17. Gilbert, S. F. Bearing crosses: a historiography of genetics and embryology. *Am. J. Med. Genet.* **76**, 168–182 (1998).
18. Bateson, W. Evolutionary faith and modern doubts. *Science* **40**, 1412–1415 (1922).
19. Desmond, A. & Moore. J. *Darwin: the Life of a Tormented Evolutionist* (Norton, New York, 1991).
20. Boore, J. L., Lavrov, D. V. & Brown, W. M. Gene translocation links insects and crustaceans. *Nature* **392**, 667–668 (1998).
21. Aguinaldo, A. M. *et al.* Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**, 489–493 (1997).
22. De Rosa, R. *et al. Hox* genes in brachiopods and priapulids and protostome evolution. *Nature* **399**, 772–776 (1999).
23. Shimamura, M. *et al.* Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* **388**, 666–670 (1997).
24. Gould, S. J. *Ontogeny and Phylogeny* (Harvard Univ. Press, Cambridge, Massachusetts, 1977).
25. Maxam, A. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA* **74**, 560–564 (1977).
26. Gehring, W. J. The genetic control of eye development and its implications for the evolution of the various eye-types. *Int. J. Dev. Biol.* **46**, 65–73 (2002).
27. Halder, G., Callaerts, P., & Gehring, W. J. Induction of ectopic eyes by targeted expression of the *eyeless* gene in *Drosophila*. *Science* **267**, 1788–1792 (1995)
28. Pineda, D. *et al.* Searching for the prototypic eye genetic network: *Sine oculis* is essential for eye regeneration in planarians. *Proc. Natl Acad. Sci. USA* **97**, 4525–4529 (2000).
29. Wawersik, S. & Maas, R. L. Vertebrate eye development as modeled in *Drosophila*. *Hum. Mol. Genet.* **9**, 917–925 (2000).
30. Salvini-Plawin, L. V. & Mayr, E. Evolution of photoreceptors and eyes. *Evol. Biol.* **10**, 207–263 (1977).
31. Erwin, D. H. The origin of bodyplans. *Am. Zool.* **39**, 617–629 (1999).
32. Pichaud, F. & Desplan, C. *Pax* genes and eye organogenesis. *Curr. Opin. Genet. Dev.* **12**, 430–434 (2002).
33. Gilbert, S. F. *Developmental Biology* 7th edn (Sinauer Associates, Sunderland, Massachusetts, 2003).
34. Aitkenhead, M. *et al.* Paracrine and autocrine regulation of vascular endothelial growth factor during tissue differentiation in the quail. *Dev. Dyn.* **212**, 1–13 (1998).
35. Eremina, V. *et al.* Glomerular-specific alterations of VEGF-A expression lead to distinct congenital and acquired renal diseases. *J. Clin. Invest.* **111**, 707–716 (2003).
36. Lenski, R. E., Ofria, C., Pennock, R. T. & Adami, C. The evolutionary origin of complex features. *Nature* **423**, 139–144(2003).
37. Galant, R. & Carroll, S. B. Evolution of a transcriptional repression domain in an insect Hox protein. *Nature* **415**, 910–913 (2002).
38. Ronshaugen, M., McGinnis, N. & McGinnis, W. Hox protein mutation and macroevolution of the insect body plan. *Nature* **415**, 914–917 (2002).
39. Carroll, S. B., Weatherbee, S. D. & Langeland, J. A. Homeotic genes and the regulation and evolution of insect wing number. *Nature* **375**, 58–61 (1995).
40. Weatherbee, S. D. *et al.* Ultrabithorax function in butterfly wings and the evolution of insect wing patterns. *Curr. Biol.* **9**, 109–115 (1999).
41. Merino, R. *et al.* The BMP antagonist Gremlin regulates outgrowth, chondrogenesis and programmed cell death in the developing limb. *Development* **126**, 5515–5522 (1999).
42. Gaunt, S. J. Conservation in the *Hox* code during morphological evolution. *Int. J. Dev. Biol.* **38**, 549–552 (1994).
43. Burke, A. C., Nelson, A. C., Morgan, B. A. & Tabin, C. *Hox* genes and the evolution of vertebrate axial morphology. *Development* **121**, 333–346 (1995).
44. Averof, M. & Patel, N. H. Crustacean appendage evolution associated with changes in *Hox* gene expression. *Nature* **388**, 682–686 (1997).
45. Harris, M. P., Fallon, J. F. & Prum, R. O. Shh–BMP2 signaling module and the evolutionary origin and diversification of feathers. *J. Exp. Zool. Part B Mol. Dev. Evol.* **294**, 160–176 (2002).
46. Yu, M., Wu, P., Widelitz, R. B. & Chuong. C. M. The morphogenesis of feathers. *Nature* **420**, 308–312 (2002).
47. Cohn, M. J. & Tickle, C. Developmental basis of limblessness and axial patterning in snakes. *Nature* **399**, 474–479 (1999).
48. Ferkowicz, M. J. & Raff, R. A. *Wnt* gene expression in sea urchin development: heterochronies associated with the evolution of developmental mode. *Evol. Dev.* **3**, 24–33 (2001).
49. Kuratani, S., Nobusada, Y., Horigome, N. & Shigetani, Y. Embryology of the lamprey and evolution of the vertebrate jaw: insights from molecular and developmental perspectives. *Philos. Trans. R. Soc. Lond. B* **356**, 1615–1632 (2001).
50. Nijhout, H. F. Development and evolution of adaptive polyphenisms. *Evol. Dev.* **5**, 9–18 (2003).
51. Waddington, C. H. in *Evolution (Soc. Exp. Biol. Symp. VII)* (eds Brown, R. & Danielli, J. F.) 186–199 (Cambridge Univ. Press. Cambridge, UK, 1953).
52. Schmalhausen, I. I. *Factors of Evolution: the Theory of Stabilizing Selection* (Univ. of Chicago Press, Chicago, 1949).
53. Shapiro, A. M. Seasonal polyphenism. *Evol. Biol.* **9**, 259–333 (1976).
54. Brakefield, P. M. *et al.* Development, plasticity, and evolution of butterfly eyespot patterns. *Nature* **384**, 236–242 (1996).
55. West-Eberhard, M. J. Phenotypic plasticity and the origins of diversity. *Annu. Rev. Ecol. Syst.* **20**, 249–278 (1989).
56. West–Eberhard, M. J. *Developmental Plasticity and Evolution*. Oxford University Press (2003).
57. Rockman, M. V. & Wray, G. A. Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* **19**,1991–2004 (2002).
58. Pigliucci, M. *Denying Evolution: Creationism, Scientism, and the Nature of Science* (Sinauer Associates, Sunderland, Massachusetts, 2002).
59. Haldane, J. B. S. Foreword. in *Evolution (Soc. Exp. Biol. Symp. VII)* (eds Brown, R. & Danielli, J. F.) ix–xix (Cambridge Univ. Press, Cambridge, UK, 1953).

Acknowledgements

This Perspective is based on a talk originally presented as a lecture to the Society for Developmental Biology (SDB) at its annual meeting in 2002. I wish to thank the Education Committee of the SDB for the opportunity to write it and to Kenneth Miller and Sean Carroll for their helpful comments. Funding was from the National Science Foundation and from Swarthmore College, Pennsylvania.

Online links

DATABASES

The following terms in this article are linked online to:

FlyBase: http://flybase.bio.indiana.edu
Distal-less | *Ubx*
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
Pax6
Swiss-Prot: http://www.expasy.ch
BMP2 | BMP4 | Gremlin | Sonic hedgehog

FURTHER INFORMATION

Evolution, development, and creationism — a supplement: http://zygote.swarthmore.edu/Darwin

Access to this interactive links box is free online.

OPINION

Vertebrate gene predictions and the problem of large genes

Jun Wang, ShengTing Li, Yong Zhang, HongKun Zheng, Zhao Xu, Jia Ye, Jun Yu and Gane Ka-Shu Wong

To find unknown protein-coding genes, annotation pipelines use a combination of *ab initio* gene prediction and similarity to experimentally confirmed genes or proteins. Here, we show that although the *ab initio* predictions have an intrinsically high false-positive rate, they also have a consistently low false-negative rate. The incorporation of similarity information is meant to reduce the false-positive rate, but in doing so it increases the false-negative rate. The crucial variable is gene size (including introns) — genes of the most extreme sizes, especially very large genes, are most likely to be incorrectly predicted.

We live in the halcyon days of large-scale DNA sequencing. Each release of a sequenced genome is accompanied by a list of genes, many of which are computer predictions. Experimental confirmation in the form of sequenced transcripts of full-length cDNAs is extensive for mice[1,2], less so for humans[3] and non-existent for pufferfish[4]. For invertebrate genomes, cDNAs are less important because the genes are smaller and easier to predict. Nevertheless, the many fruitfly[5] and nematode[6] cDNAs that were produced after their genomes were sequenced have been invaluable in finding residual errors in the definition of exon boundaries. As it is difficult to get cDNAs that are expressed transiently, or at low levels, in specific tissues and at specific developmental stages, predicting genes will remain an integral part of DNA sequence analysis for the foreseeable future. Therefore, it is imperative for the biologists who use gene-prediction programs to understand what they can and cannot do. Although some features of these programs are better than is commonly thought, others are worse. It is tempting to dismiss the programs as being inherently unreliable (see BOX 1 for a note on fluctuations in gene number in the human genome), but, in fact, they fail for specific reasons that can be understood with minimal jargon and without delving into algorithmic minutiae.

ANNOTATION PIPELINES have been comprehensively reviewed[7]. Every pipeline incorporates information from known genes. No pipeline ever substitutes a predicted gene for a known

Box 1 | How many genes are there in the human genome?

On 30 May 2003, at Cold Spring Harbor Laboratory, New York, the winner of the human gene sweepstakes was finally announced[38]. Lee Rowen, from the Institute for Systems Biology, Seattle, won with a wager of 25,947 genes. Hers was the lowest wager from among the more than 460 that had been placed. The official count was announced to be 24,847 — substantially lower than the 30,000–40,000 that had been estimated in February 2001, with the initial analyses of the draft sequence for the human genome. Were the original estimates too high, or was this latest estimate too low? A little appreciated fact was that 24,847 represented the number of genes for which the organizers felt they had the best supporting evidence, on the basis of sequence similarity to known genes or proteins in the vertebrate databases. It was a low estimate, and they confessed to the media that this was probably not the final answer. So, the number of genes in the human genome remains unknown.

one. However, one of the main justifications for spending enormous amounts of money on genome sequencing is to identify new genes for which there is only partial or no previous information. This is done by using a combination of *AB INITIO* GENE PREDICTION — a statistical process that finds protein-coding genes[8] — and similarity to experimentally confirmed genes or proteins. In the Ensembl pipeline[9], which was used for the human and mouse genomes, the *ab initio* programs are called upon first. Then, to reduce the high incidence of FALSE POSITIVES (FPs), the resulting gene predictions are 'fixed' by the incorporation of similarity information. Ensembl keeps only those exons that show sequence similarity to a gene or protein in the vertebrate databases, not necessarily the entire gene prediction. Here, we evaluate this process, step by step, on the basis of our analyses of a set of 7,485 REFSEQ[10] full-length human cDNAs, all of which have perfect BLAT[11] alignments in the human genome sequence. Sequences were downloaded from the Santa Cruz genome browser[12]. The cDNA sequences were dated August 2002, and the genome sequences were dated June 2002.

If gene-prediction programs were to work well on any particular gene set, they would be expected to work best on RefSeq genes, because these are the genes that they have been trained on, as opposed to unknown genes. However, even for this idealized gene set, the programs are not perfect, and analysing these imperfections can be instructive. Our analysis is meant to be as realistic as possible — for example, the entire chromosome is studied, rather than a pre-selected region with the RefSeq genes already excised. Intergenic sequences are only considered indirectly and are not modelled explicitly, as their identity is not known. In plant genomes, intergenic sequences are known to be nested clusters of long terminal repeat (LTR) retrotransposons[13], yet transposable elements are almost never found in the introns of plant genes[14]. Given what we now know, it would be intellectually dishonest to model intergenic sequences as either random sequences or transposable elements.

Vertebrate genomes contain many pseudogenes, particularly single-exon processed pseudogenes[15], and large numbers of these are often incorrectly predicted to be genes. For example, many such errors were found in the re-annotation of human chromosome 22 (REF. 16). As the difference between a pseudogene and a real gene can be a single base pair, removing pseudogenes is too difficult a task for the initial gene prediction. Here, we only focus on whether the exon boundaries are correctly defined, as this is the fundamental problem for *ab initio* gene prediction.

Box 2 | Intrinsic trade-offs in gene-prediction programs

Gene prediction is an intrinsically statistical process. It searches for patterns that are correlated with protein-coding sequences, but the correlation is only true in a statistical sense, and there is no reason to expect perfection. Different programs search for different patterns. Programs such as GenScan and FgeneSH search for patterns of codon usage that are specific to protein-coding sequences. Newer programs, such as TwinScan and SGP2, search for sequence conservation in a related species, as well as codon usage. The problem with all statistical-detection algorithms is that there is no guarantee that every instance of the desired pattern is the result of a protein-coding sequence. Even a random non-coding sequence can come up positive, and the longer that sequence is, the more likely it is that this will happen. Hence, there is always a false positive (FP) rate. Conversely, there is no guarantee that every instance of a protein-coding sequence will result in a detectable pattern, and there is always a FALSE-NEGATIVE (FN) RATE . Many methods have been invented to reduce the FP rates. The Ensembl annotation system does this by eliminating any exon that is not similar to a gene or protein in the vertebrate databases. The problem with these efforts to reduce FP rates is that they always increase FN rates. Trade-offs such as these are common in all branches of experimental science, from biology to physics.

Ab initio gene predictions

Ab initio gene predictions rely on two classes of sequence information: SIGNAL TERMS and CONTENT TERMS. Signal terms refer to short sequence motifs (such as splice sites, branch points, polypyrimidine tracts, start codons and stop codons) that are found in almost all eukaryotic genes. For the smaller eukaryotic genomes, such as that of yeast, signal terms contain almost enough information to define the genes; however, for the vertebrate genomes, in which intron sizes can reach hundreds of kilobases, signal terms are inadequate. Exon detection must rely on the content terms, which refer to the patterns of codon usage that are unique to a species, and allow coding sequences to be distinguished from the surrounding non-coding sequences by a statistical detection algorithm (see BOX 2 for a discussion on the intrinsic limitations of all statistical detection algorithms). The use of patterns of codon usage to detect exons has a few caveats. First, the programs must be taught what the codon-usage patterns look like, by presenting them with a TRAINING SET of known coding sequences, and a new training set is needed for each species. Second, untranslated regions (UTRs) at the ends of the genes cannot be detected, although most programs can identify polyadenylation sites. Third, non-protein-coding RNA genes cannot be detected, although there are specialized programs that will attempt this[17]. Fourth, and finally, none of these programs can detect alternatively spliced transcripts.

Here, we focus on two of the most popular *ab initio* programs: GenScan (the default used by Ensembl)[18] and FgeneSH[19]. These two programs share the same overall strategy; the main difference is that GenScan places more emphasis on the content terms, whereas FgeneSH places more emphasis on the signal terms. It might be expected that the programs will encounter problems with OUTLIER GENES that have radically different characteristics from those in the training set. Perhaps, among genes with restricted expression patterns that are not in the databases, there might be different codon-usage patterns. Such an assertion is almost impossible to disprove, but considering the many thousands of vertebrate cDNAs that have been sequenced, it is far more reasonable to assume that the sampling is sufficiently representative. Systematic problems with the *ab initio* predictions can be shown to be functions of variables that are not directly related to codon-usage patterns. We consider different

indicators of performance, as a function of all potentially relevant variables (see BOX 3 for a more detailed justification of our non-uniform histogram-binning methods), to find out if any of these measurements indicators of performance are severely affected at extreme values of these variables.

A worst-case example, which highlights most of the gene-prediction problems that are commonly encountered, is illustrated in FIG. 1. Perhaps the most obvious problem is the high incidence of FP and false negative (FN) errors. We define the FP rate as the probability that a base that is predicted to be protein-coding is in fact not known to be coding, as determined by the cDNA alignment (see BOX 4 for a subtle point on the different definitions of FP). Conversely, we define the FN rate as the probability that a protein-coding base is not predicted to be coding. Previous assessments computed what are called PER-BASE PAIR RATES (per-bp rates), which do not require the reading frames to be correct. However, in most cases, having the correct reading frame is crucial. We therefore show PER-AMINO ACID RATES (per-aa rates) that take this into account. The overall difference is small — insisting that the reading frames are correct only increases the error rates by 1% — but it is reassuring to have actually checked this. We plotted FPs and FNs as a function of potentially relevant variables incorporating various aspects of gene structure and sequence content, including GENE SIZE, exon size, number of exons and GC content. We discovered that the most important variable is the gene size.

Problems of gene size

Gene size is defined as the size of the unspliced protein-coding transcript, without the 5′ and 3′ UTRs, but including all of the introns between the start and stop codons. Gene size is sometimes defined as the coding region without the introns — here, we refer to this property as CDS SIZE. The mean gene size in our data set of RefSeq genes is 46.1 kb (52.8 kb if UTRs are included). This is unlikely to be the correct mean for the human genome because we only accepted perfect BLAT alignments, and many larger genes were rejected as a result of trivial single-base discrepancies. Considering that gene size varies by three orders of magnitude, while number of exons per gene and CDS size vary by only a factor of three (FIG. 2), it is clear that large genes are primarily attributable to large introns, not to more exons or larger exons mostly introns. At the other extreme, small genes of less than 1 kb in size are usually single-exon genes. On the basis of how these FP and FN rates vary as a function of gene size (FIG. 3), it is clear that large multiple-exon genes and small single-exon genes both present problems for GenScan and FgeneSH, but for different reasons.

Large genes. As gene size increases, FP rate goes up, but FN rate does not. This is to be expected, given the nature of the gene-prediction algorithms, which search for characteristics of the coding sequence that are only correct in a statistical sense and are not expected to be valid in every considered sequence (BOX 2). There is always some probability of an FP (or FN) error and this increases with the length of the sequence that is under consideration. The difference is that for FP rates the sequences under consideration are the introns, whereas for FN rates they are the exons. If there is a sufficiently large intron, every *ab initio* program will predict an exon where there is not one. An increase in FP rate with an increase in gene size is an intrinsic property of all *ab initio* programs. By contrast, although there is a small increase in CDS size as a function of gene size, it cancels out after normalizing to per-aa rates, and, as a result, FN rate is not sensitive to increases in gene size. Any FP exon is highly likely to contain triplets that are interpreted as stop codons; this leads to premature termination of the predicted gene and subsequent gene fragmentation, where several gene fragments are predicted instead of the one gene. This is a particularly serious problem for genes that are more than 100 kb in size (FIG. 4). Conversely, predicted genes composed of predominately FP exons are likely to have small predicted sizes. When the predicted gene size is below 1 kb (GenScan), or below 10 kb (FgeneSH), it can be expected that most of the exons will be FP exons (FIG. 5). This simple 'rule of thumb' can be used to filter out the obviously incorrect predictions.

Box 3 | Measuring performance against a continuous variable

It is common practice to summarize the performance of a gene-prediction program by one or two measures such as SENSITIVITY and SPECIFITY. This approach is appealing in its simplicity, but it can obscure the possibility that there are classes of genes for which the performance is especially good, or especially bad. To detect these differences, it is necessary to consider how performance varies as a function of some continuous variables that describe the properties of the gene set. We discovered that the most important variable is gene size (including introns). The idea is to compute average performance in groups of genes that are clustered on the basis of their size. Most plotting software will divide a chosen axis into uniformly sized histogram 'bins'. This is inappropriate if the gene-size distribution is non-uniform, and there are far more genes in the middle than at the ends. We wanted to see how the performance degrades at the tails of this distribution. There is no one ideal bin size. Choose too small a bin, and the averages become too noisy at the ends, where there are not enough genes; with too large a bin, the ends get subsumed into a single bin, which obscures any potentially interesting trends. The solution is to put a constant number of genes into each bin, and to use non-uniform bin sizes, adjusting the number of bins to get a targeted number of genes. For this analysis, the target is 150 genes.

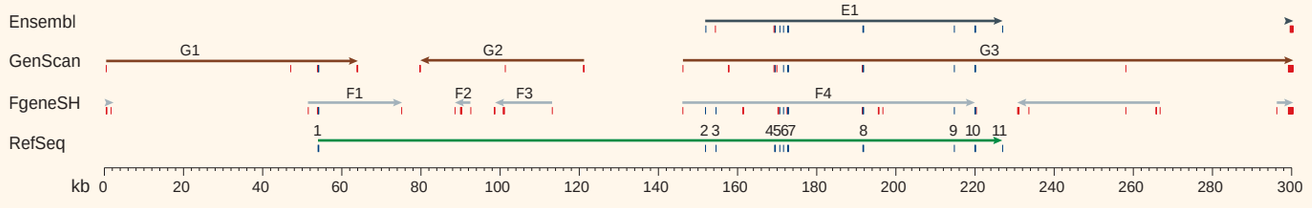

Figure 1 | **Actual versus predicted exons in a known gene: TEA domain family member 1 (SV40 transcriptional enhancer factor on human chromosome 11).** Correctly predicted exons are coloured blue, whereas incorrectly predicted exons are coloured red. Arrows indicate the direction of transcription for each predicted gene. RefSeq indicates that there are 11 exons, labelled 1–11, which span a genomic region of 173 kb. FgeneSH has four predictions that overlap with this gene, labelled F1–F4. GenScan has three predictions that overlap with this gene, labelled G1–G3. Note that F2, F3 and G2 are false-positive exons, in the large 98 kb first intron of this gene. OVER-PREDICTION is another problem, as exemplified by G1 and G3, which did not terminate correctly at the start and stop codons. Most of these problems are fixed in the Ensembl prediction, labelled E1, but even so, it failed to identify the first exon. Approximately half of this genomic region is incorrectly annotated as a 'gene desert', because of one large intron.

As large genes are such a problem for *ab initio* prediction, it has to be considered whether there is any biological significance to a gene being so big. For example, are the large genes concentrated in specific functional classifications? We divided our data set into groups on the basis of gene size, and classified functions using Gene Ontology[20]. No significant differences were observed. However, tissue specificity, as estimated by the presence of at least one expressed sequence tag (EST) in the human databases, is correlated to gene size (FIG. 6). The criterion was that at least 80% of the EST had to match the cDNA sequence, with no concern for how many ESTs matched the cDNA. We wanted to find the probability that a specific gene is expressed in that particular tissue, regardless of its expression level. For terminally differentiated cells, such as those of the brain, large genes are expressed at least as often, and sometimes more so, than small genes. By contrast, for fast-dividing cells, such as those found in carcinomas, large genes are expressed less often. The long transcription times that are required for large genes, typically 7 hours per megabase[21], means that there is not enough time to complete the transcription before the next mitosis in fast dividing cells.

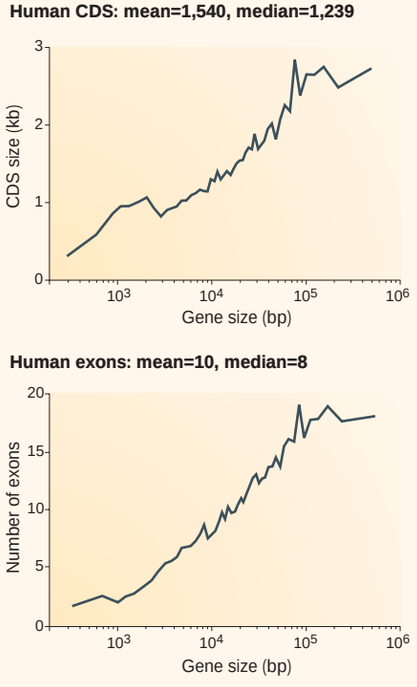

Figure 2 | **Correlation between gene size and intron size.** Although gene size varies by three orders of magnitude, the number of exons per gene and the cDNA size (CDS size) vary by only a factor of three. Exon size is essentially a constant, and most of any increase in gene size is the result of an increase in intron size.

Box 4 | Distinction between false positives and over-predictions

One of the more amusing problems with the gene-prediction programs is that they do not know how to quit when they are ahead. Gene boundaries are not well defined. It is common for exons to be predicted 3′ of the stop codon, or 5′ of the start codon. These are obviously erroneous exon predictions, to the extent that they do not belong to the gene in question, but should they be treated as false positives (FPs)? We do not think that it is fair to do so, because some of these exons might represent a coding sequence in an adjacent gene. It is impossible to prove otherwise until we have all the cDNAs, but this is unlikely to happen any time soon, if ever. Hence, we use the term over-prediction to refer to exons that lie entirely outside the region of the genome that is defined by the cDNA alignment, but belong to a prediction that does overlap with this region. This is treated as a separate phenomenon, which is distinct from FP exons that have some overlap with this region. In practice, the important difference between these two phenomena is that FPs are sensitive to gene size, but over-predictions are not.

Small genes. Single-exon genes that are smaller than 1 kb present a different problem, as FP and FN rate both increase within the limits of small genes. It might be thought that these would be the easiest genes to predict as, without the introns, this problem is similar to finding genes in bacteria, in which accurate programs such as GeneMark[22] rely largely on open reading frame (ORF) detection. In fact, small genes are intrinsically difficult to detect, partly because of the lack of splicing signals on either side of the single exon, but mostly because of the decreasing signal-to-noise ratios as the size of the coding region decreases. In vertebrates, this problem is further complicated by the abundance of single-exon-processed pseudogenes, which are commonly mistaken for real genes. As most vertebrate genes have many more than one exon, this problem can be considered to be of low significance. In our data set, 4.5% of the genes are single exon, and this is likely to be an overestimate, because even RefSeq can be contaminated by pseudogenes. By contrast, for invertebrate and plant genomes, single-exon genes make up more of the gene set and this problem cannot be so readily dismissed.

Previous reviews of *ab initio* gene prediction did not consider gene size to be the most relevant variable. Simple averages over the total gene set were reported, and only exon size and GC content were considered[23]. We looked at these variables too, but did not find them to be as informative as gene size. Certainly, small exons are troublesome (FIG. 3), but in most cases, the fact that a few dozen bases are mis-specified is insignificant, compared with the more serious problem of large genes. Moreover, many of these earlier analyses used a data set with a mean gene size of only 5–10 kb[24], because when these studies were done, large megabase-sized genomic contigs were not available. It is interesting that mean gene sizes have increased as the human genome sequence has neared completion. For the initial analyses of the draft sequence, mean gene size was only 27 kb, but two years later, it was 51–59 kb in the finished chromosomes 14, 20 and 21 (REF. 25).

Sequencing errors. How might sequencing errors affect the accuracy of gene predictions? To address this issue, we introduced random single-base substitution and insertion/deletion (indel) errors into the human genome sequence, and then ran the *ab initio* programs again (TABLE 1). Generally, substitution errors do not significantly affect FP rate, but for substitution errors of more than 10^{-3} errors per bp, there is a marked increase in FN. Indel errors are less tolerated, as they affect FP at rates of more than 10^{-3}, and FN at rates of more than 10^{-4}. Nevertheless, these results indicate that the present standard for a 'finished' sequence, set by the public consortium as 10^{-4}, is more than adequate for gene identification purposes.

Over-prediction of genes

Gene boundaries are often poorly defined, in that the predicted gene does not terminate at the start and stop codons. Most assessments of *ab initio* gene prediction confuse this problem with FPs, but we believe that it should be treated as a distinct phenomenon that we call over-prediction (BOX 4), because, unlike FPs, the probability of over-prediction is independent of gene size (FIG. 7). Over-predictions usually result from a failure to detect the first and last exons that contain the start and stop codons, repectively. For FgeneSH, 43% of the 5′ over-predictions and 92% of the 3′ over-predictions result from a missing first and last exon. Similarly, we found for GenScan that 45% of the 5′ over-predictions and 94% of the 3′ over-predictions result from a missing first and last exon. Even when the start codon is correctly detected, the program might simply decide not to terminate there, as was the case for 52% of the 5′ over-predictions in FgeneSH and 50% in GenScan. It is likely that the absence of a polyadenylation site renders the 5′

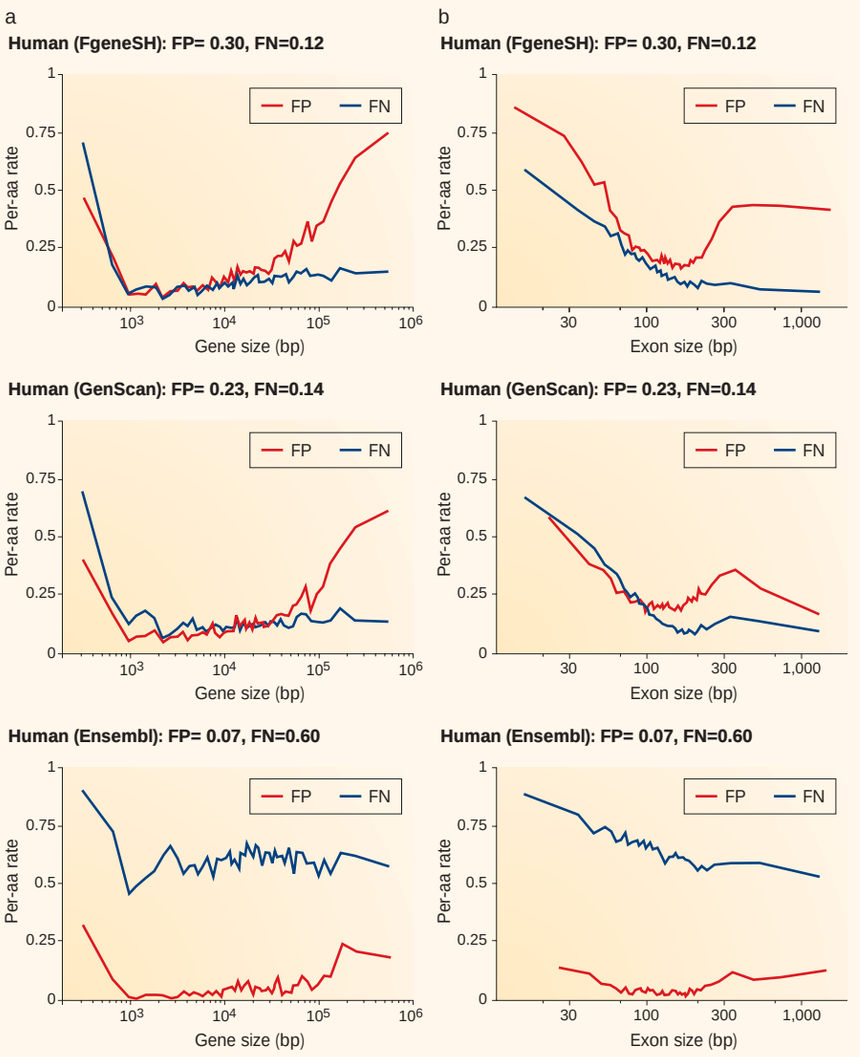

Figure 3 | **Size dependencies for false-positive and false-negative rates.** Error rates are depicted as a function of the size of the actual gene, from the alignment of RefSeq cDNAs to genome (**a**), or the size of the exon (**b**). Note that two different exon sizes are used: false positive (FP) uses predicted exon size, but false negative (FN) uses actual exon size. Per-amino acid (per-aa) rates check if the reading frame is correctly assigned, as well as if the correct nucleotides are predicted to be coding.

boundary prediction more difficult. A few over-predictions result from frame-shift errors that render the start and stop codons unrecognizable. Terminal exons are difficult to detect, because they are bounded by only one splice site, instead of two. Moreover, the detectable protein-coding portion is often a small fraction of some larger UTR-containing exon. For example, in our RefSeq gene set, detectable exons smaller than 20 bp comprised 5.7%, 5.7% and 0.3% of start-containing, stop-containing and internal exons, respectively.

Where there is an over-prediction, the gene finder must keep going until it finds a start or stop codon. Often, it travels through a distance that is comparable to the mean gene size. In FgeneSH, the mean (median) over-prediction distances for the 5′ and 3′ ends are 22.0 kb (10.9 kb) and 20.9 kb (10.9 kb), respectively. These distances are even larger in GenScan, at 39.7 kb (21.3 kb) and 37.1 kb (20.4 kb) for the 5′ and 3′ ends, respectively. If another gene lies within this over-prediction region, the *ab initio* program will often merge the two genes together. It is difficult to determine how often this has happened when a full set of cDNAs is not available and it is not known which genes are adjacent to each other. What can be said is that the over-prediction probability establishes an upper limit for the likelihood that two adjacent genes have been merged into one prediction. We believe that over-prediction of a first gene might hinder the detection of a neighbouring gene, and although the likelihood of this happening is difficult to estimate, it too is bounded above by the over-prediction probability.

Incorporation of similarity

When considering the problems of FP errors in large genes, and over-predictions in genes of any size, it is not surprising that many biologists are frustrated by the output of the *ab initio* programs[26,27]. To address these concerns, Ensembl incorporates similarity information into its pipeline to reduce the incidence of FP errors and over-predictions. However, this has consequences. If we only accept those genes that are similar to genes already in the databases, we preclude ourselves from identifying new genes. This goes against the main reason for sequencing genomes — to find new genes. The key question is to what extent the FP errors and over-predictions are being exchanged for FN errors once similarity has been incorporated (BOX 2). We simulated what would have happened if the RefSeq gene had been 'unknown', by removing anything from the vertebrate databases that had more than 90% amino-acid identity to the RefSeq gene.

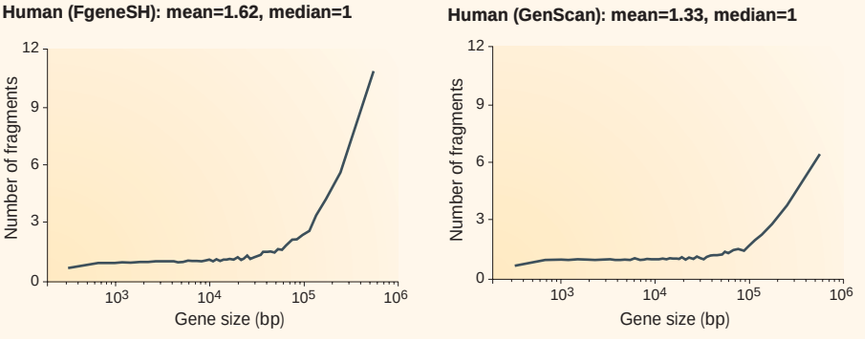

Figure 4 | **Size dependency for gene fragmentation problem.** The number of predicted gene fragments as a function of the size of the actual gene. Every prediction that overlaps with the actual gene is counted.

Note that the 90% rule is only intended to remove sequences for the gene in question, not those homologues that Ensembl might have found. Indeed, the set of all genes that are homologous to our reference-gene set show far less similarity than 90%, with an asymmetric distribution that peaks around ~30–40%.

This simulation can be done in a completely realistic way, because the source code for Ensembl is freely available. The 90% rule has simply to be inserted into their code and applied at the point just before it searches the vertebrate databases. With a GenScan prediction, a BLAST search is carried out on this modified database. Using the same code as Ensembl, an entirely new gene model is built by aligning the best hits back to the genome. It is found that FPs are reduced to an overall rate of 7%, but at the cost of a substantially larger FN rate (FIG. 3). Over-prediction rates are reduced to 2% at both 5′ and 3′ ends (data not shown). It is likely that Ensembl also eliminates FPs that are predicted in intergenic sequences.

When considering FN errors, it is important to establish whether they are randomly distributed or concentrated in specific genes. This is a crucial distinction between GenScan and Ensembl. To show this, for each gene the probability of a COMPLETE MISS (CM) was computed, which we define as the failure to detect even 100 bp of the coding region. Only 5% of the genes are completely missed in GenScan, compared with 44% in Ensembl (FIG. 8). This significant increase in CM rate for Ensembl is owing to those genes that have no remaining homologues in the databases, after the 90% rule is applied. These are known as 'simulated unknown' genes. Their exact number is a function of our simulation parameters. Although this number decreases as the databases expand, it is unlikely to ever reach zero, because there will always be some genes with restricted expression patterns that are entirely missing from the databases.

FIGURE 1 illustrates that the presence of a large intron causes Ensembl to miss the first exon, and creates the illusion that about half of this region is a 'gene desert'. Hence, we define a false desert (FD) as the amount of a genome region, as delineated by the cDNA alignment, that is not covered by a gene prediction. In GenScan, FD rate is small and independent of gene size, but in Ensembl, FD rate is large and increases with gene size, particularly above 100 kb (FIG. 8). This size dependency arises from the interspersed presence of the FP exons in the large introns of the large genes, which makes it difficult for Ensembl to recognize the few true exons that are correctly detected by GenScan. An intrinsic property of this

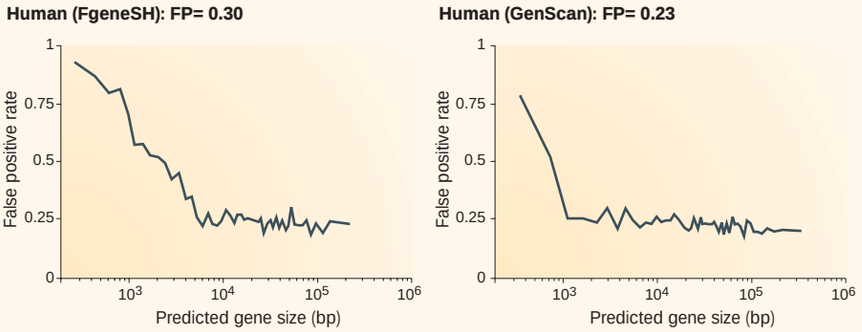

Figure 5 | **Detection of erroneous predictions using gene size.** If a gene is truly unknown, then so is its actual gene size; however, this figure shows that a small predicted gene size is a useful indicator of potentially erroneous predictions.

Glossary

AB INITIO GENE PREDICTION
The identification of protein-coding genes in genomic sequence, using no prior knowledge other than the signal and content terms.

AMYGDALA
An almond-shaped neurostructure that is involved in the production and response to non-verbal signs of anger, avoidance, defensiveness and fear.

ANNOTATION PIPELINES
A series of computer procedures that is used to identify the biological contents of a sequenced genome. Gene finding is only the first of many steps. Subsequent steps might include the identification of homologous genes, the assignment of biological function and so on.

CDS SIZE
The size of the spliced transcript, excluding introns. As gene-prediction programs do not detect untranslated regions, we do not include them in this definition.

COMPLETE MISS
(CM). The probability that less than 100 bp of the protein-coding sequence of a gene is correctly predicted.

CONTENT TERMS
Patterns of codon usage, which are unique to each species, that allow protein-coding sequences to be distinguished from surrounding non-coding sequence.

FALSE DESERT
(FD). A fraction of a sequence of a gene, including its introns which is not covered by any of the gene predictions.

FALSE NEGATIVE
(FN). The probability that a segment that is known to code for protein is not correctly predicted to be coding, specified as a per-base pair or per-amino acid rate.

FALSE POSITIVE
(FP). The probability that a segment that is predicted to code for protein is not in fact known to be coding, given as a per-base pair or per-amino acid rate. Note that we only count those exons that have some overlap to the region of the genome that is defined by the cDNA alignment. Exons that lie outside this region are relegated to the over-predictions.

GENE SIZE
The size of the unspliced transcript, including introns. As gene-prediction programs do not detect untranslated regions, we do not include them in this definition.

OUTLIER GENES
Genes the sequence characteristics of which are sufficiently outside the normal range to create problems for *ab initio* gene prediction.

OVER-PREDICTION
Predicted exons that lie entirely outside the region of the genome that is defined by the complementary DNA alignment, but which are part of a prediction that has some overlap with this region. Note the distinction between this and false positives.

PER-AMINO ACID RATE
(Per-aa rate). In computing FPs and FNs, this is the method in which we also insist that the correct amino acids are predicted, which requires that the reading frame is correctly assigned.

PER-BASE PAIR RATE
(Per-bp rate). In computing FPs and FNs, this is the method in which we only ask that the correct nucleotides are predicted, without checking if the reading frame is correctly assigned.

REFSEQ
The division of GenBank that is devoted to full-length reference sequences for experimentally confirmed genes.

SENSITIVITY
A measure of prediction that is equivalent to one minus the false-negative rate.

SERIAL ANALYSIS OF GENE EXPRESSION
(SAGE). A quantitative expression assay that is based on tags that are 10–20 bp in length, which are derived from mRNAs.

SIGNAL TERMS
Short sequence motifs, such as splice sites, branch points, polypyrimidine tracts, start codons and stop codons, that are used to detect exon boundaries.

SPECIFICITY
A measure of prediction that is equivalent to one minus the false-positive rate.

TRAINING SET
A set of known protein-coding sequences that is used to teach the *ab initio* gene-prediction program what the codon-usage patterns look like for a given species.

combined *ab initio* and similarity approach to gene prediction is that even when the presence of a gene is correctly detected, it is possible that only a small piece of it is annotated. Often, the missing portions are those with the large introns. Consider the re-annotation of human chromosome 22 (REF. 16), which benefited from many new cDNAs. Some genes were lost, some genes were gained and many genes were 'fixed'. In the end, the total number of protein-coding genes was virtually unchanged (from 545 to 546), but the sum of the gene sizes increased from 13.0 to 18.6 Mb — a 43% increase.

The sceptical reader might ask if this observed increase in FD rates at gene sizes above 100 kb was the result of size dependent biases in the vertebrate databases, as opposed to the difficulty of filtering out all the FP exons. We therefore ran our Ensembl simulations again, using the RefSeq genes to query the databases, instead of the GenScan predictions. The resulting rates for FP, FN and CM were unchanged. The only significant change was in FD rate which remained around 50%, regardless of the gene size (data not shown). This confirms that the observed size dependency was the result of how Ensembl interacts with GenScan, and that it is always the large genes that suffer.

Common misconceptions

It is understandable that many biologists associate FPs with gene predictions, as it likely that, at some point in their career, an FP error would have wasted their time. FN errors are only missed opportunities. Moreover, the combined *ab initio* and similarity method of gene prediction is a recent approach. The idea that FNs are now a big problem, rather than FPs, has not yet been widely acknowledged; neither has the fact that a correctly detected gene might be only partially annotated. However, the genome annotators have made no efforts to hide this. For example, in human chromosome 20 (REF. 28), the gene predictions were divided into known, new and putative genes, on the basis of the extent of the experimental confirmation. The mean gene sizes for the three categories were reported to be 51.3, 25.1 and 9.1 kb, respectively. Predicted genes that lacked confirmation from a full-length cDNA had gene sizes that were, at best, roughly half of what they should have been. This is consistent with the FN, CM and FD rates in our Ensembl simulations, and justifies, in retrospect, our seemingly arbitrary 90% rule.

Recent experiments have indicated that there is a substantial FN problem in the human genome annotations. However, the interpretations focused on the initial gene count estimates of 30,000–40,000 (REFS 3,29). There was no differentiation between missing genes and missing exons from genes that were at least partially detected. In one example, SERIAL ANALYSIS OF GENE EXPRESSION (SAGE) experiments[30] found many new exons that were not in the annotations, but their criterion for declaring that an exon belongs to a new gene was that it be found more than 5 kb away from an annotation. As the mean intron size is 5 kb, and as annotations tend to fragment near big introns, this criterion is clearly inadequate. Microarray experiments[31] are more difficult to evaluate because of their low signal-to-noise ratios, which make the detection of individual exons impossible. Taken cautiously, they also report a massive FN problem. The growing consensus from the analysis of the large number of full-length mouse cDNAs gathered by FANTOM[32] is that the number of protein-coding genes will be ~35,000. This is within the predicted range from the initial annotations of the human genome, but it is a long way off the figure of 24,847 Ensembl annotated genes that was taken as the official count for the human gene sweepstakes (BOX 1).

There is a deeper problem associated with the fact that FDs are found mostly in the largest genes. Reports of 'gene deserts' date from the initial annotation of human

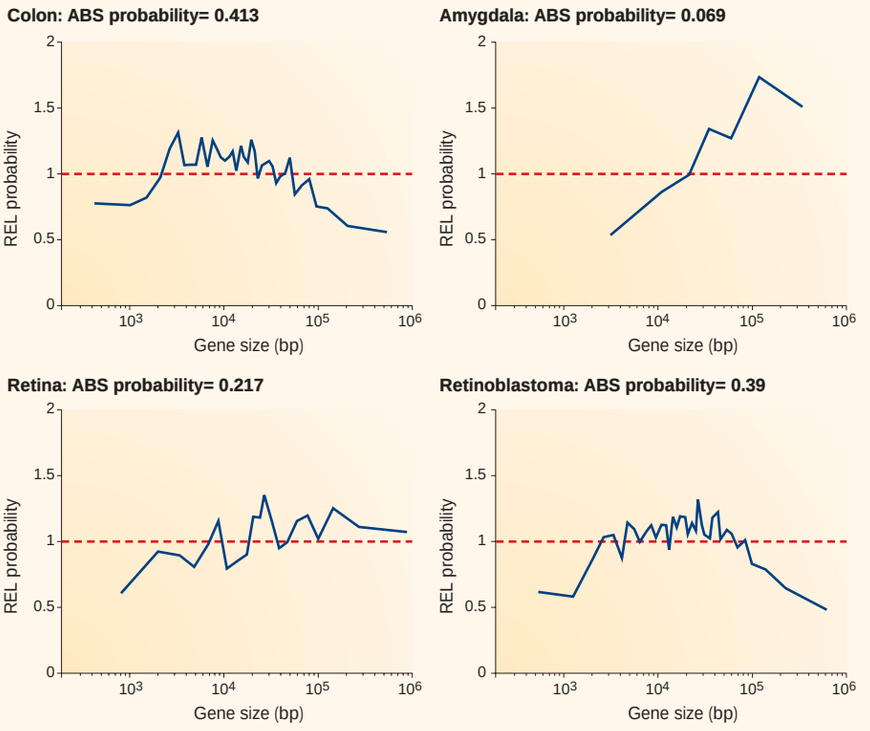

Figure 6 | **Size dependency in tissue-specific expression.** The absolute (ABS) probability is the likelihood that at least one matching expressed-sequence tag (EST) is found in a tissue, averaged over that subset of genes. Given a range of gene sizes, we define the relative (REL) probability as the likelihood that at least one matching EST is found in a tissue, averaged over that subset of genes and divided by the absolute probability. Each data point incorporates as many genes as necessary to get ~75 genes with a matching EST. Tissues depicted are colon, AMYGDALA, retina and retinoblastoma.

Table 1 | **FgeneSH predictions with simulated sequencing errors**

Sequencing errors per bp	Substitution		Insertion/deletion	
	FP	FN	FP	FN
10^{-2}	0.32	0.29	0.58	0.66
10^{-3}	0.30	0.14	0.33	0.23
10^{-4}	0.30	0.12	0.30	0.14
0	0.30	0.12	0.30	0.12

False positive (FP) and false negative (FN) rates are computed for a broad range of single-base substitution and insertion/deletion errors. Overall FP and FN rates are calculated across the entire set of genes, as opposed to calculating a mean of the per-gene rates.

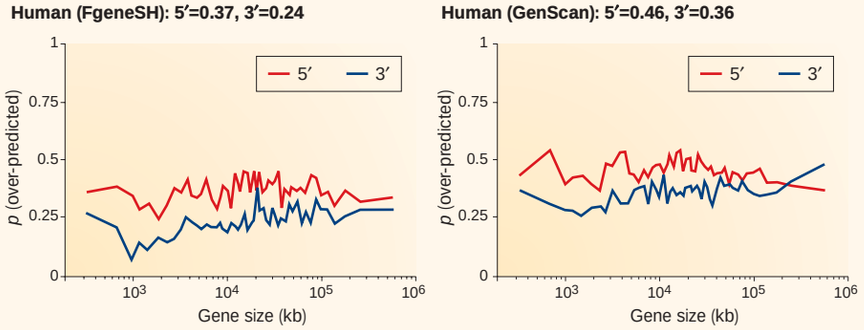

Figure 7 | **Size independence of the over-prediction problem.** The probability (*p*) that a gene is over-predicted at the 5′ or 3′ end (predicted gene boundary extends beyond the actual gene) is shown as a function of the size of the actual gene.

chromosome 21 (REF. 33), through the initial annotation of the draft genome and up to the present day. However, the possibility must be considered that some of these 'deserts' are either CM genes with no clear homologues in the vertebrate databases, or partially annotated large genes that are missing large fragments as a result of the FD problem. If this was extrapolated to its logical limit, the need for large amounts of intergenic sequence becomes questionable, as a small number of large genes would be enough to account for most of the gene deserts (see BOX 5 for a discussion on how this misconception has led to talk of mysterious 'dark genes').

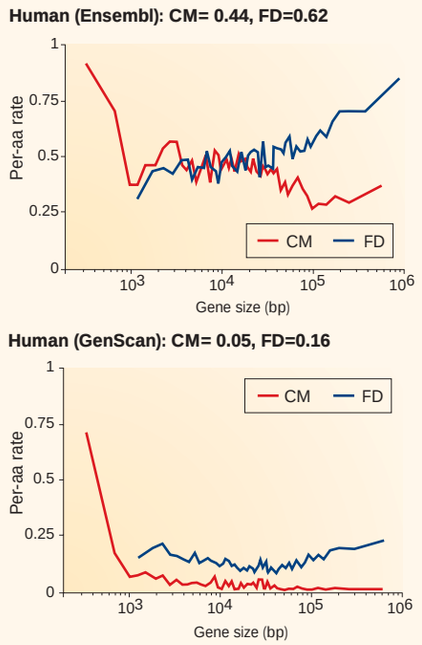

Figure 8 | **Complete and partial failure to detect a gene.** A complete miss (CM) is a gene in which less than 100 bp of its protein-coding sequence is correctly predicted. After eliminating CM genes, a FALSE DESERT (FD) is the fraction of a gene sequence, including its introns, that is not covered by any of the gene predictions. We plotted CM and FD rates as a function of the size of the actual gene.

Epilogue

Given the imperfect nature of the gene-prediction process, what should a potential user do? It depends on which errors will be most detrimental: FPs or FNs. If FPs are troublesome for a particular application, Ensembl annotations should be used, as their overall FP rates are a mere 7%. It should be remembered, however, that some genes might be missing and that, even when a gene is predicted, it is possible that only a small portion of it is described by the annotations, particularly if it is a large gene. Conversely, if FNs are more troublesome, the raw *ab initio* predictions can be used, which are available from the Santa Cruz site[12]; only 5% are completely missed at the gene level. It is useful to remember that the predictions are fragmentary, with many FPs, especially in the large genes. Extreme caution should be exercised when the predicted gene size is unusually small. The threshold varies depending on the program — it is 1 kb for GenScan, but closer to 10 kb for FgeneSH. If all that is needed is a few hundred bases of reliable coding sequence, for example, to be used as a probe in an expression array, it would be recommended to choose from the middle of the predicted gene. Finally, when searching for the rest of a gene, where a small piece is already known, remember that the mean gene size is at least 50 kb and that megabase-sized genes are not unheard of. The search should not be abandoned after only a few kilobases.

As more genomes are sequenced, a newer *ab initio* gene-prediction method, based on a combination of cross-species comparisons and the original codon-usage ideas[34], will become more popular. FP rates are reduced, without having to resort to comparisons against known genes or proteins. It is a genuine improvement, but it is not a panacea. Fundamentally, the process is just as statistical in nature as GenScan and FgeneSH (BOX 2). By combining sequence conservation with codon usage, statistical power is increased, but this does not eliminate the problem of FP errors in large genes. It is certainly possible that the problem will be shifted to larger genes, but will it be enough to make these programs reliable? The preliminary results indicate that it will not.

Some of these cross-species *ab initio* gene-prediction algorithms have been tested on the human and mouse genomes. It is difficult to do a direct comparison with the results presented here, because of the different definitions of performance that are used by different authors. For example, TwinScan[35] reported an improvement in nucleotide specificity, from 29.57% for GenScan to 44.14% for TwinScan. Superficially, specificity is equivalent to one minus the FP rate, but on closer inspection, it is found that the TwinScan analysis did not separate falsely predicted exons in the gene

Box 5 | Intergenic sequence and the problem of 'dark genes'

Although the concept of an intergenic sequence is certainly valid, an important point that is commonly forgotten is that it is not easy to prove that a particular sequence is intergenic. Just because the present programs cannot find a gene in a sequence does not prove that there is no gene there. Nevertheless, the fact that an estimated one-third to two-thirds of the human genome has no detectable genes has led to talk of mysterious 'dark genes' that might one day be found there, as an analogy to the dark matter that astrophysicists famously refer to. We believe that resorting to such terminology is premature when there is an obvious potential solution. Many genes are completely missed, and even when the presence of a gene is correctly detected, a large proportion of its content might not be annotated (especially if a gene is >100 kb). This warps our estimates of the amount of intergenic sequence. After all, a single mispredicted 500 kb gene is equal to a hundred mispredicted 5 kb genes. We have previously estimated[39] that genes that are larger than 100 kb constitute 16.5% of all protein-coding genes, on the basis of the total number of genes, but 70.4% of the gene set, on the basis of the sum of their gene sizes. The implication is that the underlying justification for 'dark genes' can be eliminated partly by adding back genes that were completely missed, and partly by adding back the large introns in those large genes that were only partially annotated.

region from over-predicted exons outside the gene region (BOX 4). But this is not the main point: the point is that their predictions are still far from perfect. SGP2 (REF. 36) reports were similar. Experimental verification of the predicted genes from these two programs found 1,019 new mammalian genes[37]. However, without meaning to trivialize this result, 1,019 new genes is a 'drop in the ocean' in relation to 30,000 or 40,000 genes.

Some would argue that the cell itself is not looking at patterns of codon usage or at cross-species conservation when it transcribes and splices a gene. It must be using a set of deterministic rules that could be followed for gene prediction. The problem is that no one knows what these rules are, particularly for the large genes. It is astonishing that genes that are more than one megabase in size can be processed at all. So, until the molecular mechanisms of transcription and splicing are better understood, statistical approaches to gene prediction will continue to dominate, and biologists must learn to appreciate their limitations.

Jun Wang, ShengTing Li, Yong Zhang, HongKun Zheng, Zhao Xu, Jia Ye, Jun Yu and Gane Ka-Shu Wong are at the Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China.

Jun Wang, Jun Yu and Gane Ka-Shu Wong are at the James D. Watson Institute of Zhejiang University, Hangzhou Genomics Institute, Key Laboratory of Bioinformatics of Zhejiang Province, Hangzhou 310007, China.

Yong Zhang is at the College of Life Sciences, Peking University, Beijing 100871, China.

Jun Yu and Gane Ka-Shu Wong are at the UW Genome Center, Department of Medicine, University of Washington, Seattle, Washington 98195, USA.

J. W., S. T. L. and Y. Z. contributed equally to this work.
Correspondence to G. K.-S. W.
e-mail: gksw@genomics.org.cn
doi:10.1038/nrg1160

1. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
2. Okazaki, Y. *et al*. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
3. Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
4. Aparicio, S. *et al*. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
5. Misra, S. *et al*. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol.* **3**, 0083.1–0083.22 (2002).
6. Reboul, J. *et al*. *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genet.* **34**, 35–41 (2003).
7. Stein, L. Genome annotation: from sequence to biology. *Nature Rev. Genet.* **2**, 493–503 (2001).
8. Zhang, M. Q. Computational prediction of eukaryotic protein-coding genes. *Nature Rev. Genet.* **3**, 698–709 (2002).
9. Hubbard, T. D. *et al*. The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
10. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
11. Kent, W. J. BLAT — the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
12. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
13. Bennetzen, J. L. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell* **12**, 1021–1029 (2000).
14. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).
15. Harrison, P. M. & Gerstein, M. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318**, 1155–1174 (2002).
16. Collins, J. E. *et al.* Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13**, 27–36 (2003).
17. Eddy, S. R. Computational genomics of noncoding RNA genes. *Cell* **109**, 137–140 (2002).
18. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
19. Salamov A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
20. Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433 (2001).
21. Tennyson, C. N., Klamut, H. J. & Worton, R. G. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nature Genet.* **9**, 184–190 (1995).
22. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
23. Rogic, S., Mackworth, A. K. & Ouellette, F. B. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11**, 817–832 (2001).
24. Burset, M. & Guigo, R. Evaluation of gene structure prediction programs. *Genomics* **34**, 353–367 (1996).
25. Heilig, R. *et al*. The DNA sequence and analysis of human chromosome 14. *Nature* **421**, 601–607 (2003).
26. Ashburner, M. A biologist's view of the *Drosophila* genome annotation assessment project. *Genome Res.* **10**, 391–393 (2000).
27. Claverie, J. M. Do we need a huge new centre to annotate the human genome? *Nature* **403**, 12 (2000).
28. Deloukas, P. *et al*. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).
29. Venter, J. C. *et al*. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
30. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nature Biotechnol.* **20**, 508–512 (2002).
31. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
32. Okazaki, Y. & Hume, D. A. A guide to the mammalian genome. *Genome Res.* **13**, 1267–1272 (2003).
33. Hattori, M. *et al*. The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
34. Ureta-Vidal, A., Ettwiller, L. & Birney, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature Rev. Genet.* **4**, 251–262 (2003).
35. Flicek, P., Keibler, E., Hu, P., Korf, I. & Brent, M. R. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res.* **13**, 46–54 (2003).
36. Parra, G. *et al.* Comparative gene prediction in human and mouse. *Genome Res.* **13**, 108–117 (2003).
37. Guigo, R., *et al*. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl Acad. Sci. USA* **100**, 1140–1145 (2003).
38. Pearson, H. Geneticists play the numbers game in vain. *Nature* **423**, 576–576 (2003).
39. Wong, G. K., Passey, D. A. & Yu, J. Most of the human genome is transcribed. *Genome Res.* **11**, 1975–1977 (2001).

Acknowledgements
We thank E. Eyras at the Sanger Center, UK, for explaining the details of the Ensembl procedures to us. This work was sponsored by the Chinese Academy of Sciences, Commission for Economy Planning, Ministry of Science and Technology, National Natural Science Foundation of China, Beijing Municipal Government, Zhejiang Provincial Government and Hangzhou Municipal Government. Some of this work was also supported by the National Human Genome Research Institute.

Online links

FURTHER INFORMATION
BLAST: http://www.ncbi.nlm.nih.gov/BLAST
BLAT: http://genome.ucsc.edu/cgi-bin/hgBlat
Ensembl: http://www.ensembl.org
FANTOM: http://fantom.gsc.riken.go.jp
FgeneSH: http://www.softberry.com/berry.phtml?topic=gfind
Gene Ontology: http://www.geneontology.org
GeneMark: http://opal.biology.gatech.edu/GeneMark
GenScan: http://genes.mit.edu/GENSCAN.html
RefSeq: http://www.ncbi.nlm.nih.gov/RefSeq
SGP2: http://www1.imim.es/software/sgp2
TwinScan: http://genes.cs.wustl.edu
UCSC Human Genome Browser:
http://genome.ucsc.edu/cgi-bin/hgGateway
Access to this interactive links box is free online.