Ab initio vertebrate gene predictions — the problem of big genes

Jun Wang,[1,2,*] ShengTing Li,[1,*] Yong Zhang,[1,3,*] HongKun Zheng,[1] Zhao Xu,[1] Jia Ye,[1] Jun Yu,[1,2,4] and Gane Ka-Shu Wong.[1,2,4,¶]

[1]*Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China.*

[2]*James D. Watson Institute of Zhejiang University, Hangzhou Genomics Institute, Key Laboratory of Bioinformatics of Zhejiang Province, Hangzhou 310007, China.*

[3]*College of Life Sciences, Peking University, Beijing 100871, China.*

[4]*UW Genome Center, Department of Medicine, University of Washington Seattle, WA 98195, USA.*

[*]*These authors contributed equally to this work.*

[¶]Corresponding author. E-MAIL **gksw@genomics.org.cn**.

Preface

To find unknown protein-coding genes, annotation pipelines use a succession of *ab initio* gene prediction and similarity to experimentally confirmed genes (and proteins). Here we show that, although the *ab initio* predictions have an intrinsically high false positive rate, they have a consistently low false negative rate. Incorporation of similarity information is meant to reduce this false positive rate, but in doing so it increases the false negative rate. The crucial variable is gene size (with introns) — the worst errors are always found at the most extreme sizes, and especially in the biggest genes.

We live in the halcyon days of large-scale DNA sequencing. Accompanying each release of any sequenced genome is a list of genes, a large fraction of which are computer predictions. Experimental confirmation in the form of sequenced transcripts in full-length cDNAs, instead of in incomplete expressed-sequence-tags (ESTs), is extensive for mouse (Waterston, et al. 2002; Okazaki, et al. 2002), less so for human (Lander, et al. 2001), and non-existent for pufferfish (Aparicio, et al. 2002). For invertebrate genomes, cDNAs are less critical because the genes are smaller and easier to predict, but nonetheless the many fruitfly (Misra, et al. 2002) and nematode (Reboul, et al. 2003) cDNAs produced after the genomes were sequenced have been invaluable in finding residual errors in the definition of exon boundaries. Given how difficult it is to get cDNAs that are expressed transiently, or at low levels, in specific tissues and developmental stages, predicted genes will remain a fixture of DNA sequence analysis for the indefinite future. Therefore it is imperative for the biologists who use these gene predictions to understand what these programs can and cannot do. By some measures, they are better than commonly thought, and by others they are worse. It is tempting to dismiss these programs as being inherently unreliable (Box 1), but in fact, they fail for specific reasons that can be understood with a minimum of jargon and without delving into algorithmic minutiae.

ANNOTATION PIPELINES have been comprehensively reviewed (Stein 2001). Every pipeline will of course incorporate information from known genes, and no pipeline would ever substitute a predicted gene for a known one. However, one of the main justifications for spending the enormous amounts of money that are being spent on genome sequencing is to identify new genes for which there is only partial or no previous information. This is done using a combination of *ab initio* gene prediction, a statistical process to find protein-coding genes (Zhang 2002), and similarity to experimentally confirmed genes or proteins. In the *Ensembl* pipeline (Hubbard, et al. 2002), 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, the resulting gene predictions are "fixed" by incorporation of similarity information. *Ensembl* keeps only those exons that exhibit 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, based on our analyses of a set of 7,485 REFSEQ (Pruitt

and Maglott 2001) full-length human cDNAs with a perfect *BLAT* (Kent 2002) alignment in the human genome sequence. All of the requisite sequences were downloaded from the Santa Cruz genome browser (Kent, et al. 2002). The cDNA sequences were dated August 2002, and the genome sequences were dated June 2002.

One would expect that, if these programs work particularly well on any sequence, they should work best on RefSeq, because these are the genes over which they have been tuned, as opposed to truly unknown genes. However, even for this idealized gene set, the programs are by no means perfect, and analyzing these imperfections can be instructive. Notice that in running these programs, we tried to be as realistic as possible. For example, we study the entire chromosome, not a pre-selected region with the RefSeq gene already excised. Intergenic sequences are only considered indirectly, but they will not be modeled explicitly, because no one knows what they are. In plant genomes, they are known to be nested clusters of LTR retrotransposons (Bennetzen 2000), but in vertebrate genomes, the situation is more complicated. Retrotransposons are found in abundance inside the introns of vertebrate genes, even as they are not found inside the introns of plant genes (Yu, et al. 2002). Given what we now know, it would be intellectually dishonest to model intergenic sequences as either random sequences or transposons.

Pseudogenes are another issue. Vertebrate genomes are riddled with pseudogenes, particularly single-exon processed pseudogenes (Harrison and Gerstein 2002), and large numbers of these are often incorrectly predicted as genes. For example, many such errors were found in the re-annotation of human chromosome 22 (Collins, et al. 2003). Because the difference between a pseudogene and a real gene can be a single base pair, removing pseudogenes is too lofty a goal for the initial gene prediction. Here, we will only consider whether or not the exon boundaries are correctly defined, because that is the fundamental problem for *ab initio* gene prediction.

Ab initio gene predictions

Ab initio gene predictions rely on two classes of sequence information, which are called signal and content terms. Signal terms refer to those short sequence motifs (such as

the splice sites, branch points, poly-pyrimidine 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; but for the vertebrate genomes, where 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 their surrounding non-coding sequences by a statistical detection algorithm (Box 2). Reliance on codon usage has implications. First, the programs must be taught what these 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 try (Eddy 2002). Finally, none of these programs can detect alternatively spliced transcripts.

We focused on two of the most popular *ab initio* programs, *GenScan* (Burge and Karlin 1997) and *FgeneSH* (Salamov and Solovyev 2000). The former is the default used by *Ensembl*. These two programs share the same overall strategy, and the main difference is that *GenScan* places more emphasis on the content terms, while *FgeneSH* places more emphasis on the signal terms. One might expect the programs to stumble on outlier genes with radically different characteristics from the training sets. Perhaps, among genes with restricted expression patterns that are not in the databases, there are rather different codon usage patterns. Such an assertion is just about impossible to disprove, but considering the many thousands of vertebrate cDNAs that have been sequenced, it is far more reasonable to assume that we have a sufficiently representative sampling. To the extent that there are systematic problems with the *ab initio* predictions, we can show that they are functions of variables that are not directly related to codon usage patterns. What we will do is consider different measurements of performance, as a function of all potentially relevant variables (Box 3), and then see if any of these measurements of performance are severely degraded at extreme values of these variables.

A worst-case example, containing 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 false positive and false negative errors. We define the false positive (FP) rate as the probability that a base that is predicted to be coding does not correspond to a protein-coding base, as determined by cDNA alignments (Box 4). Conversely, we define the false negative (FN) rate as the probability that a protein-coding base is not predicted to be coding. Previous assessments computed what we would call "per-bp" (per-base pair) rates, which do not require that the reading frames be correct. However, for most people, having the correct reading frame is critical. We therefore show "per-aa" (per-amino acid) rates that take this into account. The overall difference is small. Insisting that the reading frames be correct only increases the error rates by 1%. But, it is a conceptually important point, and reassuring to know. We plotted FP and FN 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. In the end, we discovered that the single 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 with all of the introns between the start and stop codons. In some minds, gene size refers to the coding region without the introns — we will refer to this as "CDS size". The mean gene size in our data set of RefSeq genes is 46.1 Kb. If we include the UTRs too, it would be 52.8 Kb. 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 a trivial single-base discrepancy. Considering that a change of three orders of magnitude in gene size is accompanied by less than one of magnitude change in the number of exons (Fig. 2), it is clear that big genes in human are mostly attributable to big introns. At the other extreme, small genes of less than 1 Kb in size are usually aligned as single-exon genes. Based on how these FP and FN rates vary as a function of gene size (Fig. 3), it is clear that big multiple-exon genes and small single-exon genes both present problems for *GenScan* and *FgeneSH*, but for different reasons.

Big genes. As gene size increases, FP goes up, but FN does not. This makes sense, given the nature of the gene prediction algorithms, which search for characteristics of the coding sequence that are only correct in a statistical sense (Box 2). Hence, there is always some probability of an FP (or FN) error and this increases with the length of the sequence under consideration. The difference is that, for FP rates, the sequence under consideration is the introns, but for FN rates, it is the exons. Given a sufficiently large intron, every *ab initio* program will predict an exon where there is none. An increase in FP with increases in gene size is an intrinsic property of all *ab initio* programs. In 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 is not sensitive to increases in gene size. So, thinking of the FP problem again, it is clear that an FP exon is highly likely to contain triplets that get interpreted as stop codons. This leads to premature termination of the predicted gene, and subsequently to gene fragmentation, where multiple gene fragments are predicted instead of the one gene. This is a particularly serious problem at gene sizes over 100 Kb (Fig. 4). By the same logic, predictions of predominately FP exons are likely to have a small gene size. When the predicted gene size is below 1 Kb (*GenScan*), or 10 Kb (*FgeneSH*), one should expect that most of the exons will be FP exons (Fig. 5). This simple rule-of-thumb can be used to filter out the worst predictions.

Considering that big genes are such a problem for *ab initio* prediction, one has to wonder if there is any biological significance to a gene being so big. For example, are the big genes concentrated in specific functional classifications? We divided our data set into groups based on gene size, and classified functions using *Gene Ontology* (Gene Ontology Consortium 2001). No significant differences were observed. However, tissue specificity, as estimated by the presence of at least one 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 to 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, big genes are expressed at least as often, and sometimes more so, than small genes. In contrast, for fast dividing cells, such as those in carcinomas, big genes are expressed less often. This is

consistent with the long transcription times that are required for big genes, typically 7 hrs per megabase (Tennyson, et al. 1995). For fast dividing cells, there is not enough time to complete the transcription of a big gene before the next mitosis.

Small genes. Single-exon genes smaller than 1 Kb present a different problem. FP and FN both increase in the limit of small genes. One might think that these would be the easiest genes to predict since, without the introns, this problem is similar to finding genes in bacteria, where reliable programs such as *GeneMark* (Lukashin and Borodovsky 1998) 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. Since most vertebrate genes have many more than one exon, this problem has been relegated to a lower priority. In our data set, 4.5% of the genes are single-exon, and this is likely to be an over-estimate, because even RefSeq can be contaminated by pseudogenes. In contrast, for invertebrate and plant genomes, single-exon genes are a larger fraction 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 relevant variable. Simple averages over the total gene set were reported, and to the extent that dependences were studied, only exon size and GC content were considered (Rogic, et al. 2001). 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 miscalled is insignificant, compared to the more serious problem of big genes. Moreover, many of these earlier analyses used a data set with a mean gene size of only 5 to 10 Kb (Burset and Guigo 1996), because in those days, large megabase-sized genomic contigs were not available. It is interesting how mean gene sizes have increased as finishing of the human genome sequence has progressed. For the initial analyses of the draft sequence, mean gene size was only 27 Kb, but two years later, it was 51 to 59 Kb in finished chromosomes 14, 20, and 21 (Heilig, et al. 2003).

Sequence errors. One could legitimately ask how sequence errors might affect the accuracy of the gene predictions. To address this issue, we introduced random single-base substitution and insertion-deletion errors into the human genome sequence, and then ran the *ab initio* programs again (Table 1). In general, substitution errors do not significantly affect FP, but for substitution errors of more than 1E-3, there is a marked increase in FN. Insertion-deletion (indel) errors are less tolerated, given that they affect FP at indel rates of more than 1E-3, and FN at rates of more than 1E-4. Nonetheless, these results indicate that the current standard for "finished" sequence, set by the public consortium to be 1E-4, is more than adequate for gene identification purposes.

Over-prediction of genes

Quite often, the gene boundaries are poorly defined, in the sense that the predicted gene does not terminate at the start and stop codons. Most assessments of *ab initio* gene prediction confuse this problem with FP, but we believe it should be treated as a distinct phenomenon that we call "over-prediction" (Box 4) because, unlike FP, the probability of over-prediction is independent of gene size (Fig. 7). Over-predictions usually result from a failure to detect the terminal exons that contain the start and stop codons. For *FgeneSH*, 43% of the 5' over-predictions, and 92% of the 3' over-predictions, result from a missing terminal exon. Similarly, we found for *GenScan* that 45% of the 5' over-predictions, and 94% of the 3' over-predictions, result from a missing terminal exon. Even when the start codon is correctly detected, the program may 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' boundary problem 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 goes for a substantial fraction of the 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 should lie within this over-prediction distance, the *ab initio* program will often merge the two genes together, into a single prediction. It is difficult to determine how often this has happened because we do not have all the cDNAs, and so we do not know which genes are adjacent to each other. What we can say is that the over-prediction probability establishes an upper bound for the likelihood that two adjacent genes are merged into one prediction. We believe that over-prediction of a first gene may hinder the detection of a neighboring second 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

Considering the problems of FP errors in big genes, and over-predictions in genes of any size, it is not surprising that many biologists are frustrated by the output of these *ab initio* programs (Ashburner 2000; Claverie 2000). To address these concerns, *Ensembl* incorporates similarity information in its pipeline to reduce the incidence of FP errors and over-predictions. However, this has consequences. If we will only accept those genes that look like something already in the databases, we preclude ourselves from identifying new genes. This makes a mockery of one of the main reasons for why we sequence genomes, which is to find new genes. One has to wonder, to what extent are the FP errors and over-predictions being exchanged for FN errors when we incorporate similarity (Box 2)? Here, we will simulate what would have happened if the RefSeq gene had been "unknown", by removing from the vertebrate databases anything with more than 90% amino acid identity to the RefSeq gene. Notice that the 90% rule is only intended to remove sequences for the gene in question, not those homologs that *Ensembl* would have latched onto. Indeed, the set of all homologs to our reference gene set exhibits far less similarity than 90%, with an asymmetric distribution peaked closer to 30 or 40%.

This simulation can be done in a completely realistic manner, because the source code for *Ensembl* is freely distributed. One merely has to hack into their code at the point just before it searches the vertebrate databases, and apply our 90% rule. Given a *GenScan* prediction, one performs a *BLAST* search on this modified database. Using the exact same code as *Ensembl*, one builds an entirely new gene model, by aligning the best hits back to the genome. What we find is that FP is 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 very likely that *Ensembl* also eliminates false positives predicted in intergenic sequence. Getting back to the FN problem, it is important to ask if the FN errors are randomly distributed among genes, or concentrated in specific genes. It turns out that this is a crucial distinction between *GenScan* and *Ensembl*. To demonstrate, we computed for each gene the probability of a complete miss (CM), which we define as the failure to detect even 100 bp of the coding region. A mere 5% of the genes are CM in *GenScan*, but 44% are CM in *Ensembl* (Fig. 8). This massive increase in CM for *Ensembl* is due to those genes that have no remaining homolog in the databases, after the 90% rule is applied. They are simulated "unknown" genes. Their exact number is a function of our simulation parameters. Although this number will shrink as the databases expand, it is not likely to ever be zero, because there will always be a few genes with restricted expression patterns that are entirely missing from the databases.

Figure 1 illustrates that the presence of a big intron caused *Ensembl* to miss the first exon, and created the illusion that about half of this region is a "gene desert". Hence we define a false desert (FD) as the fraction of a cDNA-defined genome region that is not covered by a gene prediction. In *GenScan*, FD is small and independent of gene size, but in *Ensembl*, FD 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 within the big introns of the big genes, which makes it so difficult for *Ensembl* to recognize the few true exons that are correctly detected by *GenScan*. It is an intrinsic property of this combined *ab initio* and similarity approach to gene prediction that even when the presence of a gene is correctly detected, it is possible that only a small piece of it is annotated. Often missing are those portions with the big introns. Consider the re-annotation of human chromosome

22 (Collins, et al. 2003), which benefited from many new cDNAs. Some genes were lost, some genes were gained, and many genes were "fixed". In the end, their total number of protein-coding genes was basically unchanged, from 545 to 546, but the sum of the gene sizes increased from 13.0 to 18.6 Mb, a 43% increase.

One could reasonably ask if this observed increase in FD rates at gene sizes above 100 Kb was due to some size dependent biases in the vertebrate databases, as opposed to the difficulty of seeing past all these 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 essentially unchanged. The only real change was in FD, which hovered 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 big genes that suffer.

Common mis-perceptions

It is understandable why false positives are the first thing many biologists think of when they think about gene predictions. At some point in their life, some FP error wasted their time. FN errors are only missed opportunities. Moreover, the combined *ab initio* and similarity approach to gene prediction is a very recent trend. The idea that false negatives are now a big problem, not false positives, has not sunk into the collective consciousness, let alone the concept that even when a gene is correctly detected, it is possible that only a small piece of it is annotated. To their credit, the genome annotators have made no efforts to hide this fact. For example, in human chromosome 20 (Deloukas, et al. 2001), the gene predictions were divided into known, novel, and putative genes — based on the extent of 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 it justifies in retrospect our seemingly arbitrary 90% rule.

Recent experiments have indicated that there is indeed a substantial false negative problem in the human genome annotations. However, their interpretations focused on the initial gene count estimates of 30,000 to 40,000 (Lander, et al. 2001; Venter, et al. 2001). They did not distinguish between missing genes and missing exons from genes that are at least partially detected. In one example, SERIAL-ANALYSIS-OF-GENE-EXPRESSION (SAGE) experiments (Saha, et al. 2002) found many novel exons that were not in the annotations, but their criterion for declaring that an exon belongs to a novel gene was that it be found more than 5 Kb from an annotation. Given that the mean intron size is 5 Kb, and the fact that annotations tend to fragment near big introns, this criterion is clearly inadequate. The microarray experiments (Kapranov, et al. 2002) are more difficult to evaluate, because of their low signal-to-noise ratios, which make the detection of individual exons impossible. Taken with a grain of salt, they too report a massive false negative problem. The growing consensus from the analysis of the large number of full-length mouse cDNAs gathered by *FANTOM* (Okazaki and Hume 2003) is that the number of protein-coding genes will be about 35,000. This is within the predicted range from the initial annotations of the human genome, but it is a far cry from the 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 false deserts concentrate in the biggest genes. Continuing reports of "gene deserts", dating from the initial annotation of human chromosome 21 (Hattori, et al. 2000), through the initial annotation of the draft genome, and well into today, must be tempered against the possibility that some of these so-called deserts are either CM genes with no clear homologs in the vertebrate databases, or partially annotated big genes that are missing big pieces as a result of the FD problem. If one were to extrapolate this to its logical limit, the need for large amounts of intergenic sequence becomes questionable, since a relatively small number of very big genes would be enough to account for most of the gene deserts (Box 5).

Epilogue

Given the imperfect nature of the gene prediction process, what should a potential user do? It depends on how one feels about false positives versus false negatives. If false

positives are troublesome for a particular application, use the *Ensembl* annotations. After all, their overall FP rates are a scant 7%. One should however remember 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 big gene. Conversely, if false negatives are more troublesome, one can use the raw *ab initio* predictions, which are available from the Santa Cruz site (Kent, et al. 2002). After all, at the gene level, only 5% are completely missed. Simply remember that the predictions are fragmentary, with many false positives, especially in the big genes. Extreme caution should be exercised when the predicted gene size is unusually small. The actual threshold varies with the program. It is 1 Kb for *GenScan*, but closer to 10 Kb for *FgeneSH*. If all that you need is a few hundred bases of reliable coding sequence, say for use as a probe in an expression array, it would be wiser 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 at all 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 older *ab initio* ideas (Ureta-Vidal, et al. 2003), 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 every bit as statistical in nature as were *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 big genes. It is certainly possible that the problem will be shifted to bigger genes, but will it be enough to make these programs reliable? The preliminary results say no.

Some of these cross-species *ab initio* gene prediction algorithms have been tested on the human and mouse genomes. It is difficult for us to do a direct comparison with the results presented here, because of the different definitions of performance that are used in different papers. For example, *TwinScan* (Flicek, et al. 2003) reported an improvement in nucleotide specificity, from 29.57% for *GenScan* to 44.14% for *TwinScan*. Superficially, specificity is equivalent to one minus FP, but upon closer inspection, we realized that the

TwinScan analysis did not separate falsely predicted exons in the gene region from over-predicted exons outside the gene region (Box 4). But that is not really the point. The point is that their predictions are still far from perfect. *SGP2* (Parra, et al. 2003) reported much the same thing. Experimental verification of the predicted genes from these two programs netted 1,019 new mammalian genes (Guigo, et al. 2003). We do not mean to trivialize the result, but we would be remiss in our duties if we did not point out that 1,019 new genes is a drop-in-the-ocean relative 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 we can follow for gene prediction. Although we agree with this assessment, the problem is that no one knows what these rules are, particularly in the big genes. When you think about it, it is astonishing that genes over a megabase in size can be processed at all. Thus 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.

Acknowledgments

We thank Eduardo Eyras at the Sanger Center 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.

References

S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J.M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**: 1301-1310.

M. Ashburner. 2000. A biologist's view of the Drosophila genome annotation assessment project. *Genome Res.* **10**: 391-393.

J.L. Bennetzen. 2000. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell* **12**: 1021-1029.

C. Burge and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**: 78-94. http://genes.mit.edu/GENSCAN.html.

M. Burset and R. Guigo. 1996. Evaluation of gene structure prediction programs. *Genomics* **34**: 353-367.

J.M. Claverie. 2000. Do we need a huge new centre to annotate the human genome? *Nature* **403**: 12-12.

J.E. Collins, M.E. Goward, C.G. Cole, L.J. Smink, E.J. Huckle, S. Knowles, J.M. Bye, D.M. Beare, and I. Dunham. 2003. Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13**: 27-36.

P. Deloukas, L.H. Matthews, J. Ashurst, J. Burton, J.G. Gilbert, M. Jones, G. Stavrides, J.P. Almeida, A.K. Babbage, C.L. Bagguley, et al. 2001. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**: 865-871.

S.R. Eddy. 2002. Computational genomics of noncoding RNA genes. *Cell* **109**: 137-140.

P. Flicek, E. Keibler, P. Hu, I. Korf, M.R. Brent. 2003. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res.* **13**: 46-54.

Gene Ontology Consortium. 2001. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**: 1425-1433. http://www.geneontology.org.

R. Guigo, E.T. Dermitzakis, P. Agarwal, C.P. Ponting, G. Parra, A. Reymond, J.F. Abril, E. Keibler, R. Lyle, C. Ucla, et al. 2003. 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.

P.M. Harrison and M. Gerstein. 2002. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318**: 1155-1174.

M. Hattori, A. Fujiyama, T.D. Taylor, H. Watanabe, T. Yada, H.S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.K. Choi, et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405**: 311-319.

R. Heilig, R. Eckenberg, J.L. Petit, N. Fonknechten, C. Da Silva, L. Cattolico, M. Levy, V. Barbe, V. de Berardinis, A. Ureta-Vidal, et al. 2003. The DNA sequence and analysis of human chromosome 14. *Nature* **421**: 601-607.

T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al. 2002. The *Ensembl* genome database project. *Nucleic Acids Res.* **30**: 38-41. http://www.ensembl.org.

P. Kapranov, S.E. Cawley, J. Drenkow, S. Bekiranov, R.L. Strausberg, S.P. Fodor, and T.R. Gingeras. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**: 916-919.

W.J. Kent. 2002. *BLAT* — the BLAST-like alignment tool. *Genome Res.* **12**: 656-664.

W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. *Genome Res.* **12**: 996-1006. http://genome.ucsc.edu.

E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

A.V. Lukashin and M. Borodovsky. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**: 1107-1115.

S. Misra, M.A. Crosby, C.J. Mungall, B.B. Matthews, K.S. Campbell, P. Hradecky, Y. Huang, J.S. Kaminker, G.H. Millburn, S.E. Prochnik, et al. 2002. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biol.* **3**: research0083.1-0083.22.

Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi, H. Bono, S. Kondo, I. Nikaido, N. Osato, R. Saito, H. Suzuki, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**: 563-573.

Y. Okazaki and D.A. Hume. 2003. A Guide to the Mammalian Genome. *Genome Res.* **13**: 1267-1272.

G. Parra, P. Agarwal, J.F. Abril, T. Wiehe, J.W. Fickett, R. Guigo. 2003. Comparative gene prediction in human and mouse. *Genome Res.* **13**: 108-117.

H. Pearson. 2003. Geneticists play the numbers game in vain. *Nature* **423**: 576-576.

K.D. Pruitt and D.R. Maglott. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**: 137-140. http://www.ncbi.nih.gov/RefSeq.

J. Reboul, P. Vaglio, J.F. Rual, P. Lamesch, M. Martinez, C.M. Armstrong, S. Li, L. Jacotot, N. Bertin, R. Janky, et al. 2003. C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34**: 35-41.

S. Rogic, A.K. Mackworth, and F.B. Ouellette. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11**: 817-832.

S. Saha, A.B. Sparks, C. Rago, V. Akmaev, C.J. Wang, B. Vogelstein, K.W. Kinzler, and V.E. Velculescu. 2002. Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20**: 508-512.

A.A. Salamov and V.V. Solovyev. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* **10**: 516-522. http://www.softberry.com/berry.phtml?topic=gfind.

L. Stein. 2001. Genome annotation: from sequence to biology. *Nat. Rev. Genet.* **2**: 493-503.

C.N. Tennyson, H.J. Klamut, and R.G. Worton. 1995. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat. Genet.* **9**: 184-190.

A. Ureta-Vidal, L. Ettwiller, and E. Birney. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4**: 251-262.

J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al. 2001. The sequence of the human genome. *Science* **291**: 1304-1351.

R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

G.K. Wong, D.A. Passey, and J. Yu. 2001. Most of the human genome is transcribed. *Genome Res.* **11**: 1975-1977.

J. Yu, S. Hu, J. Wang, G.K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296**: 79-92.

M.Q. Zhang. 2002. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* **3**: 698-709.

Figure captions

Figure 1: Actual versus predicted exons in a known gene: TEA domain family member 1 (SV40 transcriptional enhancer factor; GenBank NM_021961), found on human chr-11. Correctly predicted exons are colored blue, while incorrectly predicted exons are colored

red. Arrowed lines indicate the direction of transcription for each predicted gene. RefSeq indicates that there are 11 exons, labeled 1 to 11, spanning a genomic region of 173 Kb. *FgeneSH* has 4 predictions that overlap with this gene, labeled F1 to F4. *GenScan* has 3 predictions that overlap with this gene, labeled G1 to G3. Notice how F2, F3, and G2 are entirely FP exons, within the big 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, labeled E1, but even so, it failed to identify the first exon. About half of this genomic region is incorrectly annotated as a "gene desert", all because of one big intron.

Figure 2: Correlation between gene size and intron size. The number of exons per gene increases by less than one order of magnitude as the gene size increases by 3 orders of magnitude. Given that the exon size is not increasing, most of the increases in gene size must be attributed to increases in intron size.

Figure 3: Size dependencies for false positive (FP) and false negative (FN) rates. Error rates are depicted as a function of the size of the actual gene (from alignment of RefSeq cDNAs to genome) or the size of the exon. Notice that two different exon sizes are used. FP uses predicted exon size, but FN uses actual exon size.

Figure 4: Size dependency for gene fragmentation problem. Here, we show 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.

Figure 5: Detection of erroneous predictions using gene size. If a gene is truly unknown, then so is its actual gene size, but this figure demonstrates that a small predicted gene size is a useful indicator of potentially erroneous predictions.

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

Figure 7: Size independence of the over-prediction problem. The probability 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.

Figure 8: Complete and partial failure to detect a gene. A complete miss (CM) is a gene where fewer than 100 bp of the coding region is correctly predicted. After eliminating CM genes, we compute the false desert (FD) rate as the fraction of the gene region, defined by cDNA alignments, that is mistakenly thought to be a "gene desert". We show CM and FD as a function of the size of the actual gene.

Table Captions

Table 1: *FgeneSH* predictions with simulated sequencing errors. 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 computed across the entire set of genes, as opposed to computing a mean of the per gene rates.

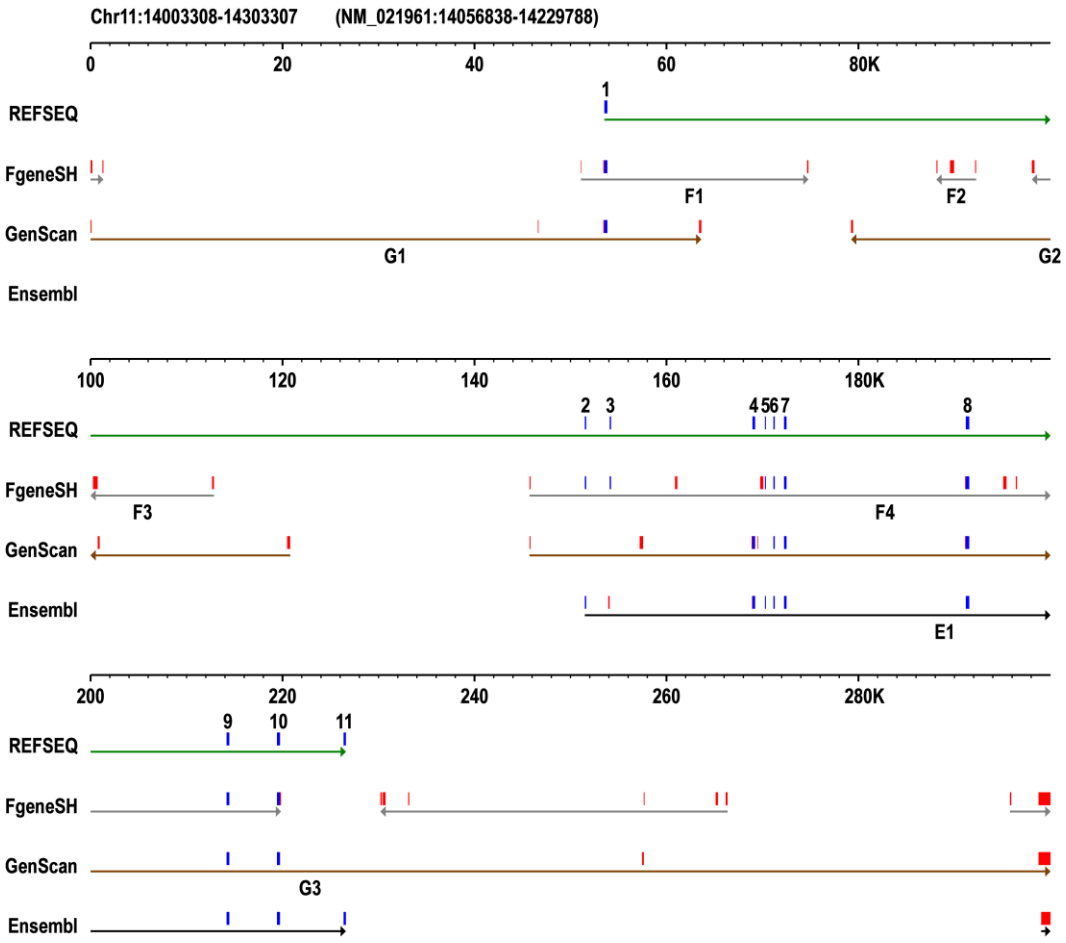
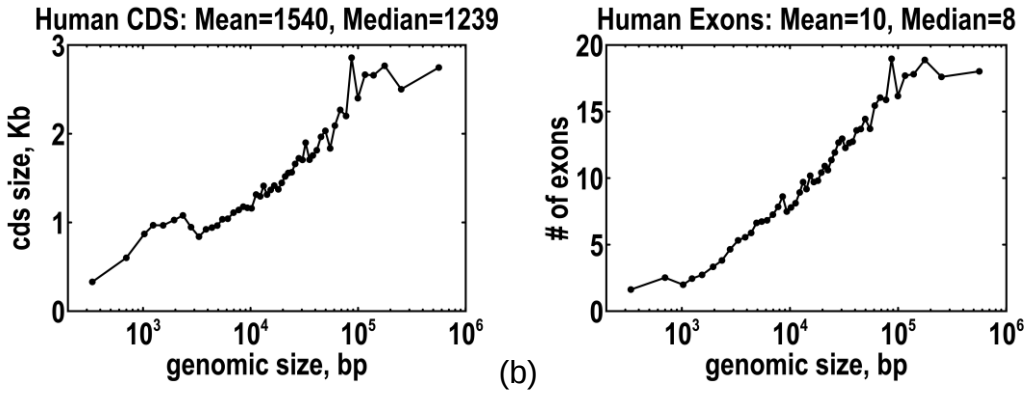
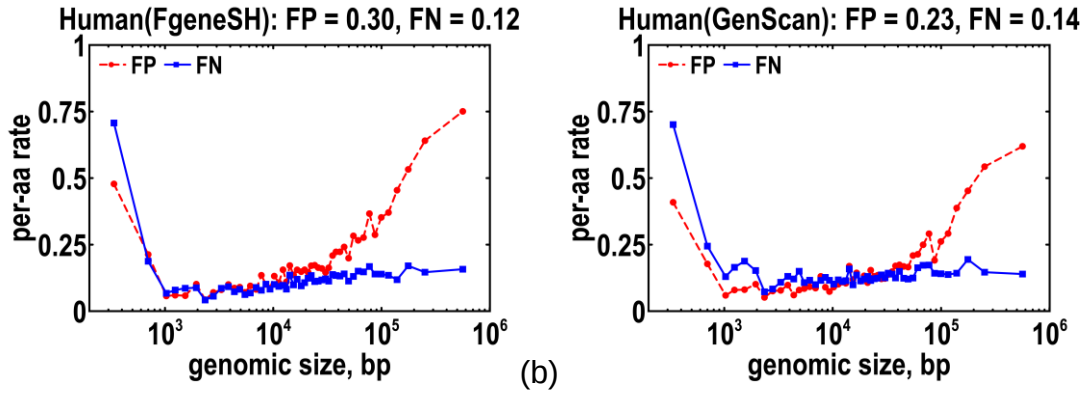

Chr11:14003308-14303307 (NM_021961:14056838-14229788)
0
20
40
60
80K
REFSEQ
1
FgeneSH
F1
F2
GenScan
G1
G2
Ensembl
100
120
140
160
180K
REFSEQ
2 3
4 5 6 7
8
FgeneSH
F3
F4
GenScan
Ensembl
E1
200
220
240
260
280K
REFSEQ
9
10
11
FgeneSH
GenScan
G3
Ensembl

Figure 1.



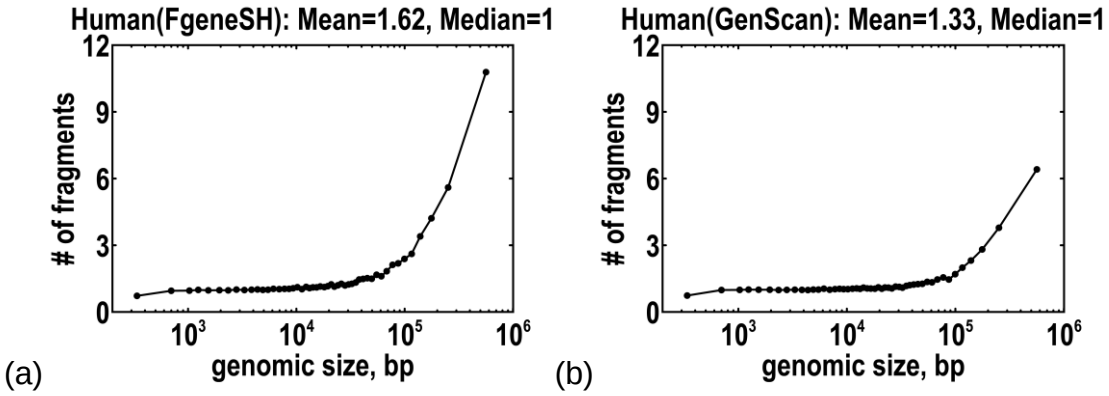
(a) (b)

Figure 2.



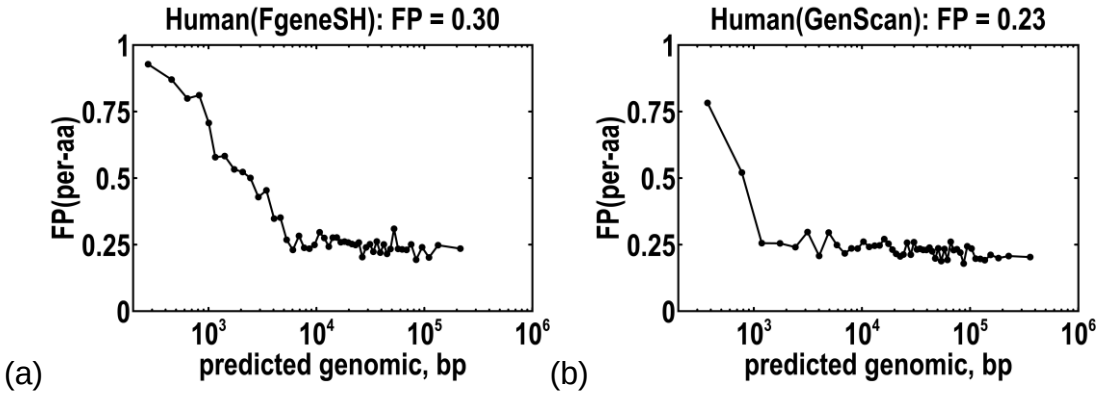
(a) (b)

Figure 3.



(a) (b)

Figure 4.



(a) (b)

Figure 5.

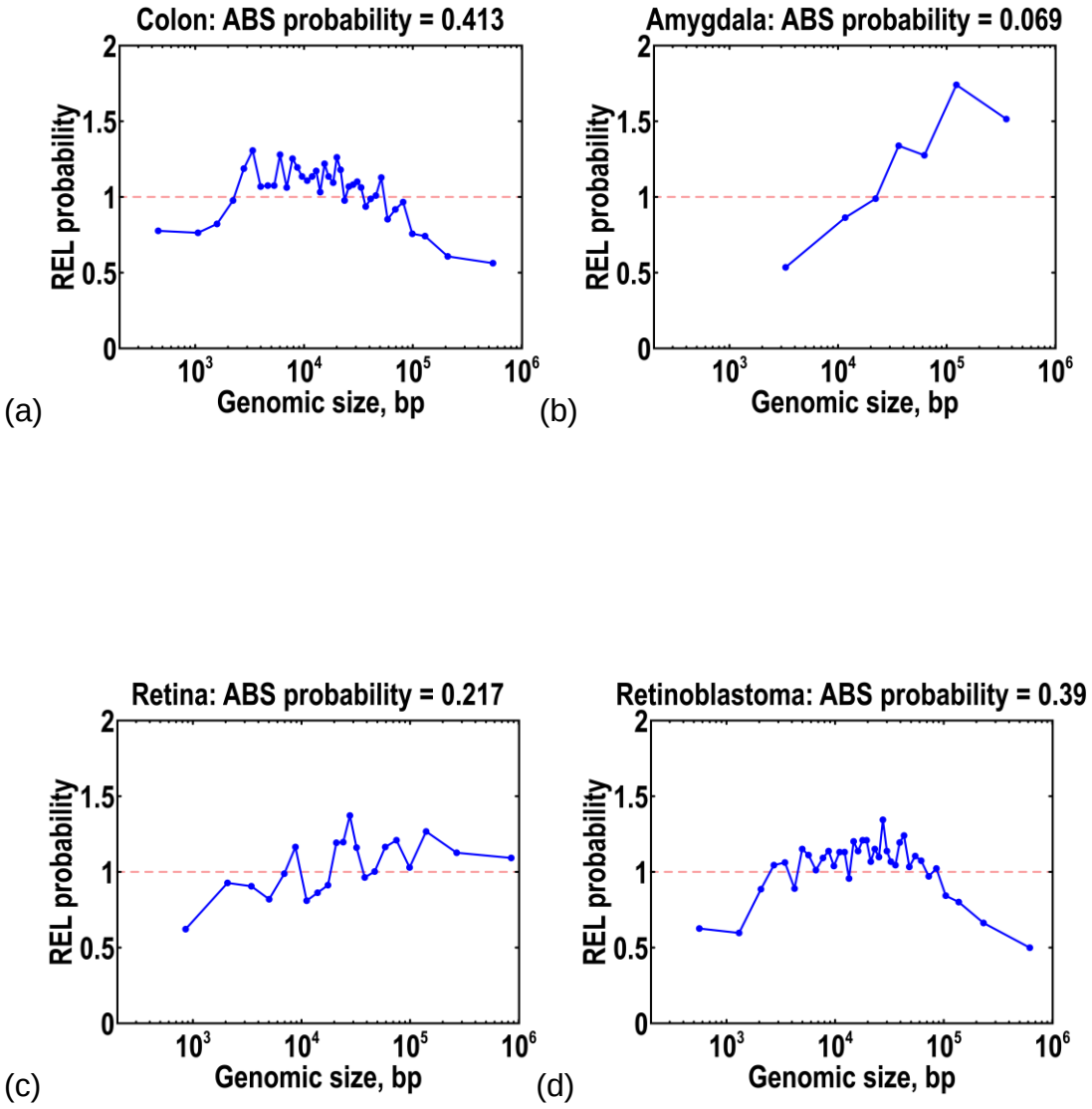

Colon: ABS probability = 0.413
REL probability
Genomic size, bp
(a)
Amygdala: ABS probability = 0.069
REL probability
Genomic size, bp
(b)
Retina: ABS probability = 0.217
REL probability
Genomic size, bp
(c)
Retinoblastoma: ABS probability = 0.39
REL probability
Genomic size, bp
(d)

Figure 6.

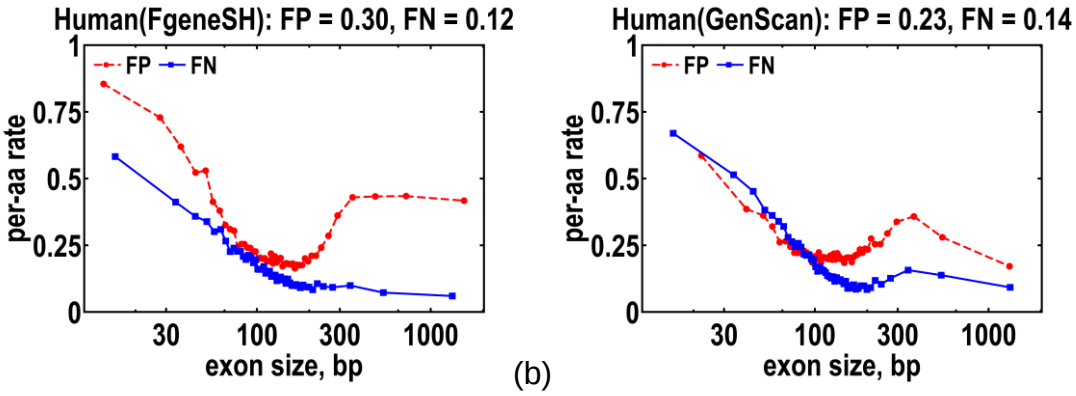


Figure 7.

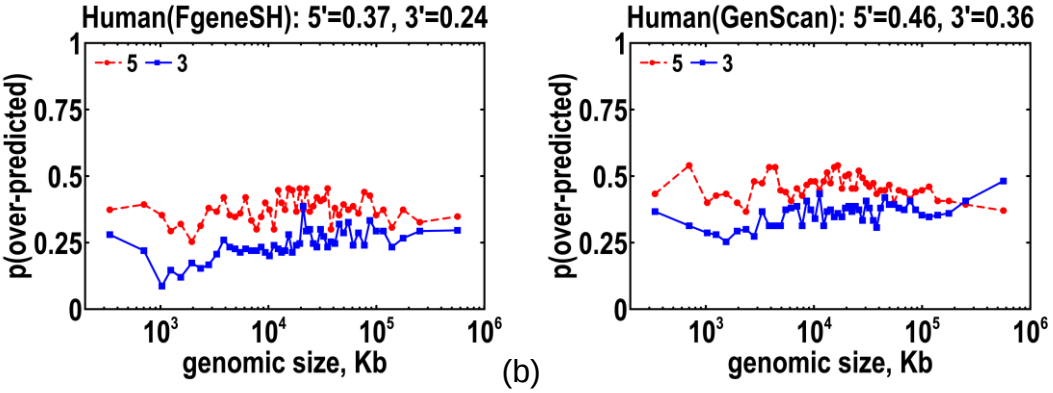


Figure 8.

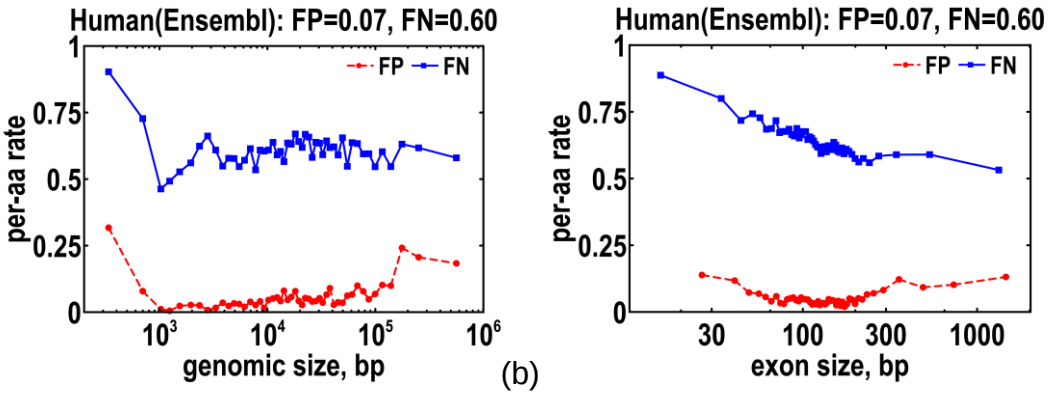


Figure 9.

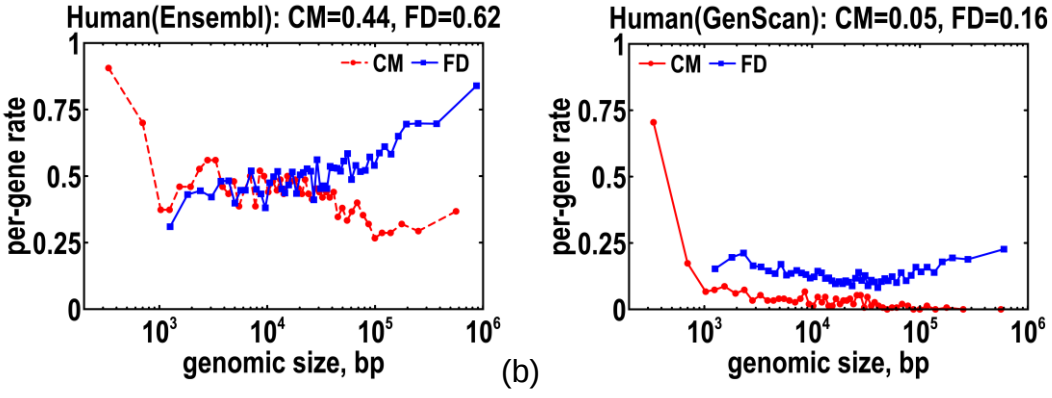


Figure 10.

Table 1.

Error	Substitution		Insertion-Deletion	
Rate	FP	FN	FP	FN
1E-02	0.32	0.29	0.58	0.66
1E-03	0.30	0.14	0.33	0.23
1E-04	0.30	0.12	0.30	0.14
Zero	0.30	0.12	0.30	0.12

Box 1: So, how many genes are there in the human genome?

On May 30th of 2003, at Cold Spring Harbor Laboratory in New York, the winner of the human gene sweepstakes was finally announced (Pearson 2003). Lee Rowen, from the Institute for Systems Biology in 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, a substantially lower number than the 30,000 to 40,000 that had been estimated in February of 2001, with the initial analyses of the draft sequence for the human genome. Were these original estimates too high, or was this latest estimate too low? A little appreciated fact was that 24,847 represented that number of genes for which the organizers felt they had the best supporting evidence, based on sequence similarity to known genes or proteins in the vertebrate databases. It was a lowballed estimate, and they confessed to the media that this was probably not the final answer. In truth, the number of genes in the human genome remains unknown.

Box 2: Intrinsic tradeoffs in gene prediction: Squeezing a balloon

Gene prediction is an intrinsically statistical process. It searches for patterns that are correlated with protein-coding sequence, 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 like *GenScan* and *FgeneSH* search for patterns of codon usage that are specific to protein-coding sequence. Newer programs like *TwinScan* and *SGP2* search for sequence conservation in a related species, in addition to codon usage. The problem with ALL statistical detection algorithms is that there is no guarantee that every instance of the desired pattern is due to protein-coding sequence. Even random non-coding sequence can come up positive, and the longer that sequence is, the more likely this will happen. Hence there is always some false positive (FP) rate. Conversely, there is no guarantee that every instance of protein-coding sequence will result in a detectable pattern, and there is always some false negative (FN) rate. Many ways 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. Tradeoffs like these are quite common in all branches of experimental science, from biology to physics. It is often called "squeezing a balloon", and it even extends to everyday life, for which, the economist Milton Friedman famously said, there is no such thing as a free lunch.

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 numbers. 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 something like this, 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 single most important variable is gene size (with introns). The idea is to compute average performance in groups of genes clustered based on size. Most plotting software will divide a chosen axis into uniformly sized bins. This is a bad idea if the gene size distribution is non-uniform, and there are far more genes in the middle than at the tails. What we wanted to see is how the performance degrades at the tails, and there is no one ideal bin size. Choose too small a bin, and the averages become too noisy at the tails, where there are not enough genes. Too large a bin, and the tails get subsumed into a single bin, thus obscuring any potentially interesting trends. The trick is to put a constant number of genes into each bin, and the solution is to use non-uniform bin sizes, adjusting the number of bins to get a targeted number of genes (say 150).

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? We do not think that it is fair to do so, because some of these exons may be due to 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 defined by the cDNA alignment, but belong to a prediction that does overlap with this region. These are treated as a separate phenomenon, distinct from false positive exons that have some overlap with this region. In practice, the important difference between these two phenomena is that false positives are sensitive to gene size, but over-predictions are not.

Box 5: Intergenic sequence and the problem of "dark genes"

Although the concept of intergenic sequence is certainly valid, an important point that is lost in many publications is that it is not so easy to prove that a particular sequence is intergenic. Just because the current programs cannot find a gene in a sequence does not prove that there is no gene there. Nevertheless, the fact that an estimated 1/3 to 2/3 of the human genome has no detectable genes has lead to talk of "dark genes", for whatever it is that might be found there, in analogy to the dark matter that astrophysicists famously talk about. We believe that it is premature to talk about these things when an obvious solution is sitting right in front of our eyes. As we show in this review, many genes are completely missed, and even when the presence of a gene is correctly detected, a large fraction of its genomic extent may not be annotated, especially for big genes above 100 Kb. This skews our perception of the amount of intergenic sequence. After all, a single miscalled 500 Kb gene is akin to a hundred miscalled 5 Kb genes. We have previously estimated (Wong, et al. 2001) that genes above 100 Kb constitute 16.5% of all protein-coding genes, based on the number of genes, but 70.4% of the gene set, based on the sum of their gene sizes. The implication is that one can eliminate the underlying justification for dark genes, partly by adding back genes that were completely missed, and partly by adding back the big introns in those big genes that were only partially annotated.

Box 6: Glossary of Terms (in order of appearance in manuscript)

Signal terms. Short sequence motifs like splice sites, branch points, poly-pyrimidine tracts, start codons, and stop codons that are used to detect exon boundaries.

Content terms. Patterns of codon usage that are unique to each species and allow protein-coding sequence to be distinguished from surrounding non-coding sequence.

***Ab initio* gene prediction.** A program that identifies protein-coding genes in genomic sequence, using no prior knowledge other than the signal and content terms.

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

CDS size. The size of the spliced transcript, without introns. Since gene prediction programs do not detect UTRs, we do not include them in this definition either.

Gene size. The size of the unspliced transcript, with introns. Since gene prediction programs do not detect UTRs, we do not include them in this definition either.

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

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-bp or per-aa rate.

Over-prediction. Predicted exons that lie entirely outside the region of the genome defined by the cDNA alignment, but which are part of a prediction that has some overlap with this region. Notice the distinction between this and false positives (above).

Sensitivity. This is equivalent to one minus the false negative rate.

Specificity. This is equivalent to one minus the false positive rate.

Per-bp rate. In computing FP and FN, if all that is required is that the base pairs be correctly labeled as coding or non-coding, we call that a "per-bp rate".

Per-aa rate. In computing FP and FN, if getting the reading frame right is necessary, we call that a "per-aa rate", meaning that we correctly called the amino acids.

Complete miss (CM). The probability that fewer than 100 bp (*i.e.*, a typical exon) of the protein-coding region in a known gene is actually predicted as coding.

False desert (FD). Considering the region of the genome defined by the cDNA alignment, the fraction of this region that is not covered by any gene prediction.