

Computational comparison of two draft sequences of the human genome

John Aach*†, Martha L. Bulyk‡, George M. Church†, Jason Comander§, Adnan Derti†|| & Jay Shendure*†

† The Lipper Center for Computational Genetics, ‡ Program in Biophysics, † Department of Genetics and § Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA

|| Program in Bioinformatics, Boston University, Boston, Massachusetts 02215, USA

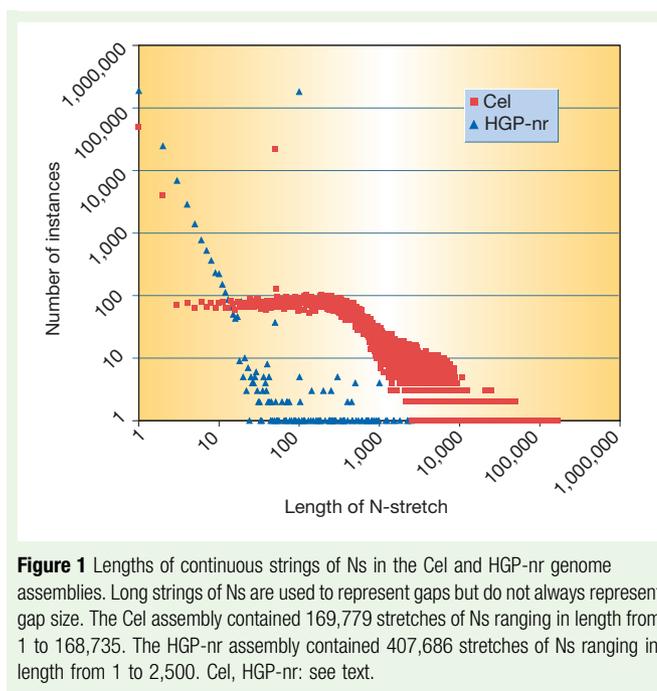
* These authors contributed equally to this work

We are in the enviable position of having two distinct drafts of the human genome sequence. Although gaps, errors, redundancy and incomplete annotation mean that individually each falls short of the ideal, many of these problems can be assessed by comparison. Here we present some comparative analyses of these drafts. We look at a number of features of the sequences, including sequence gaps, continuity, consistency between the two sequences and patterns of DNA-binding protein motifs.

The two draft sequences of the human genome were generated by the Human Genome Project (HGP)¹ and Celera Genomics². Genome sequencing entails direct sequencing of DNA fragments and assembly of the fragment sequences into larger units on the basis of their overlaps (shotgun assembly). The HGP used a hierarchical mapping and sequencing approach, involving generation of a series of overlapping clones that cover the entire genome and shotgun sequencing of each clone. The genome sequence was reconstructed by assembling the fragments on the basis of sequence overlap and mapping and chromosomal position information on the clones. Celera Genomics used a whole-genome shotgun sequencing approach, without generating a series of overlapping clones, but also incorporated HGP information where available.

We worked with three versions of genome sequence. (1) Data from the individual large-insert clones sequenced by the publicly sponsored HGP is available from the National Center for Biotechnology Information (NCBI; <http://www.ncbi.nlm.nih.gov>), and is denoted here as HGP-all. HGP-all comprises 4.8 gigabases (Gb) from 34,084 large-insert clones (identified as phase 0 (raw) to phase 3 (finished) sequence) and is highly redundant because it contains sequences from a collection of overlapping clones, as well as other sources whose overlaps have not been fully resolved through assembly. (2) Several genome-wide assemblies have been produced by merging the data in HGP-all. We studied an assembly comprising 2.9 Gb in 6,094 sequences—designated here as HGP-nr ('nonredundant')—in which clearly identifiable redundancies were eliminated. Several other genome assemblies have also been generated; these assemblies are not analysed here. (3) We used Celera Genomics' 'Human Genome D'², which represents 2.9 Gb in 54,061 sequences, denoted here as Cel (which we obtained through an academic licence to the Celera database). We did not analyse their larger unassembled databases of 23.1 million human fragments and 2.8 million polymorphic variants. NCBI has made rapid progress towards annotating HGP-nr for genes on the basis of their RefSeq database³, which is a manually curated collection of mRNA sequences from known genes, but at the time of writing (December 2000) the Cel sequence had no directly linked annotation.

Cel and HGP-nr use long strings of Ns to indicate gaps in assemblies (Fig. 1). Although the two draft genome sequences are similar in size, HGP-nr contains fewer unidentified bases than Cel (0.65% versus 8.7%, respectively; the 1.0-Gb HGP-all phase 3 sequence contains even fewer, 0.005%), largely because gaps are annotated differently in the two sequences. Thus, when one removes the unidentified bases (Ns), the amount of specified nucleotide sequence is 2.84 Gb for HGP-nr and 2.66 Gb for Cel. HGP-nr



contains 181,079 strings of 100 Ns to represent gaps but contains strings of up to 2,500 Ns; Cel contains 21,684 strings of 50 Ns but contains strings of up to 168,735 Ns. Figure 2 shows the sizes of continuous segments of sequence not broken by strings of 20 or more Ns (ungapped sequences). Whereas Cel uses the lengths of very long strings of Ns to represent the estimated sizes of large gaps, HGP-nr generally simply uses strings of 100 Ns and does not use string length to represent gap size. Therefore the smaller continuous strings of Ns in HGP-nr does not imply that gaps in HGP-nr are smaller than in Cel.

As another indicator of the continuity of the assemblies, we examined the ten genes in the RefSeq database with the largest messenger RNAs and tested whether their coding sequences could be found on single contigs, the largest continuous sections of sequence (possibly interrupted by N strings) generated by sequence assembly (Fig. 2). We used the BLAST search algorithm⁴ to locate the best contig matches for the first and last 500 base pairs (bp) of the coding regions of these ten genes. Six of these genes in HGP-nr, and seven in Cel, had both ends on the same contig. In HGP-nr, *ACF7* and *RYR2* had ends in different contigs, while *NEB* and *MUC2*

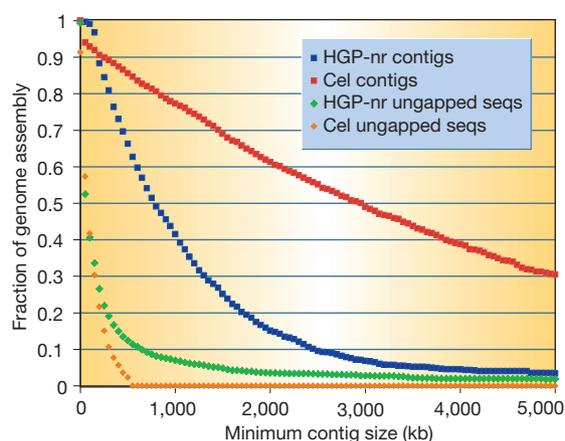


Figure 2 Cumulative histogram of the fraction of assembled genome sequence present in contigs and in ungapped continuous sequences that exceed a given length in the HGP-nr and Cel assemblies. Contigs are the largest continuous segments of sequence generated by assemblies. The Cel assembly consisted of 2.9 Gb in 54,061 contigs. The HGP-nr assembly consisted of 2.9 Gb in 6,094 contigs. Contigs may contain gaps represented by strings of Ns (Fig. 1). Here ungapped continuous sequences (ungapped seqs) are counted as maximal stretches of sequence that do not contain strings of 20 or more Ns. Cel, HGP-nr: see text.

matched a contig for one end but a matching sequence could not be located for the other. In Cel, *ACF7* and *MUC2* had ends in different contigs, and a smaller coding region (*TNNT2*) appeared to have one end in a Cel contig and the other in a stretch of Ns and an unfinished fragment. These results indicate that both sequence assemblies have limited ability to build contigs covering the longest stretches of DNA that cells themselves can transcribe continuously.

Comparing unique stretches of sequence

Oligonucleotides that occur only once in the genome can be useful for making specific probes for genomic DNA or messenger RNA, or as the 3' ends of specific primers for polymerase chain reaction (PCR) amplification. They can also provide a high-level statistical view of how much sequence content the two draft genome sequences have in common. We analysed stretches of 15 nucleotides, referred to as 15-mers. In 6 Gb of genomic sequence (with both strands considered), the number of occurrences of an arbitrary 15-mer (of which there are 4^{15} or slightly more than one billion) can be estimated as following a Poisson distribution with mean of about 6 (see Supplementary Information). Disregarding splice junctions, a typical mRNA sequence of 2000-bp should therefore contain about five unique 15-mers and have a 99% likelihood of having at least one. Using a computer algorithm to analyse the two genome assemblies, we considered every possible 15-mer and determined whether it occurred 0, 1 (unique) or multiple times. Sequence and assembly errors, gaps, redundancies, splice junctions and polymorphisms all affect whether a given 15-mer will be counted as unique. We therefore refer to 15-mers seen only once as candidate unique 15-mers (cu-15s). Assuming a combined rate of sequence error and polymorphism of 0.1% per base, we estimate a cu-15 false positive rate of 0.14% and a cu-15 false negative rate of 9.0% (see Supplementary Information). A false positive is a multiply occurring 15-mer that is detected as unique because variations have made all but one occurrence appear different; a false negative is a unique 15-mer that is detected as multiple because of variations in similar 15-mers.

We found 169,609,634 cu-15s in the 2.9-Gb Cel sequence and 160,311,078 in the 2.9-Gb HGP-nr sequence. Of the cu-15s in the two sequences, 19,270,620 are found only in Cel and 17,527,980 are

found only in HGP-nr; therefore, nearly 11% of cu-15s in each sequence are not shared with the other. Because the analysis is expected to have a false negative rate of 9% for each database, this suggests that the true amount of sequence present in one database and not the other is about 0.14% for both HGP-nr and Cel (see Supplementary Information). Cel and HGP-nr therefore contain similar amounts of unique sequence, and most unique sequences are common to both. We also looked for unique 15-mers representing 10,292 mRNA sequences from the RefSeq database³ (see Supplementary Information); we found 2,526,912 when compared against Cel and 2,372,185 when compared against HGP-nr. Again the Cel and HGP-nr results are comparable.

By analysing cu-15s in coding sequences we found additional differences between the Cel and HGP-nr draft genome assemblies. For instance, the coding regions for the genes *LOC63301* and *GR3*, which are 99% identical over their 1,113 bases, are annotated in HGP-nr as being located on two distinct chromosomes: contig NT_008902 (chromosome 10) and contig NT_023464 (chromosome 6), respectively. However, we found that these sequences are contained on only a single Cel contig (GA_x2HTBL4-GEJJ:1..500000) and, consistent with this, we found 396 cu-15s in the Cel database shared by both sequences. Analysis of the HGP-nr sequence using 'BLAST2 sequences' (ref. 5) indicates that the contigs share a larger region of sequence identity, with the first 10,652 bases of NT_023464 being 99% identical to a region inside NT_008902. This difference could arise from an error in Cel assembly, caused by inability to distinguish two regions with very high sequence identity, or from insufficient coverage of the regions by high quality sequence reads. Alternatively, an HGP-nr error could arise from assembly limitations or insufficient high quality coverage of the single region endpoints, or from erroneous mapping of the single region to two chromosomes.

Searches for DNA-binding protein motifs

Compared with other organisms^{6,7}, the characterization of DNA-binding protein motifs in humans has been limited until recently by the paucity of sequence data for noncoding DNA. Although the draft human sequence has overcome this limitation, the large size of the human sequence and the existence of active regulatory sites located far from the genes they regulate⁸ pose new challenges to binding-site analysis. To assess how much they are enriched in the neighbourhood of genes in humans, we searched for motifs for two DNA-binding proteins (EGR-1 and CRX) in 4-kb upstream sequences of 3,352 genes and compared their abundances with those found in random 4-kb sequences and available positive control upstream sequences known or likely to contain binding sites (see Supplementary Information). We focused on HGP-nr sequence rather than Cel for upstream sequence because more annotation for genes was available, but used both HGP-all phase 3 and Cel for random sequences.

The human zinc finger DNA-binding protein EGR-1 (homologous to mouse Zif268) is induced by growth factors and nerve cell depolarization and is involved in cellular proliferation and differentiation⁹. We developed weight matrices that variously took into account a recognized nine-base consensus binding site (GCG[G/T]GGGCG¹⁰), *in vitro* selection data indicating a modified consensus and specific interactions with flanking bases¹¹, and our own double-stranded DNA array binding data, which indicates that the middle three bases of the site do not independently affect binding specificity¹². Figure 3a shows the results of a search for EGR-1 sites using matrix M1EGR-1, which looks for nine-base sites and uses a scoring threshold that allows only sites with small deviations from the consensus (see Supplementary Information). We found that upstream regions were significantly enriched for EGR-1 sites compared with random sequences, but not significantly different from 17 positive controls (Table 1). However, EGR-1 sites are GC-rich and therefore could occur more frequently by chance in

CpG islands upstream of genes. We therefore used a second matrix, M2EGR-1, which took into account the flanking bases of the nine-base site and compensated for the local GC content of potential sites when calculating scores (see Supplementary Information). The results of this search still indicate over-representation of EGR-1 sites in upstream regions (Fig. 3a, Table 1).

But not all DNA binding motifs show this pattern. We also searched for binding motifs for the photoreceptor homeobox transcription factor¹³ CRX. We generated a matrix, M1CRX, for CRX which indicated that CRX binding motifs are significantly under-represented in upstream regions compared with random sequence (Fig. 3b, Table 1). We are currently inspecting 67 CRX-regulated genes as potential positive controls. The only two inspected in detail so far, *RHO* (which encodes rhodopsin) and *PDC* (which encodes phosducin), are found to have three and two upstream sites, respectively. Although each of these is significantly higher than the number of sites found in the 3,352 upstream regions or random controls, the number of positive controls examined is too low to draw general conclusions (see Supplementary Information). These preliminary results suggest that DNA binding protein motifs have different statistical representation in upstream regions

compared with other sequence, but that over-representation in these regions does not necessarily indicate regulation by the protein. We are continuing our investigations with larger sets of upstream sequences (see Supplementary Information).

Although there are similar average numbers of DNA-binding sites in random sequences from Cel and HGP-all, they are significantly different (Table 1). This probably reflects subtle differences in the sequences that could be detected because such large samples were compared (83,800 for HGP-all and 33,520 for Cel). The differences between upstream sequence and random sequence were significant regardless of whether random sequence came from HGP-all or Cel.

Discussion

The HGP-nr and Cel draft genome assemblies are similar in size, contain comparable numbers of unique sequences (which we analysed as unique 15-nucleotide stretches, cu-15s), and exhibit similar statistics for sample candidate DNA protein-binding motifs. Some differences emerge at a detailed level. Sequence content aside, the assemblies are also packaged differently. Contigs in each exhibit different size and gap distributions (Figs 1 and 2), and HGP presents more stages of assembly by providing four phases of sequence data compared with the single Cel Human Fragments database. More annotation is also available at HGP. We expect all these differences to diminish as assemblies become more complete and comprehensive.

The complete human genome presents us with challenges on several levels. First, the analysis of 3 Gb of human sequence requires increased computer resources compared with analysis of smaller genomes, although this need can be addressed to an extent through modest computational goals and careful planning. For instance, a suffix array is a fast and convenient way of finding unique sub-sequences in a set of sequences¹⁴, but in our hands it requires 12 bytes of RAM per base pair of sequence when the number of sequences is large. The computer program described above required less than 300 Mbytes of RAM and could analyse cu-15s in the 4.8-Gb HGP-all in eight hours on a high-end conventional PC (see Supplementary Information). A second challenge is the fact that the genome sequences are constantly being updated, so programs that analyse them must be modified and rerun frequently to include the latest information and to accommodate new packaging and annotation. But this challenge brings with it the promise of rapid progress and new opportunity.

The broadest challenge will be the development of algorithms and platforms to meet the new computational needs that will arise as a

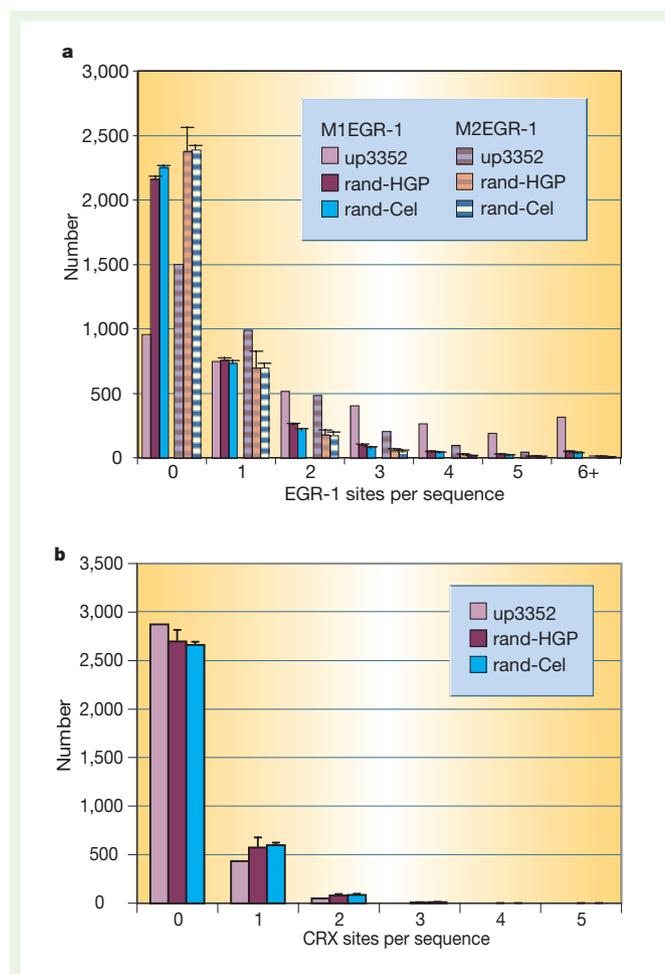


Figure 3 Numbers of DNA-binding sites per sequence from searches of upstream regions in sets of 3,352 4,000-bp sequences as determined by three different DNA-binding protein motif matrices. Sequence set abbreviations up3352, rand-HGP and rand-Cel are as in Table 1. up3352 bars represent counts of upstream sequences containing indicated numbers of sites. rand-HGP (rand-Cel) bars represent averages of counts of random sequences over 25 (10) selections of 3,352 sequences. Error bars for rand-HGP (rand-Cel) represent s.d. of the 25 (10) sets of counts. **a**, Sites per sequence for EGR-1 binding sites as determined by matrices M1EGR-1 and M2EGR-1. **b**, CRX sites as determined by matrix M1CRX.

Table 1 DNA-binding protein binding sites

Sequence set	Sites per sequence	Sequence set comparisons		
		up3352	rand-HGP	rand-Cel
Matrix M1EGR-1				
up+c	2.35 ± 1.80	0.8140	1.4 × 10 ^{-7*}	2.2 × 10 ^{-10*}
up3352	2.21 ± 2.45		0*	0*
rand-HGP	0.64 ± 1.34			1.1 × 10 ^{-25*}
rand-Cel	0.55 ± 1.17			
Matrix M2EGR-1				
up+c	1.12 ± 0.99	0.6951	0.0032*	0.0002*
up3352	1.00 ± 1.26		5.5 × 10 ^{-240*}	0*
rand-HGP	0.42 ± 0.98			0.0004*
rand-Cel	0.40 ± 0.80			
Matrix M1CRX				
up3352	0.16 ± 0.40		1.4 × 10 ^{-14*}	7.0 × 10 ^{-19*}
rand-HGP	0.22 ± 0.48			7.8 × 10 ^{-6*}
rand-Cel	0.24 ± 0.50			

Average number ± s.d. of DNA-binding protein motifs found by different motif matrices in sets of 4,000-bp genomic sequences, and comparisons of average numbers found for different sets. Motif matrices M1EGR-1 and M2EGR-1 for EGR-1 binding sites, and M1CRX for CRX binding sites, are described in the text. Sequence sets: up+c, upstream regions of 17 EGR-1 positive control genes (see Supplementary Information); up3352, upstream regions extracted from HGP-nr sequence for 3,352 genes (see text); rand-HGP, 25 sets of 3,352 4,000-bp regions randomly extracted from HGP-all phase 3 sequence; rand-Cel, 10 sets of 3,352 4,000-bp regions randomly extracted from Cel sequence. Sequence set comparisons: t-test probabilities that average sites/sequence are equal for different sequence sets. Asterisk, statistically significant comparison (t-test $P < 0.05$).

result of the human genome sequence. For instance, any single 9–11-mer will occur 1,450–23,000 times in 6 Gb of DNA by chance alone, and therefore a 9–11-base recognition sequence for a protein such as EGR-1, which admits many variations, will occur many times more often. How do we distinguish actual regulatory sites? Focusing on regions near the 5' ends of genes may provide a first order approach, but the case of EGR-1 above suggests that this may not be very specific for functional sites. A promising approach is to look at regions of the upstream sequences that are conserved in other mammals^{15,16}. Similar genome-wide comparison requirements are arising in other contexts. For instance, new mutations have accumulated in the human population at a rate of 1–100 mutations per generation over the past 5,000 generations¹⁷. This means that among the 6 billion humans alive today, there is a reasonable chance that all possible single nucleotide polymorphisms exist for each of the approximately 3 billion base pairs of the human genome. It is becoming clear that costs must be reduced to study effectively the association of traits with polymorphisms¹⁸ or haplotypes¹⁹, but moving from population associations to accurate assessment of individual variations is likely to require sequence analysis of the two copies of the three billion base pairs in any of the six billion humans who desire it.

We remain optimistic that these challenges will be met. Through remarkable miniaturization and quality improvements, the cost of genomics has been dropping roughly twofold every 18 months for decades, paralleling the trend for computing²⁰. In a few years' time we may be able to read the 6 billion DNA base pairs in a human cell almost as easily and inexpensively as we can read a similar number of bits in a CD or DVD today. Today's milestone—two human genome sequence drafts—foreshadows a future in which resequencing and comparison of entire mammalian genomes will be routine operations for biology laboratories. □

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Maglott, D. R., Katz, K. S., Sicotte, H. & Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28**, 126–128 (2000).
4. 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).

5. Tatusova, T. A. & Madden, T. L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**, 247–250 (1999).
6. McGuire, A. M., Hughes, J. D. & Church, G. M. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* **10**, 744–757 (2000).
7. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214 (2000).
8. Blackwood, E. M. & Kadonaga, J. T. Going the distance: a current view of enhancer action. *Science* **281**, 60–63 (1998).
9. Sukhatme, V. P. *et al.* A zinc finger-encoding gene coregulated with *c-fos* during growth and differentiation, and after cellular depolarization. *Cell* **53**, 37–43 (1988).
10. Christy, B. & Nathans, D. DNA binding site of the growth factor-inducible protein Zif268. *Proc. Natl Acad. Sci. USA* **86**, 8737–8741 (1989).
11. Swirnoff, A. H. & Milbrandt, J. DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell. Biol.* **15**, 2275–2287 (1995).
12. Bulyk, M. L. Development and application of microarray technologies for the highly parallel analysis of the sequence specificity of DNA binding proteins. Dissertation (Harvard Medical School, Boston, 2000).
13. Livesey, F. J., Furukawa, T., Steffen, M. A., Church, G. M. & Cepko, C. L. Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene *Crx*. *Curr. Biol.* **10**, 301–310 (2000).
14. Gusfield, D. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology* (Cambridge Univ Press, Cambridge, 1997).
15. Lipman, D. J. Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res.* **25**, 3580–3583 (1997).
16. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26**, 225–228 (2000).
17. Eyre-Walker, A. & Keightley, P. D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347 (1999).
18. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
19. Davidson, S. Research suggests importance of haplotypes over SNPs. *Nature Biotechnol.* **18**, 1134–1135 (2000).
20. Moore, G. M. Cramping more components onto integrated circuits. *Electron. Mag.* **38**, 114–117 (1965).

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. It is also available at <http://arep.med.harvard.edu>.

Acknowledgements

We thank J. Seidman, C. Cepko, S. Blackshaw and Y. Grad for information on genes considered in the paper; E. Lander, C. Venter, K. Pruitt, J. Ostell and J. Hughes for comments on this manuscript and information on genome sequences; P. Johnson and T. Chen for software assistance; and J. McPhee and M. Fanjoy for system support. We also thank the Lipper Foundation, DOE (G.M.C.), and the Paul & Daisy Soros Fellowship for New Americans (A.D.) for funding.

Correspondence should be addressed to G.M.C. (e-mail: church@arep.med.harvard.edu).