ELSEVIER

Comparative genomics as a tool for gene discovery

Aaron J Windsor and Thomas Mitchell-Olds

With the increasing availability of data from multiple eukaryotic genome sequencing projects, attention has focused on interspecific comparisons to discover novel genes and transcribed genomic sequences. Generally, these extrinsic strategies combine *ab initio* gene prediction with expression and/or homology data to identify conserved gene candidates between two or more genomes. Interspecific sequence analyses have proven invaluable for the improvement of existing annotations, automation of annotation, and identification of novel coding regions and splice variants. Further, comparative genomic approaches hold the promise of improved prediction of terminal or small exons, microRNA precursors, and small peptide-encoding open reading frames — sequence elements that are difficult to identify through purely intrinsic methodologies in the absence of experimental data.

Addresses
Max-Planck-Institut fuer chemische Oekologie, Abteilung Genetik und Evolution, Hans-Knoell-Strasse 8, D-07745 Jena, Germany

Corresponding author: Windsor, Aaron J (windsor@ice.mpg.de)

Current Opinion in Biotechnology 2006, **17**:161–167

This review comes from a themed issue on
Plant biotechnology
Edited by Nam-Hai Chua and Scott V Tingey

Available online 3rd February 2006

0958-1669/$ – see front matter

DOI 10.1016/j.copbio.2006.01.007

Introduction

The publication of the genome sequence for the yeast *Saccharomyces cerevisiae* in 1996 [1] ushered in the genomic era for the eukaryotic research community. Subsequently, the genome sequences of *Caenorhabditis elegans* [2], *Drosophila melanogaster* [3], *Arabidopsis thaliana* [4], and human [5] were published. These prototype genome projects provided biologists with the power to evaluate experimental observations in a whole-genome context. Technical innovations in molecular biology, biochemistry, and information processing precipitated by these projects have made high-throughput tools cost-effective and accessible to a wider range of investigators.

The need to improve annotation and gene identification in the prototype genome sequences, the desire to investigate natural variation and genome evolution, and the recognition of the practical limitations to gene discovery through the study of individual genomes has placed an emphasis on comparative studies. As such, there has been an expansion in the number of completed eukaryotic genome sequencing projects (~80) and ongoing projects (~500) over the past five years (Genomes OnLine Database v2.0; http://www.genomesonline.org). The sequencing of additional genomes poses novel logistical and technical issues for the processing and interpretation of sequence data. Unlike the prototype eukaryotic genome projects, extensive manual curation of next-generation sequencing projects is neither time- nor resource-effective. Beginning with the annotation of the mouse genome [6], the partial or complete automation of genome sequence curation has become the norm.

This review will explore recent advances in the prediction and refinement of gene models, the empirical validation of these models, and the identification of non-coding transcribed sequences using comparative genomic approaches. While drawing on the literature at large, the utility of the approaches will be evaluated relative to the current state of plant genomics. Table 1 summaries the methodologies presented in this review that have been formalized as discrete programs or computational pipelines and are available to the research community.

Generalities of gene discovery

De novo gene prediction frameworks are classified as either intrinsic or extrinsic. Intrinsic methodologies (Figure 1; path 'A') make gene predictions from only the information present in the individual DNA sequence analyzed. These methodologies are commonly encountered as *ab initio* tools [7–10] and are, by definition, not comparative. *Ab initio* gene prediction algorithms display high sensitivity, but a low specificity (see Glossary) in their output models; both of these parameters are directly related to the quality and extent of the training data provided to these programs. Training datasets are comprised of experimentally confirmed gene models and are most effective when established from the species being evaluated. These tools do not consistently predict gene boundaries, small exons, and atypical introns with accuracy. Further, *ab initio* methods are not generally capable of identifying small open reading frames (smORFs; see Glossary), transcribed non-coding sequences or regulatory elements. Studies applying comparative approaches to gene discovery have observed that gene models missed by *ab initio* methods display elevated *K*a/*K*s ratios (see Glossary) in interspecific comparisons, suggesting that the sensitivity of

Glossary
cDNA: DNA molecule with the complementary sequence to a transcribed RNA.
cRNA: RNA molecule with the complementary sequence to a transcribed RNA.
Expressed sequence tag (EST): incomplete sequence from a transcribed RNA.
***Ka*/*Ks*:** in interspecific sequence comparisons, a population genetics parameter used to infer neutral evolution versus selection in coding sequences on a *per*-site basis. *Ka*/*Ks* is the ratio of the number of nonsynonymous substitutions (*Ka*) to the number of synonymous substitutions (*Ks*). Values near 0, ~1 and >1 suggest selective constraint, neutrality, and adaptive evolution, respectively.
MicroRNA (miRNA): transcribed elements that regulate the expression of target genes at the post-transcriptional level.
Sensitivity: the ratio of correctly predicted features to the actual number of features present in the query sequence.
Small open reading frame (smORF): short stretches of sequence containing an in-frame start and stop codon, generally encoding peptides a few amino acids to <100 amino acids in length.
Specificity: the ratio of correctly predicted features to the total number of predicted features.

ab initio methodologies is reduced for rapidly evolving sequences [11••,12]. Although *ab initio* tools vary with regard to algorithm design (recently reviewed in [13]) and efficacy [14,15], the application of these tools by the casual investigator is straightforward.

Gene prediction methods based on extrinsic data (including expression evidence [16–20] and/or sequence similarity [21–24,25•]; Figure 1 paths 'B', 'C' and 'D') supplement *ab initio* prediction by providing improved specificity and complementary sensitivity. As the implementation of extrinsic gene prediction methods is complex, and because these methods provide the basis for gene discovery through comparative genomics, the remainder of the review will focus on these methodologies.

The application of expression data to gene discovery

Evidence-based gene discovery frameworks integrate empirical transcription and protein expression data with genome sequence to produce gene models (Figure 1; path 'C') and facilitate annotation [16]. Such data provide high specificity to gene model prediction, but sensitivity is contingent on the extent of the expression dataset(s). This property negatively impacts the identification of sequences with tightly regulated or low-abundance transcripts or of RNA species that are not translated. The incorporation of expression data from multiple species can overcome some of these limitations and allows the use of species with no or partial genome sequence in comparative analyses. For example, the analysis of the tran-

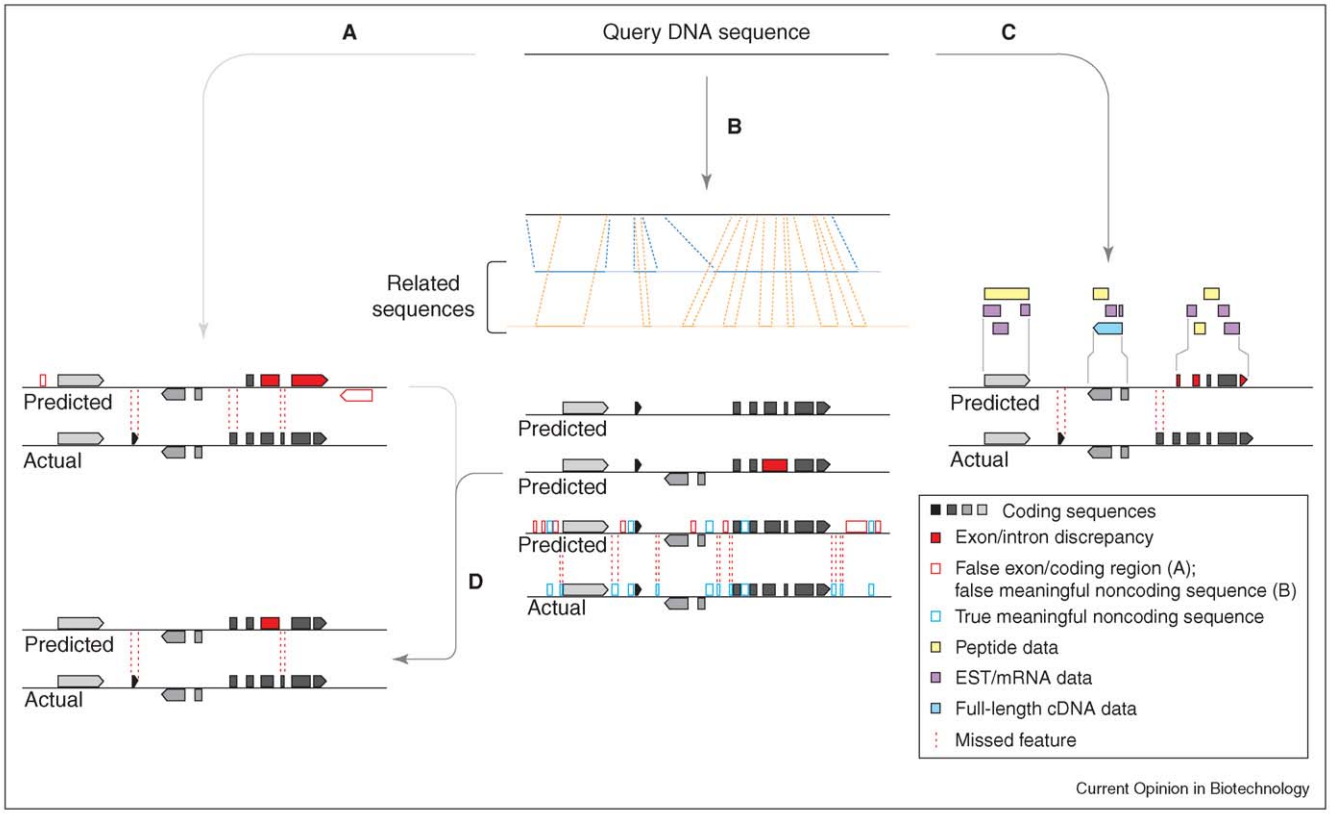
Table 1

Comparative gene discovery implementations.

	Type[a]	Comparison	Description	Reference
EAnnot	Program	Evidence-based	Similar in principle to Ensembl	[19]
Ensembl	Pipeline	Evidence-based	Developed to map transcription and expression data onto a query genome sequence. Qualifies as a comparative approach when interspecific datasets are applied	[16]
ESTMAP	Pipeline	Evidence-based	Creates gene models by mapping perfect BLASTN EST sequence hits onto a query genomic sequence	[18]
Evoprinter	Program	Similarity	Processes the output from BLAT [47] alignments. Conserved features are mapped onto a single, reference sequence and pattern identification is used to identify short, potentially degenerate sequences such as smORFs, miRNAs or regulatory domains	[25•]
Exofish	Pipeline	Similarity	Developed to predict human gene models using genome sequence data from *Tetraodon nigroviridis*	[48]
MIRFINDER	Pipeline	Similarity	Uses BLASTN output and subsequent filtering to identify miRNAs	[41]
Pattern Filtering	Program	Similarity	Interprets the alignment of two DNA sequences as a series of patterns. Regular patterns correspond to conserved sequences; 'noise' in the alignment is filtered. Annotation is performed manually with a second program, GeneGrabber	[24]
Projector	Program	Similarity	Produces gene predictions in a query sequence using the HMM algorithm of Doublescan [49]. Emission probabilities are modified using the annotation of the reference sequence	[50]
SLAM	Program	Similarity	Simultaneously produces gene predictions in two syntenic sequences. Displays a high degree of specificity at the expense of sensitivity. Attempts to discriminate between conserved coding and non-coding sequences	[22]
SGP2	Program	Similarity	Gene predictions are produced using the dynamic programming algorithm of GeneID [51]. Prediction probabilities are modified using similarity to a reference protein sequence dataset identified by TBLASTN analysis	[23]
Twain	Program	Similarity	Produces gene predictions in both sequences being compared using an HMM	[52]
TwinScan	Program	Similarity	Initial gene predictions are produced using the HMM model of GenScan [7]. Prediction probabilities are modified using similarity to a reference genome supplied by BLASTN analysis	[21]

Abbreviation: HMM, Hidden Markov Model.
[a] A pipeline is a semi- or fully automated serialized arrangement of computational analyses.

Figure 1



Intrinsic (light gray arrows) and extrinsic (dark gray arrows) methods for gene prediction. Thick black lines represent query sequence data. Watson–Crick-strand coding sequences are indicated above or below sequence strands. Path (A), *ab initio* gene prediction algorithms model gene content with data from the query sequence itself. These methods can miss features such as smORFs and small introns. Exons are also missed, but *ab initio* methods can erroneously identify exons or whole coding sequences. Generally, these methods are not applicable to the prediction of functional non-coding sequences. Path (B), similarity-based gene prediction. These methods are comparative and incorporate data from the alignment of one or more syntenic DNA sequences. Similarity-based methods display improved sensitivity and specificity for coding and non-coding sequences over *ab initio* methods. The ability to predict genes or conserved features is a function of the number of sequences compared, the evolutionary distance of these sequences, and the degeneracy and size of the features in the homologous sequences. Path (C), evidence-based gene prediction. These methods can be computational or experimental and display high specificity but low sensitivity. The efficacy of the prediction is contingent on the quality/extent of available expression data. Path (D), combinatorial approaches. In the example presented, similarity evidence is combined with an *ab initio* prediction to improve the overall prediction of gene content.

scriptome and proteome datasets for the kingdom Plantae suggests that ~19 000 gene functions are encoded in the green plant lineage [26••]. Nearly 6500 of these are encoded by orphan genes, novel genes that cannot be definitively assigned to a characterized homolog or gene family. Using a technique coined 'Proteogenomic Mapping', Jaffe and colleagues [27] have revised the annotation of *Mycoplasma pneumoniae* with corrections to existing gene models and the incorporation of several previously unidentified coding sequences. To identify novel transcribed sequences in *A. thaliana*, Yamada *et al.* [28] have used *A. thaliana* full-length cDNA data and expressed sequence tag (EST; see Glossary) data from *A. thaliana*, *Brassica*, rice and wheat to pinpoint transcribed, unannotated genomic regions. Approximately 1400 new gene models were ascribed to intergenic regions; the transcription of a majority of these has been supported using whole-genome cRNA (see Glossary) arrays or reverse transcriptase polymerase chain reaction (RT–PCR).

Sequence similarity applied to gene discovery

Similarity-based methods for gene discovery assume that the evolution of functional sequences is constrained by selection and spurious sequences are free to evolve neutrally. Thus, sequences that are conserved in interspecific comparisons are more likely to be biologically meaningful. Two recent studies [29,30•] have attempted to determine the minimum number of genome sequences that are required for the identification of conserved regions. Modeling suggests that the number of required interspecific

contrasts is inversely proportional to the physical size of the feature in question. Comparisons between evolutionarily close species will have a greater rate of false-positives for conserved regions and, although increasing genome number increases information, this information is subject to diminishing returns. These conclusions are broadly applicable, but based on postulates from mammalian comparative genomics. Highly dynamic genomes, such as those of plants, may violate the basic assumptions of orthology and copy number inherent in these models. To this end, the works of Vandepoele and van der Peer [26••] and of Xiong *et al.* [31] have demonstrated gene loss in *A. thaliana* and *Oryza sativa* as well as the presence of species- or lineage-specific gene families in virtually every angiosperm family.

The power of similarity-based gene discovery at a genome scale is demonstrated by recent work in microbial genomes. Kellis and colleagues [32] have applied homology to annotation improvement and gene discovery in *S. cerevisiae*, a species with one of the simplest and best characterized genomes. Using BLASTN [33], the group aligned the *S. cerevisiae* genome to three additional *Saccharomyces* species to identify open reading frames (ORFs) with orthologous sequences in each of the additional genomes. These ORFs were then classified as biologically meaningful by the absence of stop codon interruptions in each of the orthologous clusters. While validating or refining existing annotations, the method also recognized several novel candidate coding sequences in regions previously classified as intergenic and identified ~40 smORFs. Similarly, the genomes of three varieties of *Cyptococcus neoformans*, a fungal pathogen displaying a more complex exon–intron organization than *S. cerevisiae*, were compared [34•] using an implementation of TwinScan (Table 1). Approximately 200 new gene models were produced by the analysis with 80% of these being confirmed at the transcriptional level by RT–PCR. To specifically identify smORFs in *S. cerevisiae*, Kessler *et al.* [35] applied a subtractive approach where genomic regions containing previously annotated ORFs were purged from the subject dataset. Candidate smORFs with homologs in other fungal species were bioinformatically identified from the reduced dataset and a subset of candidates (117) was analyzed for transcription and homology in an expanded range of organisms. More than two-thirds of the candidate smORFs assayed displayed evidence of transcription and one candidate, smORF2, had a homolog in every species examined, including human. Thus, comparative approaches are able to identify previously unknown coding elements even in relatively simple, well-characterized genomes.

Microbial studies benefit from the simplicity of the organisms in question and from the availability of multiple suitable and fully sequenced genomes for comparative analyses. Currently, higher eukaryotes generally lack some or all of these advantages. Nonetheless, comparative approaches are proving very effective in the identification of novel genes in complex genomes. Using TwinScan, 256 new gene models have been predicted in a whole-genome comparison of *C. elegans* and *Caenorhabditis briggsae* [11••]. Of these, the transcription of 146 models could be confirmed. To identify genes with perfectly conserved structure, Dewey and coworkers [36•] compared the mouse and human genomes using SLAM (Table 1) and extended their analysis to the recently completed rat genome. Approximately 3700 nearly perfect ortholog sets were observed with 924 of these representing novel genes.

The lack of available completed genome sequences and the evolutionary distances associated with completed genome sequences impose limits to whole-genome comparative approaches among plants. The creative application of similarity-based analyses, however, has allowed the identification of novel coding sequences. Several thousand conserved unannotated regions were recently recognized in *A. thaliana* relative to the partial genome sequence of *Brassica oleracea* [37•,38•]. In these approaches, conserved genomic regions, as identified by TwinScan, with physical proximity in the *A. thaliana* reference genome were chained together to produce novel gene models. Approximately 300 resultant models were subsequently assayed for transcription, which was detected in 27% of cases. Interestingly, some gene models for which transcription was not detected displayed conservation relative to a third, more distantly related, genome *O. sativa* [38•]. In a similar analysis, an implementation of Exofish (Table 1) was optimized for *O. sativa–A. thaliana* comparisons [39]. Observed islands of sequence conservation were likewise chained to produce gene models, thus allowing the creation of more than 3500 new gene annotations. Nearly a third of these new models have been supported by comparison to *A. thaliana* full-length cDNAs.

Similarity-based approaches are also being applied successfully to the identification of transcribed non-coding genes such as microRNA (miRNA; see Glossary) precursors and their target sequences [40–42,43••]. Among these studies, differences in miRNA precursor organization between plant and animal systems highlight the benefits of homology over intrinsic methods for recognizing these elements. Further, the application of multiple phylogenetic comparisons [43••] has demonstrated the existence of taxon-specific elements and losses — features of genome evolution that may not be identified through intrinsic methods or pairwise comparison. Evidence for widespread adaptive evolution in non-coding sequences of *Drosophila* [44••] and the implications of this observation for speciation and local adaptation, however, indicate that conservation is only one criterion for the identification of biologically meaningful non-coding sequences.

Recent studies have demonstrated the utility of merging evidence-based and similarity-based approaches to gene discovery. To improve the rat genome annotation, Wu *et al.* [45] combined Ensembl and rat–human TwinScan predictions to identify coding regions that would otherwise have been missed by Ensembl alone. The resultant rat gene models were then validated by RT–PCR and by comparison with putative homologs in the Human Gene Mutation Database. In similar work, novel mouse genes were identified via an initial Ensembl analysis followed by mouse–human TwinScan and mouse–human SGP2 (Table 1) analyses [12]. Of the gene models obtained, 62% were supported by transcriptional data. Of the supported gene models, 76% were predicted by both TwinScan and SGP2 indicating that, although similar in principle, TwinScan and SGP2 analyses complement each other.

Current implementations of similarity-based gene prediction, although effective, are still limited with regard to whole-genome analyses. Approaches directly utilizing homology based on BLAST and related algorithms may lack the sensitivity to detect short and/or degenerate genome features. Approaches such as TwinScan, SGP2, and Projector (Table 1) make use of underlying *ab initio* predictions and thus rely on training datasets for their execution. Training datasets are potentially limiting among next-generation genome sequencing projects and can introduce ascertainment bias in relation to unidentified genes with atypical organization or structural components. Further, all of these analyses are inherently pairwise; multiple comparisons require *post hoc* integration of individual comparisons. Recently, algorithms such as Pattern Filtering [24] and Evoprinter [25•] (Table 1) have been developed that display improved sensitivity over other prediction methods for features such as smORFs or even regulatory sequences. Neither algorithm makes use of training data. While Pattern Filtering has only been applied to pairwise analyses, Evoprinter has been specifically developed for multiple comparisons and includes a second function, Evodifference, which identifies species- or lineage-specific sequence losses.

Conclusions

Comparative approaches are proving their value for gene discovery and annotation improvement. Current results indicate that the availability of additional genome sequences and application of combinatorial approaches will further improve efficacy. Although recent studies have used transcription or protein expression data to support novel gene models, little attention has been focused on the functions of the coding regions identified; only one paper reviewed here used mutational and complementation analyses to demonstrate a phenotype [35]. The integration of technologies that transfer protein function between species [46] into comparative genomic frameworks may partially alleviate this deficiency; however, the importance of functional characterization cannot be ignored. In the absence of such analyses, the line between 'prediction' and 'discovery' will remain blurred.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al.*: **Life with 6000 genes**. *Science* 1996, **274**:563-567.
2. *C. elegans* Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology**. *Science* 1998, **282**:2012-2018.
3. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al.*: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**:2185-2195.
4. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408**:768-815.
5. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.
6. Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**:520-562.
7. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**:78-94.
8. Guigo R, Knudsen S, Drake N, Smith T: **Prediction of gene structure**. *J Mol Biol* 1992, **226**:141-157.
9. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders**. *Bioinformatics* 2004, **20**:2878-2879.
10. Korf I: **Gene finding in novel genomes**. *BMC Bioinformatics* 2004, **5**:59-68.
11. •• Wei C, Lamesch P, Arumugam M, Rosenberg J, Hu P, Vidal M, Brent MR: **Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions**. *Genome Res* 2005, **15**:577-582.

This paper demonstrates the advantage of comparative analysis over *ab initio* gene model prediction. The study shows that the new genes identified by TwinScan are more divergent than previously annotated models, suggesting that population genetic parameters such as *Ka*/*Ks* may be informative in gene prediction frameworks.

12. Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C *et al.*: **Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes**. *Proc Natl Acad Sci USA* 2003, **100**:1140-1145.
13. Wang Z, Chen Y, Li Y: **A brief review of computational gene prediction methods**. *Genomics Proteomics Bioinformatics* 2004, **2**:216-221.
14. Pavy N, Rombauts S, Dehais P, Mathe C, Ramana DVV, Leroy P, Rouze P: **Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences**. *Bioinformatics* 1999, **15**:887-899.
15. Yao H, Guo L, Fu Y, Borsuk LA, Wen T-J, Skibbe DS, Cui X, Scheffler BE, Cao J, Emrich SJ *et al.*: **Evaluation of five *ab initio* gene prediction programs for the discovery of maize genes**. *Plant Mol Biol* 2005, **57**:445-460.
16. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T *et al.*: **The Ensembl genome database project**. *Nucleic Acids Res* 2002, **30**:38-41.
17. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression**. *Science* 1995, **270**:484-487.

18. Milanesi L, Rogozin IB: **ESTMAP: a system for expressed sequence tags mapping on genomic sequences**. *IEEE Trans Nanobioscience* 2003, **2**:75-78.

19. Ding L, Sabo A, Berkowicz N, Meyer RR, Shotland Y, Johnson MR, Pepin KH, Wilson RK, Spieth J: **EAnnot: a genome annotation tool using experimental evidence**. *Genome Res* 2004, **14**:2503-2509.

20. Brendel V, Xing L, Zhu W: **Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus**. *Bioinformatics* 2004, **20**:1157-1169.

21. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction**. *Bioinformatics* 2001, **17**:S140-S148.

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

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

24. Moore JE, Lake JA: **Gene structure prediction in syntenic DNA segments**. *Nucleic Acids Res* 2003, **31**:7271-7279.

25. • Odenwald WF, Rasband W, Kuzin A, Brody T: **EVOPRINTER, a multigenomic comparative tool for rapid identification of functionally important DNA**. *Proc Natl Acad Sci USA* 2005, **102**:14700-14705.

This study demonstrates that multiple species comparison can achieve near single nucleotide resolution in the identification of conserved features, both coding and noncoding.

26. •• Vandepoele K, Van de Peer Y: **Exploring the plant transcriptome through phylogenetic profiling**. *Plant Physiol* 2005, **137**:31-42.

A comparative approach using transcriptome data from within Plantae. The paper effectively highlights the technical challenges associated with plant comparative genomics that arise from segmental and tandem duplications. Approximately 19 000 gene functions are identified; total numbers are comparable for both monocots and eudicots. The paper provides evidence of genome reduction in both *A. thaliana* and *O. sativa*.

27. Jaffe JD, Berg HC, Church GM: **Proteogenomic mapping as a complementary method to perform genome annotation**. *Proteomics* 2004, **4**:59-77.

28. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M *et al.*: **Empirical analysis of transcriptional activity in the *Arabidopsis* genome**. *Science* 2003, **302**:842-846.

29. Cooper GM, Brudno M, Nisc CSP, Green ED, Batzoglou S, Sidow A: **Quantitative estimates of sequence divergence for comparative analyses of Mamm genomes**. *Genome Res* 2003, **13**:813-820.

30. • Eddy SR: **A model of the statistical power of comparative genome sequence analysis**. *PLoS Biol* 2005, **3**:e10.

An excellent statistical treatment of the efficacy of multiple species comparisons as applied to gene or sequence feature discovery. The article is well written and clear, even for the non-statistically minded.

31. Xiong Y, Liu T, Tian C, Sun S, Li J, Chen M: **Transcription factors in rice: a genome-wide comparative analysis between monocots and eudicots**. *Plant Mol Biol* 2005, **59**:191-203.

32. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements**. *Nature* 2003, **423**:241-254.

33. Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

34. • Tenney AE, Brown RH, Vaske C, Lodge JK, Doering TL, Brent MR: **Gene prediction and verification in a compact genome with numerous small introns**. *Genome Res* 2004, **14**:2330-2335.

This study uses the fungal pathogen *C. neoformans* to demonstrate the accuracy that TwinScan can achieve in the prediction of atypical introns. Further, this work shows that TwinScan performance improves when empirical models of intron length are substituted for the geometric probability of intron lengths used in the standard TwinScan/GenScan implementations.

35. Kessler MM, Zeng Q, Hogan S, Cook R, Morales AJ, Cottarel G: **Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome**. *Genome Res* 2003, **13**:264-271.

36. • Dewey C, Wu JQ, Cawley S, Alexandersson M, Gibbs R, Pachter L: **Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat**. *Genome Res* 2004, **14**:661-664.

This study uses SLAM to identify nearly 3700 orthologous triples in the mammalian species. Novel human gene models identified in this study display restricted transcription and tissue specificity, demonstrating an advantage of similarity-based gene prediction over evidence-based prediction.

37. • Ayele M, Haas BJ, Kumar N, Wu H, Xiao Y, Van Aken S, Utterback TR, Wortman JR, White OR, Town CD: **Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in *Arabidopsis***. *Genome Res* 2005, **15**:487-495.

This study, and the following reference, use the partial genome sequence of *B. oleracea* to identify unannotated, transcribed coding sequences in *A. thaliana*, thereby demonstrating that a substantial proportion of coding sequences remain unidentified in this model system.

38. • Katari MS, Balija V, Wilson RK, Martienssen RA, McCombie WR: **Comparing low coverage random shotgun sequence data from *Brassica oleracea* and *Oryza sativa* genome sequence for their ability to add to the annotation of *Arabidopsis thaliana***. *Genome Res* 2005, **15**:496-504.

See annotation for [37•].

39. Castelli V, Aury J-M, Jaillon O, Wincker P, Clepet C, Menard M, Cruaud C, Quetier F, Scarpelli C, Schachter V *et al.*: **Whole genome sequence comparisons and 'full-length' cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation**. *Genome Res* 2004, **14**:406-413.

40. Lau NC, Lim LP, Weinstein EG, Bartel DP: **An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans***. *Science* 2001, **294**:858-862.

41. Bonnet E, Wuyts J, Rouze P, Van de Peer Y: **Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes**. *Proc Natl Acad Sci USA* 2004, **101**:11511-11516.

42. Jones-Rhoades MW, Bartel DP: **Computational identification of plant microRNAs and their targets, including a stress-induced miRNA**. *Mol Cell* 2004, **14**:787-799.

43. •• Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RHA, Cuppen E: **Phylogenetic shadowing and computational identification of human microRNA genes**. *Cell* 2005, **120**:21-24.

In this study, phylogenetic shadowing of 10 primate species was applied to identify key structural features of genomic miRNA regions. The analysis was extended to human–mouse and human–rat comparisons for the identification of ~1000 candidate miRNAs, nearly 700 of which are conserved among vertebrates.

44. •• Andolfatto P: **Adaptive evolution of non-coding DNA in Drosophila**. *Nature* 2005, **437**:1149-1152.

This study not only identifies conserved non-coding sequences in *Drosophila*, but provides evidence that a substantial proportion of non-coding DNA is more diverged than predicted under neutrality. Thus, sequence signatures of adaptive evolution also provide a signal for the identification of biologically meaningful non-coding sequences.

45. Wu JQ, Shteynberg D, Arumugam M, Gibbs RA, Brent MR: **Identification of rat genes by TWINSCAN gene prediction, RT-PCP, and direct sequencing**. *Genome Res* 2004, **14**:665-671.

46. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han J-DJ, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs**. *Genome Res* 2004, **14**:1107-1118.

47. Kent WJ: **BLAT: The BLAST-like alignment tool**. *Genome Res* 2002, **12**:656-664.

48. Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quetier F *et al.*: **Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence**. *Nat Genet* 2000, **25**:235-238.

49. Meyer IM, Durbin R: **Comparative *ab initio* prediction of gene structures using pair HMMs**. *Bioinformatics* 2002, **18**:1309-1318.

50. Meyer IM, Durbin R: **Gene structure conservation aids similarity based gene prediction**. *Nucleic Acids Res* 2004, **32**:776-783.

51. Parra G, Blanco E, Guigo R: **GeneID in Drosophila**. *Genome Res* 2000, **10**:511-515.

52. Majoros WH, Pertea M, Salzberg SL: **Efficient implementation of a generalized pair hidden Markov model for comparative gene finding**. *Bioinformatics* 2005, **21**:1782-1788.