

22. Barrett, A. J., Rawlings, N. D. & Woessner, J. F. *Handbook of proteolytic enzymes* (Academic, San Diego, CA, 1999).
23. Bergerat, A. *et al.* An atypical topoisomerase II from archaea with implications for meiotic recombination. *Nature* **386**, 414–417 (1997).
24. Stein, D. B. & Searcy, D. G. Physiologically important stabilization of DNA by a prokaryotic histone-like protein. *Science* **202**, 219–221 (1978).
25. Hixon, W. G. & Searcy, D. G. Cytoskeleton in the archaeobacterium *Thermoplasma acidophilum*? Viscosity increase in soluble extracts. *BioSystems* **29**, 151–160 (1993).
26. Smith, P. F., Langworthy, T. A. & Smith, M. R. Polypeptide nature of growth requirement in yeast extract for *Thermoplasma acidophilum*. *J. Bacteriol.* **124**, 884–892 (1975).
27. Frishman, D., Mironov, A., Mewes, H. W. & Gelfand, M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.* **26**, 2941–2947 (1998).
28. Frishman, D. & Mewes, H. W. PEDANTIC genome analysis. *Trends Genet.* **13**, 415–416 (1997).
29. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
30. Koretke, K. K., Russell, R. B., Copley, R. R. & Lupas, A. N. Fold recognition using sequence and secondary structure information. *Proteins Struct. Funct. Genet.* **37**, 141–148 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank P. Forterre and P. Lopez for helping to define the origin of replication; M. Boicu and C. Czoppelt for sequencing; G. Mannhaupt for annotating part of the ORFs; I. Echabre for preparing template DNA and sequencing; and B. Marshall for developing software for gene cluster analysis and for data management.

Correspondence and requests for materials should be addressed to W.B. (e-mail: baumeist@biochem.mpg.de). The *Thermoplasma acidophilum* genome sequence has been deposited in the EMBL database (accession number AL139299).

An SNP map of the human genome generated by reduced representation shotgun sequencing

David Altshuler*†, Victor J. Pollara*, Chris R. Cowles*, William J. Van Etten*, Jennifer Baldwin*, Lauren Linton* & Eric S. Lander*‡

* Whitehead Institute/MIT Center for Genome Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA

† Diabetes Unit, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA

‡ Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Most genomic variation is attributable to single nucleotide polymorphisms (SNPs), which therefore offer the highest resolution for tracking disease genes and population history^{1–3}. It has been proposed that a dense map of 30,000–500,000 SNPs can be used to scan the human genome for haplotypes associated with common diseases^{4–6}. Here we describe a simple but powerful method, called reduced representation shotgun (RRS) sequencing, for creating SNP maps. RRS re-samples specific subsets of the genome from several individuals, and compares the resulting sequences using a highly accurate SNP detection algorithm. The method can be extended by alignment to available genome sequence, increasing the yield of SNPs and providing map positions. These methods are being used by The SNP Consortium, an international collaboration of academic centres, pharmaceutical companies and a private foundation, to discover and release at least 300,000 human SNPs. We have discovered 47,172 human SNPs by RRS, and in total the Consortium has identified 148,459 SNPs. More broadly, RRS facilitates the rapid, inexpensive construction of SNP maps in biomedically and agriculturally important species. SNPs discovered by RRS also offer unique advantages for large-scale genotyping.

To discover SNPs, several copies of each locus must be sampled from a population and compared for sequence differences. Two methods have been described. The first, locus-specific polymerase chain reaction (PCR) amplification (LSA), requires the synthesis of oligonucleotide primers for each locus, limiting it to regions of known sequence and making it expensive for large-scale approaches. Moreover, LSA produces diploid genotypes; this requires identification of SNPs as heterozygotes, which is technically challenging. The second, whole-genome shotgun⁷, sequences random clones from the genomes of many individuals. It does not require previous knowledge of genomic sequence nor PCR, and provides haploid genotypes. Whole-genome shotgun is inflexible, however, requiring several-fold coverage of the genome before SNPs are discovered. For greater flexibility and efficiency, we sought to use 'reduced representations'—reproducibly prepared subsets of the genome, each containing a manageable number of loci to facilitate re-sampling. This approach presented three key challenges: creating reduced representations, finding orthologous matches and obtaining sufficient accuracy for automated calling of SNPs.

Many properties could be used to prepare reduced representations (for example, binding to a particular protein or functioning as an origin of replication). One of the simplest is to purify restriction fragments in a given size range. For example, *Bgl*II sites occur on average every ~3,100 base pairs (bp) in human DNA, which means that restriction fragments with length between 500 and 600 bp should number ~26,000 and comprise ~0.5% of the human genome. Computer analysis of 517 megabases (Mb) of finished human genomic sequence (17% of the estimated 3.1 gigabases) yielded 3,847 *Bgl*II fragments in this range—within 10% of the expected value. Thus, SNPs could be discovered by mixing DNA from many individuals, preparing a library of appropriately sized restriction fragments, and randomly sequencing clones. In this example, 52,000 sequences would provide, on average, twofold coverage of each locus, yielding thousands of SNPs.

We developed rules to align only those sequences representing the same genomic locus, excluding spurious matches arising from repeats. These rules eliminated known repeats, partial alignments that failed to extend across the entire sequence, matches showing excessive sequence divergence (compared with orthologous loci), and sequences re-sampled more often than expected for a single-copy locus (see Methods). Re-sequencing experiments confirmed that these rules successfully eliminated most spurious matches owing to paralogous repeats.

In comparing single-pass sequences for SNPs, base-calling errors can dominate the low rate of true polymorphism. With true SNPs occurring at a rate of 1 in 1,300 bp, base-calling errors must be less than 1 in 52,000 to achieve less than 5% false positive SNPs. Computer programs that estimate sequence accuracy (such as PHRED^{8,9}) judge only a small fraction of single-pass bases to be this accurate. However, we noted that many base-calling 'errors' occur adjacent to easily detected artefacts (that is, compressions and stops), or are attributable to poor alignment. We hypothesized that bases surrounded by perfectly aligned, consistently high-quality sequence (termed 'good neighbourhoods') might be more accurate than predicted by PHRED. We defined a neighbourhood quality standard (NQS) to identify such bases (see Methods).

To test the NQS, we examined 3.8 Mb of single-pass human DNA sequence obtained from bacterial artificial chromosomes (BACs), and compared the base calls to the highly accurate finished sequence of each BAC. Because BAC DNA is clonal, there are no polymorphisms, and any 'SNPs' represent base-calling errors. As expected, PHRED quality scores accurately reflect the overall likelihood of error (see below). Critically, base calls within 'good neighbourhoods' were much more accurate than predicted by PHRED: 85% of bases with PHRED scores more than 20 satisfied the NQS, and displayed an error rate of 1 in 36,000. In contrast, 15% of such bases failed the NQS, and these had a 40-fold higher error rate of 1 in

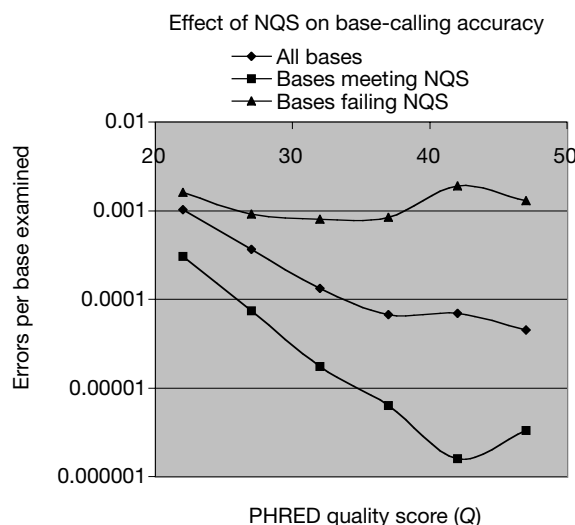


Figure 1 Impact of quality criteria on error rates. Data are plotted according to the PHRED *Q* score of each base^{8,9} and reported in bins of five PHRED *Q* units; only base substitution errors were counted. As previously reported⁹, overall PHRED scores accurately predict the observed rates of base-calling errors; however, bases meeting the NQS display

substantially lower error rates than are predicted by their PHRED scores. Although the effect is proportionally greatest for high PHRED scores, the bulk of errors avoided are found in bases with lower PHRED scores (that is, those with the highest error rates).

825 bp. Improved accuracy was observed across all PHRED scores (Fig. 1). The NQS can be used for automated SNP identification from single-pass sequence from RRS, genomic alignment, or expressed sequence tags^{10–12}. Its accuracy exceeds current goals for finished human sequence (less than 1 error in 10,000 bases) without several-fold coverage of each base; thus, the approach can potentially improve the efficiency and accuracy of genomic sequencing.

As a pilot, we pooled genomic DNA from ten individuals, digested it with *Bgl*III, separated the fragments by size by using agarose gel electrophoresis, excised a narrow slice adjacent to 564 bp, and cloned the fragments. Initial sequencing confirmed a tight distribution of insert lengths (Fig. 2a), and we sequenced 18,720 clones. After applying the alignment and base-calling rules above, 10,588 sequences were matched to one or more other clones, generating 14,400 paired reads that clustered into 3,356 ‘cliques’. The effective complexity of this library (the reciprocal of the chance that two randomly chosen clones derive from the same locus) is roughly 12,000 unique inserts, consistent with expectations. The distributions of clique sizes and SNPs within cliques were similar to those predicted under composite Poisson sampling, assuming a constant rate of nucleotide diversity, neutral alleles and a constant-sized population (Fig. 2b–d). These distributions, however, had longer than expected tails (Fig. 2b, d), which probably reflects cliques derived from repeats. Such cliques were withheld from further analysis; whether they represent unique loci will be determined by alignment to the complete human genome sequence. Sequence comparison of remaining cliques identified 1,750 candidate SNPs, with a heterozygosity (π) of 7.2×10^{-4} (consistent with previous studies^{13–16}).

Although RRS does not require prior knowledge of genome sequence, it can be extended by genomic alignment (GA). When RRS data were compared with 517 Mb (~16%) of finished human genomic sequence, 16% of clones matched a genomic contig, and SNPs were discovered at an equivalent rate ($\pi = 7.2 \times 10^{-4}$). Draft-quality genomic data can be used by applying the NQS to the submitted PHRAP quality scores (data not shown). Moreover, each SNP discovered by GA is immediately mapped to its location in the genome. Thus, the efficiency of SNP discovery and mapping is substantially increased by the availability of genomic sequence.

On the basis of these and related efforts at The Sanger Centre¹⁷ and Washington University in St. Louis (J. McPherson, personal

communication), The SNP Consortium (TSC) adopted RRS to identify 300,000 human SNPs over 2 years. TSC projects use the anonymous NIH diversity panel of 24 individuals¹⁸, from which we constructed 20 libraries using 4 different enzymes (*Bgl*III, *Hind*III, *Xba*I and *Eco*RI). We had identified 47,172 SNPs (Table 1) at the time of this writing, with the overall total of the TSC being 148,459, which shows that RRS is robust and scaleable. Each SNP is mapped by comparison with public human genome sequence (L. Stein, personal communication). Consistent with the estimated fraction of the human genome available at the time of analysis, 66% (97,983/148,459) of SNPs currently match a BAC clone; this proportion will rise as human genomic sequence accumulates throughout this year. Mapped SNPs are placed freely in the public domain at (<http://snp.cshl.org/> and <http://www.ncbi.nlm.nih.gov/SNP>).

To verify the quality of candidate SNPs, we re-sequenced 183 loci in each of the individuals used in library construction. For 6 out of 183 (3%) loci, all individuals appeared ‘heterozygous’, indicating that the sequence variation was between several copies of a repetitive locus, rather than between different individual’s copies of an orthologous locus. These will soon be easily eliminated by comparison with the full genome sequence. At the remaining 177 loci, 95% (216/227) of candidate SNPs were confirmed as polymorphic, demonstrating the accuracy of the techniques. We also validated 22 out of 24 (92%) candidate SNPs discovered solely by genomic alignment of RRS reads. Independent efforts have confirmed the high accuracy and heterozygosity of TSC SNPs. Using single-base extension to type over 1,000 TSC SNPs in the TSC discovery panel, essential identical validation results were obtained (M. Boyce-Jacino, personal communication). Furthermore, more than 80%

Table 1 Summary of SNP discovery

| | |
|-------------------------------------|---------------|
| Total reads passing quality filters | 653,348 |
| Failed owing to high repeat content | 123,993 (19%) |
| Passing reads | 529,355 (81%) |
| Twins | 528,937 |
| Paired bases examined for SNPs | 182,810,357 |
| Heterozygous positions observed | 142,651 |
| Heterozygosity (π) | 0.00078 |
| Cliques | 95,959 |
| Cliques passing all quality filters | 81,708 |
| Unique SNPs identified | 47,172 |
| SNPs per passing read | 1/11.2 |

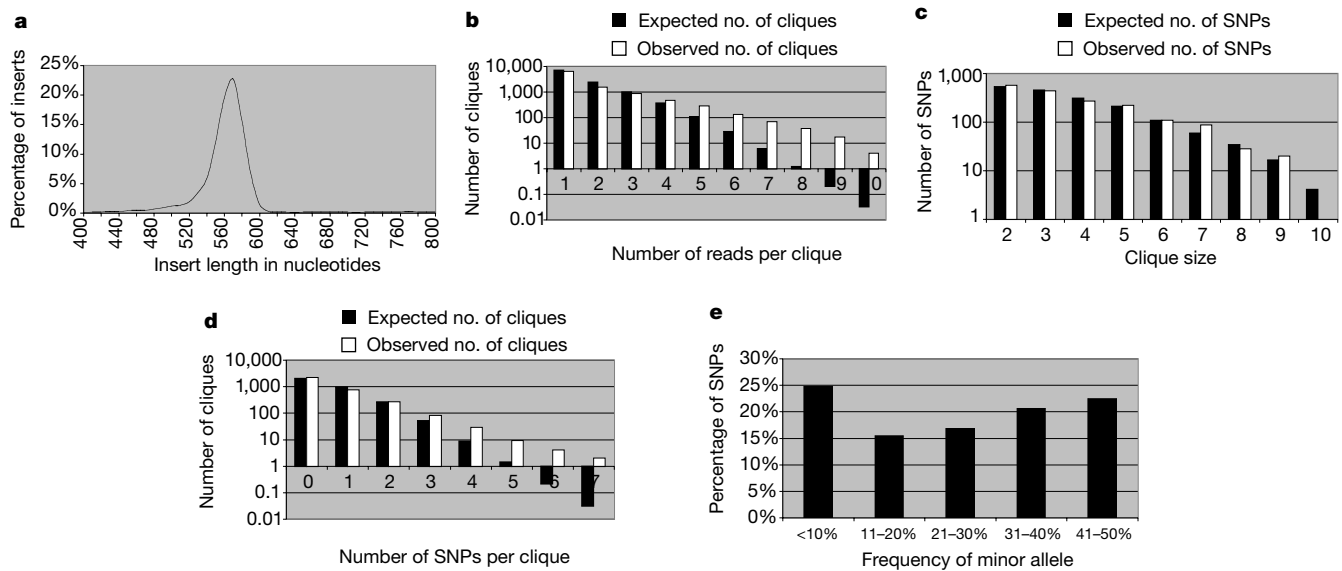


Figure 2 Pilot project data analysis. **a**, Insert size distribution. Lengths of inserts in a pilot RRS library cluster tightly around a mean of 560 bp, with 85% of reads between 535 and 585 bp. **b**, Clique size distribution. The number of cliques of each size was similar to that expected under composite Poisson sampling, with an excess of large cliques. These oversized cliques probably represent repetitive loci and were excluded from further analysis. **c**, Distribution of SNPs among cliques of different sizes. The number of SNPs found in cliques of different sizes is consistent with expectations under the neutral theory of evolution, assuming a constant rate of nucleotide diversity, neutral alleles and constant-sized population²². **d**, Distribution of loci containing different numbers of SNPs. The

number of loci with 0–7 SNPs closely matches the parameters described in **c**, with the exception of a few loci with an unexpectedly large number of SNPs. These may be attributable to inclusion of a small number of repeat sequences, or to varied levels of nucleotide diversity throughout the genome. **e**, Allele-frequency distribution of 214 validated SNPs among DNA samples used to construct the libraries (excluding two chromosomes in which the SNP was discovered). As expected, SNP discovery in a small number of chromosomes results in an allele frequency distribution that is relatively flat and thus yields a preponderance of common (>10%) alleles.

of more than 1,000 TSC SNPs were found to display high-frequency polymorphism (minor frequency > 5–10%) in independent population samples (P. Kwok, personal communication). Finally, the distribution of allele frequencies for validated SNPs in the diversity panel was roughly uniform, with a mean frequency of 24% (Fig. 2e).

We have demonstrated a new method for re-sampling loci without PCR, and for detecting SNPs with high accuracy in the resulting alignments. Although we focused on the human, RRS can be configured for any organism and scale, without the requirement for genome sequence data. Thus, RRS should be particularly useful for experimental models and agricultural species. In the human, the TSC's continuing discovery of SNPs will soon provide a map with average spacing of 3–5 kb, adequate for experimental characterization of linkage disequilibrium and haplotype-based disease association. These studies will require genotyping technology on a similar scale. In this regard, SNPs discovered by RRS offer a unique advantage: the potential for reduced representation genotyping (RRG) without locus-specific amplification. We are currently developing methods to isolate the reduced representation from each individual, generically amplify after linker ligation, and directly genotype the resulting fragments by established methods such as single-base extension¹⁹ or array hybridization¹⁴. When combined with a dense SNP map, such large-scale genotyping methods should markedly enhance attempts to unravel the genetic contribution to common diseases.

Note added in proof: Since submitting this manuscript, the number of SNPs discovered by the SNP Consortium has grown to over 350,000, of which 250,000 have been mapped to the draft sequence and are freely available at the TSC website (<http://snp.cshl.org>). □

Methods

Expected sampling of reduced representations

Given an expected average fragment size *d*, and a genome of size *G*, the number of unique

fragments (*D*) in the size range {*x*₁,*x*₂} is estimated as:

$$D = (G/d)(e^{-x_1/d} - e^{-x_2/d})$$

With *N* reads performed from a library of *D* equally represented inserts, *N*²/2*D* pairwise matches are expected. However, laboratory separations result in bell-shaped size distributions (Fig. 2a), such that the chance any insert matches another depends on where it falls in the size distribution. The expected proportion of matches is then calculated as a composite Poisson distribution after dividing the data into appropriate size bins.

Library construction and sequencing

For the RRS pilot, an equimolar mixture of 10 DNAs was created (Coriell nos 10965 (Amerindian), 10470 (African Pygmy), 11322 (Chinese), 11589 (Japanese), 13820 (Russian), 05987A (Amish), 12615 (French), 11997 (Utah), 08779A (African–American) 10540 (Melanesian)) and digested to completion with *Bgl*II (New England Biolabs). Pooled, digested DNA (100 μg) was electrophoresed using 1.2% NuSieve agarose (BioWhittaker) and a narrow band (2 mm) was excised adjacent to a 564-bp size marker (*Lambda/Hind*III). We cloned fragments into M13mp19 RFI DNA (Pharmacia) and prepared clones by standard methods. We carried out dye-primer sequencing with the M13 –21 forward primer and ABI 377 sequencers. For TSC production, we used the NIH diversity panel of 24 anonymous DNAs¹⁸. Each DNA (15 μg) was digested with *Bgl*II, *Hind*III, *Xba*I or *Eco*RI and pooled. Fragments were size separated on 1.6% SeaKemLE (BioWhittaker) agarose gels, excised in five adjacent size windows from 400 to 650 bp, and cloned into M13mp18. Sequences were obtained using standard methods and dye-primer chemistry and ABI 377 (96,960), dye-terminator chemistry and ABI 377 (22,165), or dye-terminator chemistry and ABI 3700 capillary sequencers (791,455).

Sequence alignment

We processed traces with PHRED, removed vector sequences and applied quality filters: we rejected reads with fewer than 100 NQS bases or inserts shorter than 200 bp. Passing sequences were compared by BLAST²⁰ to a library of known repeats (available on request) and rejected if more than 50% of bases matched known repeats. The remaining 'quality reads' were compared by BLAST, and pairs with > 80% identity over two-thirds of their length identified. Smith–Waterman alignment of overlapping sequence was followed by trimming of poor quality alignments at either end; if perfect alignment (10/10) did not begin within 50 bp of both ends, the pair was rejected. We identified sequence discrepancies meeting NQS and counted them. If more than 1% of NQS bases appeared polymorphic, the pair was rejected, as this level of diversity is uncommonly observed in comparison of orthologous loci from different individuals (Table 1; and refs 13–16, 21). Pairs passing these rules were clustered into cliques using transitivity. We compared clique sizes with the expected distribution under composite Poisson sampling for single-copy

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.

- Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
- Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
- Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
- Hasbacka, 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).
- Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
- Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 15173–15177 (1999).
- Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7**, 401–409 (1997).
- Ewing, B. & Green, P. Base-calling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
- 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).
- Gu, Z., Hillier, L. & Kwok, P. Y. Single nucleotide polymorphism hunting in cyberspace. *Hum. Mutat.* **12**, 221–225 (1998).
- Marth, G. T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nature Genet.* **23**, 452–456 (1999).
- 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).
- Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
- Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
- Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
- Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
- Mullikin, J. C. *et al.* An SNP map of human chromosome 22. *Nature* **407**, 516–523 (2000).
- 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).
- Landegren, U., Nilsson, M. & Kwok, P. Y. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. *Genome Res.* **8**, 769–776 (1998).
- 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).
- Cambien, F. *et al.* Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.* **65**, 183–191 (1999).
- 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. Bruskiwich, 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. Lambie, C. F. Langford, J. Macarthy, V. Mall, A. Moreland, E. K. Overton-Larty, M. T. Roney, 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)