# Genome annotation past, present, and future: How to define an ORF at each locus

Michael R. Brent

*Genome Res.* 2005 15: 1777-1786
Access the most recent version at doi:10.1101/gr.3866105

# Genome annotation past, present, and future: How to define an ORF at each locus

Michael R. Brent

*Laboratory for Computational Genomics and Department of Computer Science, Washington University, St. Louis, Missouri 63130, USA*

**Driven by competition, automation, and technology, the genomics community has far exceeded its ambition to sequence the human genome by 2005. By analyzing mammalian genomes, we have shed light on the history of our DNA sequence, determined that alternatively spliced RNAs and retroposed pseudogenes are incredibly abundant, and glimpsed the apparently huge number of non-coding RNAs that play significant roles in gene regulation. Ultimately, genome science is likely to provide comprehensive catalogs of these elements. However, the methods we have been using for most of the last 10 years will not yield even one complete open reading frame (ORF) for every gene—the first plateau on the long climb toward a comprehensive catalog. These strategies—sequencing randomly selected cDNA clones, aligning protein sequences identified in other organisms, sequencing more genomes, and manual curation—will have to be supplemented by large-scale amplification and sequencing of specific predicted mRNAs. The steady improvements in gene prediction that have occurred over the last 10 years have increased the efficacy of this approach and decreased its cost. In this Perspective, I review the state of gene prediction roughly 10 years ago, summarize the progress that has been made since, argue that the primary ORF identification methods we have relied on so far are inadequate, and recommend a path toward completing the Catalog of Protein Coding Genes, Version 1.0.**

The 10 years since *Genome Research* began publication bracket a complete era of genome research—an era of stunning successes and nagging loose ends, promise exceeded and promise as yet unfulfilled. The years 1996–2005 were characterized by tremendous optimism and productivity. In 1996, the sequencing of the human genome was scheduled to be completed in 2005 (Collins and Galas 1993). Driven by competition, automation, and technology, the genomics community far exceeded its own sequencing ambitions. But there was another goal that we have not yet reached—the genome was to provide a "parts list" for the human and other major model organisms. The parts turned out to be more varied than anticipated, and we have learned wonderful things about the biology and history encoded in genome sequences (Waterston et al. 2002; Gibbs et al. 2004). But the most fundamental parts on anyone's list, then and now, must be the complete set of translated open reading frames (ORFs) and the exon–intron structures from which they are assembled. (I will use the term ORF to denote the complete exon–intron structure of the protein coding region of any mature mRNA. Thus, a primary transcript that is alternatively spliced may represent more than one ORF.) After sequencing (Lander et al. 2001; Venter et al. 2001), completing (Collins et al. 2003; The International Human Genome Sequencing Consortium 2003), and finishing (International Human Genome Sequencing Consortium 2004) the human genome, we do not have even one complete, correct ORF for each human gene locus. In fact, we do not have a complete correct ORF for each locus in the genome of any higher eukaryote.

Among the things we have learned by analyzing mammalian genomes are the incredible abundance of alternatively spliced RNAs (Modrek and Lee 2002; Sorek et al. 2004; Kapranov et al. 2005) and retroposed pseudogenes (Torrents et al. 2003; Zhang and Gerstein 2004), as well as the importance of microRNAs, siRNAs, and other non-coding RNAs in gene regulation (Zamore and Haley 2005). We must ultimately work out all the functional alternative splices of mRNAs with their untranslated regions (UTRs), *cis*-regulatory sites, and basic functional categories. We must also identify all the functional non-coding transcripts, their *cis*-regulatory sites, and their basic functional categories (The ENCODE Project Consortium 2004). However, the achievement of these goals appears to be far in the future. As a concrete and achievable, if somewhat arbitrary, milestone along that path, I will focus on identifying at least one complete ORF for each protein-coding gene.

It is abundantly clear that the methods we have been using to identify ORFs for most of the last 10 years are inadequate for finishing the job. In this perspective, I argue that we cannot rely on any of the following to get us through the home stretch of ORF identification:

- obtaining EST or mRNA sequences from randomly selected cDNA clones,
- aligning expressed sequences to loci other than those from which they were transcribed, e.g., to the loci of gene family members or orthologs in other species,
- sequencing more genomes, or
- annotating manually by using human curators.

All of these things are valuable, but none of them is likely to get us to a new, higher plateau in the quest for a complete ORF at each protein coding locus. Instead, we will have to rely on large-scale PCR amplification of specific cDNAs followed by sequencing of the amplicons. To amplify cDNAs, we need reasonably accurate, though not necessarily perfect, gene predictions to use for PCR primer design. The further a prediction is from a true gene structure, the greater the likelihood that PCR primers designed for it will fail. Each failure increases the cost per gene

**E-mail brent@cse.wustl.edu; fax (314) 935-7302.**
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3866105.

identified and may reduce the completeness of the resulting collection of cDNA sequences. This method is feasible and in use today (Guigó et al. 2003; Dike et al. 2004; Wu et al. 2004; Eyras et al. 2005; Wei et al. 2005) but we must continue to drive its cost down by (1) continuing the steady and significant improvements in de novo gene prediction that have occurred over the last 10 years and (2) optimizing and automating both the informatics and wet lab components of large-scale RT-PCR.

In the end, success in translating genome to ORFeome will take the same route as success in sequencing itself—investment in technology development, process optimization, and improved automation. Of course, transcripts that are completely unexpressed except in very specific circumstances will tend to be missed, but we can use these high-throughput methods to make a qualitative leap in the completeness of our ORF annotation.

To provide historical context for the argument outlined above, I will first review the state of the major gene prediction methods roughly 10 years ago, when *Genome Research* began publishing. The second section below provides a brief summary of the progress that has been made in the last decade. The third section presents the argument that the methods used for most of the last 10 years are not suitable for the end stages of ORF identification. The final section spells out some details of the recommended path toward understanding the most basic products of a genome.

## Foundations of the present era

Over the last 10 years, we have relied on three fundamental methods for identifying ORFs in genomic sequence: (1) sequencing randomly selected cDNA clones and aligning the sequences to their genomic sources; (2) finding ORFs that could produce proteins similar to proteins that are already in databases; and (3) finding ORFs de novo, without reference to cDNA sequences or their conceptual translations. Each of these methods came of age in the middle 1990s.

### *Aligning cDNA and protein sequences*

In 1996, Hillier et al. reported sequencing 280,000 human ESTs, thereby increasing the size of GenBank's human EST collection by a factor of six (Hillier et al. 1996; Wolfsberg and Landsman 1997). The ongoing scale-up of EST sequencing created a demand for tools to align these sequences to their genomic sources. While local alignment tools like BLAST can give an approximate answer, determining splice sites accurately in the presence of sequencing error requires algorithms that incorporate a stronger model of the biological processes by which pre-mRNAs are spliced. The simplest approach is to allow "intron gaps" of unbounded length with no gap extension penalty, so long as they begin with GT and end with AG—the dinucleotides that bound 99% of all known introns. EST_GENOME (Mott 1997) implemented this simple approach using an algorithm that is optimal (guaranteed to find the highest scoring alignment). Finding the optimal alignment is computationally demanding, but given the inherent difficulty of the problem, the algorithm used in EST_GENOME is quite efficient. More recent programs (Florea et al. 1998; Wheelan et al. 2001) added heuristics that speed up the computation but may not always return the absolute highest-scoring alignment. They also added some detail to the scoring system by incorporating an intron-length penalty (Florea et al. 1998) or a more accurate system of splice-site scoring (Wheelan et al. 2001). There was no clear winner in terms of accuracy. Lacking a theory about how to set the parameters appropriately for a given degree of similarity between cDNA and genome, the programs were nearly always used with the default parameters. Perhaps as a result, each system was most accurate for some species and some similarity levels. But the fact that EST_GENOME returns the best alignment according to a simple, clear scoring scheme lends it a unique appeal. Its scoring scheme turned out to be adequate, in the sense that one could not obtain alignments that were significantly more accurate for a broad range of situations, given the sequence quality and computing power available at the time.

As an example of the ambiguities that arise in cDNA-to-genome alignment, consider a short cDNA segment that can be aligned as the 3′ end of a long exon with mismatches (Fig. 1A) or as a short independent exon without mismatches (Fig. 1B). Traditional spliced alignment programs, such as EST_GENOME (Mott 1997) will make this decision based on somewhat arbitrary match/mismatch scores and intron penalties that depend on whether the intron begins with GT and ends with AG. A small fraction of introns is known to be bounded by GC-AG (~1%), AT-AC (~0.15%), and even GA-AG (1–3 known cases) (Brackenridge et al. 2003). Although some AT-AC introns are spliced via the U12 spliceosome (Tarn et al. 1995; Tarn and Steitz 1996), the initial and terminal dinucleotides do not determine whether the U2 or U12 spliceosome is used (Sharp and Burge 1997). In any case, EST_GENOME does not differentiate among GC-AG, AT-AC, or any other intron boundaries except GT-AG. Thus, it will create introns starting with TT under certain circumstances, even though there is no convincing evidence that such introns exist. Compromises like this are necessary when the TT may result from error in sequencing a GT. But when the quality of the genome sequence is high enough, the probability that an intron will start with an apparent TT approaches zero. This illustrates how the best approach to cDNA-to-genome alignment depends on the quality of the sequences involved.

A related gene prediction approach is to align protein sequences or profiles from existing databases to a genome sequence (Birney et al. 1996; Gelfand et al. 1996; Birney and Durbin 1997; Birney et al. 2004b). Because most "protein" sequences in the databases are derived by conceptual translation of cDNAs, and because the alignment algorithms for cDNA and protein sequences are similar, it is tempting to treat cDNA alignment and protein alignment as a single approach to annotation. However, there is a difference in conception and typical application. Most cDNA alignment programs are intended primarily for aligning sequences to the genomic locus from which they were transcribed, although these programs have been used for cross-species alignments (Florea et al. 1998; Wheelan et al. 2001). Protein-oriented alignment programs, on the other hand, are intended for more distant relationships, such as discovering new members of a known protein family or discovering homologs in a new species.

Philosophically, these are very different approaches. The evidence that a cDNA sequence provides about the exon–intron structure from which it is assembled is much more direct than the evidence that a protein sequence provides about the loci of putative homologs. Cross-locus protein aligners must accept a significant degree of mismatch between the protein to be aligned and the target locus, which can lead to difficulty in distinguishing between functional homologs and nontranscribed pseudogenes (Birney et al. 2004b). Systems for aligning high-quality cDNA sequences to their genomic sources, on the other hand, can require an almost perfect match, which often helps to distinguish their true loci from nontranscribed pseudogenes.

GeneWise (Birney and Durbin 1997, 2000; Birney et al. 2004b) is certainly the most important protein-to-genome alignment program, since it forms a central part of the highly influential Ensembl gene annotation pipeline. The accuracy of GeneWise is determined primarily by the degree of similarity between the protein and the locus to which it is aligned. Thus, several investigators have tried to estimate its accuracy as a function of protein-to-genome similarity. Guigó and colleagues (2000) used the *P*-value of the BLASTP alignment between the protein and the gene locus. The *P*-value reflects both the length and percent identity of the alignment, but not the fraction of the true gene covered by it. Using their Semi-Artificial Gene set, which was created by concatenating single gene sequences separated by random "intergenic" sequence, they found that GeneWise exceeded the accuracy of Genscan in exact exon prediction only when aligning highly similar proteins ($P < 10^{-50}$). When aligning proteins of moderate similarity ($10^{-50} < P < 10^{-6}$), its overall accuracy was similar to that of Genscan, but GeneWise was more conservative—it missed more of the real exons, but a higher proportion of those it predicted turned out to be exactly right.

Instead of using the *P*-value, Birney et al. (2004b) measured similarity using percent identity and percent of the target gene covered by the protein. Their most similar category of proteins was 85%–95% identical to the target locus and aligned to within 20 amino acids of the target's start and stop codons. On a single-gene test set, and excluding terminal exons, they found that GeneWise had 93% exact exon specificity and 75% exact exon sensitivity. Thus, while most of the exons it predicted were correct, it failed to correctly predict one fourth of the real exons. However, two caveats regarding this test are in order. First, it is impossible to *know* when running GeneWise whether the protein alignment is close enough to the ends of the target to meet the requirements of this accuracy level, since the genes in the target sequence are unknown. If alignments are not selected on length, GeneWise has only 40% exact exon sensitivity using proteins in the 85%–95% identity range. Second, this single gene set is likely impoverished for pseudogenes, which are a major source of false positives for GeneWise.

### *De novo gene prediction*

De novo gene prediction, in its modern form, also appeared in the mid-1990s. Stormo and Haussler (1994) described the first Generalized Hidden Markov model (GHMM) for gene prediction. GHMMs are mathematical models that can be used to define probabilities for all possible exon–intron annotations on a given sequence. An accurate GHMM for gene finding will assign high probabilities to correct annotations and low probabilities to incorrect annotations. GHMMs differ from ordinary HMMs in that the log probabilities, or scores, of exons and introns can depend globally on the entire sequence of the exon or intron. In ordinary HMMs, the scores of features must be the sums of the scores of individual bases within the feature. For example, the ability to compute scores from the whole feature makes it possible to create general, nonlinear models of the lengths of exons and introns. Kulp et al. (1996) were the first to use the term GHMM in the context of gene finding, the first to describe a fully general mathematical framework for all GHMM models, and the first to implement and test a GHMM-based computer program for gene finding. Burge and Karlin (1997) developed Genscan, a GHMM-based gene prediction program that could predict multiple and partial genes on both strands. This ability made Genscan suitable for annotating ORFs in anonymous DNA sequence such as that produced by sequencing random BAC clones. As the first GHMM suitable for annotating anonymous genome sequence, Genscan defined the state of the art in both technology and accuracy. Eventually, programs like Genie were enhanced to predict multiple genes on both strands (Reese et al. 2000), but Genscan remained one of the most accurate and most widely used programs for many years.

### *Combining prediction methods*

The genome annotations that are most visible to the public and most widely used are created by combining predictions from all of these methods. In the mid-1990s, the results of cDNA alignments, protein alignments, and de novo predictions were integrated by human experts and were often followed by RT-PCR and sequencing experiments to test the predicted exon–intron structures (Ansari-Lari et al. 1996, 1997) (RT-PCR is PCR amplification of cDNAs made by reverse transcription from RNA). Automated "pipelines" for integrating evidence, such as Ensembl (Birney et al. 2004a; Curwen et al. 2004), OTTO (Venter et al. 2001), and the National Center for Biotechnology Information (NCBI) pipeline were not developed until shortly before the publication of the draft human genome sequence (1999 or 2000).

## Trajectory of improving accuracy

### *Aligning cDNA sequences*

The accuracy of prediction systems based on aligning cDNA or protein sequence depends on the sequences that are available for alignment as well as the algorithms used to align them. There can be no doubt that both the quantity and quality of expressed sequences have improved dramatically in the last ten years. For example, the human EST database has gone from 415,000 sequences in 1997 to over 6 million in 2005. Several projects, including the Mammalian Gene Collection (MGC) (http://mgc.nci.nih.gov/) (Furey et al. 2004; The MGC Project Team 2004) have produced finished sequences from large collections of cDNA clones that appear to contain a complete ORF. Indeed, the MGC collection now contains at least one cloned transcript from about 13,000 human and 12,000 mouse gene loci. These sequences have been subjected to extremely rigorous quality control so that most produce 100% identical alignments to the reference genome, except for silent discrepancies and known polymorphisms (http://genes.cse.wustl.edu/mgc/) (Furey et al. 2004).

There have been improvements in alignment algorithms,

```
A  ACTACCTCATGAGGTGGCgtga....atagtacgcgtaa.....ttagCACTTCTGGGGCCCA
   ||||||||||||| | ||>>>>>>>>>> 15907 >>>>>>>>>>>>>|||||||||||||||
   ACTACCTCATGAGTACGC..............................CACTTCTGGGGCCCA

B  ACTACCTCATGAGgtggcgtga....atagTACGCgtaa.....ttagCACTTCTGGGGCCCA
   |||||||||||||>>>>> 12584 >>>>>|||||>>>> 3323 >>>|||||||||||||||
   ACTACCTCATGAG.................TACGC.............CACTTCTGGGGCCCA
```

**Figure 1.** (*A*) A fragment of an alignment of a cDNA to genomic sequence containing 2 mismatches and a 15,907-bp intron. (*B*) Another alignment of the same two sequences containing no mismatches and a 5-bp exon in the intron of alignment A. All introns are bounded by canonical GT-AG splice sites.

too. Traditional cDNA-to-genome alignment programs do not explicitly model the probability of mismatches in the correct alignment (Fig. 1A) as compared with the probability of an additional intron in the correct alignment (Fig 1B). In fact, mismatches in correct alignments are either sequencing errors or differences between the reference genome and the genome from which the cDNA was transcribed. (Occasionally, they may also result from post-transcriptional events such as RNA editing.) Thus, the probabilities of these events depend on both the sequence quality and the rate of polymorphism (for within-species alignments) or divergence (for cross-species alignments). On the other hand, the probability of an additional intron depends on the frequency of introns in the species at hand.

A new generation of cDNA-to-genome alignment programs models all these things using pair hidden Markov models with parameters estimated from the specific cDNA collection and the genome sequence to be aligned (M. Arumugam and M.R. Brent, in prep.). For example, such systems can easily model the fact that sequencing errors are much less likely when aligning an MGC cDNA sequence to the finished human genome than when aligning a single-pass EST sequence to the draft dog genome. For 70%–80% of high-quality cDNA sequences, these more precise models will result in the same alignment as a program like EST_GENOME. In many of the remaining cases, however, they produce better alignments. For example, they are better able to distinguish small exons from sequencing errors. This accuracy improvement is made possible by the availability of very high quality sequences to align and the availability of sufficient computing power to run the pairHMM algorithms in a reasonable amount of time.

### *Single–genome de novo gene prediction*

Perhaps the best indicator of how the accuracy of de novo gene prediction has changed in the last ten years is the change in how accuracy is measured. Shortly before the publication of Genscan, Burset and Guigó (1996) published the first comprehensive comparison of vertebrate prediction programs. Their test set consisted of 570 genomic sequences no longer than 50 Kb, each with a single gene on the positive strand. Burset and Guigó found that the most accurate de novo system of that time, FGENEH (Solovyev et al. 1994), predicted just 61% of the known exons correctly, GeneID (Guigó et al. 1992) got 51%, and all the rest got well below 50% of exons right. The fraction of ORFs predicted correctly was not reported, presumably because it was near zero.

Genscan (Burge and Karlin 1997) represented a breakthrough in accuracy that led to a long, slow shift in the evaluation paradigm. When tested on Burset and Guigó's single-gene set, it predicted 78% of the exons correctly, compared with just 61% for the best previous system. Furthermore, Burge and Karlin (1997) reported that Genscan predicted 43% of the ORFs in that test correctly. They also presented an analysis of Genscan's predictions on a contiguous sequence of 117 Kb containing multiple experimentally determined gene structures (Ansari-Lari et al. 1996). Although the number of genes was too small for reliable estimation of accuracy, it is interesting that only one of eight Genscan-predicted ORFs matched the annotation exactly (12%), though most were quite similar.

As test sets became more realistic, estimates of Genscan's accuracy at predicting complete human ORFs dropped. Guigó et al. (2000) published an evaluation based on simulated human genomic sequence that they created by concatenating single-gene sequences padded by randomly generated pseudo-intergenic sequence. In this new test set, only 2.3% of the nucleotides were protein-coding, much closer to the overall average (now thought to be under 2%) than the 15% in Burset and Guigó's 1996 set. As expected, they found that prediction accuracy is much lower on contiguous sequences with typical coding density than on single-gene sequences with high coding density. However, they did not report the percentage of ORFs predicted correctly. Korf et al. (2001) tested Genscan on 7.6 Mb of mouse genome consisting of 68 contiguous sequences with an average length of 112 Kb. In this test, Genscan predicted only about 15% of annotated ORFs exactly—much lower than the 43% reported for Burset and Guigó's single-gene set. However, even this estimate turned out to be optimistic. When Genscan was finally evaluated on the entire human genome, it predicted a correct ORF at only 10% of loci containing a known ORF (9% of known ORFs, Flicek et al. 2003).

Currently, gene prediction programs are used primarily for whole genome annotation. As described above, their accuracy when evaluated on a whole genome is typically much lower than their accuracy when evaluated on isolated genes or artificially concatenated sets of single genes. Even whole chromosomes can be deceptive. For example, human chromosome 22, besides being the smallest autosome, is also unusually gene dense, with smaller than average introns and intergenic regions and above average GC content. Most gene prediction programs, including Genscan, tend to perform best on high GC, gene-dense regions. Thus, evaluation on chromosome 22 systematically overestimates the accuracy of most systems. In the current environment, the minimal standard for evaluation of gene prediction programs must be based on whole genome annotation runs. Some may argue that, since we do not know all the exon-intron structures for the human or any other genome, we cannot know the accuracy of a prediction set for the whole genome. This is true, but it should not be an impediment to evaluating whole genome annotations. Sensitivity estimates based on the subset of genes whose structures are known should be an unbiased estimate of sensitivity on all genes, to the extent that the sets of known genes and unknown genes do not differ in ways that greatly affect accuracy. While it is possible that unknown genes are radically different from known genes in this way, there is no reason to believe that they are. Specificity will be systematically underestimated when the predictions are compared to known genes rather than to all genes. Under the same assumption described above, dividing by the fraction of genes that are known (or the fraction of exons that are known, for exon-level specificity) corrects the underestimate. The exact value of that correction factor does not matter when comparing the specificities of two programs—the one with the higher raw estimate will also have the higher corrected estimate. Another approach, which seems to always give qualitatively similar results, is to use only gene predictions that overlap known genes by at least one nucleotide when computing specificity (Wei et al. 2005).

Determining gene boundaries is one of the most challenging aspects of ORF prediction—much more so than predicting the boundaries of exons with splices on both sides—and so it is also the area in which the potential for improvement is greatest. Many improvements to gene prediction algorithms have a large effect on accuracy as measured by exact ORF prediction, even though they have little effect on the accuracy of exon prediction. Thus, it is critical to include measures of exact ORF prediction in comparative evaluations of gene prediction programs.

Statistics on the exact-ORF accuracies of programs are important, but there is a legitimate argument that the value of these programs is not in predicting known genes but in predicting novel genes. Thus, the most convincing evaluation of a program or a set of programs is the extent to which its novel predictions can be verified experimentally. The trend toward publishing experimental evaluations of prediction sets (Wu et al. 2004; Brown et al. 2005; Eyras et al. 2005; Wei et al. 2005) therefore represents a significant step forward for the field of gene prediction.

### *Dual- and multi-genome de novo predictors*

Historically, Genscan represents the apogee in the arc of improving accuracy for mammalian gene predictors using a single genomic sequence as their only input (but see Stanke and Waack 2003 for improvements on *Drosophila*). No other system robustly outperformed Genscan on large, contiguous human genomic sequences until the advent of dual-genome de novo systems, which use alignments between two genomes as a rough indicator of which nucleotides are under negative selection and hence are likely to have a function that contributes to fitness. Several such systems required that orthologous stretches of mouse and human DNA be identified in advance, or were tested only on single-gene sets that were preselected to have clearly identifiable orthologs (Bafna and Huson 2000; Batzoglou et al. 2000; Alexandersson et al. 2003). However, TWINSCAN (Korf et al. 2001; Flicek et al. 2003) and SGP2 (Parra et al. 2003) could both be run on entire human chromosomes, using alignments generated by simple, robust procedures. Neither program requires that an ortholog be present in the other genome—they can just as easily exploit similarity to paralogs, or even fragments of several genes from the same family. Flicek et al. (2003) reported that TWINSCAN was able to predict 14% of known ORFs in the human genome correctly. Although SGP2 was tested only on chromosome 22, which is unusually gene dense compared with the whole genome, the accuracies of the two systems were similar. Over the last few years, incremental improvements to TWINSCAN, along with better training and testing sets, have improved its accuracy to the point where it can predict a correct ORF at about 25% of human loci with known ORFs.

A new level of accuracy was achieved this year by N-SCAN, a version of TWINSCAN with a new, phylogenetic conservation model that is capable of considering alignments among multiple genomes (Gross and Brent 2005, 2006). N-SCAN is able to predict a correct ORF at about 35% of human loci with known ORFs. It is also notably more accurate than previous systems at the exon level, predicting 85% of known human coding exons correctly, whereas previous systems predicted fewer than 75% correctly. Furthermore, it is the first program to accurately predict the boundaries of long introns—it correctly predicts about 50% of introns in the 50- to 100-Kb length range.

Many of the challenges of de novo gene prediction that have been observed over the years remain challenges today. Even the best prediction programs tend to split and fuse genes, and they have difficulty accurately predicting stop codons and especially start codons. They only predict a single isoform at each locus, even though a large fraction of human genes are alternatively spliced. Yet there has been enormous progress. We have moved from predicting a correct ORF at one tenth of the human loci to predicting a correct ORF at one third. We can now predict long introns (Gross and Brent 2005), and we can predict spliced 5′ UTRs with reasonable accuracy (Brown et al. 2005). Gene predictors are generally more accurate on more compact genomes such as those of *Caenorhabditis elegans* and *D. melanogaster* (GeneFinder: P. Green, unpubl.; Burge and Karlin 1997), but the last ten years have seen substantial progress there, too (Stanke and Waack 2003; Gross and Brent 2005; Wei et al. 2005).

### *Combining prediction methods*

In the run-up to the initial publications on the human genome, it became clear that manual integration of evidence from various prediction methods would not be fast enough to provide an analysis of the entire genome in a reasonable amount of time. As an alternative, several automated "pipelines" for integrating evidence, such as OTTO (Venter et al. 2001), Ensembl (Birney et al. 2004a; Curwen et al. 2004), and the NCBI pipeline were developed. OTTO was used primarily by the team at Celera Genomics that developed it. Ensembl annotations have been used to produce the primary gene sets in many of the publications that describe the first analysis of a new vertebrate genome sequence (e.g., Lander et al. 2001; Aparicio et al. 2002; Waterston et al. 2002; Gibbs et al. 2004; Hillier et al. 2004). The NCBI annotation pipeline is also very influential because its predictions appear in GenBank as RefSeq mRNAs and proteins with "XM" and "XP" accessions, respectively. The NCBI pipeline was originally based on GenomeScan (Yeh et al. 2001), an enhancement of Genscan that modifies the scores of potential exons depending on whether they have high-scoring alignments to proteins in the databases. More recent versions of the NCBI pipeline use an unpublished method called Gnomon (http://www.ncbi.nlm.nih.gov/genome/guide/build.html#gene). Although every pipeline works differently, both Ensembl and NCBI rely heavily on aligning protein sequences generated from one gene to the genomic loci of other genes, either within or between species. In the following discussion, I will focus on Ensembl as a representative of such annotation pipelines.

Most Ensembl gene predictions are ultimately created by GeneWise, a protein-alignment program, although Genscan is used to help identify the best proteins to align from other species (Curwen et al. 2004). Thus, GeneWise has determined the structures of a large fraction of the predicted genes used in the initial analyses of numerous vertebrate genomes. The goal in these analyses was to obtain a conservative set of predicted exons and genes—one containing few false positives. Genscan, the best available de novo gene predictor until 2003, predicts numerous false positive exons, so choosing a protein alignment method such as GeneWise made sense. Even TWINSCAN, as published in 2003, predicted only 75% of known exons correctly, as compared to 85% for Ensembl (Flicek et al. 2003).

Given the recent progress in de novo gene prediction, it is worth asking whether GeneWise is still more accurate, or even more conservative, than the best de novo predictors. A direct comparison would be most informative, but the data set that Birney et al. (2004b) used to assess the accuracy of GeneWise has been lost. However, some inferences can be drawn by comparing published results on different test sets. Gross and Brent (2005) reported that N-SCAN, when run on the whole human genome, predicts 85% of all known exons correctly; estimated specificity based on the assumption that the genome contains 200,000 exons is 86%. Considering only internal exons, as in Birney et al. (2004b), N-SCAN's estimated exon-level specificity rises to 93%.

Considering that these numbers are based on whole-genome annotations, whereas the estimated 93% for GeneWise is based on a single-gene set, it is likely that predictions based on protein homology are no longer more "conservative" (i.e., specific) than de novo predictions except when the protein is nearly identical to the target locus. Furthermore, the sensitivity of de novo methods is much higher.

A direct comparison was recently made among integrated annotation pipelines as part of the E-GASP community evaluation (Guigó and Reese 2005; http://genome.imim.es/gencode/workshop2005.html). Pipeline predictions were compared to manual annotation by the HAVANA group (see below) on the human ENCODE regions (The ENCODE Project Consortium 2004). Among entrants were Ensembl and a simple ad hoc pipeline with two stages. The first stage was to align full-ORF cDNA sequences from the MGC project and the RefSeq collection using a cDNA aligner called Pairagon, which is based on a strong prior model designed for highly accurate sequences (M. Arumugam and M.R. Brent, in prep.). The second stage, applied to regions not covered by cDNA alignments, was to exploit BLAT alignments of spliced human ESTs to guide de novo predictions by N-SCAN. These predictions, made by N-SCAN_EST (C. Wei and M.R. Brent, in prep.), are generally consistent with the EST alignments but may extend them with additional exons or link nonoverlapping ESTs into a single transcript. The results showed that the Pairagon+N-SCAN_EST pipeline was substantially more specific than Ensembl, and equally sensitive, in both the exact exon and exact transcript measures. Ensembl predicted more transcripts per locus than Pairagon+N-SCAN_EST, so it could predict at least one correct transcript at a higher percentage of the loci where it made predictions (gene specificity). Thus, Pairagon+N-SCAN_EST, which uses only human cDNA sequence aligned to its native locus, is at least as accurate as Ensembl, which uses cross-locus and cross-species protein alignment. However, cross-species protein homology is likely to contribute more to the annotation of species for which fewer cDNA sequences are available.

A more recent approach to integrating predictions is to score each potential exon using a weighted combination of evidence from alignment-based predictions and de novo predictions (Allen et al. 2004). The weights are derived from estimates of the accuracy of each prediction source. Thus, if several predictors that have proven accurate in the past agree on an exon, it will receive a high score. In the case of disagreement among predictors the score will generally be lower, but more weight will be given to more accurate predictors. This approach performs well in practice, especially when there are multiple evidence sources with roughly similar accuracy—empirically, it seems that different methods make different errors. Indeed, JIGSAW, a descendent of COMBINER (Allen et al. 2004), was slightly more sensitive than the Pairagon+N-SCAN_EST pipeline at the exon level in the EGASP evaluations, although Pairagon+N-SCAN_EST was more accurate in predicting exact ORFs. To achieve the accuracy it did, JIGSAW was run on 13 sources of evidence using genome coordinates provided by the UCSC browser, including Ensembl, RefSeq, Genscan, SGP, TWINSCAN, Human mRNAs, TIGR Gene Index, and UniGene (Wheeler et al. 2004).

Manual annotation has also progressed over the last 10 years. In 2000, the *Drosophila* community held an annotation "jamboree," in which fly biologists and bioinformaticians gathered at Celera Genomics for two weeks to create an initial annotation of the *Drosophila* genome (Pennisi 2000). This annotation has since been systematically revised and updated (Misra et al. 2002; Drysdale and Crosby 2005). In 2002, the Sanger Institute held two Human Annotation Workshops (known as Hawk meetings). A number of groups involved in human annotation gathered at these meetings and compared their annotations on designated sequences to "define a standard of annotation" and "draw up guidelines to help achieve the standard" (http://www.sanger.ac.uk/HGP/havana/hawk.shtml). These meetings led to the annotation standards that are used by the Sanger Institute's Human and Vertebrate Analysis project (HAVANA, http://www.sanger.ac.uk/HGP/havana/docs/guidelines.pdf). HAVANA annotators integrate information from alignments of expressed sequences and de novo predictions by Genscan and FGENESH (http://www.sanger.ac.uk/HGP/havana/). The HAVANA team has annotated human chromosomes 1, 6, 9, 10, 13, 20, 22, and X; mouse chromosomes 2, 4, 11, and X; and the entire zebrafish genome. Their annotation of the human ENCODE regions appears to have been quite complete, since RT-PCR and sequencing experiments have verified only a small handful of exons not annotated by HAVANA (R. Guigó, pers. comm.).

## The current turning point

### *Limits of sequencing random cDNA clones*

Improving the accuracy of annotations based on expressed sequences depends, to a large extent, on improving the collection of sequences that are available to align. The vast majority of ESTs and cDNA sequences currently in databases were obtained by sequencing clones selected at random from cDNA libraries. However, this method has been found to saturate well short of the full gene set (The MGC Project Team 2004). For example, the MGC project sequenced 5′ ESTs from more than 110 human and 80 mouse cDNA libraries and screened them for clones that appeared likely to contain a complete ORF not already in the collection. Promising clones were then sequenced to high accuracy. This produced full-ORF clones at approximately 13,000 human and 12,000 mouse gene loci—about 50%–60% of all the genes, according to current estimates. Aligning the human ESTs to the genome produced a total of about 62,000 nonoverlapping clusters, but most of these appeared not to include the 5′ end of an ORF. EST projects for other animals have yielded qualitatively similar results (e.g., Wei et al. 2005), although the number of clones sequenced has generally been less. Thus, we cannot expect to complete the annotation of any animal genome by simply sequencing deeper into cDNA libraries.

### *Limits of protein alignment*

Some genes for which we cannot obtain a full-ORF cDNA sequence can nonetheless be annotated by aligning homologous proteins to the genome. However, it appears that this approach is no more accurate than de novo prediction except when the aligned protein is nearly identical to the one encoded by the target locus. Improvements in protein alignment methods—particularly the use of alignment models that do not accept frame shifting errors—may extend this accuracy horizon somewhat. In the end, though, annotation by protein alignment will be limited by the cDNA collection from which most "proteins" are derived.

### *Limits of combiners*

Systems that combine predictions from many sources seem to provide at least a slight edge over the best single source. However, the accuracy of these systems is limited by the accuracy of the underlying prediction sources.

### *Limits of manual annotation*

Inspection by human curators seems to be an effective method of integrating evidence from alignments of expressed sequences, alignments among genomes, and de novo predictions. Among the strengths of the human curators is the ability to detect suspicious annotations, such as pseudogenes. At bottom, though, their accuracy is still limited by the accuracy of the evidence they are given to integrate. Furthermore, manual annotation is time consuming and may not be updated rapidly in response to new evidence. Most importantly, though, it is very expensive compared with automated integration of evidence. As a result, it seems unlikely that extensive manual annotation will be done beyond the genomes of *D. melanogaster*, human, zebrafish, and possibly mouse.

### *Limits of comparative genomics*

Although the accuracy of de novo gene prediction has improved dramatically since 1996, one major source of expected improvement has not yet panned out. It has been widely anticipated that the availability of multiple genomes within a fairly narrow phylogenetic clade would lead to dramatic improvements in gene prediction accuracy. By aligning multiple mammalian genomes, we should be better able to characterize the patterns of selection operating on small sections of genomes. These patterns should be indicative of specific functions. For example, regions that contain a number of substitutions all separated by multiples of three are more likely to be coding, since the third position of a codon can often be changed without changing the amino acid it encodes. The expectation that multi-genome alignments would lead to more accurate gene prediction than alignments among two genomes is quite reasonable and such improvements may yet be achieved. However, no method for improving gene prediction accuracy by using multi-genome alignments has yet been found, despite several serious efforts. For example, EXONIPHY (Siepel and Haussler 2004), an exon prediction system based on a phylogenetic generalization of HMMs, does not exceed the accuracy of dual-genome systems like TWINSCAN or SGP2 in exact exon prediction, although it does exceed them by a few percent in nucleotide specificity (Gross and Brent 2005). N-SCAN, which is based on a different phylogenetic generalization of generalized HMMs, represents a substantial improvement in accuracy over TWINSCAN and SGP2. However, running N-SCAN on multi-genome alignments has not yet produced results that are substantially better than those obtained by running it on only two genomes (Gross and Brent 2005).

There are several possible reasons for the failure, so far, to achieve substantial accuracy improvements by using multi-genome alignments. It may be that we do not yet have the right combinations of genomes sequenced to sufficiently high quality—draft sequence may not be good enough. Or, it may be that we simply cannot align these genomes precisely enough to draw accurate inferences about selection. If these are the reasons, then finished sequences from more mammals, especially primates (Boffelli et al. 2003) may lead to the anticipated improvements. Another potential explanation is that exons and splice sites conserved throughout the mammalian lineage may be less common than originally thought. Anecdotally, it has been observed that alignments among many mammalian genomes often show that a given exon does not appear in one of the species. On the other hand, it may be that the designers of de novo gene prediction algorithms have simply not been clever enough to come up with the right methods yet. In any case, it seems that we cannot count on the availability of more genome sequences to yield substantial accuracy improvements in the immediate future.

## Looking forward

It is my conviction that a finished genome sequence should reveal the set of ORFs it encodes. Therefore, I believe we must develop a cost-effective technology for translating a genome to a set of exon–intron structures and the proteins they encode. The outlines of this technology are now becoming clear, but its cost must still be reduced through automation and optimization.

The current gold standard of evidence for gene structures is cDNA sequence aligned to the genomic locus from which it was transcribed. This leaves something to be desired, in that one must still infer the exon-intron structure by alignment and the protein product by conceptual translation of the most likely-looking open reading frame. Both of these inferences are subject to error, so one might hope for confirmation by direct experimental evidence. However, there is as yet no economical, high-throughput method for obtaining such evidence. In particular, there is no analog of RT-PCR for proteins—an economical method of directly amplifying or purifying hypothesized, low-abundance proteins. Since we must rely on computational inference of protein products that aren't easily picked up by high-throughput proteomics, it is possible that incorrectly processed pre-mRNAs, such as those with retained introns, would yield incorrect inferences about functional proteins. The best approach to flagging such cases may be to screen for cDNAs that are likely candidates for nonsense-mediated decay—those with splice junctions more than 50–55 nt 3′ of the inferred ORF (Lejeune and Maquat 2005).

The most efficient way to obtain cDNA sequence for every protein-coding gene is to combine standard EST sequencing, gene prediction, and RT-PCR using primers designed to amplify predicted transcripts. A small to moderate collection of ESTs should be developed first by the standard method—sequencing randomly selected cDNA clones. This will produce sequence from transcripts that are relatively abundant, and will completely determine the exon-intron structures of abundant transcripts that are shorter than two read lengths (currently about 1400–1800 bp). The cost per transcript will remain relatively low as long as a fairly high proportion of sequences produced are new. By calculating the number of clones that must be sequenced to obtain a new EST and multiplying by the cost per clone one can estimate the cost per new cDNA read. When this cost exceeds the estimated cost per new read by RT-PCR, EST sequencing should be stopped. The resulting ESTs should be aligned to the genome using cDNA-to-genome alignment tools based on strong models of gene structure, and those that do not align well should be discarded or set aside for manual inspection if time permits. High-quality EST alignments that overlap one another must then be grouped together and computational techniques used to determine which groups are likely to contain a complete ORF. Those that do form the core set of genes in the annotation.

Once the core set has been determined, the rest of the genes must be identified by a series of RT-PCR and sequencing steps, starting with the most confident predictions and progressing toward the less confident (Fig. 2). Considering the analyses described above, predictions based on cross-locus and cross-species protein alignments are more reliable than de novo predictions only when the aligned protein is highly similar to the predicted one (probably >95% identity). Such predictions should be used to design primers for the first round of RT-PCR and sequencing experiments. After each RT-PCR and sequencing step, the resulting cDNA sequences should be aligned, grouped, and sorted by completeness of the predicted ORF as described above. Aligning the experimental sequences to the genome may confirm parts of the predicted gene structure, but it may also reveal errors in other parts of the predicted structure.

The updated set of full-ORF gene structures can now be used to train a de novo gene prediction algorithm. Typically, clusters of genomes within a clade are sequenced at once, so it is usually possible to use dual- or, potentially, multi-genome de novo prediction methods. The EST alignments that do not cover a full ORF can be used to guide the prediction algorithms, which will predict complete structures that are consistent with the alignments, but may extend them with additional exons and/or link several ESTs together into a single predicted transcript (C. Wei and M.R. Brent, in prep.) The unconfirmed regions of predictions that extend or link EST alignments can then be tested in the next round of RT-PCR. After aligning the resulting sequences to the genome, the gene structures they define can be used as additional examples for retraining the gene predictor and as additional guidance around which the gene predictor can build models. If this process is taken to convergence, where all gene models have been tested, the result will be an annotation of exon–intron structures that is more complete than any we have now and that is fully verified by native cDNA sequences.

Several variants of this approach are also being developed. One is to use Rapid Amplification of cDNA Ends (RACE) PCR, a method in which a universal primer at one end of the cDNA is paired with a single gene-specific primer inside the predicted cDNA. Certain RACE methods selectively amplify 5′ complete mRNAs with a 7 methyl guanine cap, allowing amplification of the 5′ end without knowing a sequence in the 5′ end. Only the sequences of one or more internal exons are needed for the design of the gene specific primer. Since only one exon needs to be predicted correctly, this method can be more sensitive than ordinary RT-PCR. Specificity is often a problem with RACE, but this can be ameliorated by a second round of PCR using a nested pair of universal and gene specific primers. McCombie and colleagues (Dike et al. 2004) have done this using mouse predictions by TWINSCAN (Flicek et al. 2003) and GenomeScan (Yeh et al. 2001), while Gingeras and colleagues have done it using genomic tiling arrays for both exon prediction and sequencing of RACE-PCR products (Cheng et al. 2005; Kapranov et al. 2005).

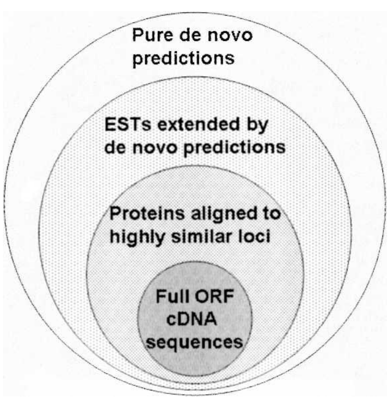


**Figure 2.** Recommended preference order for gene structure predictions. Alignments of full ORF cDNA sequences to the loci from which they were transcribed should take precedence. At loci where there is no full ORF cDNA alignment, alignments of highly similar proteins (probably >95% identity) should be used. Third best is a gene structure that is partially determined by an EST aligned to its native locus, possibly extended to a full ORF by de novo prediction. At loci where none of these are available, pure de novo predictions using dual- or multi-genome prediction algorithms should be used.

Of course, some transcripts are expressed transiently during development, or only under rare environmental conditions. We can increase the number of detectable transcripts by pooling RNA from many tissues. Cloning artifacts can be reduced by amplifying reverse-transcriptase products directly rather than using cloned cDNA libraries and by sequencing PCR products directly rather than sequencing clones. But there will still be rare transcripts that cannot be verified by a high-throughput annotation system. In the end, these will have to be identified on a case-by-case basis using traditional biochemical or genetic approaches. Nonetheless, we can use high-throughput methods to get much closer than we have so far to determining the most basic elements on the parts list of an organism.

To make this vision a reality, we must bring the cost of the RT-PCR and sequencing experiments down as far as possible. This means relying on end-to-end automation. Much of the necessary automation consists of software pipelines for selecting predictions to test, designing primer pairs to test them, and analyzing the resulting sequences to determine new gene structures. The physical processes of setting up PCR and sequencing reactions must also be optimized and automated. Finally, the accuracy of the gene predictions will be a central determinant of the cost and completeness of the resulting annotation. Prediction errors may lead to one or more PCR experiments that fail to amplify their targets and produce no useful sequence, thus raising the cost per transcript annotated. Therefore, we must continue to improve the accuracy of gene prediction by developing more complete and more realistic models of the signals in the genome sequence that guide the transcription and processing of mRNA.

The genomics community is used to rapid progress and headline-making excitement, so the temptation to "declare victory and move on" is understandable. I have heard it said numerous times that the identification of protein-coding genes is well understood, and the real challenges now are identifying transcription factor binding sites, non-coding RNA genes, and other exciting sequence elements. While these are important challenges, we must resist the temptation to leave the identification of protein-coding genes incomplete while we chase after the hottest new features. We must not forget that the defining characteristic of genomics is the all-out effort to view an organism globally by analyzing data sets that are as complete as we can possibly make them.

## Acknowledgments

I am grateful to Paul Flicek for help with analyzing the EGASP evaluation results and to Mark Diekhans for analysis of the MGC cDNA sequences. M.R.B. is supported in part by R01 HG02278, R01 AI051209, and U01 HG003150 from the National Institutes of Health; in part by grant DBI-0501758 from the National Science Foundation; and in part by National Cancer Institute funds for the Mammalian Gene Collection project under Contract No. N01-CO-12400.

## References


Alexandersson, M., Cawley, S., and Pachter, L. 2003. SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.* **13:** 496–502.

Allen, J.E., Pertea, M., and Salzberg, S.L. 2004. Computational gene prediction using multiple sources of evidence. *Genome Res.* **14:** 142–148.

Ansari-Lari, M.A., Timms, K.M., and Gibbs, R. 1996. Improved ligation-anchored PCR strategy for identification of 5′ ends of transcripts. *BioTechniques* **21:** 34–38.

Ansari-Lari, M.A., Shen, Y., Muzny, D.M., Lee, W., and Gibbs, R.A. 1997. Large-scale sequencing in human chromosome 12p13: Experimental and computational gene structure determination. *Genome Res.* **7:** 268–280.

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

Bafna, V. and Huson, D.H. 2000. The conserved exon method for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8:** 3–12.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10:** 950–958.

Birney, E. and Durbin, R. 1997. Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 56–64.

———. 2000. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10:** 547–548.

Birney, E., Thompson, J.D., and Gibson, T.J. 1996. PairWise and SearchWise: Finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.* **24:** 2730–2739.

Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., et al. 2004a. An overview of Ensembl. *Genome Res.* **14:** 925–928.

Birney, E., Clamp, M., and Durbin, R. 2004b. GeneWise and Genomewise. *Genome Res.* **14:** 988–995.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391–1394.

Brackenridge, S., Wilkie, A.O., and Screaton, G.R. 2003. Efficient use of a 'dead-end' GA 5′ splice site in the human fibroblast growth factor receptor genes. *Embo. J.* **22:** 1620–1631.

Brown, R.H., Gross, S.S., and Brent, M.R. 2005. Begin at the beginning: Predicting genes with 5′ UTRs. *Genome Res.* **15:** 742–747.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Burset, M. and Guigó, R. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353–367.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.

Collins, F. and Galas, D. 1993. A new five-year plan for the U.S. Human Genome Project. *Science* **262:** 43–46.

Collins, F.S., Green, E.D., Guttmacher, A.E., and Guyer, M.S. 2003. A vision for the future of genomics research. *Nature* **422:** 835.

Curwen, V., Eyras, E., Andrews, T.D., Clarke, L., Mongin, E., Searle, S.M., and Clamp, M. 2004. The Ensembl automatic gene annotation system. *Genome Res.* **14:** 942–950.

Dike, S., Balija, V.S., Nascimento, L.U., Xuan, Z., Ou, J., Zutavern, T., Palmer, L.E., Hannon, G., Zhang, M.Q., and McCombie, W.R. 2004. The mouse genome: Experimental examination of gene predictions and transcriptional start sites. *Genome Res.* **14:** 2424–2429.

Drysdale, R.A. and Crosby, M.A. 2005. FlyBase: Genes and gene models. *Nucleic Acids Res.* **33:** D390–D395.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.

Eyras, E., Reymond, A., Castelo, R., Bye, J.M., Camara, F., Flicek, P., Huckle, E.J., Parra, G., Shteynberg, D.D., Wyss, C., et al. 2005. Gene finding in the chicken genome. *BMC Bioinformatics* **6:** 131.

Flicek, P., Keibler, E., Hu, P., Korf, I., and Brent, M.R. 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.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.

Furey, T.S., Diekhans, M., Lu, Y., Graves, T.A., Oddy, L., Randall-Maher, J., Hillier, L.W., Wilson, R.K., and Haussler, D. 2004. Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing. *Genome Res.* **14:** 2034–2040.

Gelfand, M.S., Mironov, A.A., and Pevzner, P.A. 1996. Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci.* **93:** 9061–9066.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Gross, S.S. and Brent, M.R. 2005. Using multiple alignments to improve gene prediction. In *9th Annual International Conference, RECOMB 2005* (eds. S. Miyano et al.), pp. 374–388. Springer, Boston.

———. 2006. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13:** (in press).

Guigó, R. and Reese, M.G. 2005. EGASP: Collaboration through competition to find human genes. *Nat. Methods* **2:** 575–577.

Guigó, R., Knudsen, S., Drake, N., and Smith, T. 1992. Prediction of gene structure. *J. Mol. Biol.* **226:** 141–157.

Guigó, R., Agarwal, P., Abril, J.F., Burset, M., and Fickett, J.W. 2000. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10:** 1631–1642.

Guigó, R., Dermitzakis, E.T., Agarwal, P., Ponting, C., Parra, G., Reymond, A., Abril, J.F., Keibler, E., Lyle, R., Ucla, C., et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci.* **100:** 1140–1145.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S., and Gingeras, T.R. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15:** 987–997.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1:** S140–S148.

Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4:** 134–142.

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

Lejeune, F. and Maquat, L.E. 2005. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr. Opin. Cell Biol.* **17:** 309–315.

The MGC Project Team. 2004. The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC). *Genome Res.* **14:** 2121–2127.

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

Modrek, B. and Lee, C. 2002. A genomic view of alternative splicing. *Nat. Genet.* **30:** 13–19.

Mott, R. 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13:** 477–478.

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

Pennisi, E. 2000. Ideas fly at gene-finding jamboree. *Science*

**287:** 2182–2184.
Reese, M.G., Kulp, D., Tammana, H., and Haussler, D. 2000. Genie—Gene finding in *Drosophila melanogaster*. *Genome Res.* **10:** 529–538.
Sharp, P.A. and Burge, C.B. 1997. Classification of introns: U2-type or U12-type. *Cell* **91:** 875–879.
Siepel, A.C. and Haussler, D. 2004. Computational identification of evolutionarily conserved exons. In *RECOMB*. ACM, San Diego.
Solovyev, V.V., Salamov, A.A., and Lawrence, C.B. 1994. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **22:** 5156–5163.
Sorek, R., Shamir, R., and Ast, G. 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20:** 68–71.
Stanke, M. and Waack, S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2:** II215–II225.
Stormo, G.D. and Haussler, D. 1994. Optimally parsing a sequence into different classes based on multiple types of evidence. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2:** 369–375.
Tarn, W.Y. and Steitz, J.A. 1996. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell* **84:** 801–811.
Tarn, W.Y., Yario, T.A., and Steitz, J.A. 1995. U12 snRNA in vertebrates: Evolutionary conservation of 5′ sequences implicated in splicing of pre-mRNAs containing a minor class of introns. *RNA* **1:** 644–656.
Torrents, D., Suyama, M., Zdobnov, E., and Bork, P. 2003. A genome-wide survey of human pseudogenes. *Genome Res.* **13:** 2559–2567.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Wei, C., Lamesch, P., Arumugam, M., Rosenberg, J., Hu, P., Vidal, M., and Brent, M.R. 2005. Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. *Genome Res.* **15:** 577–582.
Wheelan, S.J., Church, D.M., and Ostell, J.M. 2001. Spidey: A tool for mRNA-to-genomic alignments. *Genome Res.* **11:** 1952–1957.
Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., et al. 2004. Database resources of the National Center for Biotechnology Information: Update. *Nucleic Acids Res.* **32:** D35–D40.
Wolfsberg, T.G. and Landsman, D. 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* **25:** 1626–1632.
Wu, J.Q., Shteynberg, D., Arumugam, M., Gibbs, R.A., and Brent, M.R. 2004. Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. *Genome Res.* **14:** 665–671.
Yeh, R.F., Lim, L.P., and Burge, C.B. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11:** 803–816.
Zamore, P.D. and Haley, B. 2005. Ribo-gnome: The big world of small RNAs. *Science* **309:** 1519–1524.
Zhang, Z. and Gerstein, M. 2004. Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev.* **14:** 328–335.

## Web site references

http://www.ncbi.nlm.nih.gov/genome/guide/build.html#gene; NCBI's description of its automated gene annotation pipeline.
http://www.sanger.ac.uk/HGP/havana/hawk.shtml; Human Annotation Workshops (Hawk).
http://www.sanger.ac.uk/HGP/havana/havana.shtml; Human and Vertebrate Analysis and Annotation (HAVANA) group at the Sanger Institute.
http://www.sanger.ac.uk/HGP/havana/havana.shtml; HAVANA annotation guidelines.
http://mgc.nci.nih.gov/; Mammalian Gene Collection (MGC) Official Home Page.
http://genes.cse.wustl.edu/mgc/; additional information on MGC clones from the Brent Lab MGC page.
http://www.sanger.ac.uk/HGP/havana/docs/guidelines.pdf; annotation guidelines used by the Sanger Institute's manual annotation group.
http://www.sanger.ac.uk/HGP/havana/; schematic of the predictions used by The Sanger Institute's manual annotation group.
http://genome.imim.es/gencode/workshop2005.html; ENCODE Gene Prediction Workshop — EGASP/2005.
http://www.genome.gov/11006929; Announcement of the completion of the human genome project.
http://genes.cse.wustl.edu/wei-2005b/; Web site for N-SCAN_EST paper and software.
http://genes.cse.wustl.edu/Arumugam-2006/; Web site for Pairagon paper and software.