Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome

David G. Wang, *et al.*

Science **280**, 1077 (1998);
DOI: 10.1126/science.280.5366.1077

env sequences in each specimen were examined by phylogenetic analysis.

11. The methods described by G. H. Learn *et al.* [*J. Virol.* **70**, 5720 (1996)] were used to align DNA sequences (with the use of CLUSTALW plus manual adjustment), calculate genetic distances (with the use of DNADIST, using the maximum likelihood method), evaluate potential sample mixups, construct neighbor joining trees, and perform bootstrap analyses (1000 replicates). Sequence regions that could not be unambiguously aligned were removed from subsequent analyses. Each sequence was compared for phylogenetic relatedness to the entire set of published and available unpublished laboratory HIV database sequences. If after this analysis the viral sequences from a mother and an infant appeared as a monophyletic group on a phylogenetic tree, they were judged to be phylogenetically linked or to have a common ancestor not shared by sequences from any other individuals evaluated. Issues regarding the assignment of phylogenetic linkage are discussed in greater detail by Learn *et al*.
12. L. M. Frenkel *et al.*, at www.sciencemag.org/feature/data/974996.shl.
13. R. Liu *et al.*, *Cell* **86**, 367 (1996).
14. L. M. Frenkel *et al*., unpublished data.
15. M.-L. Newell *et al.*, *Lancet* **347**, 213 (1996).
16. P. Palumbo, J. Skurnick, D. Lewis, M. Eisenberg, *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* **10**, 436 (1995).
17. A. McMichael, R. Koup, A. J. Ammann, *N. Eng. J. Med.* **334**, 801 (1996).
18. E. C. Holmes *et al.*, *J. Infect. Dis.* **167**, 1411 (1993).
19. T. Liu *et al.*, *J. Immunol.* **154**, 3147 (1995).
20. A. Hoffenbach *et al.*, *ibid.* **142**, 452 (1989).
21. G. Schochetman, S. Subbarao, M. L. Kalish, in *Viral Genome Methods*, K. W. Adolph, Ed. (CRC Press, Boca Raton, FL, 1996), pp. 25–41.
22. E. L. Delwart, M. P. Busch, M. L. Kalish, J. W. Mosley, J. I. Mullins, *AIDS Res. Hum. Retrovir.* **11**, 1181 (1995).
23. C. H. Contag *et al.*, *J. Virol.* **71**, 1292 (1997).

24. We thank J. Conroy for performing PCR assays; E. Abrams, M. S. Orloff, R. C. Reichman, L. M. Demeter, J. S. Lambert, R. Dolin, R. Sperling, D. Shapiro, G. McSherry, and the Ariel Project and ACTG 076 investigators for critical patient specimens; D. Swofford for use of computer program PAUP*, version 4.0.0d63; and C. B. Wilson and K. K. Holmes for editorial contributions. This work was supported by grants from the Pediatric AIDS Foundation (500153–10-PGT, 50366–14-PGR, 55516-ARI, 55529-ARI, 55525-ARI, 55532-ARI, 55526-ARI, 55531-ARI, and 55522-ARI), the U.S. Public Health Service (UO1–27658, AI32910, AI27757, and AI35539), and the Foster Foundation.

22 December 1997; accepted 26 March 1998

Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome

David G. Wang, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno,
Peter Young, Ron Sapolsky, Ghassan Ghandour,
Nancy Perkins, Ellen Winchester, Jessica Spencer,
Leonid Kruglyak, Lincoln Stein, Linda Hsie,
Thodoros Topaloglou, Earl Hubbell, Elizabeth Robinson,
Michael Mittmann, Macdonald S. Morris, Naiping Shen,
Dan Kilburn, John Rioux, Chad Nusbaum, Steve Rozen,
Thomas J. Hudson, Robert Lipshutz,* Mark Chee,
Eric S. Lander*

Single-nucleotide polymorphisms (SNPs) are the most frequent type of variation in the human genome, and they provide powerful tools for a variety of medical genetic studies. In a large-scale survey for SNPs, 2.3 megabases of human genomic DNA was examined by a combination of gel-based sequencing and high-density variation-detection DNA chips. A total of 3241 candidate SNPs were identified. A genetic map was constructed showing the location of 2227 of these SNPs. Prototype genotyping chips were developed that allow simultaneous genotyping of 500 SNPs. The results provide a characterization of human diversity at the nucleotide level and demonstrate the feasibility of large-scale identification of human SNPs.

Although the Human Genome Project still has tremendous work ahead to produce the first complete reference sequence of the human chromosomes, attention is already focusing on the challenge of large-scale characterization of the sequence variation among individuals (*1*). This genetic diversity is of interest because it explains the basis of heritable variation in disease susceptibility, as well as harbors a record of human migrations.

The most common type of human genetic variation is the SNP, a position at which two alternative bases occur at appreciable frequency (>1%) in the human population. There has been growing recognition that large collections of mapped SNPs would provide a powerful tool for human genetic studies (*1*, *2*). SNPs can serve as genetic markers for identifying disease genes by linkage studies in families, linkage disequilibrium in isolated populations, association analysis of patients and controls, and loss-of-heterozygosity studies in tumors (*1*, *2*). Although individual SNPs are less informative than currently used genetic markers (*3*), they are more abundant and have greater potential for automation (*4*, *5*).

We performed an initial survey to identify SNPs by using conventional gel-based DNA sequencing to examine sequence-tagged sites (STSs) distributed across the human genome. STSs are short genomic sequences that can be amplified from DNA samples by means of a corresponding polymerase chain reaction (PCR) assay. From among 24,568 STSs used in the construction of a physical map of the human genome at the Whitehead Institute for Biomedical Research/MIT Center for Genome Research (*6*, *7*), an initial collection of 1139 STSs was chosen (*8*). These STSs contained a total of 279 kb of genomic sequence (*9*), with one-third from random genomic sequence and two-thirds from 3′-ends of expressed sequence tags (3′-ESTs) and primarily representing untranslated regions of genes. Each STS was amplified from four samples (*10*): three individual samples and a pool of 10 individuals (thereby permitting allele frequencies to be estimated among 20 chromosomes). The PCR products were subjected to single-pass DNA sequencing based on fluorescent-dye primers and gel electrophoresis; sequence traces were compared by a computer program followed by visual inspection (*11*). Candidate SNPs were declared when two alleles were seen among the three individuals, with both alleles present at a frequency greater than 30% in the pooled sample. The term "candidate SNP" is used because a subset of such apparent polymorphisms turn out to be sequencing artifacts, as discussed below.

The survey identified 279 candidate SNPs, distributed across 239 of the STSs. This corresponds to a rate of one SNP per 1001 base pairs (bp) screened and an observed nucleotide heterozygosity of $H = 3.96 \times 10^{-4}$ (Table 1). Expressed sequences (3′-ESTs) showed a lower polymorphism rate than random genomic sequence (with

D. G. Wang, C.-J. Siao, P. Young, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, E. Robinson, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA.
J.-B. Fan, A. Berno, R. Sapolsky, G. Ghandour, L. Hsie, T. Topaloglou, E. Hubbell, M. Mittmann, M. S. Morris, N. Shen, R. Lipshutz, M. Chee, Affymetrix, Incorporated, 3380 Central Expressway, Santa Clara, CA 95051, USA.
E. S. Lander, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA, and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed.

the difference falling just short of statistical significance at $P = 0.057$, one-sided), consistent with greater constraint within genic sequences. The ratio of transitions to transversions was 2:1. Although the dinucleotide CpG makes up only about 2% of the sequence surveyed, nearly 25% of the SNPs occurred at such sites with the substitution almost always being $C \leftrightarrow T$. Cytosine residues within CpG dinucleotides are the most mutable sites within the human genome, because most are methylated and can spontaneously deaminate to yield a thymidine residue (*12*). In addition to the single-base substitutions, 23 insertion-deletion polymorphisms were also found (with all but eight involving a single base), corresponding to a frequency of one per 12 kb surveyed.

Gel-based resequencing was satisfactory for the initial screen, but we sought a more streamlined approach for a larger scale SNP identification. One such approach involves hybridization to high-density DNA probe arrays (*13*). Such "DNA chips" can be produced with parallel light-directed chemistry to synthesize specified oligonucleotide probes covalently bound at defined locations on a glass surface or "chip" (*14*). A target DNA sequence of length *L* can be screened for a polymorphism by hybridizing a biotin-labeled sample to a variant detector array (VDA) of size 8*L* (Fig. 1). For each position on both strands, the array has four 25-nucleotide oligomer probes complementary to the sequence centered at the position. The four differ only in that the central (13th) position is substituted by each of the four nucleotides. Homozygotes (AA) for the expected sequence should hybridize more strongly to the perfectly complementary probe than to the three probes containing a central mismatch. The presence of an SNP would be expected to give rise to a different hybridization pattern, with homozygotes (BB) showing strong hybridization to an alternative base and heterozygotes (AB) showing strong hybridization to two probes. The VDA thus signals the presence of a sequence variation (by a change in the hybridization pattern) and, in many cases, indicates the nature of the change (by a gain of signal at a specific mismatch probe). VDAs have been used for mutation detection of small, well-studied DNA targets [such as 387 bp from the human immunodeficiency virus–1 genome, 3.5 kb from the breast cancer–associated *BRCA1* gene, and 16.6 kb from the human mitochondrion (*13, 15*)] in large numbers of samples. In this setting, the normal hybridization pattern can be characterized with precision and single-base substitutions detected with high accuracy.

In this project, we used VDAs in a large-scale survey. A total of 16,725 STSs covering 2 Mb of human DNA were selected, with one-third from random genomic sequence and two-thirds from 3′-ESTs. The survey used 149 distinct chip designs, each containing 150,000 to 300,000 features. The STSs were examined in seven individuals, representing about 14 Mb of genomic sequence. For each chip, the corresponding STSs were amplified from an individual, pooled together, labeled with biotin, hybridized, and stained (*16*), and the resulting hybridization patterns were compared by a computer program followed by visual inspection (*17*). At each position, samples were classified as homozygous for the expected sequence, homozygous for an alternative sequence, or heterozygous.

A collection of 2748 candidate SNPs were identified, corresponding to a rate of one per 721 bp surveyed and an observed nucleotide heterozygosity of 4.58×10^{-4} (Table 1). The number of STSs containing SNPs was 2299. The SNPs had a mean heterozygosity of 33%, with the minor allele having a mean frequency of 25%. SNPs were found less often in 3′-ESTs than in random genomic sequence ($P < 0.023$, one-sided), consistent with greater constraint in genic regions.

The nucleotide heterozygosity rate was indistinguishable from the estimate obtained from gel-based sequencing ($P > 0.12$, two-sided test), as was the ratio of transitions to transversions and the proportion of SNPs occurring at CpG dinucleotides. SNPs were detected at a higher frequency in the chip-based survey because more samples were surveyed (seven versus three individuals). The observed increase of 38.8% (1/721 versus 1/1001) agreed closely

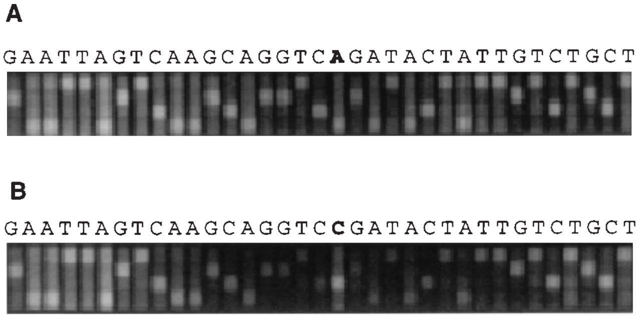
A
GAATTAGTCAAGCAGGTCAGATACTATTGTCTGCT
B
GAATTAGTCAAGCAGGTCCGATACTATTGTCTGCT

Fig. 1. SNP screening on chips. (**A**) Small portion of a VDA for an STS hybridized with the expected target sequence. Chip features in each column are complementary to successive overlapping 25-nucleotide oligomer subsequences, with the central base substituted by A, C, G, or T in the four rows. Variations from the expected sequence can usually be detected by examination of the most intense signal in each column. (**B**) The same VDA was hybridized with sequence containing an SNP (A→C) at position 19. The hybridization signal is now stronger at an alternative base at this position. It is also weaker at the surrounding positions (for example, positions 12 to 18 and 20 to 26), because probes at these positions are designed to be complementary to the A allele at the SNP and mismatch with the C allele.

Table 1. Results of SNP screening.

Variable	Gel-based sequencing			Chip-based detection		
	All STSs	STSs from 3′-EST sequences	STSs from random genomic sequence	All STSs	STSs from 3′-EST sequences	STSs from random genomic sequence
No. of STSs screened	1,139	705	434	16,725	12,649	4,076
Total bases screened	279,165	186,524	92,641	1,981,030	1,324,320	656,710
No. of candidate SNPs found	279	161	118	2,748	1,749	999
SNP frequency (*K*)	1/1001	1/1159	1/785	1/721	1/757	1/657
Heterozygosity (*H*) ($\times 10^{-4}$)	3.96 ± 0.38	3.42 ± 0.43	5.04 ± 0.67	4.58 ± 0.15	4.36 ± 0.18	5.02 ± 0.28
No. of STSs containing SNPs	239	137	102	2,299	1,515	784
% transitions among SNPs	67%	67%	67%	70%	70%	71%
% SNPs occurring within CpG	24%	23%	25%	24%	25%	22%
θ, based on *H*	3.96×10^{-4}			4.58×10^{-4}		
θ, based on *K*	4.33×10^{-4}			4.36×10^{-4}		

with expectation under classical population genetic theory (*18*). This result has implications for the choice of sample size for an SNP survey (*19*).

We estimated the error rates in the gel-based and chip-based surveys. The false-positive rate was estimated by carefully confirming candidate SNPs found in each survey by using thorough multipass sequencing (*20*): 12% of 220 candidate SNPs found in the chip-based survey and 16% of 120 candidate SNPs found by single-pass gel-based sequencing were false positives. The false-negative rate was estimated by considering a subset of STSs that had been included in both surveys: these STSs yielded 55 SNPs (all carefully confirmed to eliminate false positives), of which eight (15%) were missed by single-pass gel-based resequencing and seven (13%) were missed by the chip-based survey. Many of the errors were due to random factors, in that they were eliminated simply by repeating the original experiment. However, some were reproducible artifacts that could be eliminated only by changing the detection protocol (for example, by using dye terminators rather than dye primers in gel-based sequencing). The gel-based sequencing and chip-based analysis had similar rates of accuracy—with a false positive and false negative being found roughly every 5000 to 10,000 bases, or about 10% of the true SNP frequency. The accuracy largely reflects the particular implementation of the technologies in a high-throughput setting and could be increased at the expense of assay optimization.

Although the two surveys yielded comparable accuracy, the survey based on VDAs required considerably less laboratory work than gel-based resequencing. Both approaches required amplifying target loci. The gel-based approach then required a sequencing reaction and electrophoresis on each individual locus, whereas the chip-based approach allowed targets totaling 30 kb to be pooled into a single labeling reaction and hybridized (*21*).

The SNP collection from the two surveys was supplemented by two directed approaches based on public databases. First, we collected reports from the literature of common variants in gene coding regions. We were able to confirm 120 of 143 cases tested by virtue of detecting two alleles in our screening panel; the remainder may be true polymorphisms but simply monomorphic in the individuals tested. Second, the GenBank database contains multiple entries for some ESTs. Such entries were compared to identify single-nucleotide differences, which might reflect either common polymorphisms or sequencing errors in single-pass EST sequencing. We tested 200 such apparent differences and confirmed the presence of an SNP in 94 cases. These two directed approaches thus yielded an additional 214 SNPs.

The project has thus identified 3241 candidate SNPs to date. Confirmation (*22*) has so far been obtained for 1477 SNPs and is expected to yield ~2900 true SNPs. All information about the SNPs has been deposited on the Whitehead/MIT Center for Genome Research Web site (www.genome.wi.mit.edu) and will be updated with results of additional surveys and confirmation tests. The information is also being deposited in the GenBank database.

For SNPs to be useful in human genetic studies, they must be assembled into maps showing their chromosomal location. To create a third-generation map based on SNPs, we used whole-genome radiation-hybrid (RH) mapping (*6*, *7*, *23*), which infers the position of loci based on co-retention in a panel of human-on-hamster cell lines; it has become a primary method for constructing maps of the human genome (*6*, *7*).

The current RH map of the human genome is anchored by a scaffold of 1036 genetic markers from an earlier genetic map consisting of simple sequence length polymorphisms (SSLPs) (*7*). SNPs can be integrated with respect to the earlier genetic map by determining their position on the RH map. We have localized 1880 STSs, containing 2227 of the 3241 candidate SNPs, on the RH map and thereby relative to the human genetic map (Fig. 2 and Table 2). SNPs are not evenly distributed among chromosomes or within chromosomes because most were derived from ESTs, which are known to have an uneven distribution (*6*, *7*). SNP-containing STSs are present at a mean spacing of 2.0 centimorgans (cM) across the genome (*24*), and the map contains 58 intervals greater than 10 cM. The genetic distances on the map must be regarded as approximate because they are based on interpolation from distances in the RH map. It will be desirable to reestimate these distances on the basis of direct linkage analysis in the CEPH families, as high-throughput genotyping for the complete SNP collection becomes feasible.

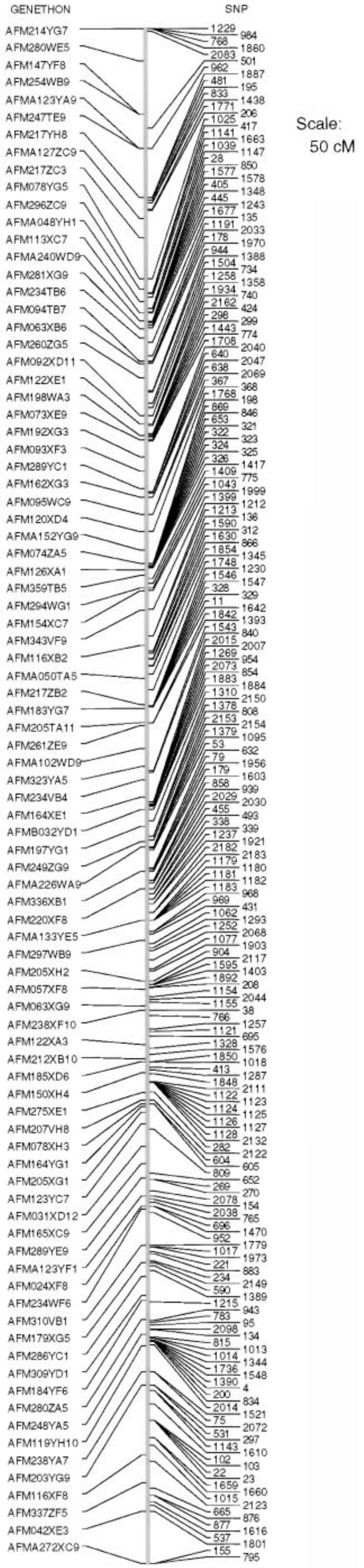

Fig. 2. A portion of the SNP genetic map (showing human chromosome 1). The full map is available on the Whitehead Institute Web site (www.genome.wi.mit.edu). Positions are based on genetic distances in centimorgans. Genetic positions of SNPs were inferred by localizing them relative to framework markers by RH mapping and then interpolating distances from centirays (on the RH map) to centimorgans (on the genetic map). Framework marker names are given in full. SNP names are named with the prefix WIAF (for example, WIAF-17), but the prefix is dropped and only the number is shown in the figure.

We next developed an efficient method for large-scale genotyping of SNPs based on extending the use of DNA chips from SNP discovery to SNP genotyping (5). We synthesized genotyping chips containing "genotyping arrays" for each SNP to be tested. Each genotyping array consists of two short VDAs corresponding to the two alternative alleles (Fig. 3). The presence of an allele should be reflected in strong hybridization to the corresponding resequencing array. PCR assays were designed for the region containing each SNP (25), with the goal of being robust and mutually compatible: the amplification targets were all small (typically, a few nucleotides around the polymorphic site), the primers all had similar calculated melting temperatures, and constant sequences were added to the 5′-ends of the forward and reverse primers to facilitate batch labeling of pooled PCR products. Each assay was tested to ensure that it amplified a single fragment from genomic DNA.

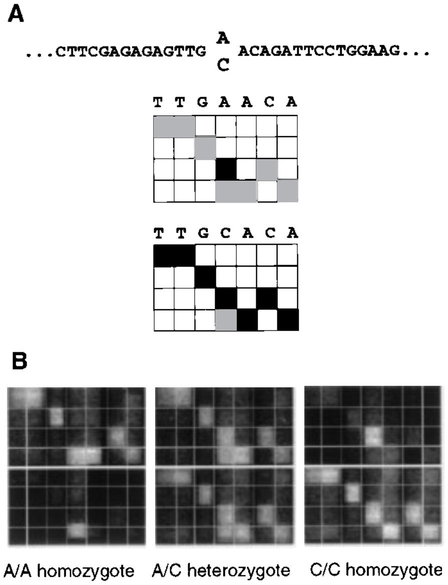

Fig. 3. Genotyping chips. (**A**) Schematic diagram of genotyping array for an SNP, consisting of two VDAs to study seven nucleotides centered around the SNP. The top and bottom arrays are designed to be complementary to the allelic sequences containing A and C, respectively. Probes perfectly matching the A and C alleles are shown in gray and black, respectively. A genotyping array for the complementary strand was also used but is not shown. (**B**) Hybridization signal for a genotyping array probed with samples from three individuals with respective genotypes AA, AC, and CC.

The most complex genotyping chip tested contained genotyping arrays for 558 candidate SNPs identified in the chip-based survey. Initially, the 558 loci were separately amplified, pooled, labeled, and hybridized to the chip. To determine whether each locus could be reliably read, we defined a formal detection test: loci passed if, for each of three individuals tested, the expected DNA sequence could be successfully read on both strands for one or both alleles. In all, 98% of the loci passed this detection test (with the remaining 2% failing as a result of weak hybridization or cross-hybridization).

We next sought to decrease substantially the sample preparation required to genotype large numbers of SNPs, as required to perform a genome scan. We developed a protocol based on multiplex PCR in which primer pairs from many different loci are combined in a single reaction (26). Although it is typically difficult to combine many PCR assays, the approach worked well for our SNP assays: 92% of the 558 loci passed the detection test when amplification was performed in 24 sets of ~23 loci; 90% passed when amplified in 12 sets of ~46 loci; 85% passed when amplified in 6 sets of ~92 loci; and 50% passed when amplified in a single set of 558 loci. The success appears to have resulted from a combination of factors, including the small size of the amplification targets, optimization of amplification conditions, and the presence of the constant sequence at the 5′-ends of the primers (27). It may be possible to salvage the unsuccessful assays by grouping them into additional multiplex sets or by redesigning the assays.

Multiplex amplification of sets of 46 loci was used in subsequent experiments because it decreased the number of reactions by a factor of 46 while allowing the vast majority (512/558) of loci to be assayed. The procedure was further tested in 39 individuals and was quite consistent: 96% of the 512 loci could be successfully read in 100% of individuals tested and the remainder in nearly all individuals.

We next developed a genotyping algorithm for each SNP. Loci were declared to pass a cluster test if the hybridization pat-

Table 2. Chromosomal distribution of genetic markers.

Chromosome	No. of framework markers used from 5264-marker Genethon genetic map	Genetic distance in cM, on Genethon genetic map	No. of SNPs	No. of STSs	Avg. distance between STSs (cM)	No. of intervals >10 cM
1	88	293	236	201	1.5	3
2	91	277	177	149	1.9	2
3	78	233	160	133	1.8	1
4	54	213	98	87	2.4	4
5	65	198	86	72	2.8	4
6	71	201	158	118	1.7	4
7	57	184	119	94	2.0	1
8	45	166	135	108	1.5	3
9	40	167	106	88	1.9	3
10	53	182	85	78	2.3	1
11	58	156	105	92	1.7	1
12	43	169	108	91	1.9	3
13	23	118	57	45	2.6	4
14	31	129	83	64	2.0	3
15	29	110	76	67	1.6	0
16	36	131	70	64	2.0	0
17	23	129	94	87	1.5	2
18	29	124	52	43	2.9	4
19	22	110	58	53	2.1	2
20	34	97	58	49	2.0	2
21	18	60	29	26	2.3	2
22	12	58	31	28	2.1	1
X	36	191	46	43	4.4	8
Total	1036	3699	2227	1880	2.0	58

terns seen in a test set of 39 individuals fell into distinct clusters, corresponding to the possible genotypes (*28*). These clusters could then be used to assign genotypes for further samples (*29*).

The cluster test was applied to the ~500 candidate SNPs that worked well under multiplex amplification conditions: 75% passed the cluster test, and careful resequencing demonstrated that all such loci were true polymorphisms. The cluster test thus provides reliable confirmation of an SNP. The remaining 25% failed the cluster test, and resequencing revealed that half were false positives in the SNP screen and half were true polymorphisms (with the poor discrimination on the chip typically due to one allele hybridizing more weakly than the other). Thus, 88% of the candidate SNPs proved to be true polymorphisms, and 86% of true SNPs passed the cluster test.

To test the reproducibility and accuracy of the genotyping method, we genotyped a set of 91 loci (passing the cluster test) in three individuals by performing chip-based genotyping on six separate occasions over a 2-month period. The correct genotypes were independently determined by thorough gel-based resequencing. The genotyping-chip assay assigned a genotype in 98% of cases (1613/1638), and this assignment proved correct in 99.9% (1611/1613) of these cases. The loci were also genotyped in two complete CEPH families. The genotypes were not independently confirmed, but they were fully consistent with mendelian segregation.

For SNPs passing the cluster test, highly accurate genotypes could thus be obtained with the simple design used here. For the remaining SNPs (14%), similar accuracy can likely be obtained but may require optimization of the genotyping array design, depending on the locus [as shown in (*5*)].

The SNP surveys provide data about human genetic diversity. Two classical measures of diversity (*30*) are *H*, the average heterozygosity per nucleotide, and *K*, the proportion of sites harboring a variation. *H* does not depend on sample size, whereas *K* increases with the number of genomes surveyed. For a population at equilibrium, the neutral theory of evolution relates *H* and *K* to the classical population genetic parameter $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per nucleotide. (θ can be thought of as twice the number of new mutations per generation arising in a population with size N_e.) Specifically, $H \approx \theta$ and $K \approx \theta\,[1^{-1} + 2^{-1} + 3^{-1} + \ldots + (n - 1)^{-1}]$, provided that θ is small. From these equations, one can estimate θ based on *H* or *K*.

The human population is not at equilibrium, but rather underwent a rapid population expansion in the last 100,000 to 200,000 years. Such population explosions tend to suppress the effects of genetic drift and thus preserve the distribution of common alleles and the value of θ. Accordingly, the value of θ is relevant to the ancestral human population before its recent expansion.

The four estimates of θ derived from *H* and *K* for the two surveys are all roughly $\theta \approx 4 \times 10^{-4}$ (Table 1). Assuming a mutation frequency of $\mu \approx 10^{-8}$ to 10^{-9}, this would suggest an effective population size of $N_e \approx 10^4$ to 10^5, which seems reasonable for the ancestral population preceding the explosion in the last 100,000 years (*31*). Strictly speaking, these estimates apply only to the European population, from which all samples were drawn. However, a preliminary survey of a more diverse sample of 31 individuals representing all major racial groups yielded a value of θ that is only 30% larger (*26*), consistent with the idea that human variation occurs primarily within rather than between racial groups (*32*).

The resources reported here represent only a first step toward a dense SNP map of the human genome. The genetic map should already be useful for family-based linkage studies, given the average spacing (2 cM) and average heterozygosity (34%) of the markers. (The heterozygosity applies to the European-derived samples studied here, but a preliminary survey of ~180 of the SNPs shows that most are also polymorphic in other groups.) It still remains to develop a suitable genotyping system, such as a 2000-SNP genotyping chip.

Large-scale screening for human variation is clearly feasible. Someday it may become possible to screen entire human genomes. In the nearer term, a key goal will be to extend SNP discovery to the protein coding regions of all human genes (roughly 120 Mb of sequence, only about 40 times more than the current study) in order to catalog the common variants that may explain susceptibility to common genetic traits and diseases (*1*).

REFERENCES AND NOTES

1. N. Risch and K. Merikangas, *Science* **273**, 1516 (1996); E. S. Lander, *ibid.* **274**, 536 (1996); F. S. Collins, M. S. Guyer, A. Chakravarti, *ibid.* **278**, 1580 (1997).
2. L. Kruglyak, *Nature Genet*. **17**, 21 (1997).
3. SNPs have only two alleles and are less informative than typical multi-allelic simple sequence length polymorphisms (SSLPs). This disadvantage can be offset by using a greater density of SNPs: a genome scan with 1000 well-spaced SNPs, for example, will extract about the same linkage information as the current standard of 400 well-spaced SSLPs (*2*).
4. B. J. Conner *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **80**, 278 (1983); U. Landegren, R. Kaiser, J. Sanders, L. Hood, *Science* **241**, 1077 (1988); D. Y. Wu *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2757 (1989); R. K. Saiki *et al*., *ibid.*, p. 6230; A.-C. Syvanen *et al*., *Genomics* **8**, 684 (1990); D. A. Nickerson *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **87**, 8923 (1990); K. J. Livak *et al*., *Nature Genet*. **9**, 341 (1995); M. T. Roskey *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **93**, 4724 (1996).
5. M. T. Cronin *et al*., *Hum. Mutat*. **7**, 244 (1996).
6. T. J. Hudson *et al*., *Science* **270**, 1945 (1995).
7. G. D. Schuler *et al*., *ibid.* **274**, 540 (1996).
8. STSs with the largest sizes were used in the gel-based screen, and the remaining STSs, having somewhat smaller sizes, were used in the subsequent chip-based screen.
9. The genomic sequence screened (279 kb) is the sum of the distances between the primer sites of the STSs successfully resequenced.
10. The individuals surveyed were chosen from Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees K104, K884, and K1331 from the Amish, Venezuelan, and Utah populations, respectively. The SNP survey by gel-based sequencing examined three unrelated individuals (K104-1, K884-2, K1331-1) and a pool of 10 individuals (K104-13, -14, -15, -16; K884-15, -16; K1331-12, -13, -14, -15). The SNP survey by chip-based analysis examined seven unrelated individuals (K104-1, -16; K884-2, -15, -16; K1331-12, -13).
11. STSs were amplified with their corresponding PCR primers as described (*6*), except that the forward primer was modified to include the M13 –21 primer site (5′-TGTAAAACGACGGCCAGT-3′) at its 5′-end. The resulting PCR products were subjected to dye-primer sequencing (*33*), with products detected on an ABI377 or ABI373 fluorescence sequence detector. Possible sequence variations were detected by the ABI Sequence Navigator software package, which suggests potential heterozygotes by identifying nucleotide positions at which a secondary peak exceeds a selected threshold (50%). Such apparent variations were then visually inspected to compare the patterns seen among the several individuals.
12. D. N. Cooper and M. Karwczak, *Hum. Genet*. **85**, 55 (1990).
13. M. Chee *et al*., *Science* **274**, 610 (1996); M. J. Kozal *et al*., *Nature Med*. **2**, 753 (1996).
14. S. P. A. Fodor *et al*., *Science* **251**, 767 (1991); A.-C. Pease *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5022 (1994). The current generation of technology allows fabrication of 1.28 cm by 1.28 cm arrays of ~320,000 distinct oligonucleotides, each residing in a "feature" of ~20 μm by 25 μm and containing $>10^7$ copies of the probe.
15. J. G. Hacia *et al*., *Nature Genet*. **14**, 441 (1996).
16. STSs were amplified with their corresponding PCR primers as described (*6*). PCR products intended for hybridization to the same chip (typically 100 to 200 STSs from a single individual) were pooled together for subsequent processing. About 1 to 2 μg of the pooled PCR product was purified with Qiaquick purification kit (Qiagen), fragmented with deoxyribonuclease (DNase) I (Promega) and labeled with biotin with terminal deoxynucleotidyl transferase (TdT, GibcoBRL Life Technology). The purification was performed according to the manufacturer's instructions. The fragmentation was performed in a 40-μl reaction with 0.2 unit of DNase I, 10 mM tris-acetate (pH 7.5), 10 mM magnesium acetate, and 50 mM potassium acetate at 37°C for 15 min, after which the reaction was stopped by heat inactivation at 96°C for 15 min. The terminal transferase reaction was performed by adding 15 units of TdT and 12.5 μM biotin-N6-ddATP (DuPont NEN) to the preceding reaction mixture, incubating it at 37°C for 1 hour, and then heat-inactivating it at 96°C for 15 min. The labeled samples were hybridized to the chip as follows. Samples were denatured at ~96°C for 5 to 6 min and cooled on ice for 2 to 5 min. Chips were first hybridized with 6× SSPET [0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA (pH 7.4), 0.005% Triton X-100] for ~5 min and then hybridized with the denatured sample in hybridization buffer [3M tetramethylammonium chloride, 10 mM tris-HCl (pH 7.8), 1 mM EDTA, 0.01% Triton X-100, herring sperm DNA (100 μg/ml), and 200 pM control oligomer] at 44°C for 15 hours on a rotisserie at 40 rpm. Chips were washed three times with 1× SSPET, 10 times

with 6× SSPET at 22°C, and stained at room temperature with staining solution [streptavidin R-phycoerythrin (2 μg/ml) (Molecular Probes) and acetylated bovine serum albumin (0.5 mg/ml) in 6× SSPET] for 8 min. After they were stained, the chips were washed 10 times with 6× SSPET at 22°C on a fluidics workstation (Affymetrix). Hybridization to the chip was detected by using a confocal chip scanner (HP/Affymetrix) with a resolution of 40 to 80 pixels per feature and a 560-nm filter.

17. Candidate SNPs were identified by using a combination of four algorithms followed by a visual inspection. At each position, the VDA contains one “expected” probe (corresponding to the sequence from which the chip was designed) and three “variant” probes (containing a substitution in the central position). The first algorithm (base-calling) looked for positions at which, in some individuals, a variant probe gave a stronger signal than the expected probe. The second algorithm (clustering) considered the signal vector $\mathbf{s}_{ij}$ from the eight probes at position *i* (four base substitutions on both strands) in individual *j* and looked for positions *i* at which the vectors $\mathbf{s}_{ij}$ fell into multiple clusters. The third algorithm (mutant fraction) was similar but focused only on the expected probe and a single variant probe at a time (rather than all three variant probes). The fourth algorithm (footprint detection) looked for the loss of signal that occurs at the expected probes in the neighborhood of an SNP (*13, 15*). The algorithms have different sensitivities for detecting heterozygous and homozygous variations.
18. As discussed in the text below, the proportion *K* of polymorphic sites is expected to be proportional to $[1^{-1} + 2^{-1} + 3^{-1} + \ldots + (n - 1)^{-1}]$, where *n* is the number of genomes sampled. The proportion of polymorphic sites is thus expected to increase by 39.3% when the number of genomes is increased from 6 (in the gel-based survey) to 14 (in the chip-based survey). This agrees well with the observed increase of 38.8%.
19. A relatively small sample size suffices to capture much of the common variation. The sample size of 14 has a 50% chance of detecting an allele with a frequency of 5%. Doubling the proportion of variant sites identified would require increasing the number of genomes surveyed from 14 to 325, on the basis of the formula for *K*. The larger sample size will tend to identify polymorphisms with lower heterozygosity.
20. STSs were resequenced on both strands with dye-primer and dye-terminator chemistry.
21. The chip-based approach has the further advantage that long STSs can be analyzed, whereas gel-based sequencing is limited to about 600 bp. It is thus possible to use fewer PCR products to analyze a region. The current study did not take advantage of this feature because we used short STSs already available from our previous work (*6, 7*).
22. Confirmation was initially performed by multipass sequencing but is currently being done by using the clustering test on genotyping chips.
23. M. A. Walter *et al*., *Nature Genet*. **7**, 22 (1994).
24. The lowest density occurs on chromosome X, which has the lowest density of STSs and which was screened in fewer total genomes in as much as the screening panel included three males.
25. For each SNP, PCR primers were chosen with the PRIMER software package (*6*) to closely flank the polymorphic base and to have a predicted melting temperature of 57°C. Forward and reverse primers were synthesized with the T7 and T3 promoter sites (5′-TAATACGACTCACTATAGGGAGA-3′ and 5′-AATTAACCCTCACTAAAGGGAGA-3′) at their respective 5′-ends. Each PCR primer pair was individually tested to determine if it produced a single clear fragment visible by agarose gel electrophoresis and ethidium-bromide staining, as described (*6*). PCR assays passing this test were further classified as being strong or weak according to the yield of the fragment produced. Primer pairs were grouped into multiplex sets, with the sets chosen to consist of either strong assays or weak assays.
26. Multiplex PCR was performed by using multiple PCR primer pairs in a single reaction. Specifically, multiplex PCR reactions were performed in a 50-μl volume containing 100 ng of human genomic DNA, 0.1 to 0.2 μM of each primer, 1 unit of AmpliTaq Gold (Perkin-Elmer), 1 mM deoxynucleotide triphosphates (dNTPs), 10 mM tris-HCl (pH 8.3), 50 mM KCl, 5 mM $MgCl_2$, and 0.001% gelatin. Thermocycling was performed on a Tetrad (MJ Research), with initial denaturation at 96°C for 10 min followed by 30 cycles of denaturation at 96°C for 30 s, primer annealing at 55°C for 2 min, and primer extension at 65°C for 2 min. After 30 cycles, a final extension reaction was carried out at 65°C for 5 min. Because the resulting PCR products were small, it was unnecessary to fragment them (as was done for the STSs in the SNP screen). The PCR products were then labeled with biotin in a standard PCR reaction, by using T7 and T3 primers with biotin labels at their 5′-ends. The reaction was performed with 1 μl of template DNA, 0.1 to 0.2 μM labeled primer, 1 unit of AmpliTaq Gold (Perkin-Elmer), 100 μM dNTPs, 10 mM tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, and 0.001% gelatin. Thermocycling was performed with initial denaturation at 96°C for 10 min followed by 25 cycles of denaturation at 96°C for 30 s, primer annealing at 52°C for 1 min, and primer extension at 72°C for 1 min. After 25 cycles, a final extension reaction was carried out at 72°C for 5 min. The PCR products from the various multiplex reactions for an individual were then pooled together. One-tenth of the pooled sample was denatured and used for chip hybridization. Chips were hybridized, washed, stained and scanned, as above (*16*).
27. D. G. Wang, unpublished observations.
28. A classification procedure for assigning genotypes was derived for each locus on the basis of the hybridization results observed in a test population of 39 individuals. The proportions of the two alleles present in the *i*-th sample, denoted $\pi_{A,i}$ and $\pi_{B,i}$, (with $\pi_{A,i} + \pi_{B,i} = 1$) were estimated essentially by comparing the observed hybridization signal to the expected signals for the two VDAs. The values $\pi_{A,i}$ for the 39 individuals lie in the interval [0,1] and should ideally cluster near 0, 0.5, and 1.0, but other patterns might occur because of differences in hybridization intensity between the two alleles. The values were optimally clustered (*33*) with the MODECLUS procedure of the SAS software package (SAS Institute). A maximum of three nonoverlapping clusters was permitted, defined by points with a minimum separation of 0.12. A locus failed the cluster test if all the samples fell into a single cluster, if the samples gave rise to two clusters but neither corresponded to the heterozygous genotype (AB), or if too many samples (more than 9 of 39) fell outside the three optimal clusters. A locus passing the cluster test gave rise to either three clusters (genotypes AA, AB, BB) or two clusters (genotypes AA, AB or BB, AB).
29. Subsequent samples were genotyped according to the cluster in which the hybridization pattern fell, with no genotype being called for samples falling outside these predefined clusters.
30. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997); N. Takahata and Y. Satta, *Proc. Natl. Acad. Sci. U.S.A*. **94**, 4811 (1997); M. Nei and D. Graur, in *Evolutionary Biology*, M. K. Hecht, B. Wallace, G. T. Prance, Eds. (Plenum, New York, 1984), vol. 17, pp. 73–118.
31. F. J. Ayala, *Science* **270**, 1930 (1995).
32. R. C. Lewontin, in *Evolutionary Biology*, T. H. Dobzhansky, M. K. Hecht, W. C. Steere, Eds. (Appleton-Century-Crofts, New York, 1972), vol. 6, pp. 381–398.
33. M. M. DeAngelis, D. G. Wang, T. L. Hawkins, *Nucleic Acids Res*. **23**, 4742 (1995).
34. W. L. G. Koontz and K. Fukunaga, *IEEE Trans. Comp*. **C-21**, 171 (1972).

35. We thank D. Stern for construction of chip scanners used in the project, C. Chen-Cheng for computation work related to the polymorphisms among EST sequences in GenBank, T. Hawkins for sequencing of some STSs, and D. Lockhart for helpful comments on the manuscript. Supported in part by grants from Affymetrix, Millennium Pharmaceuticals and Bristol-Meyers-Squibb (to Whitehead Institute), from the National Human Genome Research Institute [to Whitehead Institute (HG00098) and Affymetrix (HG01323)] and from the National Institute of Standards and Technology [to Affymetrix (70NANB5H1031)].

5 February 1998; accepted 31 March 1998

RasGRP, a Ras Guanyl Nucleotide–Releasing Protein with Calcium- and Diacylglycerol-Binding Motifs

Julius O. Ebinu, Drell A. Bottorff, Edmond Y. W. Chan,
Stacey L. Stang, Robert J. Dunn, James C. Stone*

RasGRP, a guanyl nucleotide–releasing protein for the small guanosine triphosphatase Ras, was characterized. Besides the catalytic domain, RasGRP has an atypical pair of “EF hands” that bind calcium and a diacylglycerol (DAG)-binding domain. RasGRP activated Ras and caused transformation in fibroblasts. A DAG analog caused sustained activation of Ras-Erk signaling and changes in cell morphology. Signaling was associated with partitioning of RasGRP protein into the membrane fraction. Sustained ligand-induced signaling and membrane partitioning were absent when the DAG-binding domain was deleted. RasGRP is expressed in the nervous system, where it may couple changes in DAG and possibly calcium concentrations to Ras activation.

The cellular properties of neurons are modulated by a number of extrinsic signals, including synaptic activity, neurotrophic factors, and hormones. These signaling systems alter the intracellular concentrations of second messengers such as calcium and cyclic nucleotides, and these small molecules can regulate the activities of protein kinases (*1*). As the mechanisms linking Ras signaling to nerve function are not completely understood, we developed a cDNA cloning approach to identify proteins that enhance Ras signaling in the brain. From rat brain mRNA, we derived cDNAs that