loci in each length bin. Unexpectedly large cliques (with a less than 50% chance of being unique by comparison with expectation) were withheld.

SNP identification

A base met the NQS if its PHRED quality score was ≥ 20, and the 5 bases to either side displayed PHRED scores ≥ 15. Positions within alignments were considered high quality if both bases met the NQS and at least nine of the ten flanking base pairs were perfect matches. These values have been evaluated by detailed parametric testing, which will be presented elsewhere (V.J.P., B.V.E., D.A., E.S.L., unpublished data). To validate the rules, three finished BAC sequencing projects from the Whitehead Sequencing Center were selected (Genbank accession nos AC003950.1, AC007066 and AC004584, AC007159). External validation shows that the error rates of such finished sequences are less than 1 in 10,000 (C. Nusbaum, personal communication). Individual reads contributing to each BAC assembly were aligned to the consensus as described[8,9], and apparent discrepancies counted among positions with a PHRED quality (Q) score > 20. Because our initial focus was in discovering SNPs, rather than insertion/deletions, we counted only substitutions. For this reason, the observed error rate is lower than that predicted by the PHRED score, which includes all classes of sequence errors (see refs 8, 9). SNP identification was fully automated using the rules above; no human revision was allowed. Similar results were observed for both dye-terminator and dye-primer chemistry, and with both slab-gel and capillary sequence detectors (data not shown).

SNP validation

Loci containing candidate SNPs were amplified by PCR from each of the DNA samples used to make the RRS libraries (10 DNAs for the pilot, 24 for the subsequent libraries and genomic validations), and sequenced according to standard methods. A repeat locus was declared if all individuals appeared heterozygous at one or more positions. Of unique loci, the candidate polymorphism was considered validated if two or three unambiguous, distinguishable genotypes were observed. For SNPs discovered by alignment to finished genomic sequence, we did not have access to DNA from the individual used to construct the BAC library, and instead used the same panel of 24 individuals. Given that some SNPs are rare (see Fig. 2e), we estimate that 5–10% of true SNPs would appear monomorphic in such a sample based on sampling variation alone.

Received 9 March; accepted 19 July 2000.

1. Lander, E. S. The new genomics: global views of biology. *Science* **274,** 536–539 (1996).
2. Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278,** 1580–1581 (1997).
3. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (1996).
4. Hastbacka, J. *et al.* Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2,** 204–211 (1992); *erratum ibid.* **2,** 343 (1992).
5. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22,** 139–144 (1999).
6. Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96,** 15173–15177 (1999).
7. Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7,** 401–409 (1997).
8. Ewing, B. & Green, P. Base-calling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
9. 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).
10. Gu, Z., Hillier, L. & Kwok, P. Y. Single nucleotide polymorphism hunting in cyberspace. *Hum. Mutat.* **12,** 221–225 (1998).
11. Marth, G. T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nature Genet.* **23,** 452–456 (1999).
12. 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).
13. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22,** 231–238 (1999).
14. Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077–1082 (1998).
15. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22,** 239–247 (1999).
16. Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. *Genetics* **129,** 513–523 (1991).
17. Mullikin, J. C. *et al.* An SNP map of human chromosome 22. *Nature* **407,** 516–523 (2000).
18. 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).
19. Landegren, U., Nilsson, M. & Kwok, P. Y. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. *Genome Res.* **8,** 769–776 (1998).
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
21. Cambien, F. *et al.* Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.* **65,** 183–191 (1999).
22. Li, W. -H. *Molecular Evolution* (Sinauer Associates, Canada, 1997).

Acknowledgements

We are indebted to the staff of the Whitehead Institute/MIT Center for Genome Research Sequencing Center for high-throughput sequencing and to N. Stange-Thomann for contributions to library construction. We would like to thank B. Blumenstiel and R. Lane for library construction and SNP validation, and M. Molla, L. Friedland, J. Ireland and B. Gilman for informatics assistance. We appreciate helpful discussions with members of The SNP Consortium, as well as colleagues at the Whitehead/MIT Genome Center. D.A. is a recipient of a Howard Hughes Medical Institute Postdoctoral Fellowship for Physicians. C.R.C. is supported by the Cancer Research Fund of the Damon Runyon / Walter Winchell Foundation. This work was conducted under grants from the Wellcome Trust and The SNP Consortium to E.S.L.

Correspondence and requests for materials should be addressed to E.S.L. (e-mail: lander@genome.wi.mit.edu).

An SNP map of human chromosome 22

J. C. Mullikin, S. E. Hunt, C. G. Cole, B. J. Mortimore, C. M. Rice, J. Burton, L. H. Matthews, R. Pavitt, R. W. Plumb, S. K. Sims, R. M. R. Ainscough, J. Attwood, J. M. Bailey, K. Barlow, R. M. M. Bruskiewich, P. N. Butcher, N. P. Carter, Y. Chen, C. M. Clee, P. C. Coggill, J. Davies, R. M. Davies, E. Dawson, M. D. Francis, A. A. Joy, R. G. Lamble, C. F. Langford, J. Macarthy, V. Mall, A. Moreland, E. K. Overton-Larty, M. T. Ross, L. C. Smith, C. A. Steward, J. E. Sulston, E. J. Tinsley, K. J. Turney, D. L. Willey, G. D. Wilson, A. A. McMurray, I. Dunham, J. Rogers & D. R. Bentley

The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

The human genome sequence will provide a reference for measuring DNA sequence variation in human populations. Sequence variants are responsible for the genetic component of individuality, including complex characteristics such as disease susceptibility and drug response. Most sequence variants are single nucleotide polymorphisms (SNPs), where two alternate bases occur at one position[1–3]. Comparison of any two genomes reveals around 1 SNP per kilobase[1,3]. A sufficiently dense map of SNPs would allow the detection of sequence variants responsible for particular characteristics on the basis that they are associated with a specific SNP allele[4–6]. Here we have evaluated large-scale sequencing approaches to obtaining SNPs, and have constructed a map of 2,730 SNPs on human chromosome 22. Most of the SNPs are within 25 kilobases of a transcribed exon, and are valuable for association studies. We have scaled up the process, detecting over 65,000 SNPs in the genome as part of The SNP Consortium programme, which is on target to build a map of 1 SNP every 5 kilobases that is integrated with the human genome sequence and that is freely available in the public domain.

Single nucleotide polymorphisms (SNPs) are stable, bi-allelic sequence variants that are distributed throughout the genome, which can be assayed using high-throughput automated methods. Sequence variants have been detected previously by analysis of sequence differences in clusters of expressed sequence tags[7,8]; or by re-sequencing DNA fragments after amplification from different individuals[9–11], sometimes following prescreening[12–14]. These approaches are effective for exploring sequence variation of individual genes in depth. An alternative approach, which takes advantage of the reagents from the human genome project, is to detect SNPs in regions of overlap between bacterial clones containing sequences of independent genomes (ref. 2; and E.D. *et al.*, unpublished data). This analysis provides a valuable resource of SNPs in short sections throughout the genome, but each section is interspersed by large regions (typically 0.1–0.5 megabases (Mb)) where the sequence of only one genome is available, and where no polymorphisms can be detected.

We evaluated two large-scale sequencing strategies to identify SNPs to cover the entire human genome at a target density of 1 SNP per 5 kilobases (kb). In the reduced representation shotgun (RRS)

strategy[15], specific subsets of restriction fragments, made from an equimolar mixture of DNA isolated from unrelated individuals, are repeatedly sampled by sequencing. Reads derived from the same fragment in different genotypes are aligned into clusters, or cliques, and high-confidence sequence differences between any two reads (that is, candidate SNPs) are recorded. Experimental verification of a subset of these candidate SNPs (by re-sequencing to identify the SNP in individual samples of the original DNA panel) tests the criteria used in the computational detection of candidate SNPs. In the genomic-alignment strategy, single reads obtained by shotgun sequencing of a library of DNA fragments are aligned directly to available genomic sequence to detect the candidate SNPs. The two strategies are complementary: the RRS strategy allows detection of SNPs throughout the genome without genomic sequence, but the SNPs are not automatically mapped; whereas the genomic-alignment strategy requires genomic sequence, and provides a map location for each SNP. Both strategies are applicable to any genome.

The availability of the complete sequence of chromosome 22 (ref. 16) enabled us to compare both approaches directly. The efficiency of SNP detection for each approach was measured as the number of SNPs detected relative to the amount of new raw sequence data generated for the analysis (see Table 1). For RRS, chromosome 22 was flow-sorted from seven unrelated individuals (see Methods), and clones from a library of 1.2–2.1-kb fragments generated by *Hin*dIII digestion were sequenced. After removing highly repetitive or poor quality sequences, 4,584 reads were aligned into 1,842 clusters (1,013 with more than one read, plus 829 singletons), corresponding to a mean cluster depth (number of reads in clusters / number of clusters) of 2.49. Using a specific set of criteria (see Methods), we detected 455 candidate SNPs in the 1,013 clusters of two or more reads. Experimental verification of a subset of these candidates by re-sequencing confirmed that 95% (74/78) were true SNPs. The remaining four candidate SNPs were homozygous in all DNA samples tested, and are presumed to be sequencing errors. The efficiency of SNP detection was 1 SNP per 10.1 reads (1 SNP per 4.79 kb of raw data; see Table 1). In a separate analysis we detected an additional 13% variants comprising insertion/deletion polymorphisms. Two-thirds of these were variations in poly(A) tracts, and the remainder (4% of all variants) are potentially a valuable additional source of polymorphisms for genetic studies.

In contrast to the RRS strategy, in which the library must be sequenced to a sufficient depth to obtain clusters of multiple reads for SNP detection, genomic-alignment analysis minimally requires alignment of just one read against finished genomic sequence. The genomic-alignment strategy should therefore result in a higher efficiency of SNP detection. Like RRS, the genomic-alignment strategy should also yield high-confidence SNPs, as all human genomic sequence is finished to an accuracy of more than 99.99%, (and therefore has a quality value Q of 40; ref. 17). To test this hypothesis, we aligned the RRS reads obtained from the previous experiment (including all singletons) to the finished chromosome 22 sequence (33.4 Mb). We identified 914 candidate SNPs, and included all the SNPs found by the RRS analysis. The success rate of verification was 94% (115/122). The strategy therefore results in a twofold improvement in the efficiency of SNP detection (1 SNP per 5.0 reads compared with 1 SNP per 10.1 reads; see Table 1).

Table 1 Comparison of strategies for SNP identification

	Chromosome 22			Whole genome
	*Hin*dIII RRS	RRS-GA	Random GA	*Pvu*II RRS
SNPs detected	455	914	1,845	52,354
Reads	4,584	4,584	5,567	473,249
Clusters	1,842	1,842	5,401	188,545
Cluster depth	2.49	3.49	2.03	2.51
SNPs per kb raw data	1/ 4.79	1/2.38	1/1.39	1/3.98
SNPs per read	1/10.1	1/5.0	1/3.0	1/9.0

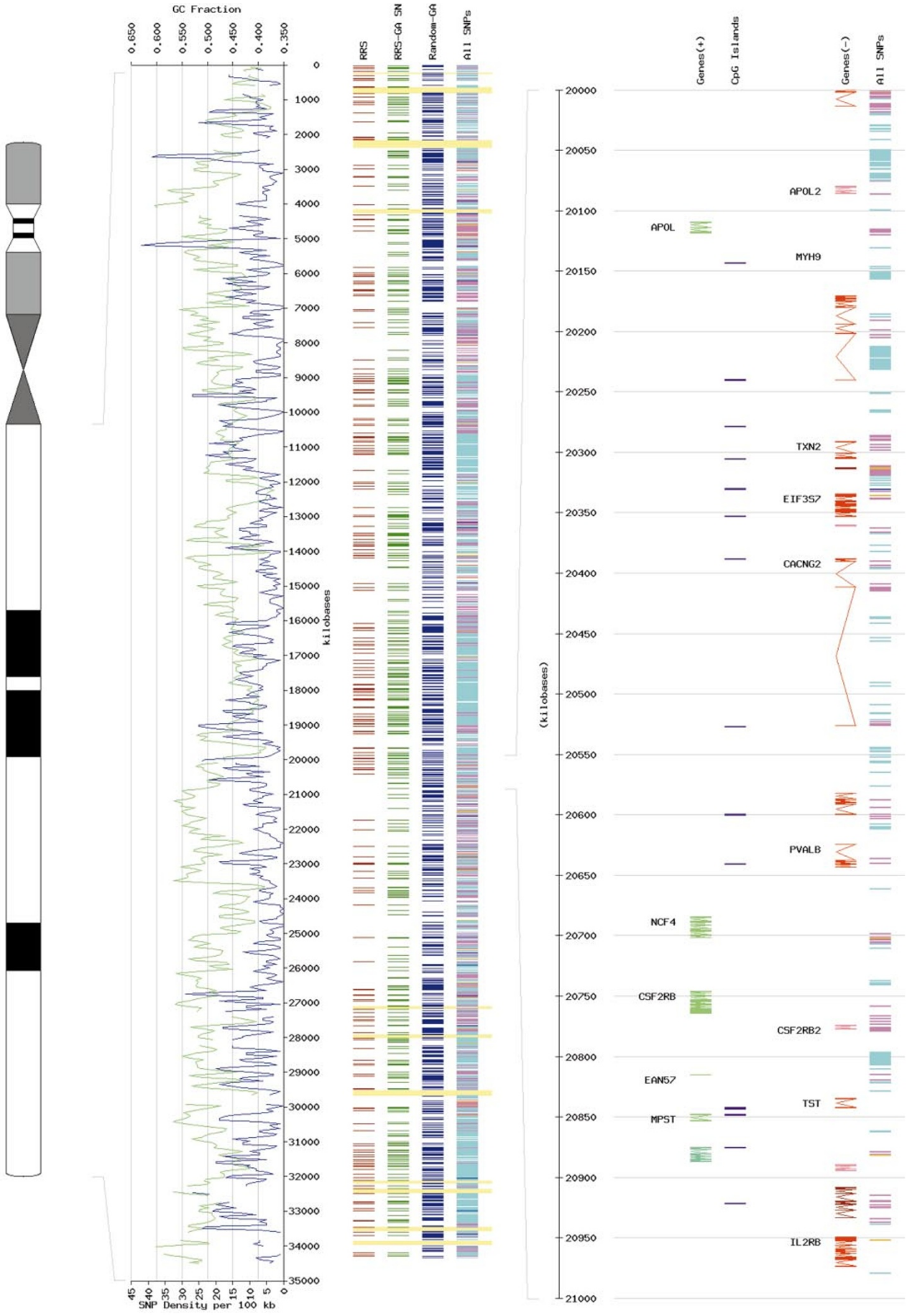
Summary of SNP detection by different strategies. Clusters, the number of unique genomic loci represented in the clusters obtained (including those with single reads); cluster depth, (number of reads in clusters)/(number of clusters); genomic sequence is counted as a single read per cluster where applicable (columns 2 and 3); SNPs per read, SNPs detected per read analysed; SNPs per kb raw data, an exact measure of relative efficiency, in SNPs per kilobase (at *Q*23) of raw data at *Q*23 used in the analysis. The total amount of *Q*23 raw sequence data used in each analysis was 2.18 Mb, 2.18 Mb, 2.57 Mb and 208.2 Mb, respectively. GA, genomic alignment.

The genomic-alignment approach should be most efficient if each new sequence read aligns to a separate section of the genome. We therefore constructed a library of randomly sheared fragments from flow-sorted chromosome 22 DNA, and aligned 5,567 high-quality reads to the finished genomic sequence. We detected 1,845 candidate SNPs, and the verification success rate was 97% (33/34). The SNP detection rate in this analysis is 1 SNP per 1,391 bases of raw sequence data of $Q \geq 23$ (or 1 SNP per 3.0 reads). Genomic-alignment analysis of the random shotgun sequences (compared with genomic-alignment analysis of the RRS sequences) thus provided a further 1.6-fold improvement in the efficiency of SNP detection.

A total of 2,730 different SNPs (that is, 1 per 12 kb of chromosome 22 sequence) were identified during these studies. The position of each SNP was determined by aligning its flanking sequence to the sequence of chromosome 22. The plots in Fig. 1 illustrate the distribution of SNPs detected by each method. Chromosome 22q contains at least 545 transcribed genes (3,632 exons) on the basis of the reported annotation. 1,043 of the SNPs (38%; magenta bars in Fig. 1) detected in this study lie either inside or within 5 kb of a transcribed exon in the current annotated set. In all, 1,771 (65 %) of the SNPs are within 25 kb of an exon, and may be informative in association studies, depending on the extent of linkage disequilibrium in genomic regions[5,18,19]. From the current set of annotated transcribed exons, 37% (1,333) of them have at least one SNP within 5 kb, and 84% (3,039) have at least one SNP from the present set within 25 kb. This study therefore already provides SNPs that are sufficiently close to most of the transcribed regions to be in linkage disequilibrium with possible functional variants, and this coverage will improve when a more dense SNP map is produced for the whole genome.

On the basis of the results of this study, and studies at the Whitehead Institute and the Genome Sequencing Center, St Louis, The SNP Consortium (TSC: a consortium of academic and pharmaceutical companies, see http://snp.cshl.org) initiated a programme with the goal of generating a freely available public resource of 300,000 SNPs, of which at least 150,000 would be mapped by April 2001. As part of this work, we have extended our SNP identification programme by application of the RRS strategy to a whole-genome library of *Pvu*II fragments of 0.925–1.250 kb, constructed using DNA isolated from 24 unrelated individuals (the same panel used by all TSC participants; see Methods). Up to April 2000, sequencing to a cluster depth of 2.51 has yielded 52,354 candidate SNPs. This corresponds to a detection rate of 1 SNP per 9.0 reads. The verification success rate for these SNPs was 97% (122/128). Candidate SNPs in three additional loci were heterozygous in all DNA samples tested, and these loci are presumed to be low-copy repeats. The success rate of SNP verification concurs with the results of the pilot study, and confirms the feasibility of the RRS approach to detect SNPs in the whole genome. Furthermore, alignment of the *Pvu*II RRS data with the 500 Mb finished genomic sequence that is available resulted in identification of a total 19,192 SNPs (including 9,727 not detected by RRS). On the basis that roughly 2,700 Mb of the genome sequence becomes available, the number of SNPs found by genomic-alignment analysis of this set of RRS reads would be twofold higher than by RRS analysis alone, in agreement with the results of the chromosome 22 study.

Production of the working draft sequence (comprising mapped bacterial clones each sequenced to a depth of ≥ threefold in bases of $Q \geq 20$; ref. 20) of the human genome is well advanced (http://www.nhgri. nih.gov/HGP). Genomic-alignment analysis of RRS or



random shotgun sequences using the draft would provide a high yield of candidate SNPs. The sequence reads are assembled using the program PHRAP (P. Green, personal communication), which provides a combined quality score for each base of the consensus sequence. This PHRAP quality score can be used for identification of candidate SNPs by genomic-alignment analysis using the working draft, in the same way as using finished sequence, by requiring a minimum PHRAP quality score (Q) of 40. After aligning the sequence data from the *Pvu*II RRS library against a sample of unfinished sequence with variable quality scores, a set of candidate SNPs was selected using a PHRAP Q value of ≥ 40. The verification success rate was 99% (244/246). This confirmed the validity of using genomic-alignment analysis with the working draft sequence of the genome to identify (and map) candidate SNPs.

The initial goal of the TSC was to identify 300,000 SNPs (1 per 10 kb) by April 2001. Our work shows that the goal can be achieved using the RRS strategy and the resources currently available in the consortium. To our benefit, the acceleration in the human genome sequencing programme over the past 12 months will provide the opportunity to perform genomic-alignment analysis on all the sequence data, thus providing a substantial improvement in the yield of SNPs (at least twofold on current projections, or 600,000 SNPs) from the TSC programme. Furthermore, the majority of SNPs will be mapped directly by alignment to the genomic sequence, thus providing an SNP map (1 SNP per 5 kb on average) of the human genome. This freely available, public resource will underpin extensive genetic studies to establish the extent of linkage disequilibrium in the human genome, and to investigate the association of genes with complex disease and other phenotypic traits. □

Methods

Sequencing and alignment

Chromosome 22 was flow-sorted from individual lymphoblastoid cell lines derived from seven individuals of north European origin, selected from the Porton Down collection of unrelated individuals collected as controls for disease association studies (Panel HRC, nos 159, 146, 184, 163, 160, 226, 193, 575 and 148). Chromosome preparations were pooled and digested with *Hin*dIII, and a gel-purified size fraction representing 1.2–2.1 kb was subcloned into the plasmid vector pUC18. We picked clones from a library of more than 100,000 transformants and sequenced them using forward and reverse primers as described[21]. Data were collected on ABI377 or 3700 machines (PE Biosystems), and analysed using the base-calling algorithm PHRED[22,23]. All sequence reads that contained more than 100 bases with a PHRED quality score (Q) ≥ 30 were selected for masking of repeats using RepeatMasker (version 21/03/99: A. Smit and P. Green, http://ftp.genome.washington.edu/RM/RepeatMasker.html). Reads with more than 80 bases of unique sequence were then assembled using a method based on Cross_Match followed by multiple sequence alignment of the reads matching unique genomic loci on chromosome 22.

The *Pvu*II RRS libraries were constructed using genomic DNA from 24 individuals containing one or more representatives of a range of ethnic groups (the DNA Polymorphism Discovery Resource M24PDR, Coriell Cell Repositories[24]), by agarose gel fractionation of a complete digest of a mixture of the DNAs, followed by subcloning of specific size fractions in pUC18. For assembly of forward and reverse sequence reads, PHRAP (P. Green, University of Washington) was used to assemble the RRS *Pvu*II library data (see Table 1).

Candidate SNPs

Criteria for selection of candidate SNPs were as follows (see also ref. 15): (1) the quality value (Q) of the SNP base (a value recorded for each base in a sequence read by the base-calling program PHRED[22,23]) was ≥ 23; and the Q value for the 5 bases on either side of the SNP was ≥ 15. (2) At least nine of the flanking ten bases matched between reads. All ten had to match for the Cross_Match (P. Green, University of Washington) based method. (3) The cluster depth was no greater than eight reads, on the basis that deeper clusters might comprise a low-copy repeat. (4) The number of candidate SNPs in a cluster was ≤ 4, on the basis that clusters with more divergent sequences might be composed of low-copy repeats (that is, recently diverged paralogous sequences, accumulating sequence differences between them). The minimum Q for the SNP base was chosen as follows: visual examination of the trace data for a randomly selected set of candidate SNPs called at $Q \geq 20$ revealed 15 out of 111 (13%) obvious sequence artefacts, 9 of which had Q values in the range from 20 to 22. When the SNP identification was rerun at $Q \geq 23$, a new randomly selected set of candidate SNPs was examined, from which only 5 out of 117 (4.3%) failures were observed. This failure rate was confirmed by the experimental verification data.

Classification of single-base substitutions

We classified single-base substitutions on the basis of transitions or transversions as follows. C to T (or G to A) transitions: 70.1% of all SNPs were possible CpG to TpG mutations (the most frequently observed single-base substitution, which is presumed to arise following deamination of 5-methylcytosine in Me-CpG; ref. 25); 29.1% of all SNPs were transversions of C/A, C/G and T/A (15.7%, 8.6% and 5.6% of all SNPs, respectively).

SNP verification

Candidate SNPs from each dataset were selected at random for experimental verification. Primer pairs were designed using PRIMER (http://www.genome.wi.mit.edu/genome_software/genome_software_index.html) and loci were amplified from the DNA samples of the individual cell lines used for the initial library construction. Polymerase chain reaction products were purified by treatment with shrimp alkaline phosphatase and exonuclease 1, sequenced from both ends, and the data for each SNP assembled and examined in a GAP4 database.

Data access

All data including sequence reads, candidate SNPs and verification information were submitted to the Data Coordination Centre (DCC) of The SNP Consortium (http://snp.cshl.org). Each SNP is assigned a unique identifier (for example, TSC0137673) and released in the public domain when given a map position either by alignment to mapped genomic sequence, or by whole-genome radiation hybrid mapping. SNP information is available at the above web site (from the home page, click on 'data', then 'object search', and enter 'TSC0137673', and search); and also in dbSNP (http://www.ncbi.nlm.nih.gov/SNP).

Received 10 March; accepted 19 July 2000.

1. Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077–1082 (1998).
2. 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).
3. Kwok, P. Y., Deng, Q., Zakeri, H., Taylor, S. L. & Nickerson, D. A. Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs. *Genomics* **31,** 123–126 (1996).
4. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (1996).
5. Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278,** 1580–1581 (1997).
6. Chakravarti, A. Population genetics—making sense out of sequence. *Nature Genet.* **21,** 56–60 (1999).
7. 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).
8. Picoult-Newberg, L. *et al.* Mining SNPs from EST databases. *Genome Res.* **9,** 167–174 (1999).
9. 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).
10. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22,** 231–238 (1999).
11. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22,** 239–247 (1999).
12. Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K. & Sekiya, T. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc. Natl Acad. Sci. USA* **86,** 2766–2770 (1989).
13. Ganguly, A., Rock, M. J. & Prockop, D. J. Conformation-sensitive gel electrophoresis for rapid detection of single-base differences in double-stranded PCR products and DNA fragments: evidence for solvent-induced bends in DNA heteroduplexes. *Proc. Natl Acad. Sci. USA* **90,** 10325–10329 (1993).
14. O'Donovan, M. C. *et al.* Blind analysis of denaturing high-performance liquid chromatography as a tool for mutation detection. *Genomics* **52,** 44–49 (1998).
15. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407,** 513–516 (2000).
16. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
17. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9,** 1–4 (1999).
18. Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96,** 15173–15177 (1999).
19. Kruglyak, L. Prospects for whole–genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22,** 139–144 (1999).
20. Bouck, J., Miller, W., Gorrell, J. H., Muzny, D. & Gibbs, R. A. Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Res.* **8,** 1074–1084 (1998).
21. The Sanger Centre & The Washington University Genome Sequencing Centre. Toward a complete human genome sequence. *Genome Res.* **8,** 1097–1108 (1998).

◀ **Figure 1** Single nucleotide polymorphism map of human chromosome 22. The four columns of coloured bars represent the SNP map determined by each analysis. See Table 1 for SNP totals corresponding to the three columns 'RRS', 'RRS-GA' and 'Random GA'; the column 'All SNPs' contains the non-redundant set of 2,730 SNPs. GA, genomic alignment. Gold and magenta bars denote SNPs that are located inside or within 5 kb of an exon, respectively. One region of the chromosome (20,000–21,000 kb; coordinates as in ref. 16; see also http://www.sanger.ac.uk/HGP/Chr22) is enlarged to show the positions of individual SNPs relative to annotated exons and putative CpG islands.

22. 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).
23. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
24. 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).
25. Holliday, R. & Grigg, G. W. DNA methylation and mutation. *Mutat. Res.* **285,** 61–67 (1993).

Acknowledgements

We are grateful for the contribution of all the members of the sequencing and supporting teams, and members of the informatics group of the Sanger Centre (see http://www.sanger.ac.uk). We thank the Wellcome Trust and The SNP Consortium for financial support, and members of the participating centres for helpful discussions.

Correspondence and requests for materials should be addressed to D.R.B. (e-mail: drb@sanger.ac.uk).

A vertebrate globin expressed in the brain

Thorsten Burmester*†, Bettina Weich‡, Sigrid Reinhardt§ & Thomas Hankeln†‡

** Institutes of Zoology, ‡ Molecular Genetics, Biosafety Research and Consulting, and § Physiological Chemistry, Johannes Gutenberg University Mainz, D-55099 Mainz, Germany*
† These authors contributed equally to this work

Haemoglobins and myoglobins constitute related protein families that function in oxygen transport and storage in humans and other vertebrates[1,2]. Here we report the identification of a third globin type in man and mouse. This protein is predominantly expressed in the brain, and therefore we have called it neuroglobin. Mouse neuroglobin is a monomer with a high oxygen affinity (half saturation pressure, $P_{50} \approx 2$ torr). Analogous to myoglobin, neuroglobin may increase the availability of oxygen to brain tissue. The human neuroglobin gene (*NGB*), located on chromosome 14q24, has a unique exon–intron structure. Neuroglobin represents a distinct protein family that diverged early in metazoan evolution, probably before the Protostomia/Deuterostomia split.

Globins are porphyrin-containing proteins that bind oxygen reversibly and are therefore important in the respiratory system of living species[1]. They have been found in many taxa, including bacteria, plants, fungi and animals[3]. Two types of globins have been described in vertebrates. The heterotetrameric haemoglobins transport oxygen in the blood, whereas the monomeric myoglobin of muscle cells facilitates the diffusion of oxygen to the mitochondria[4]. Although globins are among the best-investigated vertebrate proteins and several functional variants of the haemoglobin subunits are known[1], no other distinct types of globins have been identified so far in this taxon.

In the databases of anonymous mouse and human complementary DNAs (expressed sequence tags; ESTs[5]), we found partial globin-like sequences that do not correspond to any known haemoglobin or myoglobin. We cloned and sequenced the coding regions of the human and mouse cDNAs and the genomic region of the human gene. The mouse and human gene each code for proteins of 151 amino acids (relative molecular mass 17,000; M_r 17K) that are 94% identical; this is higher than the conservation between the orthologous haemoglobins or myoglobins of these species (77–85% identity) and within the uppermost range of more than 1,100 proteins compared between man and mouse[6]. Although the proteins clearly belong to the globin superfamily, they share little amino-acid sequence similarity with vertebrate myoglobins (< 21% identity) and haemoglobins (< 25% identity), suggesting a distinct evolution and function (Fig. 1).

We analysed the expression pattern of the human gene by northern hybridization to a filter containing RNA from different tissues and developmental stages (Table 1; see Supplementary Information). We note a predominant expression in the brain with the strongest signals observed in the frontal lobe, the subthalamic nucleus and the thalamus. Globin expression was also detected by messenger RNA *in situ* hybridization in neuronal cells of mouse brain regions (Fig. 2). We therefore propose to designate this protein neuroglobin (NGB). Polymerase chain reaction with reverse transcription (RT–PCR) experiments using murine RNA confirmed that other tissues contain only minor amounts of *Ngb*

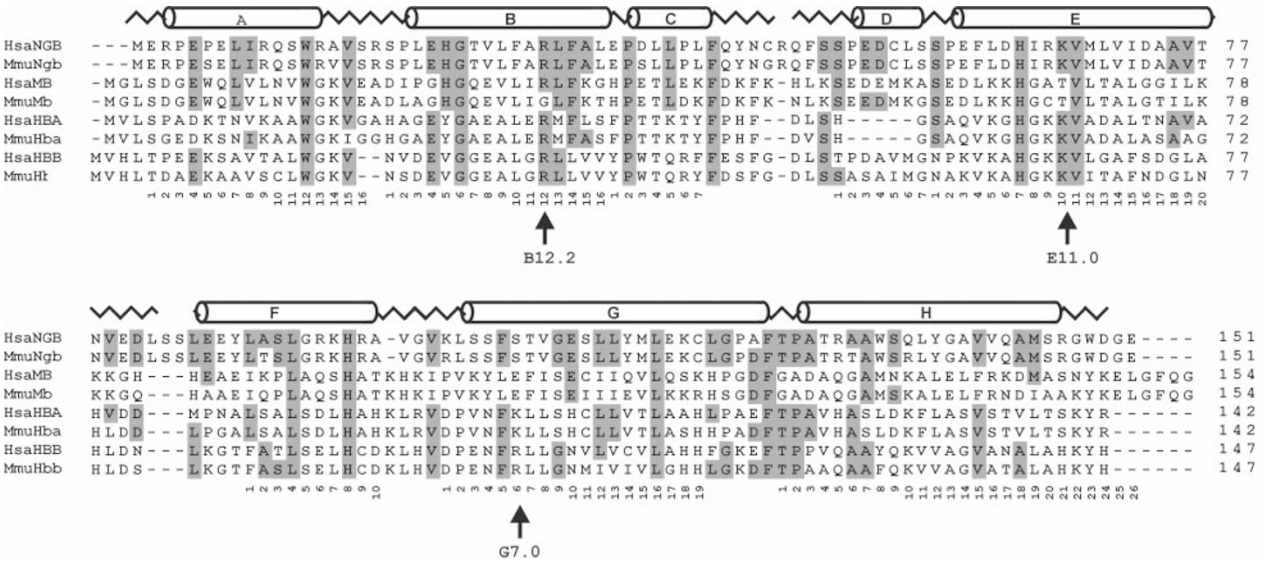


Figure 1 Comparison of human and mouse neuroglobin (HsaNGB and MmuNgb) with myglobins (HsaMB, accession number M14603; MmuMb, P04247) and haemoglobins α and β (HsaHBA, J00153; HsaHBB, M36640; MmuHba, A45964; MmuHbb, P02088). The globin consensus numbering is given below the sequences, the secondary structure of the human haemoglobin β is superimposed in the upper row. α-Helices are designated A to H, amino acids conserved between the neuroglobins and the myoglobins or haemoglobins are shaded. Intron positions in the human NGB sequence (at B12-2, E11-0 and G7-0) are indicated by arrows.