Additional data file 1

Gene-prediction software

Program	Class*	URL
BLAST [61]	4	http://blast.ncbi.nlm.nih.gov/Blast.cgi
Twinscan [62]	5	http://mblab.wustl.edu/
Sgp2 [63]	5	http://genome.imim.es/software/sgp2/
SLAM [64]	5	http://bio.math.berkeley.edu/slam/mouse/
DoubleScan [65]	5	http://www.sanger.ac.uk/Software/analysis/doublescan/
Augustus [66]	6	http://augustus.gobics.de/
GeneID [67]	6	http://genome.imim.es/software/geneid/
Genscan [68]	6	http://genes.mit.edu/GENSCANinfo.html
GlimmerHMM [69]	6	http://www.cbcb.umd.edu/software/GlimmerHMM/
GeneMark [70]	6	http://exon.gatech.edu/GeneMark/
GenomeScan [71]	7	http://genes.mit.edu/genomescan.html
N-SCAN(_EST) [72]	7, 5	http://mblab.wustl.edu/
Fgenes [h/h++] [73]	7,6	
BLAT [74]	8	http://genome.ucsc.edu/FAQ/FAQblat
Procrustes [75]	8	no longer available
GeneWise [76]	8	ftp://ftp.ebi.ac.uk/pub/software/unix/wise2/
EST_GENOME [77]	8	ftp://ftp.sanger.ac.uk/pub/pmr/est_genome.4.tar.Z
Exonerate [78]	8	http://www.ebi.ac.uk/~guy/exonerate/
GMAP [79]	8	http://www.gene.com/share/gmap/
Pairagon [80]	8	http://mblab.wustl.edu/software/pairagon/
Exogean [81]	9	http://www.biologie.ens.fr/dyogen/
Aceview [82]	9	http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/
Pairagon + N-SCAN_EST [80]	9	http://mblab.wustl.edu/
Ensembl [83]	9	http://www.ensembl.org/
GNOMON [23]	9	http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml
UCSC Genes [84]	9	http://genome.ucsc.edu/
GLEAN [85]	10	http://sourceforge.net/projects/glean-gene
Jigsaw [86]	10	http://www.cbcb.umd.edu/software/jigsaw/
Evigan [87]	10	http://www.seas.upenn.edu/~strctlrn/evigan/evigan.html
GAZE [88]	10 + 2	http://www.sanger.ac.uk/Software/analysis/GAZE/
EuGene [89]	10 + 2	http://www.inra.fr/mia/T/EuGene/
Genomix [90]	10 + 5	http://www.sanger.ac.uk/Software/analysis/genomix/

*Program classes are numbered according to Figure 1 of the main text. Briefly: (2) signals; (4) similarity search; (5) multi-genome; (6) *ab initio*; (7) expression evidence; (8) spliced alignment; (9) pipeline; (10) combiner.

References

61. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
62. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction.** *Bioinformatics* 2001, **17(Suppl 1)**:S140-S148.
63. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigó R: **Comparative gene prediction in human and mouse.** *Genome Res* 2003, **13:**108-117.
64. 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.
65. Meyer IM, Durbin R: **Comparative ab initio prediction of gene structures using pair HMMs.** *Bioinformatics* 2002, **18:**1309-1318.
66. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts.** *Nucleic Acids Res* 2006, **34(Web Server issue):**W435-W439.
67. Parra G, Blanco E, Guigó R: **GeneID in *Drosophila*.** *Genome Res* 2000, **10:**511-515.
68. Burge CB, Karlin S: **Finding the genes in genomic DNA.** *Curr Opin Struct Biol* 1998, **8:**346-354.
69. Majoros WH, Pertea M, Antonescu C, Salzberg SL: **GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders.** *Nucleic Acids Res* 2003, **31:**3601-3604.
70. Lukashin AV, Borodovsky M: **GeneMark.hmm: new solutions for gene finding.** *Nucleic Acids Res* 1998, **26:**1107-1115.
71. Yeh RF, Lim LP, Burge CB: **Computational inference of homologous gene structures in the human genome.** *Genome Res* 2001, **11:**803-816.
72. Gross SS, Brent MR: **Using multiple alignments to improve gene prediction.** *J Comput Biol* 2006, **13:**379-393.
73. Salamov AA, Solovyev VV: **Ab initio gene finding in *Drosophila* genomic DNA.** *Genome Res* 2000, **10:**516-522.
74. Kent WJ: **BLAT - the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-2292R.
75. Gelfand MS, Mironov AA, Pevzner PA: **Gene recognition via spliced sequence alignment.** *Proc Natl Acad Sci USA* 1996, **93:**9061-9066.
76. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res* 2004, **14:**988-995.
77. Mott R: **EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA.** *Comput Appl Biosci:CABIOS* 1997, **13:**477-478.
78. Slater G, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinf* 2005, **6:**31.
79. Wu T, Watanabe C: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics* 2005, **21:**1859-1875.
80. Arumugam M, Wei C, Brown RH, Brent MR: **Pairagon+N-SCAN_EST: a model-based gene annotation pipeline.** *Genome Biol* 2006, **7(Suppl 1):**S5.1-S5.10.
81. Djebali S, Delaplace F, Crollius HR: **Exogean: a framework for annotating protein-coding genes in eukaryotic genomic DNA.** *Genome Biol* 2006, **7(Suppl 1):**S7.1-7.10.
82. Thierry-Mieg D, Thierry-Mieg J: **AceView: a comprehensive cDNA-supported gene and transcripts annotation.** *Genome Biol* 2006, **7(Suppl 1):**S12.11-12.14.
83. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H,

Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, *et al.*: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30:**38-41.

84. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC Known Genes**. *Bioinformatics* 2006, **22:**1036-1046.
85. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM: **Creating a honey bee consensus gene set.** *Genome Biol* 2007, **8:**R13.
86. Allen JE, Majoros WH, Pertea M, Salzberg SL: **JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions.** *Genome Biol* 2006, **7(Suppl 1):**S9.1-9.13.
87. Liu Q, Mackey AJ, Roos DS, Pereira FC: **Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction.** *Bioinformatics* 2008, **24:**597-605.
88. Howe K, Chothia T, Durbin R: **GAZE: a generic framework for the integration of gene-prediction data by dynamic programming.** *Genome Res* 2002, **12:**1418-1427.
89. Foissac S, Schiex T: **Integrating alternative splicing detection into gene prediction.** *BMC Bioinf* 2005, **6:**25-25.
90. Coghlan A, Durbin R: **Genomix: a method for combining gene-finders' predictions, which uses evolutionary conservation of sequence and intron-exon structure.** *Bioinformatics* 2007, **23:**1468-1475.