Homology-based annotation yields 1,042 new candidate genes in the *Drosophila melanogaster* genome

Shuba Gopal[1]*, Mark Schroeder[1]*, Ursula Pieper[1,2], Alexander Sczyrba[1], Gulriz Aytekin-Kurban[1], Stefan Bekiranov[1], J. Eduardo Fajardo[1], Narayanan Eswar[2], Roberto Sanchez[2], Andrej Sali[2] & Terry Gaasterland[1]

*These authors contributed equally to this work.

The approach to annotating a genome critically affects the number and accuracy of genes identified in the genome sequence. Genome annotation based on stringent gene identification is prone to underestimate the complement of genes encoded in a genome. In contrast, over-prediction of putative genes followed by exhaustive computational sequence, motif and structural homology search will find rarely expressed, possibly unique, new genes at the risk of including non-functional genes. We developed a two-stage approach that combines the merits of stringent genome annotation with the benefits of over-prediction. First we identify plausible genes regardless of matches with EST, cDNA or protein sequences from the organism (stage 1). In the second stage, proteins predicted from the plausible genes are compared at the protein level with EST, cDNA and protein sequences, and protein structures from other organisms (stage 2). Remote but biologically meaningful protein sequence or structure homologies provide supporting evidence for genuine genes. The method, applied to the *Drosophila melanogaster* genome, validated 1,042 novel candidate genes after filtering 19,410 plausible genes, of which 12,124 matched the original 13,601 annotated genes[1]. This annotation strategy is applicable to genomes of all organisms, including human.

Conservative annotation requires that each predicted gene from the genomic sequence match an expressed sequence tag (EST), a cDNA or a known protein from the organism in question[1]. Although not all genes are represented in the EST libraries for an organism[2], the EST libraries represent far more genes than the protein sequence databases. At the same time, if the predicted protein sequence translated from a predicted gene aligns significantly with protein sequences from other organisms or the same organism, the sequence conservation indicates that the gene may be real even in the absence of EST data.

The first stage of our approach applies gene identification software to parse the genomic sequence and identify all sequence features that indicate initial, internal and terminal exons in plausible genes. The second stage compares all predicted genes with all available gene and protein sequences, including ESTs from other organisms at the protein level. We predicted genes using existing algorithms implemented in GENSCAN (refs. 3,4) that operate on the genome sequence directly to assess codon usage, stop codons and potential splice sites. GENSCAN assigns exons to distinct genes that preferably start with a promoter and end with a polyadenylation site. At the expense of 'over-predicting' possible genes (false positives[5]) or predicting extra exons within true genes, this step minimizes the likelihood of missing potential genes (false negatives). The fact that GENSCAN missed 1,093 previously annotated genes indicates that ultimately all available gene-finding tools should be used when annotating a genome.

To confirm plausible genes through more remote but biologically meaningful sequence conservation, we compared them with six-frame translations of EST sequences from GenBank (DBEST; ref. 6), including ESTs from human and mouse, to all known *Drosophila* protein, cDNA and EST sequences; to proteins predicted from other complete genomes[7]; and to proteins maintained in the GenBank non-redundant protein sequence database[8] (NR) and the PDB protein structure database[9]. Table 1 lists the searches used to execute this strategy through the MAGPIE/EGRET genome annotation system[7,10,11] and the MODPIPE protein structure modeling pipeline[12,13]. Genes were classified by supporting evidence: translated ESTs and protein alignments (class 1); translated EST but not protein alignments (class 2); and protein alignments but not translated EST (class 3). Evidence class 1 is the strongest, requiring sequence similarity to both a known protein and a translated EST. Next strongest, classes 2 and 3, require translated EST sequence similarity or protein similarity, but not both. We set aside 590 novel genes that encoded putative transposon-related proteins. We counted genes as previously annotated if they matched existing genes at a high percentage identity and matched the same locations in the genome.

When sequence similarity is low, a reliable three-dimensional model of the predicted protein indicates correct gene assignment. We applied comparative protein structure modeling with model assessment[15] to all predicted proteins that aligned with proteins of known structure. Each predicted protein was subjected to fold assignment, sequence-structure alignment, model building and model assessment[12,14]. Passing the model assessment test confirms a model as 'reliable'. In studies of model genomes, 20–60% of known protein sequences have at least one segment related to known structures[12,15]. Applied to genomes, modeling yields less than 4% false positives even when 30% of all reliable models are based on matches with statistically insignificant PSI-BLAST E-values.

Table 2 shows the numbers of predicted genes in each evidence category: total plausible genes (19,410), previously annotated genes that matched plausible genes (12,124 of 13,601), genes without supporting evidence (5,654) and novel candidate genes (1,042 predicted genes that match no previously annotated gene but with supporting evidence). The total number of novel candidate genes represents a 7.6% increase over the 13,601 currently annotated genes. The total complement of predicted genes did not include 1,093 previously annotated genes[1]. Table 3 enumerates the novel candidate genes according to the classes of evidence.

Laboratories of [1]Computational Genomics and [2]Biophysics, The Rockefeller University, New York, New York, USA. Correspondence should be addressed to T.G. (e-mail: gaasterland@rockefeller.edu).

Table 1 • Computational searches on the *Drosophila* genome

Tool	Input	Target	Cutoff	Output
GENSCAN	1378 100Knt ctgs	n/a	none	19,410 genes
BLASTX	1378 100Knt ctgs	NR	≤10^{-4}	protein alignments
BLASTN	19,410 DNA seqs	GenBank DNA	≥88% id	exact mRNA matches
BLASTN	19,410 DNA seqs	*Drosophila* ESTs	≥88% id	exact EST matches
BLASTN	19,410 DNA seqs	FGENESH predicted exons	≥88% id	exact predicted exon matches
BLASTP	19,410 aa seqs	annotated *Drosophila* proteins	≥88% id	previously annotated genes
BLASTN	19,410 DNA seqs	annotated *Drosophila* genes	≥88% id[1]	previously annotated genes
TBLASTN	19,410 aa seqs	GenBank ESTs	≤10^{-4}	partial alignments
BLASTP	19,410 aa seqs	NR	≤10^{-4}	annotated protein homologs
BLASTP	19,410 aa seqs	archaea predicted proteins	≤10^{-4}	archaeal predicted protein homologs
BLASTP	19,410 aa seqs	bacteria predicted proteins	≤10^{-4}	bacterial predicted protein homologs
BLASTP	19,410 aa seqs	eukaryotic predicted proteins	≤10^{-4}	eukaryotic predicted protein homologs
PFAM	19,410 aa seqs	protein families	NA	function motifs
PROSITE	19,410 aa seqs	functional patterns	NA	function motifs
BLOCKS	19,410 aa seqs	protein families	NA	function motifs
PSIBLAST	19,410 aa seqs	NR	31 iterations	protein-specific substitution matrices (PSSMs)
MODPIPE	19,410 aa seqs	PDB	≤10^{-2}	putative structure models and/or fold assignments

Computational searches carried out on the *Drosophila* genome via MAGPIE and MODPIPE, 283,590 searches total. (Knt, 1,000 nt; ctgs, contiguous DNA sequences; seqs, sequences; percent identity cutoffs were over the portion of the sequence aligned.) The table shows the algorithm (Tool) used to evaluate sequences, the number and type of input sequences (Input), the sequences used for comparison (Target), the minimum or maximum threshold used as an acceptable resulting score (Cutoff), and the type of output (Output). (nt, nucleotides; ctgs, contigs; aa, amino acid; seqs, sequences; NR, GenBank nonredundant protein sequences; ESTs, expressed sequence tags.)

Reliable protein structure models were calculated for segments in 4,347 (32%) of proteins previously annotated in the *Drosophila* genome. An additional 1,180 (9%) previously annotated proteins had significant PSI-BLAST alignments with E-value≤0.0001 to a known protein structure, totaling 5,527 (41% of 13,601) with structural information. For the 1,042 novel candidate genes, reliable models were constructed for 168 (16%) novel protein sequences, and an additional 92 (9%) proteins had significant PSI-BLAST alignments with E-value≤0.0001 to a known protein structure, yielding 260 (25%) novel candidate genes with supporting structural evidence, of which 4 had only remote alignments (E-value>0.0001) and reliable models as evidence. Predicted genes with supporting evidence are available (http://genomes.rockefeller.edu/dm and http://guitar.rockefeller.edu/modbase).

Some candidate genes with similarity, but not identity, to annotated genes may be pseudogenes. Pseudogenes occur through cDNA recombination and genomic duplication. The former produces pseudogenes with missing introns; the latter may produce truncated or frameshifted genes. All novel candidate genes that match other genes contain introns or encode proteins with at most 30–75% similarity to other genes and thus are not recent duplicates.

The 13,601 previously annotated genes were predicted primarily by GENIE (ref. 16) and partly by GENSCAN (ref. 1) and confirmed by alignment with *Drosophila* EST, cDNA and protein sequences. GENSCAN predictions that did not match GENIE predictions, *Drosophila* ESTs or known proteins had not been analyzed further[1]. In contrast, we kept all genes predicted by GENSCAN and processed them to search for further supporting evidence. The primary difference in the approaches was the gene filtering method.

The second major difference between the approach here and the previous annotation was that we compare the protein translations of predicted genes with the protein translations of all available EST data in all six frames. The amount of translatable sequence available in the EST databases far exceeds the amount of sequence in the curated protein databases. Consequently, our search inspected a much larger pool of available protein sequence for supporting evidence for predicted genes. Additional human and mouse cDNA sequence data from high-throughput sequencing efforts[17] are likely to provide further supporting evidence for plausible genes. As the human and mouse genomes emerge, protein translations of conserved sequences will provide a rich source of additional evidence.

Comparative protein structure modeling with model assessment was the third major difference here. In the absence of strong sequence similarities, comparative modeling and assessment can provide supporting evidence for predicted genes.

The false-positive rate of the entire approach indicates how many novel candidate genes are likely to be false. The rate depends only on the filtering step. False positives would be due to sequence similarity, model assessment error, transposon-related proteins not identified as such, and pseudo-genes. For novel proteins confirmed by sequence alignments alone, the false-positive rate is quantified by the alignment E-value cutoff, 10^{-4}, 1 false positive in 10,000. The false-negative rate depends on the gene prediction tools and the completeness of protein and EST databases.

As a computational test of the novel genes, we compared them with *Drosophila* EST sequences deposited before March 2000 and with 9,090 EST sequences deposited after March 2000, the date of the previous annotation. We found 22 genes matched new sequences alone at 98% identity or higher.

Once candidate coding regions are identified computationally, the next step is high-throughput experimental confirmation. The ORF sequence tag method has been used for *Caenorhabditis elegans* to confirm 796 of 1,222 sampled from 9,888 predicted genes[18] with no further evidence and may be useful here.

Table 2 • Evidence categories for predicted genes in each of chromosomes 2, 3, 4 and X and unassembled sequence

Chromosome	2	3	4	X	U	Total
Total predicted genes	6,937	8,261	97	3,315	800	19,410
Prev. annotated genes (% predicted genes)	4,351 (62%)	5,370 (65%)	71 (82%)	2,061 (62%)	271 (34%)	12,124 (62%)
Genes with no evidence (% predicted genes)	1,939 (28%)	2,472 (30%)	11 (11%)	1,062 (32%)	170 (21%)	5,654 (29%)
Novel candidate genes (% predicted genes) (% increase ann. genes)	481 (7%)	291 (3%)	9 (9%)	131 (4%)	130 (16%)	1,042 (5%) (7.6%)
Novel transposon-related predicted genes (% predicted genes)	166 (1%)	128 (1%)	6 (0%)	61 (0%)	229 (1%)	590 (3%)
Transposon-related prev. annotated gene	100	107	9	38	130	384

For GENSCAN predicted genes in each of chromosomes 2, 3, 4 and X and for unassembled sequence (U), total number of predicted genes (Total predicted genes); predicted genes that matched previously annotated genes (Previously annotated genes); predicted genes with no supporting evidence (Genes with no evidence); predicted genes with supporting evidence and no match at all to previously annotated genes (Novel candidate genes); total number of candidate genes; number of transposon related genes. Supporting evidence was collected for each predicted gene by comparing its amino acid translation using TBLASTN with GenBank EST sequences; BLASTP with proteins predicted from bacterial, archaeal and eukaryotic genomes; BLASTP with the GenBank non-redundant protein database (NR); BLASTP with the currently annotated *Drosophila* proteins; and BLASTN with the coding sequences for the currently annotated *Drosophila* proteins.

In future work, the next step in using protein structure prediction to confirm predicted proteins will be to test alternatively spliced genes through protein structure modeling. As the protein structure databases grow, the impact of structure modeling on genome annotation will increase[19].

Our annotation of the *Drosophila* genome demonstrates the power of predicting all possible putative genes and comparing their predicted proteins with all available sequence and structure data. Conservative, stringent genome annotation maximizes the likelihood that an annotated gene is expressed. In contrast, using all available gene-finding tools maximizes the likelihood of finding rarely expressed and unique new genes. The two-stage genome annotation strategy introduced here combines the merits of both approaches to annotation. For the *Drosophila* genome, the approach confirms computationally 1,042 novel candidate genes. Even if some elements of a predicted candidate coding region are incorrect, the prediction provides a valuable starting point for further experimentation, especially if the annotation suggests functional or structural features.

Methods

Gene structure identification. We identified plausible genes with existing gene-prediction software. We used the GENSCAN gene identification software[4] with default parameters. GENSCAN predicted 19,410 DNA sequences that met minimal requirements including the following: predicted genes consist of a series of translatable open reading frames each bounded by intron-exon splice sites that preserve frame in the coding sequence without introducing stop codons; the first coding exon contains a start codon (ATG) and the last coding exon contains a stop codon; the first exon may be preceded by a predicted promoter region, and the last exon may be followed by a predicted terminator region.

Sequence comparisons. Each protein sequence translated from the predicted coding sequence for each predicted gene was compared with six-frame translations of all available expressed sequence tags and complete cDNAs, and with protein sequences from all organisms in GenBank and dbEST. The EST comparisons were carried out with TBLASTN (ref. 20), which compares amino acid query sequences with all six frame translations of each EST or cDNA sequence. Protein sequences were compared directly through BLASTP (ref. 2). Gene DNA sequences were compared with DNA sequences in GenBank, dbEST and exons predicted from FGENESH through BLASTN. FGENESH (ref. 21) exons were downloaded from the EBI web site (http://www.ebi.ac.uk). All alignments were carried out with masking for low information content and default parameters otherwise. Sequence motifs related to function were identified through PFAM (ref. 22), BLOCKS (ref. 23) and PROSITE (ref. 24). PSI-BLAST (ref. 25) was used with masking for low information content to identify remote sequence conservation and construct position-specific substitution matrices (PSSMs) for each predicted protein sequence.

Protein 3D structure model construction. Fold assignment was done by PSI-BLAST (ref. 13)). A PSSM was built by using the non-redundant protein sequence database from NCBI. The PSSM was then used to identify all representative protein structures from the Protein Data Bank (PDB (ref. 13) that matched the putative protein over at least 30 residues and with an E-value <0.01. Using the PSI-BLAST alignments, comparative protein structure models for the predicted proteins were built by satisfaction of spatial restraints, as implemented in the program MODELER[26]. The fold of the models was evaluated by a multifactorial model assessment procedure that relied on a Z-score calculated from a statistical energy function, sequence similarity on which the model was based, and model compactness[12,15].

Identification of previously annotated genes. Predicted genes were mapped to previously annotated genes using BLASTN to compare DNA sequences and BLASTP to compare protein sequences with low information content regions masked. If pairwise alignments had 88% identity or greater, the predicted gene was considered synonymous with the previously annotated gene. The 88% cutoff was chosen as a liberal lower bound on matching genes because GENIE and GENSCAN can sometimes vary in the specific exons selected, in an effort to call predicted genes novel only when there was substantial sequence divergence. Likewise, if pairwise alignments had 75% identity or greater and if the difference between percent similarity and percent identity was less than 10%, then the predicted gene was considered synonymous to the annotated gene. All genes identified as synonymous were mapped to the genomic sequence and checked for overlapping boundaries. These predicted genes are labeled 'previously annotated'.

Transposon-related genes. Coding sequences were labeled as transposon-related if the following conditions were met: (i) the protein translations aligned with a large number of other proteins within the genome, indicating a high level of internal propagation; (ii) matching protein sequence descriptions contained transposon-related keywords including "transposable," "mobile element," "retroviral," "retrovirus," "reverse transcriptase," "pol," "gag," "blastopia," and "mycopia"; or (iii) the predicted protein matched known transposon-related proteins.

Gene structure refinement. To refine predicted genes, the proteins or complete cDNAs that matched the predicted protein most closely were aligned with TBLASTN with the predicted gene and with the genomic sequence containing the gene. This comparison identified obvious missing exons or extra exons. This step also identified exons that should be reassigned from one gene

Table 3 • For chromosomes 2, 3, 4 and X and for unassembled sequence (U), novel candidate genes enumerated by type of supporting evidence

Chromosome	2	3	4	X	U	Total	% Total
1. EST & protein	145	74	4	32	24	279	27%
2. EST only, no protein	193	149	4	62	50	458	44%
3. protein, no EST	143	68	1	37	56	305	29%
Total	481	291	9	131	130	1,042	

to another and genes that should be merged. Further, this step provided data to refine exon boundaries when best-matching protein sequences were very closely conserved and indicated missing or extra amino acids[11].

Acknowledgments

We thank S. Burley, M. Vidal, J. Sorge, J. Goncalves, M. Ashburner, S. Lewis, M. Young and U. Gaul for insights and comments. This work was partially supported by the Mathers, Sinsheimer and Mallinkrodt Foundations, National Cancer Institute Health grant R33CA84699, National Institutes of Health grant P50GM62529, and the National Science Foundation grant DBI-9984882.

Received 22 December 2000; accepted 7 February 2001.

1. Adams, M.D. *et al*. The genome sequence of Drosophila melanogaster. *Science* **287**, 2185–2195 (2000).
2. Rubin, G.M. *et al.* A Drosophila complementary DNA resource. *Science* **287**, 2222–2224 (2000).
3. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
4. Burge, C.B. & Karlin, S. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8**, 346–354 (1998).
5. Reese, M.G. *et al*. Genome annotation assessment in Drosophila melanogaster. *Genome Res.* **10**, 483–501 (2000).
6. Boguski, M.S., Tolstoshev, C.M. & Bassett, D.E. Gene discovery in dbEST. *Science* **265**, 1993–1994 (1994).
7. Gaasterland, T. & Ragan, M.A. Constructing multigenome views of whole microbial genomes. *Microb. Comp. Genomics* **3**, 177–192 (1998).
8. Benson, D.A. *et al*. GenBank. *Nucleic Acids Res.* **27**, 12–17 (1999).
9. Bhat, T.N. *et al*. The PDB data uniformity project. *Nucleic Acids Res.* **29**, 214–218 (2001).
10. Deckert, G. *et al*. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353–358 (1998).
11. Gaasterland, T. *et al*. MAGPIE/EGRET annotation of the 2.9-Mb Drosophila melanogaster Adh region. *Genome Res.* **10**, 502–510 (2000).
12. Sánchez, R. & Sali, A. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc. Natl. Acad. Sci. USA* **95**, 13597–13602 (1998).
13. Sánchez, R. & Sali, A. ModBase: a database of comparative protein structure models. *Bioinformatics* **15**, 1060–1061 (1999).
14. Sánchez, R. & Sali, A. Evaluation of comparative protein structure modeling by MODELLER -3. *Proteins* **Suppl. 1**, 50–58 (1997).
15. Martí-Renom, M.A. *et al*. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
16. Reese, M.G., Kulp, D., Tammana, H. & Haussler, D. Genie—gene finding in Drosophila melanogaster. *Genome Res*. **10**, 529–538 (2000).
17. Strausberg, R.L., Feingold, E.A., Klausner, R.D. & Collins, F.S. The mammalian gene collection. *Science* **286**, 455–457 (1999).
18. Reboul, J. *et al.* Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nature Genet*. **27**, 332–336 (2001).
19. Burley, S.K. *et al*. Structural genomics: beyond the human genome project. *Nature Genet*. **23**, 151–157 (1999).
20. Altschul, S.F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
21. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res*. **10**, 516–522 (2000).
22. Henikoff, J., Henikoff, S. & Pietrokovski, S. New features of the Blocks Database servers. *Nucleic Acids Res.* **27**, 226–228 (1999).
23. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. The PROSITE database, its status in 1999. *Nucleic Acids Res*. **27**, 215–219 (1999).
24. Altschul, S.F. & Koonin, E.V. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444–447 (1998).
25. Sali, A. & Blundell, T.L. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
26. Bateman, A. *et al*. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res*. **27**, 260–262 (1999).