GENOME ANNOTATION: FROM SEQUENCE TO BIOLOGY

Lincoln Stein

The genome sequence of an organism is an information resource unlike any that biologists have previously had access to. But the value of the genome is only as good as its annotation. It is the annotation that bridges the gap from the sequence to the biology of the organism. The aim of high-quality annotation is to identify the key features of the genome — in particular, the genes and their products. The tools and resources for annotation are developing rapidly, and the scientific community is becoming increasingly reliant on this information for all aspects of biological research.

WORKING DRAFT
A 'working draft' sequence has come to mean a genomic sequence before it is finished. Working draft sequences contain multiple gaps, unrepresented areas and misassemblies. In addition, the error rate of working draft sequence is higher than the 1 in 10,000 error rate that is standard for finished sequence.

FASTA FILE
A common file format used for the storage and transfer of sequence data. It contains raw DNA or protein sequence, but no annotation information.

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA.
e-mail: lstein@cshl.org

For thousands of years, rabbis have laboured over the text of the Torah, seeking to make this cryptic, uneven and internally contradictory text into a coherent system of law, and storing this commentary into an annotated version of the text, known as the Talmud. Over time, the amount of annotation in the Talmud has greatly exceeded the original text — each line of the Torah is now surrounded by layers of commentary in an onionskin fashion.

So it is with the genome. The past decade has seen the completion of numerous whole-genome-sequencing projects, beginning with microbial genomes[1–3] and continuing with the eukaryotic species *Saccharomyces cerevesiae* (yeast)[4], *Caenorhabditis elegans* (worm)[5], *Drosophila melanogaster* (fruitfly)[6], *Arabidopsis thaliana* (mustard weed)[7], and culminating most recently with the announcement by public and private groups of WORKING DRAFT versions of the human genome[8,9]. Other genomes are either on the way or contemplated, including mouse, rat, zebrafish, pufferfish and non-human primates.

Like the torah, the genome represents a mixture of evolutionary 'intent' (insofar as natural selection can be said to intend anything) and historical accident. Although there is assuredly an underlying rhyme and reason to the genome, there is also much that is haphazard. Fragments of viral and prokaryotic genomes that infected ancestral individuals, mobile elements, pseudogenes and repetitive elements, are all traps (or opportunities) waiting to surprise the genome scientist. More importantly, principal aspects of the basic organization of the genome are still quite murky: the regulation of alternative splicing, the control of transcription, the role of intergenic material, and the function of many non-coding RNAs, to name just a prominent few.

The attention of the sequencing community is now focused on genome annotation — the process of taking the raw DNA sequence produced by the genome-sequencing projects and adding the layers of analysis and interpretation necessary to extract its biological significance and place it into the context of our understanding of biological processes. Genome annotation itself is a multi-step process, falling more or less neatly into three categories: nucleotide-level, protein-level and process-level annotation (FIG. 1).

This review surveys the various ways that genome annotation is carried out, the techniques used and the diverse sociological models that have been adopted to organize the annotators.

Nucleotide-level annotation

Yet another eukaryotic genome is complete, and someone hands you a big unannotated FASTA FILE (three billion base pairs in the case of a mammalian species, or the content of about five CD-ROMs). What now?

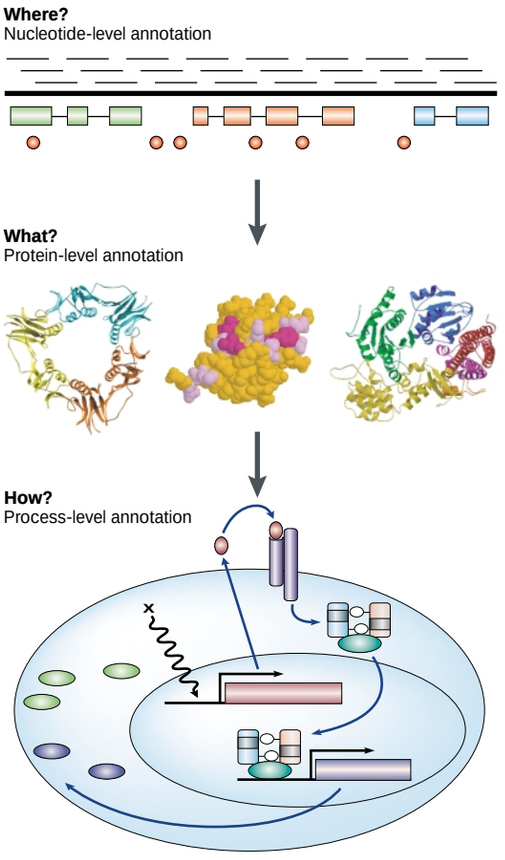

Figure 1 | **The three layers of genome annotation: where, what and how?**

Mapping. The first step in genome annotation is to identify the punctuation marks. Where are the known genes, genetic markers and other landmarks previously identified by genetic, cytogenetic or RADIATION HYBRID MAPPING? Where are the tRNAs, rRNAs and other non-translated RNAs? Where are repetitive elements? Is there evidence for ancient duplications in the genome and, if so, where are the end points of the putative duplicated regions? All these questions are really an extended form of physical mapping, attempting to convert the *terra incognita* of raw DNA sequence into a set of easily recognized landmarks and reference points.

Along with 'gene finding', the principal activity of this phase of annotation is identifying and placing all known landmarks into the genome. For example, during nucleotide-level annotation, annotators will search for known genetic markers, radiation hybrid markers and CLONE ENDS in the sequence, and place them; they thereby form bridges between the genomic sequence and pre-existing genetic, radiation hybrid and physical maps. This provides a path to connect the pre-genomic literature, which is often based on such landmarks, with post-genomic research.

RADIATION HYBRID MAPPING
An experimental technique that uses radiation-induced chromosomal breakpoints in somatic-cell hybrids to map the positions of sequence tagged sites (STSs).

CLONE ENDS
Genomic sequencing projects typically sequence the ends of bacterial artificial chromosome and plasmid clones, in addition to shotgun sequencing entire clones. During assembly, the clone end sequences are used to create a scaffold on which the genome sequence is pieced together.

Finding genomic landmarks. Finding landmarks is a relatively straightforward task. Short sequences, such as PCR-based genetic markers, can be identified rapidly using the e-PCR program[10]. Longer sequences, such as restriction-fragment length polymorphism markers, can be found using BLASTN[11], SSAHA[12] or another rapid sequence-similarity searching algorithm. (See BOXES 1 and 2 for information about the resources and software tools discussed in this article.)

In the case of the human working draft, an important bridging activity was the integration of the sequence with the cytogenetic map, much beloved by clinical geneticists and hunters of disease genes. In this case, the human cytogenetic map was integrated with the genome by systematic fluorescence *in situ* hybridization mapping of bacterial artificial chromosome (BAC) clones against metaphase chromosomes[13]. The assembly of the fly genome was aided immeasurably by *in situ* hybridization of each BAC in the physical map to polytene chromosomes, providing bridges between the cytogenetic and physical maps with an accuracy of a few hundred kilobases[6].

Gene finding. Gene finding is the most visible part of this phase. In small prokaryotic genomes, gene finding is largely a matter of identifying long open reading frames (ORFs). Even here, however, ambiguities arise if long ORFs overlap on opposite strands, and the true coding region must be sorted out. As genomes get larger, gene finding becomes increasingly tricky. The main issue is the signal-to-noise ratio. In a prokaryotic genome, such as *Haemophilus influenzae*, 85% of its 1.8-Mb genome is in coding regions. The corresponding number in yeast is not much lower, at 70%. For these genomes, 'calling genes' is an exercise in running a computer program that carries out a six-frame translation and identifies all ORFs that are longer than a chosen threshold. But even in these small genomes, finding genes has not been entirely effortless. The number of predicted yeast genes, for example, took several years to settle down, and there are still several short ORFs that have an uncertain status as bona fide genes.

In the fly and the worm, however, less than 25% of the genome is in coding regions, and the number falls to just a few per cent in humans. The process of finding genes is further complicated by the presence of splicing and alternative splicing. In the human genome, a typical exon is 150 bp and a typical intron is several kilobases, and there is no clear delineation between the intergenic regions that separate adjacent genes and the intragenic regions that separate exons. Defining the precise start and stop position of a gene and the splicing pattern of its exons among all the non-coding sequence is like finding a very small and indistinct needle in a very large and distracting haystack.

Several sophisticated software algorithms have been devised to handle gene prediction in eukaryotic genomes, including GENSCAN[14], Genie[15], GeneMark.hmm[16], Grail[17], HEXON[18], MZEF[19], Fgenes[20] GeneFinder (P. Green, unpublished data)

and HMMGene[21]. These algorithms typically consist of one or more SENSORS that attempt to adduce the presence of a gene feature from motifs or statistical properties of the DNA. For example, as transcribed regions are associated with (G+C)-rich regions, a sensor for transcriptional start sites might measure the G+C content of the region being scanned. A sensor for splice sites compares the current region to splice consensus sequences.

Some gene predictors stop with the prediction of a single feature, such as the exon predictors HEXON and MZEF. Most, however, attempt to use the output of several sensors to generate a whole gene model, in which a gene is defined as a series of exons that are coordinately transcribed. This is typically done with NEURAL NETWORKS (Grail), a RULE-BASED SYSTEM (GeneFinder) or, increasingly favoured, with a HIDDEN MARKOV MODEL (HMM) (GenScan, Genie, HMMGene, GeneMark.hmm and Fgenes). The HMM approach has the advantage of explicitly modelling how the individual probabilities of a sequence of features are combined into a probability estimate for the whole gene (FIG. 2).

Despite great progress, however, gene prediction entirely based on DNA analysis is still far from perfect. In the recent comparison of gene-prediction programs reported by Reese *et al.*[22], the authors of nearly a dozen algorithms were asked to predict genes in two well-annotated regions of the fruitfly genome. The best algorithms could achieve sensitivities (a measure of the ability to detect true positives) and specifities (a measure of the ability to discriminate against false positives) of ~95 and ~90%, respectively, when asked to predict whether a particular nucleotide is in an exon. However, accuracy dropped off rapidly if the criterion was changed to calling the boundaries of an exon correctly, and still further if the algorithm was required to predict the entire gene structure correctly. Under the latter requirements, the best gene predictors had a sensitivity of 40% and a specificity of 30%, meaning that most of the genes predicted by these programs contain errors ranging from an incorrect exon boundary to a missed or phantom exon. Between 5 and 15% of genes were missed entirely in this contest.

Although there has not yet been an equivalent comparison of gene-prediction programs on the human genome, it is safe to assume that these programs will fare less well because of the lower signal-to-noise ratio. As expected, one study showed that the GENSCAN accuracy dropped rapidly as intergenic lengths in a simulated data set increased[23].

Fortunately, we do not have to rely completely on *AB INITIO* GENE PREDICTION programs. The similarity of a region of the genome to a sequence that is already known to be transcribed is the single most powerful predictor of whether a sequence is transcribed. A nucleotide match to a cDNA, expressed sequence tag (EST), or even a BLASTX match to a gene in another species is good evidence that a region belongs to a gene.

However, the process of deriving a complete gene model from one or more sequence similarities is not nearly as straightforward as it might sound. For one thing, pseudogenes are a common feature of eukaryotic genomes. Many similarity-based gene-prediction algorithms require evidence that the gene is spliced and that the splices maintain an in-phase ORF. However, this criterion biases gene prediction against single-exon genes. And there are many additional problems: ESTs are fragmentary and might suffer from several artefacts, including contamination with genomic DNA, chimerism and lane-tracking errors during automated sequencing; cDNA sequences might contain repetitive elements that will cause spurious genomic matches; similarities to proteins in other species might suffer from evolutionary divergence or the orthologue–paralogue problem (discussed in more detail below); and the presence of alternative splicing considerably complicates the interpretation of alignments between genomic DNA, cDNAs and ESTs. More seriously, however, similarity data is never complete. Even the most comprehensive EST projects will miss low-copy-number transcripts and those transcripts that are expressed only transiently or under unusual conditions.

The current trend in gene prediction is to make as much use of sequence-similarity data as possible. The latest generation of gene-prediction algorithms, such as Grail/Exp, Genie EST and GenomeScan[24], combine *ab initio* predictions with similarity data into a single probability model. In Reese *et al.*[22], the algorithms that took similarity data into account generally outdid those that did not take them into account.

Because of the newness of the combined algorithms, however, most genome-wide gene-annotation systems so far have run sequence-similarity searches and *ab initio* gene predictors separately, then combined and reconciled the predictions later. In the case of the worm genome, this reconciliation was initially carried out by curators who manually examined each gene prediction in the context of matching ESTs and homologues from other species[5]. More recently, the process has been accelerated significantly by automated procedures for reconciling EST alignments with gene predictions (ACEmbly, J. Thierry-Mieg, unpublished data and REF. 25), and by systematically PCR-amplifying a cDNA library using primer pairs that span predicted genes[26].

In the case of the human working draft, both Celera[8] and the public sequencing consortium[9] developed automated rules-based gene-prediction systems that attempted to mimic how a human annotator might examine a sequence. The Celera system, a proprietary software package called Otto, gives sequence similarity the highest priority, drawing evidence that a region is transcribed from sequence similarities found in the RefSeq library of well-characterized human genes[27], from the Unigene set of human ESTs[28], and from SWISS-PROT[29] and other protein databases. Otto then uses GENSCAN to find and refine the splicing pattern of the predicted gene.

SENSORS
An algorithm specialized to identify a feature of a sequence, such as a possible splice site.

NEURAL NETWORK
Neural networks are analytical techniques modelled after the (proposed) processes of learning in cognitive systems and the neurological functions of the brain. Neural networks use a data 'training set' to build rules that can make predictions or classifications on data sets.

RULE-BASED SYSTEM
A type of computer algorithm that uses an explicit set of rules to make decisions.

HIDDEN MARKOV MODEL
A type of computer algorithm that represents a system as a set of discrete states and transitions between those states. Each transition has an associated probability. Markov models are 'hidden' when one or more of the states cannot be directly observed.

AB INITIO GENE PREDICTION
A class of software that attempts to predict genes from sequence data without the use of prior knowledge about similarities to other genes.

Box 1 | Resources and tools I

The following list provides brief descriptions of some of the software tools and resources mentioned in the article. Online, links are provided to these resources from this box, and from the text.

BLASTN, BLASTX,
BLASTP, PSI-BLAST . . . . . http://www.ncbi.nlm.nih.gov/BLAST/
This family of sequence-similarity search tools allows you to rapidly search a query protein or nucleotide sequence against a large database of sequences, to identify sequences that are similar to the search sequence.

Ensembl . . . . . http://www.ensembl.org
This web site, a joint project of the European Bioinformatics Institute and the Sanger Centre, seeks to make available a high-quality, consistent set of annotations on the human and mouse genomes.

e-PCR . . . . . http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi
BLAST does not work well with short sequences, such as PCR primers. The ePCR program was developed to search rapidly for a primer pair (using their sequences and known physical separation) in a large sequence, such as a genome.

FlyBase . . . . . http://www.flybase.org/
A fully realized model organism system database, presenting curated annotations on the *Drosophila melanogaster* genome, as well as rich information on the genetics of the organism, mutant strains and molecular resources.

GeneMark.hmm . . . . . http://genemark.biology.gatech.edu/GeneMark/
Another gene-prediction program that uses hidden Markov models (HMMs). The online version supports numerous eukaryotic and prokaryotic genomes.

Genie . . . . . http://www.fruitfly.org/seq_tools/genie.html
The gene-prediction program used to annotate genes in *Drosophila melanogaster*. The online version has been trained for human and *Drosophila* sequence.

GENSCAN. . . . . http://genes.mit.edu/GENSCAN.html
This is probably the most widely used gene-prediction program. It uses HMMs to predict the presence of a gene given the raw DNA sequence. The online version provides prediction services for vertebrates, *Arabidopsis thaliana* and maize.

Grail http://compbio.ornl.gov/Grail-1.3/
One of the oldest gene-prediction programs still in use, this software uses a neural network to predict genes. The online version provides gene-prediction services for human, mouse, *Arabidopsis*, *Drosophila* and *Escherichia coli* sequences.

The Human Sequencing Consortium, based on the Ensembl gene annotation system[9], took almost the reverse approach, beginning with *ab initio* gene predictions from GENSCAN and then strengthening the predictions using nucleotide and protein similarities. These predicted gene models were merged and reconciled with the output of Genie EST, and finally merged with the contents of the RefSeq library.

Although the two groups approached the gene finding problem from different directions, both gave greater weight to cDNA and EST alignments than to *ab initio* gene prediction. So, it is not too surprising that the estimates from both groups of the number of genes were very close, ~30,000.

Non-coding RNAs and regulatory regions. There is much more to the genome than coding regions. On the cutting edge of nucleotide-level annotation is the search for non-coding RNAs and transcriptional regulatory regions. Non-coding RNAs include tRNAs, rRNAs, small nucleolar RNAs and small nuclear RNAs. rRNAs can be found easily by similarity searching, but the rest are tricky, because of both their short length and their nucleotide diversity. tRNAs are amenable to *de novo* prediction through algorithms that search for characteristic structural signatures, such as hairpin formation[30–32]. The most widely used tRNA prediction program is tRNAScan-SE[32], which combines several algorithms to identify tRNAs with high accuracy in good running times. It can also distinguish active tRNAs from tRNA pseudogenes. This program was used during the recent annotation of the public human sequence to identify 497 tRNAs and 324 putative pseudogenes[9].

Other non-coding RNAs, such as telomerase RNA and the U1–12 series of spliceosome RNAs, can be identified by sequence similarity, but there are likely to be many non-coding RNAs that have not yet been identified[33]. Just coming online now are new algorithms based on identifying characteristic patterns of mismatched base pairs in cross-species alignments, for example mouse and human (S. Eddy, personal communication). Preliminary results indicate that there might be hundreds of previously unrecognized non-coding RNAs in the genome. It will be fascinating to learn the role and function of non-coding RNAs discovered in this way.

The situation is similar with regulatory regions (reviewed in REF. 34). A relatively small number of transcriptional-factor-binding sites have been identified by classical experimental methods. The sequences for these sites are available in curated databases such as TRANSFAC and PROSITE, and can be found in genomic sequence by applying similarity search methods that are specialized for short motifs. However, as with the non-coding RNAs, the number of known regulatory regions is almost certainly a small fraction of what is out there.

The development of algorithms to search for regulatory regions is a hot research topic in bioinformatics. One popular class of algorithms, exemplified by the MEME program[35], searches for motifs in nucleotide and protein sequences that occur more often than would be predicted by chance. However, the predictions from such algorithms are strongly dependent on the care with which the input sequences are chosen, and are typically used strictly as the starting point for experimental confirmation.

However, it is widely anticipated that nucleotide-level similarity comparisons among two or more related species will allow us to identify control and regulatory regions with notably greater power and accuracy. When orthologous genomic regions of two closely related species, such as *C. elegans* and *Caenorhabditis briggsae*[36,37], mouse and human[38,39], and human and chimpanzee[40], are examined, conserved blocks of nucleotide similarity upstream from the transcriptional start site are commonly found, indicating possible evolutionarily conserved regulatory regions. With the mouse genomic sequence now becoming available, there should be a great explosion in such studies over the next few years.

Box 2 | Resources and tools II

HMMER . . . . . http://hmmer.wustl.edu/
HMMER is a protein-similarity search engine. It uses HMMs to provide greater sensitivity when searching for evolutionarily distant proteins. It is more accurate than BLASTP, but notably slower.

HMMGene . . . . . http://www.cbs.dtu.dk/services/HMMgene/
Another HMM-based gene-prediction program. This one provides options for vertebrate and *Caenorhabditis elegans* sequences.

InterPro . . . . . http://www.ebi.ac.uk/proteome/
The InterPro database is on its way to becoming the prime resource for protein-level annotation.

Proteome, Inc. . . . . http://www.proteome.com/
An example of the 'cottage industry' approach to annotation. The curated databases at this site include protein databases for *C. elegans*, yeast, human and mouse.

RefSeq and LocusLink, National Center for Biotechnology Information (NCBI) . . . . . http://www.ncbi.nlm.nih.gov/
In addition to the GenBank database of nucleotide and protein sequences, and the PubMed bibliographic search service, NCBI hosts RefSeq and LocusLink, two curated databases of nucleotide- and protein-level annotation.

RepeatMasker. . . . . http://www.genome.washington.edu/uwgc/analysistools/repeatmask.htm
RepeatMasker is the most widely used tool for identifying and masking repeats in genomes. It operates using a large database of repeat consensus sequences previously identified in various eukaryotic species.

SSAHA. . . . . http://www.sanger.ac.uk/Software/analysis/SSAHA/
SSAHA is a new, fast algorithm for searching for nearly identical nucleotide sequences. It makes it possible to quickly match a partial DNA sequence to a large genome, such as in the human.

SWISS-PROT and SWISS-PROT TrEMBL . . . . . http://www.ebi.ac.uk/swissprot/
The SWISS-PROT database is a high-quality, curated database of protein sequences from all species. SWISS-PROT TrEMBL is an automated merge of SWISS-PROT with predicted proteins from genomic sequencing and other sequencing projects. A rule-based system minimizes the amount of redundancy and errors in the merged database.

UCSD Genome Browser . . . . . http://genome.ucsc.edu/
This web site, run by graduate student Jim Kent, presents an integrated view of annotations on the human genome contributed by numerous groups.

Identifying repetitive elements. Repetitive, or interspersed, elements are an important feature of eukaryotic genomes, and indeed account for a large proportion of the variation in genome size. In the human genome, the largest genome sequenced so far, 44% of nucleotides are contributed by repetitive elements[9]. The repetitive elements are derived from active transposable elements and are frequently dismissed as parasitic or 'junk' DNA.

Identifying and mapping repetitive elements actually starts well before any other annotation activities begin, because of the need to identify and exclude repetitive regions during the genome assembly process. Known repetitive elements are masked from the sequence using a program such as RepeatMasker (A. F. A. Smit and P. Green, unpublished data), or are identified and excluded by measuring the apparent coverage of the region being assembled. Once a family of repetitive elements is identified, it is relatively straightforward to find other members of the family by nucleotide-based sequence-similarity searching.

However, repetitive elements are more than a nuisance for genome assembly. The study of repeats is itself a fascinating discipline, which has revealed a virtual bestiary of repeat families, each with unique distinguishing features and survival strategies. There were several surprises involving repetitive elements during the annotation of the human genome. One of the more interesting came from the analysis of the apparent age of representatives of each family, in which age was determined by counting the mutations that had accumulated in the representative relative to the canonical sequence. This analysis[9] detected a marked and unexplained decline in transposable element activity over the past 33–50 Myr of human evolution. Another unexpected outcome of this annotation activity was the finding of a shift of older *ALU*-type transposable elements from (A+T)-rich regions, which are favoured by the transposition mechanism, to gene-rich (G+C)-rich regions — a finding that raises the possibility that Alu elements are subject to positive selection. Clearly, much is to be gained from a closer examination of this 'junk' DNA.

Mapping segmental duplications. Distinct from identifying and mapping repetitive elements is the identification of large segmental-duplications in the genome.

One of the big surprises that emerged during the sequencing of the mustard weed genome was evidence for several distinct large-scale duplication events in the organism's past[7,41]. Over 60% of the predicted ORFs in *Arabidopsis* match a paralogue somewhere else in the genome, and these duplicated genes are organized into large syntenic blocks that might be several hundred ORFs long (FIG. 3).

An analysis of the apparent age of these duplications, on the basis of sequence divergence, indicates four distinct segmental-duplication events in the genome, occurring between 100 and 200 Myr ago[41] — a period of time that spans the diversification of the ANGIOSPERMS. This helps to explain the paucity of conserved synteny between the genetic maps of rice (a MONOCOTYLEDON) and *Arabidopsis* (a DICOTYLEDON). So, the study of segmental duplications can reveal important information about the evolutionary history of a species.

The human genome itself has blocks of segmental duplication, although none so marked as mustard weed. The Celera and sequencing consortium papers indicate that 5% of the genome might be involved in ancient segmental duplications, and there is preliminary evidence that the true number might be significantly higher (E. Eichler, personal communication).

Mapping variations. The last principal activity of nucleotide-level annotation is identifying and mapping polymorphisms. Single-nucleotide polymorphisms (SNPs) have emerged as a valuable tool for genetic mapping, population genetics studies and clinical diagnosis[42].

SNPs had a prominent role in the recent annotations of the human genome[8,43]. In principle, it is easy to identify SNPs by simply aligning the genomic

ALU SEQUENCE
A dispersed, intermediately repetitive DNA sequence found in the human genome in about 300,000 copies. The sequence is about 300 base pairs long. The name Alu comes from the restriction endonuclease (*Alu*I) that cleaves it.

ANGIOSPERM
Flowering seed plant.

MONOCOTYLEDON
One of the two principal classes of flowering plant, monocots are characterized by a single cotyledon (primitive leaf) in the embryonic plant. Maize, rice, wheat and other grasses are common monocots.

sequences of two or more individuals and finding places where the sequence of one diverges from the other. In practice, SNP-calling algorithms must distinguish biological variations from those due to sequencing errors. Several algorithms have been developed to do this[12,43,44]; each uses sequence quality data to make a probability estimate.

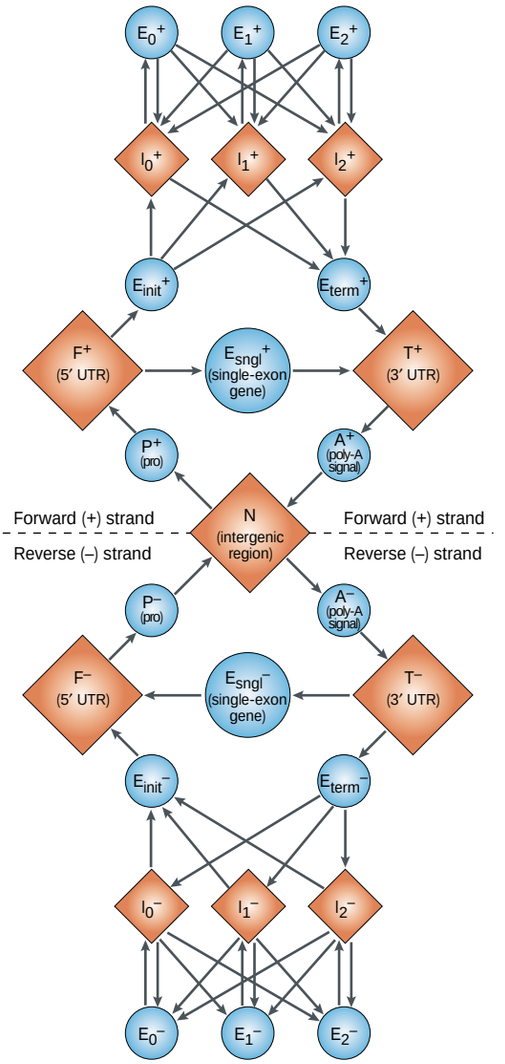

Figure 2 | **A hidden Markov model explicitly models the probabilities for the transition from one part of a gene to another.** In this model, used by the GENSCAN algorithm, each circle or diamond represents a functional unit in the gene. For example, E_{init}^+ is the initial exon and E_{term}^+ is the last. The arrows represent the probability of a transition from one part of a gene to another. The algorithm is 'trained' by running a set of known genes through the model and adjusting the weights of each transition to reflect realistic transition probabilities. Thereafter, test sequence data can be run through the model one base position at a time, and the model will read out the probability of a gene being present at that position. The states that occur below the dashed line correspond to a gene in the reversed strand, and thus are symmetric with those above the line. E, exon; I, intron; UTR, untranslated region; pro, promoter. (From REF. 14 © Academic Press Ltd (1997).)

Sequence data for SNP calling can come from various sources. In the case of *C. elegans*, SNPs have been identified by shotgun sequencing DNA from various wild-type isolates (*C. elegans* SNP data[45]) and then aligned to the reference genomic sequence. In the case of the human sequence, SNPs have been called by aligning ESTs to genomic sequence and by pooling and shotgun sequencing the genomic DNA of a panel of unrelated individuals, followed by alignment of the sequence data to the human working draft[43]. Numerous SNPs came from within the human working draft itself by identifying variations in regions of clone overlap. In the case of the Celera draft, a total of five unrelated donors were used to create the libraries used for shotgunning and assembly. In the case of the public consortium sequence, 75% the sequence was contributed by a single individual, and the polymorphisms found were presumably because of heterozygosity.

Annotation of human SNPs have yielded some unexpected results. The most intriguing finding is that the distribution of SNPs departs significantly from what would be predicted under the standard population genetics models. Some regions of the genome contain SNP hot spots and others are SNP poor. What combination of historical, structural or selective pressures are responsible for this uneven distribution, and how is it related to other factors, such as the distribution of genes and repetitive elements?

Protein-level annotation

After asking 'where', genome annotators ask 'what'. This stage of genome annotation seeks to compile a definitive catalogue of the proteins of the organisms, to name them and to assign them putative functions.

H. influenzae has 1,709 genes, yeast has ~5,600, and fly, worm, mustard weed and human have ~13,000, ~19,000, ~24,000 and >30,000, respectively. Of these gene sets, only a small fraction correspond to known, well-characterized proteins. Faced with numerous proteins of unknown function, annotators generally begin by classifying them into more manageable groups or protein families, and by using similarities to better-characterized proteins of other species.

This process sounds easier than it is in practice. The intrinsic problem comes from the nature of the evolutionary process. During the evolution of a protein family, an ancestral protein is duplicated one or more times, and the copies diverge, forming a family of related proteins known as paralogues. However, similarity in function does not inevitably follow from sharing a common ancestor, and there are many cases of two protein family members that have strikingly divergent functions. For example, the lens crystallins are derived from a family of proteins that ordinarily function as enzymes and CHAPERONINS[46,47]. Further complicating the issue is the proclivity of genes to pick up or lose functional domains during the course of evolution, creating chimeric proteins that share two or more unrelated ancestors[48].

DICOTYLEDON
One of the two principal classes of flowering plant, dicots are characterized by two cotyledons (primitive leaves) in the embryonic plant. Tomatoes, maple trees and mustard are common dicots.

CHAPERONINS
A class of ring-shaped, heat-shock proteins that have a key role in protein folding and protection from stress.

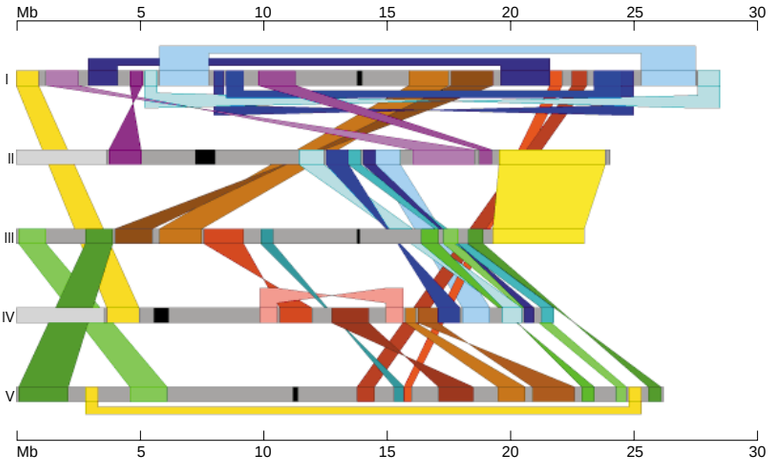

Figure 3 | **Segmental duplications.** The *Arabidopsis thaliana* genome is dominated by large regions of segmental duplication. The horizontal bars represent the five *Arabidopsis* chromosomes (I–V). Coloured bands connect similar regions. In some cases, the duplications are on the same chromosome, and in others, the duplications have been inverted (twisted bands). Light grey regions are the nucleolus organizer regions, black represents the centromeres. Scale is in megabases. (From Dirk Haase, MIPS Institute for Bioinformatics, GSF National Research Centre for Environment and Health, Munich, Germany. Kindly provided by the MIPS *Arabidopsis thaliana* database.)

The comparison of proteins between species is a rich source of functional annotation. For example, if a well-characterized yeast protein is known to be involved in the initiation of DNA replication, then it is likely that a protein predicted from the human genomic sequence that is similar to the yeast protein will have the same function. However, the nature of protein evolution again lurks to ambush the unwary annotator. The human gene might be directly descended from a common ancestor of the yeast gene, in which case it is called an orthologue, or it might be descended from a duplicated and diverged copy of the gene, in which case it is a paralogue. In this case, it would be a mistake to conflate the yeast with the human protein and assume that they have the same functional role.

Although various techniques have been developed to identify and cluster groups of orthologous proteins in an automated fashion, for example the COG system[49], many predicted proteins evade unambiguous automated classification into an orthologous group. In practice, what is typically done now is to classify predicted proteins on the basis of functional domains, folds and motifs, as well as by broad similarity to better-characterized proteins. The sorting out of protein phylogeny then follows on as a slower research activity.

A typical protein annotation pipeline will search for similarities using the BLASTP or PSI-BLAST tools[50] against several different databases of protein sequences. Among the most valuable of the whole-protein sequence collections are SWISS-PROT and SWISS-PROT TrEMBL[28]. The former is a curated collection of confirmed protein sequences (86,593 as of release 39), which have been extensively annotated and cross-referenced with other sequence and structure databases. SWISS-PROT annotations include bibliographic references, descriptions of the function and biological role of the protein, protein family assignments and pointers to structural data, if available. Because of its curated nature, the protein sequences contained in SWISS-PROT are likely to be of high reliability.

However, the accelerated rate of genome sequencing has outstripped the speed of the curators of SWISS-PROT. This gap is filled by SWISS-PROT TrEMBL, an automated translation of coding DNA sequence (CDS) entries submitted to the nucleotide databases. It acts as the 'pre-annotation' source for SWISS-PROT. TrEMBL is filtered to remove redundancy within itself and between TrEMBL and SWISS-PROT, and is then subject to a first-pass search for functional domains using the tools described below. The current version of TrEMBL (release 16) contains 489,620 sequences, an order of magnitude larger than SWISS-PROT.

A complementary approach is to search against databases of functional domains. Among the commonly used databases are PFAM[51], a collection of HMM profiles and alignments for common protein families. The current release (6.1, March 2001) contains 2,727 protein family entries, and is searchable using the HMMER software tool[32]. A typical PFAM entry is 'TAZ zinc finger', the motif that allows the cAMP-response-element-binding protein (CREB) to bind to other transcription factors, such as p53, and to regulate the initiation of transcription. Other databases commonly used during automated protein annotation include the following: PRINTS[52], a compendium of short protein motifs that captures common protein folds and domains; PROSITE[53], a database of longer protein signatures known as profiles; ProDom[54], a collection of protein domains derived from the PSI-BLAST procedure[50]; the SMART curated collection of protein domains[55]; and BLOCKS[56], a database of conserved protein regions and their multiple alignments.

The various protein family, domain and motif databases are highly overlapping, but differ in their nomenclature, their search methods and their suitability for diverse tasks. This makes it difficult to interpret the results when a predicted protein hits entries in several of the databases. Have they hit the same thing?

An important development in recent years has been an intensive effort to integrate the protein signature databases into the unified InterPro resource[57]. InterPro is a cross-referencing system for equivalent entries in the PFAM, PRINTS, PROSITE, ProDom, BLOCKS and SMART databases. This resource allows genome annotators to run a predicted protein against each of the member databases, collect the matching domains and families, and then translate each into its corresponding unique InterPro entry. The most recent InterPro release (3.0, March 2001) contains 3,591 entries that correspond to 2,628 protein families, 888 domains, and 75 repeats and post-translational modification sites. Each InterPro entry contains a brief description of the family or domain, a list of SWISS-PROT and TrEMBL proteins that contain the entry, literature references and outgoing links to the corresponding entries of the member databases.

InterPro has been used as the basis for several protein annotation projects for the yeast, worm, fruitfly, *Arabidopsis* and human genomes. In these organisms, between 40 and 50% of predicted proteins have a match to at least one InterPro entry. More than half of the predicted proteins in eukaryotes belong to novel protein families, highlighting how much still needs to be learned.

Process-level annotation

The last and, in many ways, most challenging part of genome annotation is relating the genome to biological processes. How do the building blocks of genes and proteins relate to the cell cycle, cell death, embryogenesis, metabolism, and the maintenance of health and disease? Functional annotation, as it is sometimes called, has long been a feature of genome-sequence analysis. The publication of each new genome is inevitably accompanied by a table or a pie-chart showing the distribution of proteins classified by function, for example 'metabolism' and 'cytoskeleton'. Until recently, what had been lacking in these analyses was a commonly accepted classification scheme that combined the breadth required to describe biological functions among diverse species with the specificity and depth needed to distinguish a particular protein from other members of its family. The lack of such a standard hampered the ability to relate genes that were annotated by different research groups, particularly when crossing species borders.

A breakthrough of sorts came two years ago, when three model organism databases, the *Saccharomyces* Genome Database[58], FlyBase[59] and the Mouse Genome Database[60], formed a consortium to create a Gene Ontology (GO)[61]. The GO is a standard vocabulary for describing the function of eukaryotic genes. It consists of three subparts: molecular function, biological process and cellular component. Molecular function terms describe the tasks carried out by individual gene products, such as its enzymatic activity. Biological process terms are used for broader biological goals, such as meiosis. Cellular component terms describe genes in terms of the subcellular structures they are localized to, such as organelles, as well as the macromolecular complexes they belong to, such as the ribosome.

The elegance of the GO is that it is organized as a hierarchy of terms (or more accurately, as a DIRECTED ACYCLIC GRAPH that allows a term to appear in several places in the hierarchy). More general terms, such as 'enzyme', lead to more specific terms, such as 'lyase', 'carbon–oxygen lyase', 'hydro-lyase' and 'threonine dehydratase'. This flexibility allows genes to be annotated to whatever level of specificity the current biological understanding allows. A protein that is clearly an orthologue of a threonine dehydratase in another species can be annotated with the most specific term, whereas another protein that is clearly in the lyase family, but the enzymatic activity of which has not been confirmed, can be labelled with one of the more generic terms. This design also allows the GO to become increasingly 'bushy' as more specialized terms are added to the existing hierarchy. To give a specific example of how this works, the enzyme phosphatidate cytidylyltransferase (EC 2.7.7.41), is described by the GO function 'phosphatidate cytidylyltransferase' (accession GO:0004605), the process 'phospholipid biosynthesis' (accession GO:0008654) and the component 'membrane fraction' (GO:0005624).

Recently, other model organism databases have joined the GO Consortium, including the *Arabidopsis* Information Resource[62], the *C. elegans* database WormBase[63] and the fission yeast database PomBase. Each of these genome databases is annotating the confirmed and predicted genes in its corresponding species using GO terms, allowing quick comparison and cross-referencing between them. GO has also been adopted by several groups for use in annotating human genes. GO-term annotations are a feature of the InterPro database mentioned earlier, and are used in the Proteome databases and in the RefSeq database[26].

Process-level annotation extends well beyond purely computational work. High-throughput techniques, such as transposon-mediated insertional mutagenesis (reviewed in REF. 64), microarray expression analysis[65], RNA INTERFERENCE[66], direct assay of protein-expression levels by mass spectroscopy[67], screening for protein–nucleotide interaction sites using systematic evolution of ligands by exponential enrichment (SELEX)[68] and other techniques, green-fluorescent-protein-based assays for determining the anatomic and temporal patterns of gene expression[69] and yeast two-hybrid studies[70], all provide vital clues to the role that genes and proteins have in biological processes, and provide a rich layer of annotation. Indeed, taken a step further, conventional bench research and genome annotation begin to merge. Each experiment adds an item of information to our knowledge of biology, and this in turn enhances our understanding of the genome through the genes and proteins it touches.

The sociology of genome annotation

Annotation at the nucleotide, protein and process levels form the basis of genome annotation. All that remains is marshalling the technology, manpower and computational resources to undertake the mammoth task of annotating a new genome. Inevitably, this leads our discussion away from the science of annotation to its sociology.

Organizing genome annotation efforts. Genome annotation generally follows one of several organizational models: the factory, the museum, the cottage industry and the party. Each model is better suited to a different phase of the annotation task.

During the first phase of annotation, when the primary concern is to find genes, to map variations and to identify other structural landmarks, the factory model works best. This model, best exemplified by the Ensembl project, relies on a high degree of automation. The Ensembl annotation system is built around a cluster of 500 computers and a pipeline of annotation

DIRECTED ACYCLIC GRAPH (DAG). A type of hierarchy similar to the outline of a paper in that it has headers, subheaders and sub-subheaders. The main difference from a strict hierarchy is that each topic in a DAG is allowed to have more than one parent topic.

RNA INTERFERENCE
A phenomenon in which the expression of a gene is temporarily inhibited when a double-stranded complementary RNA is introduced into the organism.

SELEX
This is an *in vitro* selection method in which very large collections of oligonucleotides can be screened for specific functions.

Figure 4 | **An example of genome annotation.** The Ensembl web site is a rich source of annotations on the human genome. This view shows the positions of predicted genes, homologous regions of the mouse, single nucleotide polymorphisms, repetitive elements, and the locations of clone ends and other structural landmarks. (From the Ensembl Web site, a collaboration between EMBL-EBI and the Sanger Centre.)

programs. A sequence entering the pipeline is run through a suite of gene-prediction programs, various nucleotide- and protein-based similarity searching algorithms, and the protein-domain search programs described above. The results are then collated by a rule-based system into the gene predictions that are published on the Ensembl Web site. This design leads to a deliberately broad but shallow 'baseline' annotation of the genome (FIG. 4). As a consequence, the bulk of the Ensembl development team are engineers and computer scientists.

In contrast to the factory is the museum model, typically seen in later phases of genome annotation, when the emphasis shifts from finding the genes to interpreting their functional roles. In the museum model, a set of curators catalogue and classify the genome in a systematic fashion, finding and correcting mistakes made by the gene-prediction and functional annotation algorithms. In contrast to the earlier phase, much of this work is done by hand, and much of it is orientated towards capturing the current and past scientific literature on the organism and integrating it with the genomic sequence.

The museum model is favoured by the model organism databases, and also applies to some extent to the groups who study the relationship and phylogeny of protein sequences. In a typical model organism database, curators work through a certain number of research papers each week, abstract the conclusions from the paper and integrate them with the database. The result is a curated 'best guess' at the functional significance of each gene in the genome, with ties back to the literature whenever possible.

In addition to genome annotations, model organism databases curate information about special aspects of the biology of their model. For example, WormBase maintains the cell lineage and neuronal wiring diagram of the nematode. FlyBase tracks the many strains and mutants that make *Drosophila* a rich experimental animal.

A variant on the museum theme is the cottage industry, an organizational model adopted with success by Proteome, Inc. (J. Gerrels, personal communication). In this model, the bulk of the curators work part-time out of their homes or labs, and are recruited from the ranks of postdoctoral fellows, graduate students and faculty.

Then there is the party model, made famous by the *Drosophila* annotation jamboree hosted by Celera and the Berkeley *Drosophila* Genome Sequencing group[71]. The party model puts leading biologists from the community into the same room together with an equal number of bioinformaticians and has them spend a solid block of time (typically a week) annotating the genome. The biologists mine the genome for their favourite gene families, and the bioinformaticians provide the tools and technical expertise to facilitate this. The jamboree is, in effect, a frontal charge on the genome.

The jamboree model was more recently used at a meeting of the FANTOM (Functional Annotation of Mouse) Group at RIKEN in September 2000, to annotate ~21,000 full-length mouse cDNA clones[72]. However, other groups have not rushed to embrace this organizational model, and whether annotation jamborees are to become a permanent feature of the genomics landscape, once the novelty of genomic sequences wears off, remains to be seen.

Publishing and sharing annotations. Genomes are typically annotated by several groups. For example, the human genome is undergoing active annotation by Celera, Ensembl, the National Center for Biotechnology Information, the computational biology group at Oak Ridge National Laboratory and others.

The manifest strength of having several groups involved is that researchers benefit from their diverse approaches. The weakness is that a diversity of sources for the annotations tends to fragment the information. A researcher seeking to compare the annotations made by one group to those of another must visit several different web sites and surmount various obstacles, including incompatible user interfaces, coordinate systems, file formats and naming conventions. Without extensive bioinformatics support, this task can be nearly impossible to carry out on more than a few genes at a time.

Fortunately, several positive trends indicate that there is light at the end of the tunnel. One is in the realm of file formats, in which a small number of

OPEN SOURCE
A type of software distribution in which the source code (the human-readable instructions) are made freely available. The Linux operating system is open source. The Microsoft Windows and Macintosh operating systems are not.

standard genome annotation mark-up languages have begun to gain general acceptance. One such language is GAME, originally developed for use in the annotation of the *Drosophila* sequence and now the underlying data transfer language of the OPEN SOURCE Apollo annotation editor, which is under development at the Sanger Centre. GAME is notable for its rich syntax for describing the experimental evidence that underlies an annotation. Another is the distributed annotation system (DAS), a lightweight annotation language intended primarily for indexing and visualization. FlyBase, Ensembl and WormBase all now make their genome annotations available in this format.

A second positive trend is the development of a rich set of open-source libraries of software tools for storing, manipulating and visualizing genome annotations (BioPerl 2001, BioPython 2001, BioJava2001 and BioCORBA 2001). If, as seems increasingly likely, new annotation projects adopt pre-existing tools rather than creating their own, the current confusion of user interface conventions, search tools and data formats, might some day become a distant memory.

Bringing annotation into the mainstream. Genome annotation is no different in many respects from other aspects of molecular biology. It, too, involves hypothesis creation, testing, refinement and publication. There are also obvious differences. Whereas the experimental results from conventional research fit easily in the size and format of a printed publication, genome annotation studies are most suited for publication in electronically accessible databases. Annotation sets are also commonly updated regularly, giving them a fluidity that is uncharacteristic of conventional publications. Annotation sets do not receive citations in MedLine.

Links

FURTHER INFORMATION e-PCR | BLASTN | SSAHA | GENSCAN | Genie | GeneMark.hmm | Grail | HEXON | MZEF | Fgenes | HMMGene | BLASTX | RefSeq | Unigene | SWISS-PROT | Ensembl | RepeatMasker | *C. elegans* SNP data | COG | BLASTP | TrEMBL | PFAM | HMMER | PRINTS | TRANSFAC | PROSITE | ProDom | PSI-BLAST | SMART | BLOCKS | InterPro | *Saccharomyces* Genome Database | FlyBase | Mouse Genome Database | Gene Ontology | The *Arabidopsis* Information Resource | WormBase | Pombase | Proteome | FANTOM | Oak Ridge National Laboratory | DAS | BioPerl 2001 | BioPython 2001 | BioJava 2001 | BioCORBA 2001 | MIPS *Arabidopsis thaliana* database

The differences between conventional biological research and genome annotation have tended to marginalize the latter. Annotation is seen as an activity that is set apart from biology, done by a special group of people who are not exactly biologists. This is unfortunate, because after the initial high-throughput stage, genome annotation becomes crucially dependent on conventional hypothesis-driven research. Tying the experimental literature to the genome is a crucial step in making genomic data useful to the general research community.

The strength of curated databases is that human curators can identify contradictions between papers, and in some cases can catch and correct mistakes by the original authors. However, the literature-driven curation model is also inefficient because it requires curators to transcribe conclusions from papers into the database. Significant time is lost trying to fill in crucial gaps in the paper, such as missing sequence accession numbers and ambiguously identified genes, and a curator is often called on to make a judgement call on the basis of incomplete information. Most curators dream about an 'annotation ready' paper in which the relevant genes, proteins and processes are identified in some sort of structured abstract.

As a community, we should consider ways to bring genome annotation into the mainstream. Biologists must become directly involved with annotation; database curators must be afforded the same respect as their counterparts on the bench. One step might be to make the curation of a gene, gene family, or other set of database records, something akin to writing an invited review article, and provide its author with a citation. Another step would be to incorporate principles of peer review into curated annotations. One also could imagine using scientific meetings as an occasion for the leading scientists in a field to review and comment on the current status of an annotation database.

As the scientific literature becomes increasingly electronic, and genome annotation becomes increasingly like the literature, it is time to think about marrying the two. This will accelerate the process of annotation and help to clear the path from sequence to biology. Indeed, although genome sequencing itself is a highly specialized activity, genomic annotation is something to which the entire life science community can contribute.

1. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
2. Fraser, C. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
3. Cole, S. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
4. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546 (1996).
   The description of how the first eukaryotic genome was sequenced and annotated.
5. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
   The sequencing and annotation of the *Caenorhabditis elegans* genome.
6. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
   The sequencing and annotation of the *Drosophila melanogaster* genome.
7. *Arabidopsis* Genomics Initiative (AGI). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
8. Venter, J. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
   A landmark paper describing the human 'rough draft' (private version) and its annotation.
9. International Human Genome Sequencing Consortium (IHGSC). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
   A landmark paper describing the human 'rough draft' (public version) and its annotation.
10. Schuler, G. Sequence mapping by electronic PCR. *Genome Res.* **7**, 541–550 (1997).
11. 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).
12. Ning, Z., Cox, A. & Mullikin, C. SSAHA: A fast search method for large DNA databases. *Genome Res*. (submitted).
13. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
14. Burge, C. & Karlin, S. Prediction of complete gene

structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
The first description of GENSCAN and one of the best introductions to hidden Markov model-based gene-prediction programs.
15. Reese, M. G., Kulp, D., Tammana, H. & Haussler, D. Genie — gene finding in *Drosophila melanogaster*. *Genome Res.* **10**, 529–538 (2000).
16. Besemer, J. & Borodovsky, M. Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.* **27**, 3911–3920 (1999).
17. Uberacher, E. & Mural, R. Locating protein-coding regions in human DNA sequences by a multiple sensor–neural network approach. *Proc. Natl Acad. Sci. USA* **88**, 11261–11265 (1991).
18. Solovyev, V., Salamov, A. & Lawrence, C. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **22**, 5156–5163 (1994).
19. Zhang, M. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA* **94**, 565–568 (1997).
20. Solovyev, V., Salamov, A. & Lawrence, C. Identification of human gene structure using linear discriminant functions and dynamic programming. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 367–375 (1995).
21. Krogh, A. Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 179–186 (1997).
22. Reese, M. *et al.* Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10**, 483–501 (2000).
The most comprehensive comparison of nucleotide-level annotation tools so far.
23. 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 (2001).
A crucial comparison of *ab initio* gene-prediction algorithms versus those based on similarity searches.
24. Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**, 803–816 (2001).
25. Kent, W. J. & Zahler, A. M. The intronerator: exploring introns and alternative splicing in *Caenorhabditis elegans*. *Nucleic Acids Res.* **28**, 91–93 (2000).
26. Reboul, J. *et al.* Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nature Genet.* **27**, 332–336 (2001).
27. Pruitt, K., Katz, K., Sicotte, H. & Maglott, D. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* **16**, 44–47 (2000).
28. Schuler, G. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75**, 694–698 (1997).
29. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
30. Le, S., Chen, J. & Maizel, J. in *Structure and Methods: Human Genome Initiative and DNA Recombination* Vol 1 (eds Sarma, R. H. & Sarma, M. H.) 127–136 (Adenine, New York, 1990).
31. Pavesi, A., Conterio, F., Bolchi, A., Dieci, G. & Ottonello, S. Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucleic Acids Res.* **22**, 1247–1256 (1994).
32. Lowe, T. & Eddy, S. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
33. Eddy, S. Noncoding RNA genes. *Curr. Opin. Genet. Dev.* **9**, 695–699 (1999).
An easily accessible introduction to the fascinating world of non-coding RNA prediction.
34. Pennacchio, L. & Rubin, E. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2**, 100–109 (2001).
35. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
36. Thacker, C., Marra, M. A., Jones, A., Baillie, D. L. & Rose, A. M. Functional genomics in *Caenorhabditis elegans*: an approach involving comparisons of sequences from related nematodes. *Genome Res.* **9**, 348–359 (1999).
37. Aamodt, E. *et al.* Conservation of sequence and function of the *pag-3* genes from *C. elegans* and *C. briggsae*. *Gene* **8**, 67–74 (2000).
38. 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).
39. Qiu, Y. *et al.* Human and mouse *abca1* comparative sequencing and transgenesis studies revealing novel regulatory sequences. *Genomics* **73**, 66–76 (2001).
40. Margarit, E. *et al.* Identification of conserved potentially regulatory sequences of the *SRY* gene from 10 different species of mammals. *Biochem. Biophys. Res. Commun.* **17**, 370–377 (1998).
A good example of how comparative genomics can be used to identify putative regulatory sequences.
41. Ku, H. M., Vision, T., Liu, J. & Tanksley, S. D. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl Acad. Sci. USA* **97**, 9121–9126 (2000).
An elegant illustration of the power of comparative genomics.
42. Brookes, A. The essence of SNPs. *Gene* **234**, 177–186 (1999).
A comprehensive introduction to the potential contribution of single nucleotide polymorphisms to the understanding of human biology.
43. The SNP Consortium. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
44. Marth, G. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nature Genet.* **23**, 452–456 (1999).
45. Koch, R. *et al.* Single nucleotide polymorphisms in wild isolates of *Caenorhabditis elegans*. *Genome Res.* **10**, 1690–1696 (2000).
46. Piatigorsky, J., Kantorow, M., Gopal-Srivastava, R. & Tomarev, S. I. Recruitment of enzymes and stress proteins as lens crystallins. *EXS* **71**, 241–250 (1994).
47. Wistow, G. Lens crystallins: gene recruitment and evolutionary dynamism. *Trends Biochem. Sci.* **18**, 301–306 (1993).
48. Henikoff, S. *et al.* Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609–614 (1997).
49. Tatusov, R., Galperin, M., Natale, D. & Koonin, E. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
50. Altschul, S. F. & Koonin, E. V. Iterated profile searches with PSI-BLAST — a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444–447 (1998).
51. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28**, 263–266 (2000).
52. Attwood, T. *et al.* PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* **28**, 225–227 (2000).
53. Hoffman, K., Bucher, P., Falquet, L. & Bairoch, A. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**, 215–219 (1999).
54. Corpet, F., Gouzy, J. & Kahn, D. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res.* **27**, 263–267 (1999).
55. Ponting, C. P., Schultz, J., Milpetz, F. & Bork, P. SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.* **27**, 229–232 (1999).
56. Henikoff, J. G., Greene, E. A., Pietrokovski, S. & Henikoff, S. Increased coverage of protein families with the BLOCKS database servers. *Nucleic Acids Res.* **8**, 228–230 (2000).
57. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37–40 (2001).
58. Cherry, J. *et al.* SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* **26**, 73–79 (1998).
59. The FlyBase Consortium. The FlyBase database of the *Drosophila* Genome Projects and community literature. *Nucleic Acids Res.* **27**, 85–88 (1999).
60. Blake, J., Eppig, J., Richardson, J., Bult, C. & Kadin, J. The Mouse Genome Database (MGD): integration nexus for the laboratory mouse. *Nucleic Acids Res.* **29**, 91–94 (2001).
A great example of a 'classic' model organism database.
61. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
62. Huala, E. *et al.* The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* **29**, 102–105 (2001).
63. Stein, L. *et al.* WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* **29**, 82–86 (2001).
This model organism database combines nucleotide-level annotation with the results of high-throughput analyses, such as RNA interference.
64. Kumar, A. & Snyder, M. Emerging technologies in yeast genomics. *Nature Rev. Genet.* **2**, 302–312 (2001).
65. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
66. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
67. Griffin, T. J. *et al.* Quantitative proteomic analysis using a MALDI quadruple time-of-flight mass spectrometer. *Anal. Chem.* **73**, 978–986 (2001).
68. Bouvet, P. Determination of nucleic acid recognition sequences by SELEX. *Methods Mol. Biol.* **148**, 603–610 (2001).
69. Gonzalez, C. & Bejarano, L. A. Protein traps: using intracellular localization for cloning. *Trends Cell Biol.* **10**, 162–165 (2000).
70. Cagney, G., Uetz, P. & Fields, S. High-throughput screening for protein–protein interactions using two-hybrid assay. *Methods Enzymol.* **328**, 3–14 (2000).
71. Pennisi, E. Ideas fly at gene-finding jamboree. *Science* **24**, 2182–2184 (2000).
72. Kawai, J. *et al*. Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).

Acknowledgements

I acknowledge R. Durbin, S. Eddy, E. Birney and A. Neuwald for helpful discussions during the preparation of this review. A portion of this work was supported by the National Human Genome Research Institute at the US National Institutes of Health.

- Now that many genome sequences are available, attention is shifting towards developing and improving approaches for genome annotation.
- Genome annotation can be classified into three levels: the nucleotide, protein and process levels.
- Gene finding is a chief aspect of nucleotide-level annotation. For complex genomes, the most successful methods use a combination of *ab initio* gene prediction and sequence comparison with expressed sequence databases and other organisms. Nucleotide-level annotation also allows the integration of genome sequence with other genetic and physical maps of the genome.
- The principal aim of protein-level annotation is to assign function to the products of the genome. Databases of protein sequences and functional domains and motifs are powerful resources for this type of annotation. Nevertheless, half of the predicted proteins in a new genome sequence tend to have no obvious function.
- Understanding the function of genes and their products in the context of cellular and organismal physiology is the goal of process-level annotation. One of the obstacles to this level of annotation has been the inconsistency of terms used by different model systems. The Gene Ontology Consortium is helping to solve this problem.
- There are several approaches to genome annotation: the factory (reliance on automation), museum (manual curation), cottage industry (exemplified by Proteome, Inc.) and party (the Celera *Drosophila* annotation jamboree).
- As more scientists come to rely on genome annotation, it will become more important for the scientific community as a whole to contribute to this continuing process.

Lincoln Stein received a dual M.D./Ph.D. degree from Harvard Medical School in 1989, and did a residency in anatomical pathology at the Brigham & Women's Hospital in Boston, Massachusetts. In 1993, he joined the Whitehead Institute Center for Genome Research where he helped develop genome maps of the human and mouse, eventually becoming the Director of Informatics there. Since 1998, he has been on the faculty of Cold Spring Harbor Laboratory, where he works on various model organism databases and software systems for genome representation and analysis.

Links
ePCR
http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi
BLASTN
ftp://ftp.ncbi.nlm.nih.gov/blast
SSAHA
http://www.sanger.ac.uk/Software/analysis/SSAHA/
GENSCAN
http://genes.mit.edu/GENSCAN.html
Genie
http://www.fruitfly.org/seq_tools/genie.html
GeneMarkHMM
http://genemark.biology.gatech.edu/GeneMark/
Grail
http://compbio.ornl.gov/Grail-1.3/
HEXON
http://searchlauncher.bcm.tmc.edu:9331/gene-finder/Help/hexon.html
MZEF
http://sciclio.cshl.org/genefinder/
Fgenes
http://searchlauncher.bcm.tmc.edu:9331/gene-finder/Help/fgenes.html
HMMGene
http://www.cbs.dtu.dk/services/HMMgene/
BLASTX
ftp://ftp.ncbi.nlm.nih.gov/blast
RefSeq
http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
UniGene
http://www.ncbi.nlm.nih.gov/UniGene/
SWISS-PROT
http://www.ebi.ac.uk/swissprot/
Ensembl
http://www.ensembl.org/
RepeatMasker
http://www.genome.washington.edu/uwgc/analysistools/repeatmask.htm
C. elegans SNP Data
http://www.genome.wustl.edu/gsc/CEpolymorph/snp.shtml
COG
http://www.ncbi.nlm.nih.gov/COG/
BLASTP
ftp://ftp.ncbi.nlm.nih.gov/blast
TrEMBL
http://www.ebi.ac.uk/swissprot/
PFAM
http://pfam.wustl.edu/
HMMER
http://hmmer.wustl.edu/
PRINTS
http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html
TRANSFAC
http://transfac.gbf.de/TRANSFAC/
PROSITE
http://www.expasy.ch/prosite/
ProDom
http://www.biochem.ucl.ac.uk/bsm/dbbrowser/jj/prodomsrchjj.html
PSI-BLAST
ftp://ftp.ncbi.nlm.nih.gov/blast
SMART
http://smart.embl-heidelberg.de/
BLOCKS
http://www.blocks.fhcrc.org/
InterPro
http://www.ebi.ac.uk/proteome/
Saccharomyces Genome Database
http://genome-www.stanford.edu/Saccharomyces/
FlyBase
http://flybase.bio.indiana.edu/
Mouse Genome Database
http://www.informatics.jax.org/
Gene Ontology
http://www.geneontology.org/
The Arabidopsis Information Resource
http://www.arabidopsis.org/
WormBase
http://www.wormbase.org/
PomBase
http://www.sanger.ac.uk/Projects/S_pombe/Pombase.shtml
Proteome, Inc.

http://www.proteome.com/
FANTOM
http://www.gsc.riken.go.jp/e/FANTOM/
Oak Ridge National Laboratory
http://www.ornl.gov/
BioXML
http://www.bioxml.org
DAS
http://biodas.org
BioPerl 2001
http://www.bioperl.org
BioPython 2001
http://www.biopython.org/
BioJava 2001
http://www.biojava.org
BioCorba 2001
http://biocorba.org/

Fig.3 legend
MIPS *Arabidopsis thaliana* database
http://mips.gsf.de/proj/thal/db/gv/rv/rv_frame.html

Box 1
BLASTN, BLASTX, BLASTP, PSI-BLAST
http://www.ncbi.nlm.nih.gov/BLAST/
Ensembl
http://www.ensembl.org
ePCR
http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi
FlyBase
http://www.flybase.org/
GeneMarkHMM
http://genemark.biology.gatech.edu/GeneMark/
Genie
http://www.fruitfly.org/seq_tools/genie.html
GENSCAN
http://genes.mit.edu/GENSCAN.html
Grail
http://compbio.ornl.gov/Grail-1.3/
HMMER
http://hmmer.wustl.edu/
HMMGene
http://www.cbs.dtu.dk/services/HMMgene/
InterPro
http://www.ebi.ac.uk/proteome/
Proteome, Inc.
http://www.proteome.com/
RefSeq and LocusLink, NCBI
http://www.ncbi.nlm.nih.gov/
RepeatMasker
http://www.genome.washington.edu/uwgc/analysistools/repeatmask.htm
SSAHA
http://www.sanger.ac.uk/Software/analysis/SSAHA/
SWISSPROT and SWISSPROT-TrEMBL
http://www.ebi.ac.uk/swissprot/
UCSD Genome Browser
http://genome.ucsc.edu/