Guide to the draft human genome

Tyra G. Wolfsberg*, Johanna McEntyre† & Gregory D. Schuler†

** Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*
† National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

There are a number of ways to investigate the structure, function and evolution of the human genome. These include examining the morphology of normal and abnormal chromosomes, constructing maps of genomic landmarks, following the genetic transmission of phenotypes and DNA sequence variations, and characterizing thousands of individual genes. To this list we can now add the elucidation of the genomic DNA sequence, albeit at 'working draft' accuracy. The current challenge is to weave together these disparate types of data to produce the information infrastructure needed to support the next generation of biomedical research. Here we provide an overview of the different sources of information about the human genome and how modern information technology, in particular the internet, allows us to link them together.

The ultimate goal of the Human Genome Project is to produce a single continuous sequence for each of the 24 human chromosomes and to delineate the positions of all genes. The working draft sequence described by the International Human Genome Sequencing Consortium was constructed by melding together sequence segments derived from over 20,000 large-insert clones[1]. All of the results of this analysis are available on a web site maintained by the University of California at Santa Cruz (http://genome.ucsc.edu). Over the next few years, draft quality sequence will be steadily replaced by more accurate data. The National Center for Biotechnology Information (NCBI) has developed a system for rapidly regenerating the genomic sequence and gene annotation as sequences of the underlying clones are revised (http://www.ncbi.nlm.nih.gov/genome/guide). Undoubtedly, others will apply a variety of approaches to large-scale annotation of genes and other features. One such project is Ensembl, a joint project of the European Bioinformatics Institute (EBI) and the Sanger Centre (http://www.ensembl.org).

Pinpointing new genes

Recent estimates have placed the number of human genes at 25,000–35,000 (refs 2, 3). More than 10,000 human genes have been catalogued in the Online *Mendelian Inheritance in Man*[4] (OMIM), which documents all inherited human diseases and their causal gene mutations. Integration of the information contained in OMIM with the working draft is facilitated by the fact that it has already been tied to reference messenger RNA sequences through a collaborative effort between OMIM, the Human Gene Nomenclature Committee and the NCBI[5,6]. As a result, the positions of many known genes have been determined by alignment of mRNAs with genomic sequences. For the remaining genes, we must currently resort to computational gene-finding methods (reviewed in refs 7, 8).

When mRNA species align differently to a genomic sequence, rhis indicates that alternative splicing has taken place. In the current set of full-length reference mRNAs, 11,174 transcripts have been sequenced from 10,742 distinct genes (2.4% of the genes have multiple splicing variants). Alignments of expressed sequence tag (EST) sequences to the working draft sequence, however, suggest that about 60% of human genes have multiple splicing variants, which has important implications for the complexity of human gene expression[1]. By their sheer numbers (currently over 2.5 million) we might expect ESTs to sample a larger fraction of splicing variants than would be the case for more traditional targeted approaches. For example, alignment of the mRNA of the membrane-bound metalloprotease-disintegrin ADAM23 (ref. 9) to the draft genome reveals that the gene consists of at least 23 exons. Of the many ESTs that also align to the *ADAM23* locus, one lacks the exon that encodes the transmembrane domain, which suggests an alternatively spliced, soluble protein. Although this is a biologically plausible conclusion, one should exercise caution when interpreting such results: ESTs are partial single-pass sequences that have been associated with a variety of artefacts, including sequencing errors and improper splicing[10,11].

Finding relatives

Genes can be found through an implied relationship to something else—for example, being a putative orthologue (related to a gene in another species). To do this, it is useful to search the genomic sequence or, preferably, its mRNA sequences and protein products, using BLAST[12]. As an example, we use the mouse *Lmx1b* gene, which encodes a LIM homeobox protein that is important in pattern development[13]. When we used the protein sequence encoded by *Lmx1b* as a query in a BLAST search against the working draft human genome sequence, the best match was to a region of 9q34, and the positions of the alignments line up with the exons of the human *LMX1B* gene (Fig. 1a).

Additional support for two genes being orthologous comes from the mouse–human homology map. Despite being separated by 200 million years of evolution, mouse and human genes often fall into homologous chromosomal regions that share a conserved gene order (synteny). In fact, the working draft sequence has helped to refine the homology map and provides inferred map positions for many mouse genes[1]. The two homeobox genes fall within a conserved syntenic block between mouse chromosome 2 and human chromosome 9 (Fig. 1b). Furthermore, the human *LMX1B* gene has been implicated in nail patella syndrome (NPS), an autosomal recessive disorder characterized by limb and kidney defects. A mouse in which *Lmx1b* has been inactivated shows a phenotype that is strikingly similar to NPS[13]. Besides providing additional support for the conclusion of orthology, this connection may provide a useful mouse model for the human disorder. In this way, information from OMIM and mouse mutants can further define human genes.

Another way to find a gene is by looking for paralogues—family members derived by gene duplication. As an example, we used human *ADAM23*, which maps to 2q33 (ref. 14). In a BLAST search against a set of proteins predicted from the draft sequence, aside from matching itself, the best match was to a peptide from chromosome 20. No ADAM family member has previously been mapped to this chromosome. The predicted protein encoded by this gene does not begin with a methionine and appears to be incomplete at its amino terminus when aligned with other family

members: this could be due to an erroneous protein prediction or a gap in the draft sequence. Computational analysis can reveal protein family domains and their relationships to three-dimensional protein structures. In this case, the putative ADAM paralogue contains both the zinc metalloprotease and disintegrin motifs characteristic of the ADAM family. Critical amino acids of the metalloprotease domain are conserved in the putative paralogue (Fig. 2b), including a trio of histidine residues in the active site (shaded yellow), which are important in complexing the zinc ion (Fig. 2a). The finding that the sequence of the new predicted ADAM member has an intact active site suggests that the predicted gene is functional, rather than being a pseudogene.

Searching by position

It is sometimes desirable to find genes by their position in the genome, rather than by sequence similarity. For example, when genetic or cytogenetic analysis has implicated a particular region in the aetiology of a disease, it is of interest to see what genes lie in the region. A natural way to describe positions in a sequence would be by base coordinates, but this is impractical for the working draft sequence, as the sequence is still being revised. Cytogenetic band nomenclature is more commonly used to describe positions in the genome, and many human diseases are linked to chromosomal deletions, amplifications and translocations[15]. However, to be useful in conjunction with the working draft, these designations must be related to the sequence. Towards this end, a consortium has integrated this information using fluorescence *in situ* hybridization (FISH) to localize BAC clones that also bear sequence tags that can be found in the draft sequence[16]. In addition to providing cytogenetic coordinates as entry points into the genome, they also provide mapped clone reagents that may be useful in further experimental work.

Another way to describe positions in the genome is relative to mapped sequence tagged site (STS) markers. This is particularly useful in positional cloning projects, where candidate regions are usually defined by polymorphic STSs used in genetic linkage analysis. For example, the breast cancer susceptibility locus *BRCA2* was originally localized by fine genetic mapping to a 600-kilobase (kb) interval on chromosome 13 centred around the STS marker D13S171 (ref. 17). STS markers from several genetic and physical maps have been localized in the working draft sequence using a procedure known as electronic PCR[18]. Thus, by simply looking up the position of D13S171, we can see the region around what is now known to be *BRCA2*, together with other features such as adjacent genes and markers, translocation breakpoints, and genetic variations.

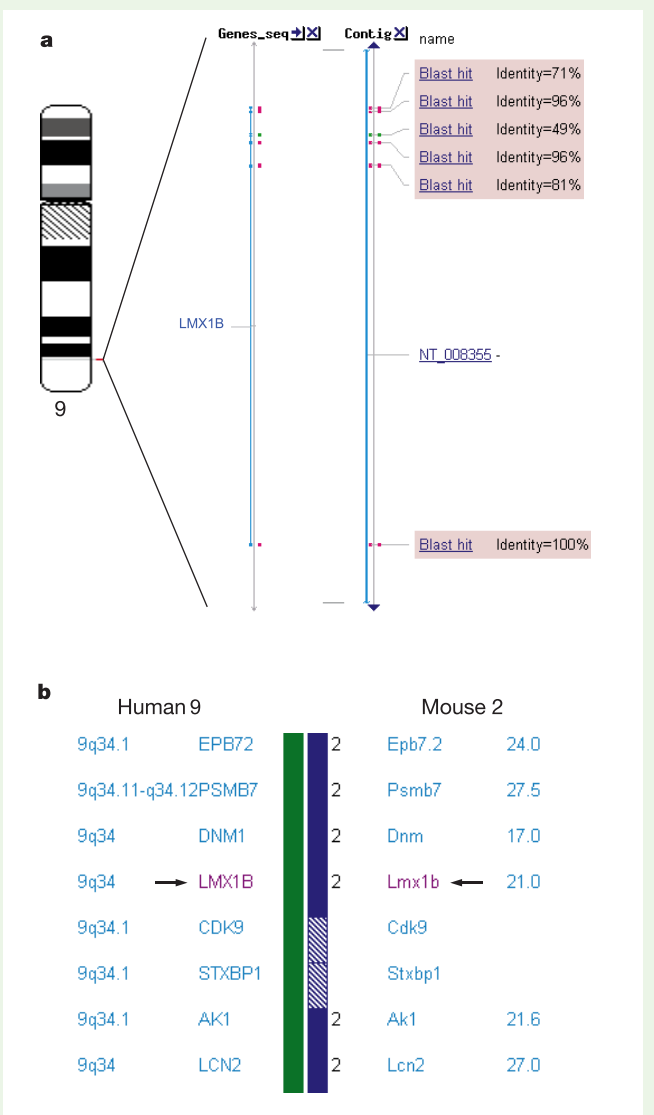

Figure 1 *Lmx1b* encodes a transcription factor that helps to control the trajectory of motor axons during mammalian limb development. **a**, The results of a search of the six-frame translation of the draft genome sequence on the NCBI site using mouse Lmx1b protein NP_034855.1 as a query and using TBLASTN with standard search parameters. The best match was to a region of chromosome 9 that contains the human *LMX1B* gene. **b**, The mouse *Lmx1b* and the human *LMX1B* genes lie within a conserved syntenic block of genes, in mouse on chromosome 2 and in human on chromosome 9. This conservation of gene order supports the theory that *Lmx1b* and *LMX1B* are orthologous.

Variations on a theme

A map of DNA sequence variations will aid our understanding of complex diseases and human population dynamics. The most common class of variation is the single nucleotide polymorphism (SNP) and the total number of SNPs in the public database (dbSNP)[19] now exceeds 2.5 million, representing 1.5 million unique SNP loci. Because database entries include flanking sequence surrounding the polymorphic base(s), it is possible to localize variations within the working draft by simple sequence alignment[20].

Histone deacetylase 3 (HDAC3) is a nucleosome-remodelling enzyme that deacetylates the lysine residues of histones, affecting transcriptional repression[21]. Its genomic region contains seven mapped SNPs near the locus, one of which falls within the coding region—a G-to-C substitution that results in the nonsynonymous substitution Arg265Pro in the protein product. The three-dimensional structure of an *Aquifex aeolicus* homologue shows that this residue is at the lip of the active-site pocket[22]. Of the two classes of eukaryotic histone deacetylase, one most often has Arg at this position and the other most often has Pro[23], a trait shared with the bacterial members of the histone deacetylase superfamily. This SNP might therefore occur at a functionally interesting site, and also gives pause for speculation: as bacterial members of the superfamily predominantly have a Pro in this position, perhaps Pro is the ancient residue at this site, and not Arg. Note that several high-throughput SNP discovery methods have been used to generate these data and not all SNPs have been rigorously validated.

Conclusions

The draft sequence provides us with the first comprehensive integration of diverse genomic resources. The mapping of ESTs, gene predictions, STSs and SNPs onto the draft sequence can enable identification of alternative splicing, orthologues, paralogues, map positions and coding sequence variations. Users should remember,

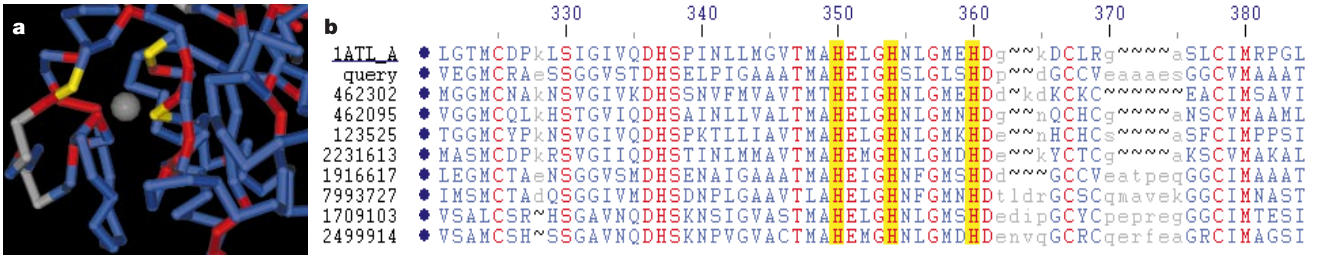

Figure 2 The ADAM23 protein sequence NP_003803.1 was used in a BLASTP search of the Ensembl confirmed peptides produced from the 5 Sep 2000 version of the working draft sequence. As of October 2000, the best match was to the peptide ENSP00000025626 derived from chromosome 2, the predicted peptide for ADAM23. The second match was to peptide ENSP00000072108, derived from chromosome 20 and falling within GenBank Acc. No. AC055771.2. We used this predicted peptide to search a database of Pfam[24] and SMART[25] protein domains that are aligned with protein structures. This Conserved Domain Database (CDD) search at NCBI resulted in hits to reprolysin and disintegrin domains. **a**, Pfam family 01421, Reprolysin, has a structure associated with it: the zinc-dependent metalloprotease Atrolysin C (PDB: 1ATL). **b**, The query ENSP00000072108 aligns with the 1ATL protein sequence and eight other ADAMs from the Pfam Reprolysin entry. In the structure and alignment, red indicates conserved residues; grey indicates non-aligned sequences. The three histidines of the metalloprotease active site, which complex with the zinc ion, are highlighted in yellow. The structure and alignment were created using the structure viewing program Cn3D.

though, that these genomic resources represent a work-in-progress, and will evolve as the genome is finished and computation methods further refined.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25,** 232–234 (2000).
3. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25,** 235–238 (2000).
4. McKusick, V. A. *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders* (Johns Hopkins Univ. Press, Baltimore, 1998).
5. Maglott, D. R., Katz, K. S., Sicotte, H. & Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28,** 126–128 (2000).
6. Pruitt, K. D., Katz, K. S., Sicotte, H. & Maglott, D. R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* **16,** 44–47 (2000).
7. Guigo, R., Agarwal, P., Abril, J. F., Burset, M. & Fickett, J. W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10,** 1631–1642 (2000).
8. Stormo, G. D. Gene-finding approaches for eukaryotes. *Genome Res.* **10,** 394–397 (2000).
9. Sagane, K., Ohya, Y., Hasegawa, Y. & Tanaka, I. Metalloproteinase-like, disintegrin-like, cysteine-rich proteins MDC2 and MDC3: novel human cellular disintegrins highly expressed in the brain. *Biochem. J.* **334,** 93–98 (1998).
10. Wolfsberg, T. G. & Landsman, D. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* **25,** 1626–1632 (1997).
11. Wolfsberg, T. G. & Landsman, D. in *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (eds Baxevanis, A. D. & Ouellette, B. F. F.) (Wiley-Liss, Inc., New York, 2001).
12. 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).
13. Chen, H. *et al.* Limb and kidney defects in Lmx1b mutant mice suggest an involvement of LMX1B in human nail patella syndrome. *Nature Genet.* **19,** 51–55 (1998).
14. Poindexter, K., Nelson, N., DuBose, R. F., Black, R. A. & Cerretti, D. P. The identification of seven metalloproteinase-disintegrin (ADAM) genes from genomic libraries. *Gene* **237,** 61–70 (1999).
15. Mitelman, F., Mertens, F. & Johansson, B. A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet.* **15,** 417–474 (1997).
16. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409,** 953–958 (2001).
17. Wooster, R. *et al.* Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12–13. *Science* **265,** 2088–2090 (1994).
18. Schuler, G. D. Sequence mapping by electronic PCR. *Genome Res.* **7,** 541–550 (1997).
19. Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* **28,** 352–355 (2000).
20. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
21. Struhl, K. Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev.* **12,** 599–606 (1998).
22. Finnin, M. S. *et al.* Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature* **401,** 188–193 (1999).
23. Leipe, D. D. & Landsman, D. Histone deacetylases, acetoin utilization proteins and acetylpolyamine amidohydrolases are members of an ancient protein superfamily. *Nucleic Acids Res.* **25,** 3693–3697 (1997).
24. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28,** 263–266 (2000).
25. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28,** 231–234 (2000).

Acknowledgements

We thank G. Marth, S. Sherry, D. Landsman, D. Church and D. Lipman for suggestions and review of the manuscript.

Correspondence should be addressed to G.D.S. (e-mail: schuler@ncbi.nlm.nih.gov).