for coding, synonymous, non-synonymous and non-coding SNPs, respectively; the corresponding values from Cargill *et al.*¹² are 0.00056, 0.00098, 0.00037 and 0.00049, respectively. The closeness of the numerical estimates confirms the validity of each study.

By considering population genetic theory and the early demographic history of the human population, our diversity values are consistent with an effective population size of 10,000 and a median mutation rate of 2.5×10^{-8} per generation³⁵. The similar diversity values for non-coding and coding DNA are not an apparent contradiction, as they fail to uncover the heterogeneity of molecular variation that arises from different constraints in different classes of sites: replacement changes in the coding and regulatory elements in non-coding sequence. Among all SNPs, the non-synonymous changes have the lowest heterozygosity. This implies a younger age, lower survival and thus stronger selection pressure at these sites. By using nucleotide sequence data in humans and primates, it has been estimated³⁶ that at least 38% of non-synonymous changes have been eliminated by natural selection in the human lineage. On comparing θ values for replacement to silent changes and assuming that all silent changes are selectively neutral (Table 2), our data suggest that we have encountered only 38% of expected replacement SNPs. In effect, these data exemplify the role of purifying selection in eliminating 62% of replacement SNPs, although genetic drift must have had a strong role as well^{13,28,29}. Consequently, these are the first polymorphism data that directly show the strong role of natural selection in shaping human gene variation.

Irrespective of how gene-based SNPs arose and are maintained, the collection of SNPs reported here are of immense value to test the hypothesis that such polymorphisms are the aetiologic cause of disease-susceptibility differences. These SNPs are of immediate value as functional variants for blood-pressure and hypertension studies. For example, one variant we identified in an individual with severe hypertension was a heterozygous premature stop codon (Q53Stop) in the gene encoding angiotensinogen (*AGT*). To assess their biological roles, it will be necessary to use these SNPs in association studies^{8,14–18} to identify variants involved in blood-pressure regulation. To enhance such studies, we have made available 2 additional collections of a total of 1,051 SNPs: we identified 39 SNPs at the cDNA level of an additional 11 genes and a collection of 138 SNPs obtained from the literature on 42 genes. We are currently developing a genotyping system using high-density arrays that can allow efficient, accurate and cost-effective assays of this entire collection^{10,11}.

These data allow specific predictions of the nature and distribution of SNPs in the estimated 75,000 human genes, that is, in a study 1,000-fold larger. Currently, over 40,000 human genes have been identified based on ESTs (refs 37,38) and there is an accelerated program to complete the human genome reference sequence, concentrating on the gene-coding segments^{34,39}. Based on a sample of 148 alleles (the size of the sample studied here), we estimate that there are close to 1 million SNPs in human genes, with approximately 500,000 being non-coding SNPs, 200,000 being silent coding SNPs and 200,000 being replacement coding SNPs.

Our study showed that of the 75 proteins encoded by the genes we screened, 83% were polymorphic at the protein level with an average heterozygosity of 17%—values that are considerably greater than classical protein studies^{40,41} and emphasize the large

degree of variation missed in those surveys. A cogent argument is that coding sequence changes are not the only candidates for functional variation and that SNPs in proximal regulatory regions, which we have minimally searched, can have large phenotypic impact, just as they do in evolution⁴². The technology and reagents are now available for assessing which gene SNPs, coding or regulatory, encode the variation in risk for common genetically complex diseases and pharmacogenetic traits.

Methods

Candidate gene selection. We identified 'candidate' genes based on known or suggested involvement in blood-pressure homeostasis and/or hypertension in one of the following biochemical pathways: renin-angiotensin, neural or hormonal pathways regulating blood pressure; regulation of vascular constriction, growth and repair; ion and other small-molecule transport pathways in the kidney; and regulation of glucose metabolism. Evidence supporting these selections was based on known aspects of blood-pressure physiology, animal models with altered blood pressure (including transgenic, knock-out or mouse and rat animal models), human genetic linkage and case-control association studies. This resulted in the formation of an on-line resource of 150 hypertension candidate genes that includes gene name, symbol, expression pattern, map location and GenBank accession numbers for genomic, cDNA or EST nucleotide sequences (M.K.H. *et al.*, manuscript submitted; <http://genome.cwru.edu/candidates/candidates.html>). We selected a subset of 75 genes for the primary polymorphism study because of their strengths as candidate genes as well as the availability of genomic sequence, obtained from GenBank records, at the time of study. GenBank records were parsed to include the 5' UTR, 3' UTR, exon and intronic sequences. Intronic sequence was limited to ~18 bp on each side of an exon for most genes. For small genes, additional tracts of intron and promoter sequence were used to increase the amount of sequence screened. The sequence lengths shown are the actual lengths scanned for polymorphism. In addition to this genomic survey, we screened an additional 11 cDNAs to identify 39 SNPs: +*ADD3*, +*EDN2*, +*GFPT1*, *NPR1*, *NPR2*, +*NPR3*, +*PTGSI*, +*SLC12A3*, +*SLC2A1*, +*SLC9A1* and *VEGFB*, of which 8 (marked by +) contained variants. As an addendum to these gene surveys, we identified a further 138 previously reported SNPs in 42 genes from the literature by a comprehensive search of PubMed and hypertension journals. A list of the total collection of 1,051 SNPs in 115 distinct genes is provided (<http://genome.cwru.edu/candidates/snps.html>) and has been submitted to dbSNP.

DNA samples analysed. In both Harare, Zimbabwe, and Tecumseh, Michigan, we ascertained 800 random individuals for blood pressure and related measurements by population-based screening; blood samples were collected under informed consent from all participating individuals in the top and bottom quartiles. We performed regression analysis in each community sample of systolic, diastolic and mean arterial blood pressure against age and sex, and derived the ranked frequency distribution of residuals. We chose 40 samples from Zimbabwe (African) and 32 samples from Michigan (European American) from the top and bottom 2.5th percentiles. We also used an additional three reference samples of Northern European descent from the CEPH (<http://www.cephb.fr/>) collection (1331-01, 1331-02 and 1331-03; Coriell Cell Repositories). Individual 1331-03 is the offspring of 1331-01 and 1331-02 and did not contribute to the total number of chromosomes screened (148). We used three chimpanzee (*Pan troglodytes*, 1868-CRL, 1847-CRL, 1857-CRL), one gorilla (*Gorilla gorilla*, 1854-CRL) and one orangutan (*Pongo pygmaeus*, 1850-CRL) sample in the primate polymorphism analysis; samples were obtained from the American Type Culture Collection. We purchased 44 human cell lines from the Coriell Cell Repository for use in cDNA screening: GM11036, GM13057, GM012005, GM012250, GM014661, GM05561, GM05963B, GM07007, GM07016, GM07050B, GM07340A, GM07426, GM11321, GM11322, GM11323, GM11324, GM11325, GM11589, GM11590, GM12137, GM13057, GM14661, GM14665, GM14667, GM14672, GM14682, GM14683, GM14698, GM14700, GM14704, GM24696, GM010923, GM012136, GM011814, GM09820, GM012251, GM06816A, GM013055, GM012548, GM012547, GM03715, GM011587A, GM010924 and GM014663.

Preparation of DNA hybridization products. We amplified each genomic region screened by the PCR in multiple segments (80 bp; 14 kb) by both

conventional and long-range PCR protocols. We pooled 205 distinct PCR products, averaging 3 kb, representing all 75 genes for each individual for each chip design; thus, only 190 kb of the total 612 kb amplified was analysed on VDAs. The experimental details of hybridization were as described¹⁰, except for the following changes: we used pooled PCR product (3–5 µg) in a 45 µl reaction, DNase I (0.5 U), TdT (30 U) and Biotin-N6-ddATP (33.3 µM). For the cDNA experiment, we prepared cDNA from each cell line by oligo dT and random hexamer priming of total cellular RNA. PCR amplification was performed using the same protocols as for genomic DNA but with cDNA (0.1 ng) as template.

SNP detection. Experimental methods for detection of candidate SNPs by comparative pattern analysis were as described¹⁰. Analysis of the chip hybridization patterns between samples, indicating potential polymorphisms, used four tests (base calling, clustering analysis, mutant fraction analysis and footprint detection) to determine the variant position and nucleotide for a set of candidate polymorphisms. Manual review of these candidates produced an empirical confidence level for those positions judged to be polymorphic. We classified SNPs as 'certain', 'likely' or 'possible' based on error rates previously determined for similar candidates. We have analysed and reported only 'certain' and 'likely' variants in this study. We re-estimated the false-positive rates for these classes in this study; positions classified as 'mismatch' are non-polymorphic sites in which all samples are homozygous for an alternative allele relative to the reference GenBank sequence on the chip. A mismatch either indicates an error in the original sequence or that the reference entry harbours a rare allele. The SNP data for all genes we studied are available (<http://genome.cwru.edu/candidates/snps.html>).

Confirmation of identified SNPs. To confirm candidate SNPs we used gel-based nucleotide sequencing of relevant PCR products from individual samples classified as either homozygous or heterozygous for an alternative (relative to the reference) base. We used dye terminator chemistry and semi-automated methods using an ABI 377 (ref. 43). We confirmed visually polymorphic sites and genotypes from sequence traces using Edit View 1.0 (Perkin Elmer) or the Consed suite of programs⁴⁴. The frequencies of confirmations reported in this study are the numbers confirmed divided by the number of successful attempts.

Nucleotide diversity and allele frequency estimation. We estimated nucleotide diversity (θ) and its standard deviation, $S(\theta)$, under the assumptions of an infinite site neutral allele model as follows:

$$\theta = K/aL, S(\theta) = \sqrt{a\theta L + b(\theta L)^2}/aL, a = \sum_{i=2}^n \frac{1}{(i-1)}, b = \sum_{i=2}^n \frac{1}{(i-1)^2},$$

where K is the number of SNPs identified in a genomic length, L base pairs and a sample of n alleles. We estimated allele frequencies by the proportion of alleles (p) in a sample and heterozygosity at each SNP as $2p(1-p)$. We also estimated nucleotide diversity as the average heterozygosity per nucleotide site (π). To test the infinite sites neutral allele model we estimated the expected allele frequency spectrum based on the formula $P(i;n-i) \approx 1/i + 1/(n-i)$, where allele frequency $p=i/n$. All calculations used $n=80$ and 68 for the Zimbabwe and Tecumseh samples, respectively, and $n=148$ for the total because the data were 97% or greater complete. The numbers of silent and replacement sites in the sequence scanned was calculated as described²⁹.

Acknowledgements

We acknowledge the assistance of N. Bringht-Twumasi and R. Keefer for technical assistance in sample preparation, PCR analyses and sequencing; M. Mittmann and E. Hubbell for chip design; A. Berno for chip data analyses; N. Patil and C. Marjoribanks for cDNA samples; M. Zwick, H. Willard, C. Langley, E. Eichler and anonymous reviewers for comments on the manuscript; and M. Chee for his efforts at the initiation of this gene screening project. This study was supported by research funds from Case Western Reserve University, University Hospitals of Cleveland, National Heart, Lung & Blood Institute (U10 HL54466) and National Human Genome Research Institute (RO1 HG01847) to A.C. This study is a component of the GenNet network of the NHLBI Family Blood Pressure Program.

Received 23 April; accepted 25 May 1999.

1. Jeffreys, A.J. DNA sequence variants in the G γ , A γ , δ - and β -globin genes of man. *Cell* **18**, 1–10 (1979).
2. Chakravarti, A. *et al.* Nonuniform recombination within the human β -globin gene cluster. *Am. J. Hum. Genet.* **36**, 1239–1258 (1984).
3. Chakravarti, A., Phillips, J.A. 3d, Mellits, K.H., Buetow, K.H. & Seeburg, P.H. Patterns of polymorphism and linkage disequilibrium suggest independent origins of the human growth hormone gene cluster. *Proc. Natl Acad. Sci. USA* **81**, 6085–6089 (1984).
4. Jorde, L. *et al.* Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. *Am. J. Hum. Genet.* **54**, 884–898 (1994).
5. Li, W.H. & Sadler, L.A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
6. 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).
7. 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).
8. Chakravarti, A. Population genetics—making sense out of sequence. *Nature Genet.* **21**, 56–60 (1999).
9. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**, 20–24 (1999).
10. Wang D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
11. Chee, M. *et al.* Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–614 (1996).
12. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
13. Chakravarti, A. It's raining SNPs, hallelujah? *Nature Genet.* **19**, 216–217 (1998).
14. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
15. Kruglyak, L. The use of a genetic map of biallelic markers in linkage studies. *Nature Genet.* **17**, 21–24 (1997).
16. Collins, F.S., Guyer, M.S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
17. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
18. Housman, D. & Ledley, F.D. Why pharmacogenomics? Why now? *Nature Biotechnol.* **16**, 492–493 (1998).
19. Ward, R. in *Hypertension: Pathophysiology, Diagnosis, and Management* (eds Laragh, J.H. & Brenner, B.M.) 81–100 (Raven Press, New York, 1990).
20. Kurtz, T.W. & Spence, M.A. Genetics of essential hypertension. *Am. J. Med.* **94**, 77–84 (1993).
21. Kaplan, N.M. *Clinical Hypertension* (Williams and Wilkins, Baltimore, 1994).
22. Jeunemaitre, X. *et al.* Molecular basis of human hypertension: role of angiotensinogen. *Cell* **71**, 169–180 (1992).
23. Smithies, O. & Maeda, N. Gene targeting approaches to complex genetic diseases: atherosclerosis and essential hypertension. *Proc. Natl Acad. Sci. USA* **92**, 5266–5272 (1995).
24. Mockrin, S.C. *Molecular Genetics and Gene Therapy of Cardiovascular Diseases* (Marcel Dekker, New York, 1996).
25. Lifton, R.P. Molecular genetics of human blood pressure variation. *Science* **272**, 676–680 (1996).
26. *The History and Geography of Human Genes* (eds Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A.) (Princeton University Press, Princeton, 1994).
27. Bird, A.P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209–213 (1986).
28. Nei, M. *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).
29. Li, W.H. *Molecular Evolution* (Sinauer Associates, Sunderland, 1997).
30. Charlesworth, B., Moran, M. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
31. 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).
32. Collins, F.S., Brooks, L.D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231 (1998).
33. Hacia, J.G. *et al.* Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nature Genet.* **18**, 155–158 (1998).
34. Venter, J.C. *et al.* Shotgun sequencing of the human genome. *Science* **280**, 1540–1542 (1998).
35. Harpending, H.C. *et al.* Genetic traces of ancient demography. *Proc. Natl Acad. Sci. USA* **95**, 1961–1967 (1998).
36. Eyre-Walker, A. & Keightley, P.D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347 (1999).
37. Wolfsberg, T.G. & Landsman, D. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* **26**, 1626–1632 (1997).
38. Gerhold, D. & Caskey C.T. It's the genes! EST access to human genome content. *Bioessays* **18**, 973–981 (1996).
39. Collins, F.S. *et al.* New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282**, 682–689 (1998).
40. Harris, H. Enzyme polymorphisms in man. *Proc. R. Soc. Lond. B. Biol. Sci.* **164**, 298–310 (1966).
41. Harris, H. & Hopkinson, D.A. Average heterozygosity per locus in man: an estimate based on the incidence of enzyme polymorphisms. *Ann. Hum. Genet.* **36**, 9–20 (1972).
42. King, M.C. & Wilson, A.C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
43. Angrist, M. *et al.* Human GFRA1: cloning, mapping, genomic structure, and evaluation as a candidate gene for Hirschsprung disease susceptibility. *Genomics* **48**, 354–362 (1998).
44. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195 (1998).