Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences

Kris Irizarry[1], Vlad Kustanovich[2], Cheng Li[3], Nik Brown[5], Stanley Nelson[2,4], Wing Wong[3] & Christopher J. Lee[1]

Single-nucleotide polymorphisms (SNPs) have been explored as a high-resolution marker set for accelerating the mapping of disease genes[1–11]. Here we report 48,196 candidate SNPs detected by statistical analysis of human expressed sequence tags (ESTs), associated primarily with coding regions of genes. We used Bayesian inference to weigh evidence for true polymorphism versus sequencing error, misalignment or ambiguity, misclustering or chimaeric EST sequences, assessing data such as raw chromatogram height, sharpness, overlap and spacing, sequencing error rates, context-sensitivity and cDNA library origin. Three separate validations—comparison with 54 genes screened for SNPs independently, verification of *HLA-A* polymorphisms and restriction fragment length polymorphism (RFLP) testing—verified 70%, 89% and 71% of our predicted SNPs, respectively. Our method detects tenfold more true *HLA-A* SNPs than previous analyses of the EST data. We found SNPs in a large fraction of known disease genes, including some disease-causing mutations (for example, the HbS sickle-cell mutation). Our comprehensive analysis of human coding region polymorphism provides a public resource for mapping of disease genes (available at http://www.bioinformatics.ucla.edu/snp).

We analysed more than 542 million base pairs of EST and mRNA sequences from Unigene[12] (release 29 March 2000), producing 48,196 candidate SNPs with a lod score greater than 3 in favour of a polymorphism as opposed to sequencing error. To test our SNPs experimentally, we selected a subset of candidates that cause RFLPs in several score ranges (lod 2–6, 6–20, >20; Fig. 1). Of 79 SNP candidates tested so far in 8–24 DNA samples, 56 showed the expected pattern of polymorphism (71%). The verification rate was lower in the 2–6 lod score range (38%) compared with higher lod scores (6–20, 69%; >20, 79%).

To further validate our results, we examined genes in which independent experimental studies have systematically searched for polymorphisms. For these genes, we determined whether any of our SNPs are not independently found by these studies. We evaluated all SNPs (lod>3) within the protein-coding regions of 54 genes independently screened by the Whitehead Institute-Affymetrix coding single nucleotide polymorphisms (WIAF-CSNP) project for polymorphisms in 40 people[9]. We mapped each of our SNPs onto the gene sequence referenced by WIAF-CSNP, and required a perfect match for its location, major nucleotide and minor nucleotide compared with a SNP reported by WIAF-CSNP. For the 54 genes analysed, 70% of our SNPs matched polymorphisms found independently by WIAF-CSNP (Table 1). Of the SNPs that were not reported by WIAF-CSNP, review of the sequencing and alignment evidence indicates that several are good candidates for real polymorphisms, and may have been missed due to experimental sampling. Several others appear to result from misclustering of paralogous sequences in the Unigene database. At least six additional SNPs reported by WIAF-CSNP within these genes were also identified by our calculations with lod scores less than 3, indicating that there are many more real SNPs in our data below our scoring threshold. We did not count these SNPs in the validation results. Our calculations also identified a number of SNPs (lod>3) in non-coding regions that were confirmed by WIAF-CSNP, but again we did not count these.

To assess the rate of SNP false positives and false negatives in a larger population sample, we examined a gene (*HLA-A*) whose

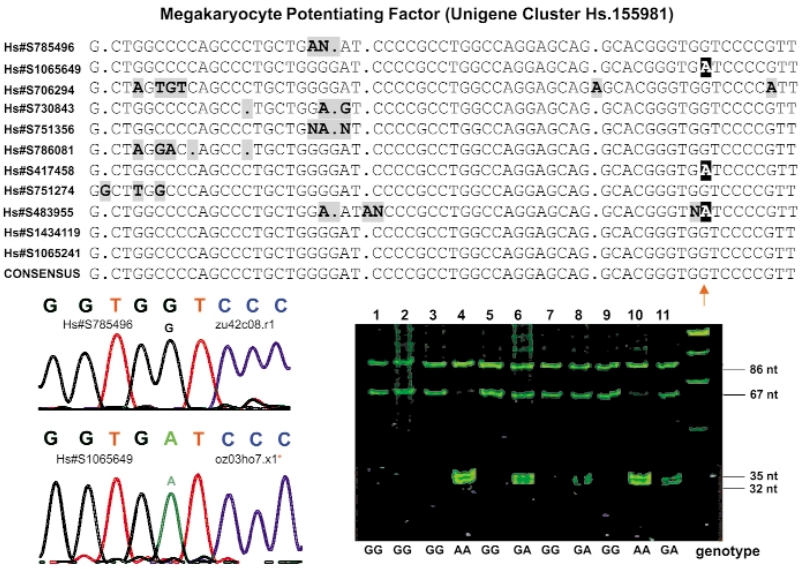

Fig. 1 Prediction and subsequent validation of a SNP in megakaryocyte potentiating factor. ***a***, Example of a predicted SNP (lod score 23.47), where some ESTs have adenine (black boxes), compared with most, which have guanine. Other sequence differences called by the program as sequencing errors are shaded grey. For clarity, only a subset of the aligned sequences are shown. ***b***, Examples of raw chromatographic trace evidence for the putative SNP. The PHRED quality scores for the A and G alleles are strong (51, 47). *Trace oz03ho7.x1 is shown reverse-complemented. ***c***, *Dpn*II RFLP analysis of a 153-bp genomic PCR product, containing 1 invariant and 1 polymorphic *Dpn*II site. Individuals homozygous for the G allele have 86- and 67-bp bands (lanes 1–3, 5, 7, 9). Individuals homozygous for the A allele have 3 bands of 86-, 35- and 32-bp (lanes 4, 10). Heterozygotes yield all four bands (lanes 6, 8, 11).

Departments of [1]Chemistry & Biochemistry, [2]Human Genetics, [3]Statistics and [4]Pediatrics, and [5]Graduate Program in Computer Science, University of California, Los Angeles, Los Angeles, California, USA. Correspondence should be directed to C.J.L. (e-mail: leec@mbi.ucla.edu).

polymorphism has been characterized very extensively. *HLA-A* has been studied thoroughly in both disease and normal populations because of its importance for transplant rejection, and over 140 alleles have been sequenced. Our screening identified 108 SNPs (lod>3) in the coding region of *HLA-A*, of which 96 match known *HLA-A* polymorphisms (Fig. 2), a verification rate of 89%. Because of the high rate of polymorphism in *HLA* genes, this rate for *HLA-A* should not be considered representative of all genes. By contrast, the CGAP SNP predictions[13], also based on the Unigene EST data, as of May 2000 predicted just 8 SNPs in *HLA-A*, missing greater than 90% of the true *HLA-A* polymorphisms verifiable in existing sequence databases.

Overall, our validation rates are similar to previous SNP detection results: 79% reproducibility of variant detection array (VDA)-detected SNPs by sequencing the same samples[10]; 55–60% reproducibility of VDA- and denaturing high-performance liquid chromatography (DHPLC)-detected SNPs (ref. 9); 63% verification of EST-based SNPs by sequencing and/or genetic bit analysis of an independent sample of 18 individuals[14]; 82% verification by RFLP of 90 individuals[13]; and 71% verification by RFLP of 24 individuals (this study). To improve accuracy, the sources of false positives need to be considered. It is difficult for sequence-based methods (like Unigene) to perfectly separate highly similar genes (>95% identity), giving erroneous 'polymorphisms'. Fortunately, the full human genome sequence should help to resolve these clusters into distinct genes. Genomic sequence can be used to separate paralogues that otherwise would yield false SNPs (ref. 15). Additionally, many dbEST sequences are from tumours, and thus some fraction of the apparent SNPs may actually be spontaneous mutations, which occur at higher rates in tumours[16].

To assess the usefulness of our SNP set for mapping of diseases and traits within the genome, we mapped our consensus gene-coding sequences and associated SNPs (lod>3) to a 1.4-Mb region of the genomic contig NT_001454 (chromosome 22). We mapped 42 genes to this region (Fig. 3), 23 of which match the locations for these genes reported by NCBI. Of our SNPs, 47 mapped to this region, within 16 of the genes. The genes in which no SNPs were predicted generally contained few ESTs (5 or less) representing only 1–3 human individuals. By contrast, only two microsatellite markers map within this 1.4-Mb interval on chromosome 22. Thus, this SNP set provides a 20-fold more dense marker set over this region, an immediate increase in map density, about one-third the level needed for single-gene resolution.

We have examined the frequency of our SNPs over a standard set of 30 positionally cloned genes[17] (Table A, see http://genetics.nature.com/supplementary_info/). Because positionally cloned genes were identified independent of any aspects of their sequence (for example, homology), they represent a pseudo-random sample of human genes. We identified candidate SNPs (lod>3) in a large fraction of these genes. Of the 30 sampled genes, 20 contain 1 or more SNPs. Moreover, these SNPs appear to be present at high frequencies suitable for mapping studies. Our calculation estimates the population allele frequency for each identified SNP, on the basis of the pattern of observations of the polymorphism across the set of catalogued libraries. Most of the candidate SNPs have an estimated heterozygote frequency of 0.1–0.5, reflecting the fact that EST-based SNP detection is biased towards the most common polymorphisms. This can be an advantage for detecting disease associations[2].

As an example, we analysed our predicted SNPs (lod>3) for *HBB* (encoding β-haemoglobin; Table B, see http://genetics.nature.com/supplementary_info/), whose polymorphism has been studied extensively in connection with human diseases, and whose protein structure and function are well understood. Our method detected six SNPs in the protein-coding region of *HBB*. Of particular interest, three SNPs are previously known mutations associated with human disease: sickle-cell anaemia (HbS, Glu6→Val; ref. 18), haemolytic anaemia (Hb Tacoma, Arg30→Ser; ref. 19) and erythrocytosis (Leu105→Phe). Of these three mutations, two were detected in shotgun EST sequences, and one from a GenBank mRNA. Thus, for genes that are represented by many ESTs in public databases, our method can actually detect disease-causing mutations in the human population. These disease-causing mutations are different from the other SNPs in their location and effect on the

Table 1 • Independent confirmation of detected SNPs

Gene	No. seq.	No. lib.	Coding region SNPs total	Coding region SNPs matches in WIAF-CSNP
AHC	21	7	0	
APOD	239	72	0	
AR	21	4	0	
AT3	60	12	2	2
BDNF	19	8	0	
CETP	20	9	1	1
CGA	129	18	0	
CNTF	5	1	0	
COMT	211	75	3	2
CYP11A	65	18	0	
CYP11B2	8	3	1	1
CYP17	47	15	1	0
DRD1	5	1	0	
DRD2	20	6	2	2
DRD5	5	2	1	0
F10	30	14	1	1
F13A1	153	45	2	2
F13B	3	1	0	
F2	37	14	1	0
F5	57	16	7	7
F9	8	2	1	1
FGA	413	29	0	
FGB	294	35	5	3
FGG	328	61	0	
FSHR	6	0	1	1
FSH	4	2	0	
GABRB1	5	1	0	
GAP43	85	29	0	
GH1	58	3	0	
GHR	33	19	0	
GNRH1	13	7	1	1
GP1BA	4	1	0	
GP9	9	5	0	
GRIN1	27	12	6	1
GRL	13	7	1	1
HCF2	75	13	1	1
HMGCR	160	50	0	
HSD3B1	58	10	1	0
HSD3B2	6	2	0	
HTR1A	4	0	0	
HTR1DB	0	0	0	
HTR1D	4	1	0	
HTR1EL	2	0	0	
HTR1E	4	0	0	
HTR2A	16	4	0	
HTR2C	9	3	0	
HTR5A	1	0	0	
HTR6	11	4	0	
IGF1	114	42	0	
ITGA2B	16	7	1	0
ITGB3	56	22	3	2
LCAT	479	137	0	
LDLR	257	94	0	
LIPC	38	7	3	3
total			46	32
percentage verified			70%	

We analysed 54 genes from the WIAF-CSNP project[9]. The number of candidate SNPs for each gene with lod score >3 that mapped to the protein-coding region is shown (total). The number which match SNPs independently detected by WIAF-CSNP (matches) and the numbers of Unigene sequences and libraries (human individuals) for each gene are also shown.

three-dimensional structure of haemoglobin. Two of the three disease mutations cause major amino acid changes (the only ones in our SNPs), and all three are located within the α-β-chain contact interface. By contrast, the other SNPs are either silent (no amino acid change) or conservative substitutions located on the protein surface distant from intermolecular contacts, where mutations can be accommodated with minimal disturbance. It may be possible to screen our large SNP database to identify a subset of SNPs more likely to produce strong effects on protein function.

Methods

SNP scoring. Trace chromatogram data of EST sequences in Unigene were obtained from Washington University (genome.wustl.edu), and processed with PHRED (ref. 20). We used ~241 million nucleotides of EST chromatogram data to train a sequencing error model conditioned on local sequence 5mer context and PHRED quality score. The conditional error rates were smoothed and interpolated in such a way that, for quality scores above 40, they converge to the rate predicted by the PHRED quality score[20,21]. As basic criteria for a predicted SNP, we require as evidence at least 1 chromatogram with PHRED quality greater than 30 (or an observation of the SNP in a finished, full-length mRNA sequence). We also limit our predictions to SNPs with estimated population allele frequency of at least 1% (95% rank confidence interval). To identify likely SNPs, single-base mismatches were reported from multiple sequence alignments produced by the programs PHRAP (P. Green, http://genome.washington.edu), Bayesian re-orienter (BRO) and partial order alignment (POA; C.J.L., manuscript in preparation) for each Unigene cluster. BRO corrects possible misreported EST orientations, whereas POA identifies and analyses non-linear alignment structures (for example, branches and loops) indicative of gene mixing/chimaeras that might produce spurious SNPs.

We calculate the likelihood-ratio of the observed sequences under a SNP model versus a pure sequencing error model:

single nucleotide polymorphisms in *HLA-A*

```
gatggccgtcatggcgccccgaaccctc[G/G/T]tcctgctactct[C/T]gggggc[C/T]ctggccc
tgacccagacctgggc[G/A]ggctc[C/T]cactccatgaggtatttc[T/A][T/A/C]cac[A/C]c
cgtgtcccggcccggc[C/A]g[C/T]gg[G/A]gagccccgcttcatcgc[C/A]gtgggctacgtgga
cgac[A/T]cgcagttcgtgc[G/A]gttcgacagcgacgccgcgagccaga[G/A]gatggagccgcgg
gcgccgtggatagagcaggag[G/A][G/A][G/T]cc[G/T]gagtattgggac[G/C][G/A]g[G/A]
a[G/C]aca[C/G]ggaa[T/A][G/C/A]tgaag[G/A]ccca[C/G][T/G]cacaga[C/T]t[G/A/C]
accgag[T/A/C]g[G/A]acctg[G/C]gga[A/T]cctgcgc[G/C]gctactacaaccagagcgagg[C/A]
cggttctcacacc[A/C/G]tccagat[A/G]atgt[A/T]tggctgcgacgtgggg[T/C]cggac[G/T]
ggcgc[T/C]tcctccgcgggtacc[A/G]caggacgc[C/T]tacgacggcaaggattacatcgcc[C/T]
tgaa[C/A]gaggacctgcgctcttgaccgcggcggacatggc[G/A]gctcaga[T/C]cacc[A/C]agc
[G/A]caagtgggag[G/A]cgg[C/T]cc[A/G]t[G/C][T/A/C]ggcggagcag[T/C][T/A/G]g
agag[C/T]ctacctgga[G/T]ggcacgtgcgtggagtggctccgcagatacctggagaacgggaaggaga
cgctgcagcgcacggac[G/C]cccccaa[G/A]ac[G/A]catatgac[T/C]caccac[G/C]c[T/C]
[G/A]gtctctgaccatga[G/A]gccac[C/T]ctgag[G/A]tgctgggccctg[A/G]gcttctaccct
gcggagatcacactgacctggcagcgggatggggaggaccagacccaggacacggagctcgtggagaccagg
cctgcaggggatgg[A/G]accttccagaagtggg[C/T][G/A][G/T/C]ctgtggtggt[G/A/C]cct
tctgga[C/G/A]aggagcagagatacacctgccatgtgcagcatgagggt[C/T]t[G/C]cccaagcccc
tcaccctgagatgggagc[C/T][G/A]tcttcccagcccacc[A/G]tccccatcgtgggcatcattgctg
gcct[G/A]gttctc[T/C]ttggagct[G/A]tg[A/T]tc[A/G]ctggagctgtggtcgctgctgt[G/A]
a[T/G]gtggaggaggaa[G/C]agctcagatagaaaaggagggag[C/T]tac[T/A]ctcaggctgcaag
ca[G/T]t[G/C]acagtgcccagggctctgat[G/A]tgtc[T/C]gtgtctctcacagcttgtaaagtgtga
```

Fig. 2 Comparison of candidate SNPs with catalogued *HLA-A* polymorphisms. SNPs with lod>3 are shown in the coding region sequence of *HLA-A*. Those matching known polymorphisms (96 total) from the *HLA-A* sequence database (http://www.anthonynolan.org.uk) are shaded grey, whereas false positives (mismatches, 12 total) are boxed.

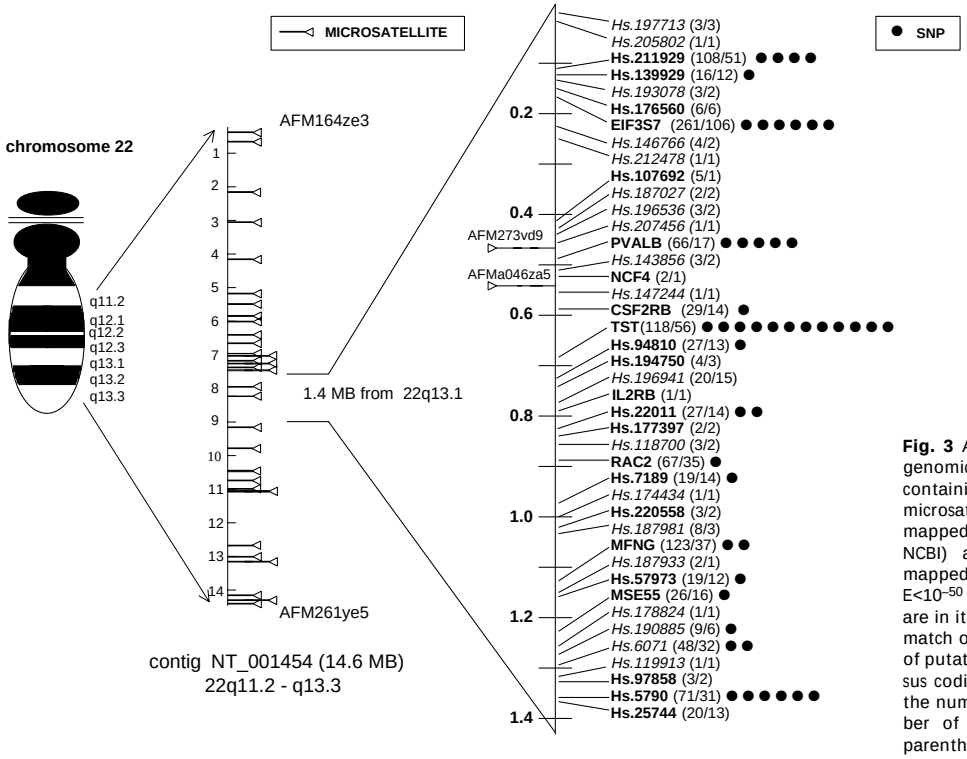

Fig. 3 A map of a 1.4-Mb region of genomic contig NT_001454 (22q13.1) containing 42 genes, 47 SNPs and 2 microsatellites. Genes previously mapped to this region (as reported by NCBI) are shown in bold; genes mapped to this region by BLAST with $E<10^{-50}$ and 97% identity or greater are in italics. We required a sequence match over a successively ordered set of putative exon(s) covering a consensus coding sequence. For each gene, the number of ESTs (n) and the number of libraries (m) are shown in parentheses (n/m).

$$SNP\ score = \frac{p(obs \mid SNP)}{p(obs \mid err)},$$

$$\text{where } p(obs \mid err) = \sum_T p(obs \mid T)p(T)$$

sums the probability of the observations over all possible 7mer 'true sequence' contexts T. This addresses possible ambiguity in the data about the true gene sequence. In the SNP model, by contrast, we only allow a single term representing the consensus sequence T^* and the 'SNP sequence' T' differing at the central nucleotide. Assuming a uniform prior $p(T) = constant$ over the possible gene sequences T, we define a log-odds score for ranking candidate SNPs:

$$p(obs \mid SNP) = \sum_T \sum_{T'} p(obs \mid T, T')p(T' \mid T)p(T) \geq p(obs \mid T^*, T')\left(\frac{1}{3}\right)p(T^*)$$

$$SNP\ score \geq \left(\frac{1}{3}\right)\frac{p(obs \mid T^*, T')}{\sum_T p(obs \mid T)}$$

$$lod = \log[p(obs \mid T^*, T')] - \log[\sum_T p(obs \mid T)]$$

This lod score does not consider the prior odds ratio against finding a SNP and thus does not give a true posterior odds ratio. To evaluate the sequencing error model, we treat each observed sequence i as an independent observation,

$$p(obs \mid T) = \prod_i p(obs_i \mid T) = \prod_i \sum_A p(obs_i, A \mid T)$$

where the summation is over a Hidden Markov Model (HMM) representing all possible alignments A of the observed sequence i to the true sequence T. The HMM sums the probability of all possible ways the observed sequence could have been emitted from the true sequence through a stochastic process of sequencing error.

The pattern of occurrence of the putative SNP over distinct library sources provides an important discriminant of genuine SNPs versus sequencing errors. Whereas sequencing errors should assort randomly over the set of libraries, a genuine SNP should follow a clustering pattern consistent with diploid genetics. Specifically, within a given cDNA library, the SNP should be observed with at least 50% frequency (from a heterozygous individual), and occasionally 100% (homozygotes).

We combined this with a Bayesian approach to estimate the population allele frequency q of a SNP, subject to the quantity and quality of sequence data available for detecting it. The probability of an individual sequence, conditional on the zygosity value for its library $z = 0, 1, 2$ (copy number of the polymorphism in the individual from whom the DNA was derived), is

$$p(obs, T^*, T', z) = \sum_A p(obs_i, A \mid T^*, T', q_z)$$

where $q_z = z/2$ is the expected frequency of the T′ SNP (for example, for a heterozygote library $q_z = 0.5$). For multiple observations from one library L, we sum over uncertainty about the true value of z in the library:

$$p(obs_{i\in L} \mid T^*, T') = \sum_{z=0,1,2} p(z) \prod_{i\in L} p(obs_i \mid T^*, T', z)$$

This allows us to introduce a dependence on the population allele frequency q assuming Hardy-Weinberg equilibrium

$$p(z \mid q) = \binom{2}{z} q^z (1-q)^{2-z}$$

We combine observations from multiple libraries:

$$p(obs \mid T^*, T', q) = \prod_L p(obs_{i\in L} \mid T^*, T', q)$$

Integrating over q,

$$p(obs \mid T^*, T') = \int_0^1 \prod_L p(obs_{i\in L} \mid T^*, T', q)\,p(q)dq$$

$$= \int_0^1 \prod_L \sum_{z=0,1,2} \binom{2}{z} q^z (1-q)^{2-z} \prod_{i\in L} \sum_A p(obs_i, A \mid T^*, T', z)p(q)dq$$

Here we used the uninformative prior $p(q) = 1$ for all q. The $p(q)$ distribution will analysed elsewhere. The posterior distribution for q is therefore

$$p(q \mid T^*, T', obs) = \frac{\prod_L p(obs_{i\in L} \mid T^*, T', q)}{\int_0^1 \prod_L p(obs_{i\in L} \mid T^*, T', q)\,dq}$$

The software is available on request.

RFLP testing. We designed unique primer sequences with Primer3 (http://www-genome.wi.mit.edu). All PCR fragments had a predicted *Dpn*II (GATC) RFLP which was scored from ethidium-bromide–stained polyacrylamide or agarose gels under standard conditions[22].

Accession numbers. dbSNP, 1508705–1556701.

Acknowledgements

We thank B. Modrek for assistance in mapping polymorphisms; P. Green for the PHRED, PHRAP and CONSED programs; K. Buetow for the CGAP SNP data; S. McGinnis for information about Unigene and E. Partsch for the sequence of pCMVSPORT. K.I. was supported by USPHS National Research Service Award GM08375. V.K. was supported by USPHS National Research Service Award GM07104. S.N. was supported by the Gwynn Hazen Cherry Memorial Laboratory. W.W. was supported by National Science Foundation grants NSF-DMS-9703918 and NSF-DBI-9904701. C.J.L. was supported by Department of Energy grant DEFG0387ER60615 and a grant from the Searle Scholars Program. Experimental SNP verification costs were supported in part by UC-Biostar grant S97106 to S.N.

Received 29 November 1999; accepted 10 July 2000.

1. Li, W. & Sadler, L.A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
2. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
3. Lai, E., Riley, J., Purvis, I. & Roses, A. A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics* **54**, 31–38 (1998).
4. 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).
5. Pennisi, E. A closer look at SNPs suggests difficulties. *Science* **281**, 1787–1789 (1998).
6. Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L. & Kwok, P.Y. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res.* **8**, 748–754 (1998).
7. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
8. Brookes, A.J. The essence of SNPs. *Gene* **234**, 177–186 (1999).
9. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
10. Halushka, M.K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–246 (1999).
11. Masood, E. Consortium plans free SNP map of human genome. *Nature* **398**, 545–546 (1999).
12. Schuler, G. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75**, 694–698 (1997).
13. Buetow, K.H., Edmonson, M.N. & Cassidy, A.B. Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genet.* **21**, 323–325 (1999).
14. Picoult-Newberg, L. *et al.* Mining SNPs from EST databases. *Genome Res.* **9**, 167–174 (1999).
15. Marth, G.T. *et al.* A general approach to single nucleotide polymorphism discovery. *Nature Genet.* **23**, 452–456 (1999).
16. Jackson, A.L. & Loeb, L.A. The mutation rate and cancer. *Genetics* **148**, 1483–1490 (1998).
17. Boguski, M.S., Tolstoshev, C.M. & Bassett, D.E.J. Gene discovery in dbEST. *Science* **265**, 1993–1994 (1994).
18. Ingram, V.M. Abnormal human haemoglobin. III. The chemical difference between normal and sickle cell haemoglobins. *Biochim. Biophys. Acta* **36**, 402–411 (1959).
19. Baur, E.W. & Motulsky, A.G. Hemoglobin tacoma—a β-chain variant associated with increased hb A2. *Humangenetik* **1**, 621–634 (1965).
20. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
21. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
22. Maeda, M. *et al.* A simple and rapid method for HLA-DQA1 genotyping by digestion of PCR-amplified DNA with allele specific restriction endonucleases. *Tissue Antigens* **34**, 290–298 (1989).