

# Phytozome: a comparative platform for green plant genomics

David M. Goodstein<sup>1,\*</sup>, Shengqiang Shu<sup>1</sup>, Russell Howson<sup>2</sup>, Rochak Neupane<sup>2</sup>, Richard D. Hayes<sup>1</sup>, Joni Fazo<sup>1</sup>, Therese Mitros<sup>2</sup>, William Dirks<sup>2</sup>, Uffe Hellsten<sup>1</sup>, Nicholas Putnam<sup>1</sup> and Daniel S. Rokhsar<sup>1,2</sup>

<sup>1</sup>U.S. Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598 and <sup>2</sup>The Center for Integrative Genomics, University of California, Berkeley, California 94720, USA

Received September 7, 2011; Accepted October 12, 2011

## ABSTRACT

The number of sequenced plant genomes and associated genomic resources is growing rapidly with the advent of both an increased focus on plant genomics from funding agencies, and the application of inexpensive next generation sequencing. To interact with this increasing body of data, we have developed Phytozome (<http://www.phytozome.net>), a comparative hub for plant genome and gene family data and analysis. Phytozome provides a view of the evolutionary history of every plant gene at the level of sequence, gene structure, gene family and genome organization, while at the same time providing access to the sequences and functional annotations of a growing number (currently 25) of complete plant genomes, including all the land plants and selected algae sequenced at the Joint Genome Institute, as well as selected species sequenced elsewhere. Through a comprehensive plant genome database and web portal, these data and analyses are available to the broader plant science research community, providing powerful comparative genomics tools that help to link model systems with other plants of economic and ecological importance.

## INTRODUCTION

Plant genome databases have grown up around different plant clades [e.g. TAIR for *Arabidopsis* (1), Gramene for grasses (2), SGN for Solanaceae (3), GDR for Rosaceae (4), LIS for legumes (5)]. This is in part due to the longstanding investment in plant genetic and physical

mapping resources and the focus of breeding programs in different agricultural crops. Comparative genomic databases that sample widely across the Viridiplantae [Phytozome, GreenPhylDB (6), Plaza (7), PlantGDB (8)] are a more recent development. These databases and associated web portals provide, at a minimum, a uniform set of tools and automated analyses across a wider range of plant genomes. In addition, those focused on green plant comparative genomics (GreenPhylDB, Plaza and Phytozome) provide putative gene families (groups of extant genes descended from a common ancestral gene) calculated at one or more speciation nodes in the plant tree of life, spanning most if not all hosted species, as well as additional gene-centric and genome-centric comparative tools. Their goal is to provide a platform for both genome-informed investigations of plant evolution, as well as a framework for transferring functional information from model plants to plants of agricultural, industrial and environmental importance.

Phytozome (<http://www.phytozome.net>), first released in 2008, provides a centralized hub that enables users with varying degrees of computational sophistication to access annotated plant gene families, to navigate the evolutionary history of gene families and individual genes, to examine plant genes in their genomic context, to assign putative function to uncharacterized user sequences and provides uniform access to plant genomics data sets consisting of complete genomes, gene and related (e.g. homologous) sequences and alignments, gene functional information and gene families, either in bulk or as the result of on-the-fly complex queries. The Phytozome web portal integrates a number of widely-used open source components [Lucene, GBrowse (9), Jalview (10), BioMart (11), mView (12) and pygr] with custom visualization code for gene family search, inspection and evaluation.

\*To whom correspondence should be addressed. Tel: +415 200 6110; Fax: +925 927 2552; E-mail: dmgoodstein@lbl.gov

## DATA SOURCES AND STANDARD ANALYSES

The v7.0 release of Phytozome contains data and analyses for 25 plant genomes, 18 of which were sequenced, assembled and partially or completely annotated at the JGI (Table 1). The gene-calling procedure for each JGI genome is described in detail in the associated genome publication, but a general overview of the JGI Plant Genome Annotation workflow is provided in [Supplementary Methods S1](#). For non-JGI genomes and annotations, assembled genome sequences and gene, transcript and peptide information is obtained in GFF or FASTA format, and subjected to consistency checking.

For non-JGI genomes, any gene symbols, database cross references, defines and experimentally-supported functional annotations (e.g. GO, EC) are also obtained. In the interests of uniformity of functional annotation, automatically generated functional annotations of non-JGI genomes are not retained. Protein-coding genes from both JGI and non-JGI genomes are then assigned PFAM domains (26), KEGG enzyme classification and KEGG Orthology assignment (27), KOG assignment (28) and Panther classification (29). Gene Ontology (GO) (30) assignments are made via pfam2GO mapping (31). All gene models and associated annotations are then loaded into Phytozome's MySQL database.

Same-species and near-species EST assemblies and Phytozome plant peptides are aligned against each genome. Each genome also undergoes whole genome alignment against a clade-informative subset of the other Phytozome genomes using the VISTA pipeline (32). Gene and alignment tracks, as well as VISTA-derived genome-wide pairwise DNA alignments are all accessible from Phytozome's GBrowse genome browser.

## GENE FAMILY CONSTRUCTION

Large scale, automated gene family construction is typically based on distance methods [Phytome (33), PlantTribes (34), InParanoid (35), OrthoMCL (36)] or, less frequently, distance-plus-character methods [OrthologID (37), TreeFam (38)], using a single peptide per locus in each genome under consideration. These distance-based methods can be broadly separated into two categories: those that implicitly (OrthoMCL) or explicitly (InParanoid) take into account the Mutual Best Hit (MBH) (39) relationship between putatively orthologous sequences and its role in setting a threshold for paralog accumulation ([Supplementary Methods S2](#)), and those that do not (Phytome, PlantTribes).

Distance-based methods have the advantage of being generally fast and scalable. Their main disadvantage lies in their reliance on a single score to characterize the evolutionary divergence of sequences, which becomes more problematic when considering species with an ancient divergence (in which case BLASTP scores tend to lose their resolving power, leading to the either the accumulation of unrelated, weakly aligning sequences into families at low significance thresholds, or the exclusion of distant but true homologs at higher significance thresholds).

Distance-plus-character-based methods use distance scores and a simple threshold to build an initial set of gene proto-families, all of whose members are more similar to each other than the threshold (e.g. OrthologID currently employs an *E*-value threshold of  $1e-20$ ). The members of each family are then included in a multiple sequence alignment (MSA), and phylogenetic trees are constructed based on discriminating residues (characters) in the MSA. The actual gene families

**Table 1.** The 25 completed plant genomes in version 7 of Phytozome

Organism	Common name	Version
<i>Aquilegia coerulea</i>	Colorado blue columbine	JGI v1.0
<i>Arabidopsis lyrata</i>	Lyre-leaved rock cress	JGI v1.0 (13)
<i>Arabidopsis thaliana</i>	Thale cress	TAIR v10 (1)
<i>Brachypodium distachyon</i>	Purple false brome	JGI /MIPS v1.0 (14)
<i>Carica papaya</i>	Papaya	ASGPB release of 2007 (15)
<i>Chlamydomonas reinhardtii</i>	Green alga	JGI assembly v4 with Augustus update 10.2 annotation (16)
<i>Citrus clementina</i>	Clementine	JGI v0.9
<i>Citrus sinensis</i>	Sweet orange	JGI/U Florida v1 assembly and v1.1 annotation
<i>Cucumis sativus</i>	Cucumber	Roche 454-XLR assembly and JGI v1.0 annotation
<i>Eucalyptus grandis</i>	Eucalyptus	JGI v1.0
<i>Glycine max</i>	Soybean	JGI Glyma1 assembly and Glyma 1.0 annotation (17)
<i>Manihot esculenta</i>	Cassava	JGI/Roche/U. Arizona v4 assembly and v4.1 annotation
<i>Medicago truncatula</i>	Barrel medic	Medicago Genome Sequence Consortium version Mt3.0
<i>Mimulus guttatus</i>	Monkey flower	JGI v1.0 release of strain IM62
<i>Oryza sativa</i>	Rice	MSU Release 6.0 (18)
<i>Physcomitrella patens</i>	Moss	JGI assembly v1.1 and COSMOSS annotation v1.6 (19)
<i>Populus trichocarpa</i>	Poplar	JGI assembly v2.0, annotation v2.2 (20)
<i>Prunus persica</i>	Peach	JGI v1.0
<i>Ricinus communis</i>	Castor bean	TIGR Release 0.1
<i>Selaginella moellendorffii</i>	Spikemoss	JGI v1.0 (21)
<i>Setaria italica</i>	Foxtail millet	JGI assembly v2.0, annotation version 2.1
<i>Sorghum bicolor</i>	Sweet sorghum	JGI v1.0 assembly, MIPS/PASA Sbi1.4 models (22)
<i>Vitis vinifera</i>	Grapevine	Genoscope March 2010 annotation on 12X assembly (23)
<i>Volvox carteri</i>	Volvox	JGI v1.0 (24)
<i>Zea mays</i>	Maize	Unfiltered protein coding models from Maizesequence.org release 5a.59 (25)

For published genomes, references are included in the version column.

correspond to the various monophyletic nodes found in the resulting trees. Phylogenetic methods are traditionally thought to be more accurate, especially when looking at anciently diverged species. However, recent work (40) on *Drosophila* and fungal species that span evolutionary distances comparable to the eudicots, has shown that a wide range of tree-building methods fail >50% of the time to produce the correct tree topology for even simple gene families, indicating the need for caution when making ortholog/paralog assignments based on gene and species tree reconciliation.

Whatever construction method is chosen, each gene family and its associated phylogenetic tree represents a hypothesis of the evolutionary history of a set of extant genes, presumed to be descendants of a single, unobservable ancestral gene. Descendants arise either via speciation (giving rise to orthologous descendants) or local or larger scale duplication events (giving rise to paralogous descendants). As orthologs are assumed to more likely share a common biological function, while paralogs are subject to both neo- and subfunctionalization (41,42), the high confidence identification of orthologs allows for the transfer of functional information from well-studied, tractable model systems (e.g. *Arabidopsis* and *Brachypodium*) to other economically or otherwise relevant plants.

Gene family construction in Phytozome uses a distance-based approach similar to the PhiGs method (43), the initial proto-family creation step used in TreeFam, with several modifications (Supplementary Methods S2). Family construction is restricted initially to a subset of core genomes, which are assumed to have relatively stable assemblies and complete structural annotations, though in some cases genomes with draft assemblies and annotations are used if the species in question is the sole representative of its clade (e.g. *Selaginella*, *Physcomitrella*, *Mimulus*). Using the assumed species tree, gene families are constructed at each evolutionary node, starting from the crown nodes [as in (44)] and moving backward in evolutionary time. At each bifurcating parent node, pairs of gene families from the two daughter nodes are combined into a parent family if they are joined by a cross-node MBH. Remaining families from the daughter nodes will be added to a parent family as paralogs if they have a hit to the parent that is stronger than the parent's best outgroup hit. This process is repeated down to the root node. MSAs from MUSCLE (45) and Hidden Markov Model (HMM) profiles from HMMER3 (46) are created for each core family. These profiles are used to 'pledge' peptides from non-core genomes into existing core families using HMMScan (46); they can also join core families if they are linked by a MBH. Non-core members can pledge to multiple families at a given node; thus the strict nesting of gene families is true for the core members only.

Figure 1 shows a typical gene family view, with the basis for each gene's membership in the family displayed in the leftmost column. A view of this family's evolutionary history (Figure 2) shows the hierarchical nesting of the core families.

The use of relatively strict significance and coverage thresholds, as well as an insistence on MBH relationships rather than simply strong similarity as the basis for seeding gene families, is intended to prevent merely similar gene families from coalescing at an inappropriate node in the tree. It also, however, biases Phytozome families towards underclustering. For this reason Phytozome includes a number of search and navigation tools, described below, to quickly bring together gene families that share overall sequence similarity or functional annotation.

## PHYTOZOME TOOLS AND VIEWS

### Text and sequence search

Genes and gene families can be retrieved from Phytozome by both keyword and sequence similarity searches. BLAST and BLAT searches of organism genomes, and BLAST searches of proteomes and gene family consensus sequences, can be used to find the genomic regions, gene transcripts, peptides and gene families most similar to a given query sequence. All gene and gene family attributes such as names, symbols, synonyms, external database identifiers, defines and functional annotation ids (e.g. PFAM00071, E.C. 1.1.1.95) are searchable, and gene families automatically inherit the attributes of their members, making it straightforward to retrieve a family of related but mostly uncurated genes as long as at least one family member is well annotated. Search can be restricted to gene families at a particular evolutionary node, and to families matching particular absence/presence phylogenetic profiles. One can also search the database of functional annotations (e.g. keywords from the descriptions of PFAM, GO, KEGG, KOG, Panther), which retrieves the set of all matching functional identifiers, and then automatically performs a second search for families marked as containing those functions.

All genes and gene families found via keyword or sequence similarity searches can be viewed individually, as described below, or first combined 'on the fly' to produce composite families, before being viewed and analyzed with the same tools used for individual families.

### Gene family and gene page views

The Gene Family view (Figure 1) provides the user with detailed information on each family and its constituent members, organized to highlight shared attributes. The default 'Genes in this family' tab displays individual family members, grouped by species and includes each member's source identifier (hyperlinked to the appropriate source database), aliases, synonyms and gene symbols, defines (where available) and a graphical view of each member's local syntenic environment. A provisional family name is provided, as well as a membership 'fingerprint' (member count for all species present at this node), and family-level KOG and KEGG-Orthology classification. The syntenic display can be replaced by a PFAM domain or gene structure (exon/intron) display. For each family member, links are provided to both a GBrowse

**Hypothetical Core Eudicot gene**  
Cluster 28726626. 17 members:

Mes Rco Ptr Mtr Gma Csa Ppe Ath Aly Cpa Csi Ccl Egr Vvi Mgu

1 1 1 1 2 1 1 1 1 1 1 1 2 1 1

Classification Find related families Align family members Get Data Display options

**KOG Class:**  
CELLULAR PROCESSES AND SIGNALING [ O ] : Posttranslational modification, protein turnover, chaperones

**KEGG Class(es):**  
000X1 : Metabolism  
000Y7 : Glycan Biosynthesis and Metabolism  
00563 : Glycosylphosphatidylinositol(GPI)-anchor biosynthesis  
K05289 : glycosylphosphatidylinositol transamidase

Genes in this family Functional Annotation Multiple Sequence Alignment Family History

	ORG	DBXREF	SYMBOL	DEFINE	DOMAINS	SYNTENY	EXONS
F	Mes	cassava4.1_002548m					
M	Rco	29917.m002001		conserved hypothetical protein			
F	Ptr	POPTR_0001s35320.1	gw1.l.2574.1	GPI transamidase component fam...			
F	Mtr	Medtr4g143270.1		Gaa1-like, GPI transamidase co...			
F	Gma	Glyma05g08990.1					
F	Gma	Glyma19g00420.1					
F	Csa	Cucsa.011280.1					
F	Ppe	ppa002024m					
F	Ath	AT5G19130.1		GPI transamidase component fam...			
F	Aly	488797	fgenes2_kg.6_19...				
F	Cpa	evm.model.supercontig_1...					
F	Csi	orange1.1g005386m					
F	Ccl	clementine0.9_003852m					
PI	Egr	Eucgr.J02243.1	Egrandis_v1_0.008...				
F	Egr	Eucgr.L01920.1	Egrandis_v1_0.004...				
F	Vvi	GSVIVT01032407001					
F	Mgu	mgv1a002326m					

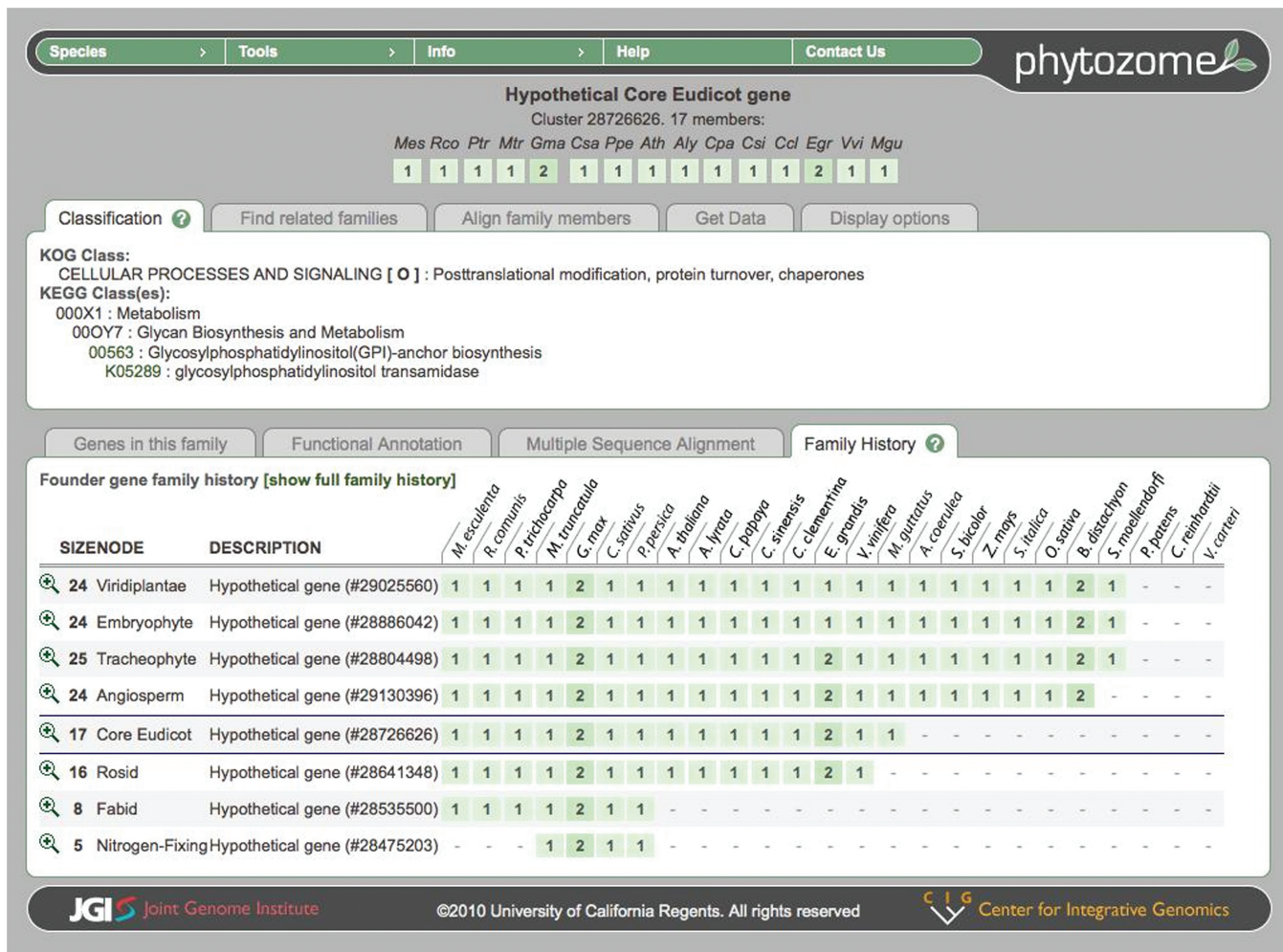
JGI Joint Genome Institute ©2010 University of California Regents. All rights reserved C I G Center for Integrative Genomics

**Figure 1.** Default view of the Gene Family page for a 17 member core eudicot family. Members are listed according to their order in the tree on the Phytozome home page. The membership class of each gene is indicated in the leftmost column (Supplementary Methods S2). For each member, Gene Page and GBrowse links are provided, as well as links to external databases (if these exist), aliases, symbols and defines. The syntenic view in the right column shows the five upstream and five downstream neighbors of each family member (who are rendered as gray icons in the middle of each syntenic row). Each syntenic segment is oriented to render family members in the same orientation (consistent with their presumed descent from a common ancestor). Gene icons sharing the same (non-white) color are all members of the same gene family at this node; this can provide syntenic support for the hypothesis of a common ancestor for family members.

view (Figure 3) of each gene in its genomic context, and a 'Gene Page' (Figure 4).

The family page is divided into a set of lower and upper tabs, roughly corresponding to 'information' and 'actions', respectively. The lower row helps users explore the consistency and evolutionary history of the family. The 'Functional Annotation' tab lists all the functional and domain annotations (e.g. PFAM, Panther, GO, KEGG, KEGG Orthology) assigned to family members, broken down by organism. Functional annotations present in all family members are highlighted. The 'MSA'

tab displays a pre-computed MUSCLE peptide alignment of all family members, which is downloadable. The family's evolutionary history can be viewed in the 'Family History' tab, where all families that are parents of, or derived from, the current family are listed. From the upper row of tabs, 'Find related families' provides a number of methods for identifying families similar to the current one: by family consensus sequence similarity, by shared functional annotation, or by shared gene membership. This is quite useful when looking for related subfamilies, or verifying that a particular combination of



**Figure 2.** Family History view of the gene family in Figure 1. All the descendants and ancestors of this core eudicot family (which is highlighted) are visible in the history view. The strict nesting of families is observed, though one needs to remember that one of the *Eucalyptus* genes in this core eudicot family is an incomplete pledge (Supplementary Methods S2), and is not present in the deeper Embryophyte and Viridiplantae ancestors.

domains is unique to a given family. ‘Align family members’ forwards family member coding or peptide sequences directly to the Jalview tool, where MSAs can be created and edited, and subsequently used to construct phylogenetic trees. ‘Get Data’ provides access to the BioMart data query tool for this family, while the family page display can be customized on the ‘Display options’ tab.

The Gene Page (Figure 4), in addition to showing single gene functional annotations and evolutionary history, includes links to alternatively spliced transcripts (if they exist), a simplified view of the gene in its genomic context (showing alternatively spliced transcripts and peptide homology tracks), direct access to genomic, transcript, coding and peptide sequences associated with this gene locus (color-coded to indicate exon/intron and UTR boundaries), and a graphical view of all other Phytozome peptides aligned [via dual affine Smith-Waterman (47)] against this gene’s peptide.

Genome-centric views are provided by GBrowse (Figure 3) for all 25 genomes currently included in Phytozome. The browsers can be accessed directly from

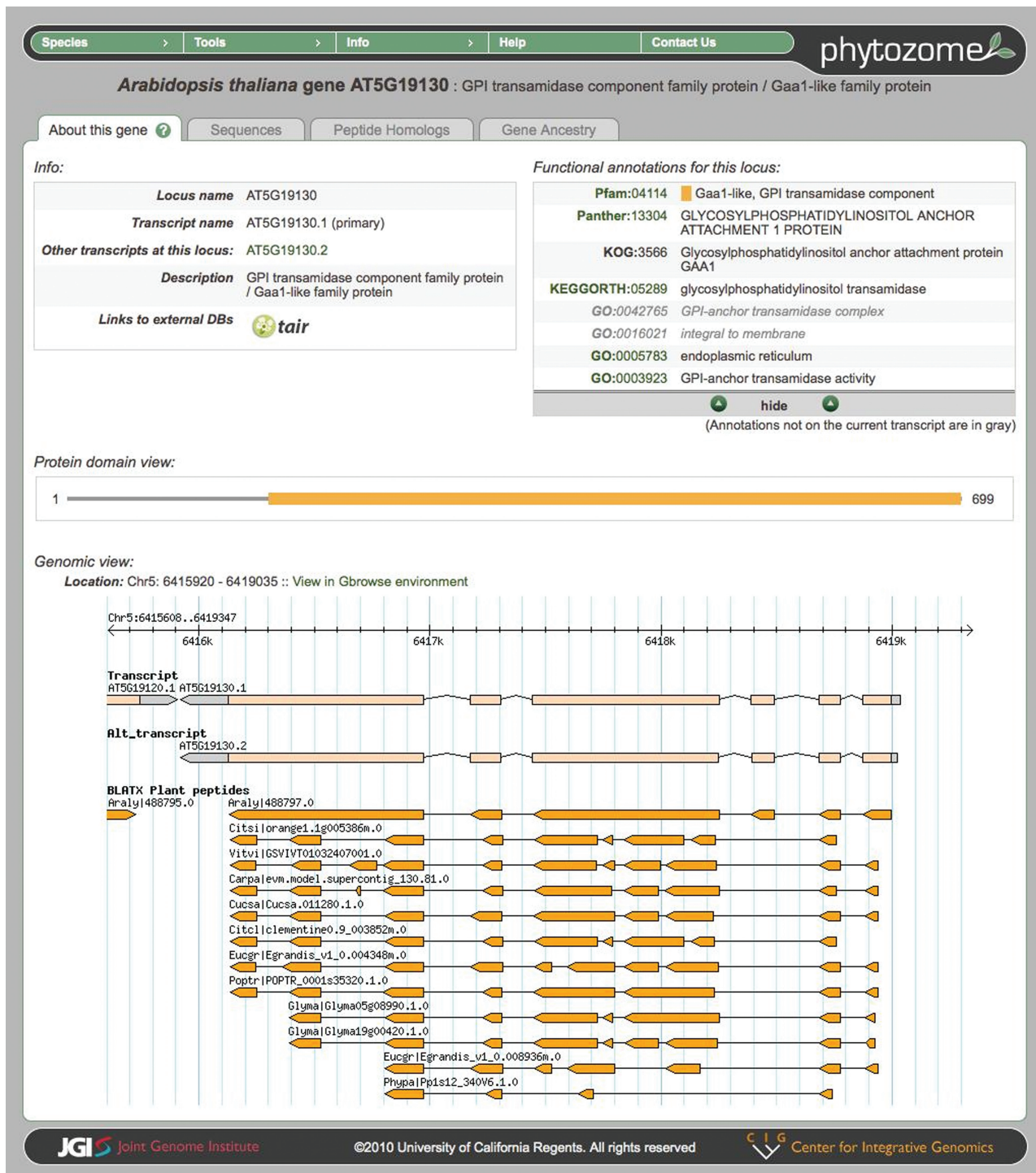
the Phytozome home page, from individual member gene links on the Gene Family or Gene page, and from the BLAST/BLAT results page for searches performed against one of the genome target databases. In the latter two cases, a zoomed-in view of the genomic region containing the selected gene (or BLAST hit) is displayed. Each browser typically displays a gene prediction track (primary and alternatively spliced transcripts), a track of homologous peptides from related species aligned against the genome, supporting EST (or EST assemblies) and one or more VISTA tracks identifying regions of this genome that are syntenic with other plant genomes included in Phytozome. All gene features are hyperlinked to their respective Gene Page, while the VISTA tracks are linked to the corresponding genomic regions in the VISTA browser.

**DATA ACCESS**

For each genome hosted at Phytozome, bulk data files are available that contain genome assembly sequence, gene structure GFF3, transcript, coding and peptide sequence



**Figure 3.** GBrowse view of the local genomic context of the poplar gene from the family in Figure 1. Primary and alternative transcripts (if present), assembled EST data and related plant peptides are shown aligned against the genome. Not shown are tracks of repetitive regions, GC content and the alignment of ESTs from related species. Interspecies whole genome alignments, displayed in the VISTA tracks, reveal the tendency towards strong genomic sequence conservation in coding regions (which are under selective pressure), which weakens as one considers more distantly related species (e.g. rice-poplar versus the more closely related eucalyptus-poplar VISTA alignments). Displayed gene models are hyperlinked to their respective gene pages.



**Figure 4.** Default view of the Gene Page for the *Arabidopsis thaliana* gene in the family of Figure 1, showing primary transcript info, functional annotations and simplified genomic context. This locus has an alternative transcript (which appears to differ primarily in its 5'-UTR). Note the strong splicing support provided by the BLATX aligned *Arabidopsis lyrata* peptide (which in actuality is also a member of this family).

in FASTA format and general annotation information (PFAM, Panther, KOG, KEGG, best rice and *Arabidopsis* homologs). For JGI genomes, we also provide repeat-masked genome assemblies, as well as supporting annotation data (e.g. the PASA EST assemblies used in gene calling).

Customized data sets consisting of gene or gene family sequences and annotations can be constructed using Phytozome's implementation of BioMart, where users can choose detailed data filters, attributes and output formats. BioMart can be accessed from the 'Get Data' tab on a gene family page (in which case the data is, by

default, initially restricted to that gene family), or directly from the Phytozome menu. It is also available at the BioMart central portal, <http://www.biomart.org>.

## PHYTOZOME SOFTWARE IMPLEMENTATION

We have made extensive reuse of available databases, software tools and data formats in our implementation of Phytozome. The Phytozome website is built on a LAMPJ stack (Linux, Apache, MySQL, php/Perl and Java). Open source visualization components of Phytozome include: Gbrowse (9), the Generic Genome Browser, from the GMOD project, for the visualization of features in their genomic context; Jalview (10), a multiple alignment viewer and editor, for the creation, detailed inspection and modification of MSAs and phylogenetic trees; BioMart (11), to enable query-based downloads of bulk data on gene families and genome annotations; BioPerl (48), for the parsing and formatting of genomic data and BLAST results; and mView (12), for the visualization of MSAs. The search system is based on the Lucene search engine (<http://lucene.apache.org/>).

## FUTURE PLANS

Phytozome content will continue to be updated at least annually, with new and updated genomes typically added in January and new feature sets released quarterly. Current plans for the January 2012 (v8) release include updates to poplar, soybean, brachypodium, maize and medicago, the first-time inclusion of the JGI genomes phaseolus (common bean) and *Capsella rubella* (an *Arabidopsis* comparator), and the externally contributed apple (49), strawberry (50) and potato (51) genomes. Version 8 is also expected to include genomic variation data (SNPs and structural variants) from the JGI and elsewhere, and expression data associated with the JGI Gene Atlas projects. Phytozome is also in the final stages of licensing for distribution to end users. We expect that the entire database and software infrastructure will be available for download by the end of 2011.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary methods S1–S2, Supplementary references [52–57].

## FUNDING

This work was performed at the U.S. Department of Energy Joint Genome Institute, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Additional funding was provided by the Gordon and Betty Moore Foundation. Funding for open access charge: Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
2. Liang,C., Jaiswal,P., Hebbard,C., Avraham,S., Buckler,E.S., Casstevens,T., Hurwitz,B., McCouch,S., Ni,J., Pujar,A. *et al.* (2008) Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res.*, **36**, D947–D953.
3. Bombarely,A., Menda,N., Tecle,I.Y., Buels,R.M., Strickler,S., Fischer-York,T., Pujar,A., Leto,J., Gosselin,J. and Mueller,L.A. (2011) The sol genomics network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Res.*, **39**, D1149–D1155.
4. Jung,S., Staton,M., Lee,T., Blenda,A., Svancara,R., Abbott,A. and Main,D. (2008) GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res.*, **36**, D1034–D1040.
5. Gonzales,M.D., Archuleta,E., Farmer,A., Gajendran,K., Grant,D., Shoemaker,R., Beavis,W.D. and Waugh,M.E. (2005) The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res.*, **33**, D660–D665.
6. Conte,M.G., Gaillard,S., Lanau,N., Rouard,M. and Perin,C. (2008) GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Res.*, **36**, D991–D998.
7. Proost,S., Van Bel,M., Sterck,L., Billiau,K., Van Parys,T., Van de Peer,Y. and Vandepoele,K. (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, **21**, 3718–3731.
8. Duvick,J., Fu,A., Muppirala,U., Sabharwal,M., Wilkerson,M.D., Lawrence,C.J., Lushbough,C. and Brendel,V. (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.*, **36**, D959–D965.
9. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
10. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
11. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart—biological queries made easy. *BMC Genomics*, **10**, 22.
12. Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.
13. Hu,T.T., Pattyn,P., Bakker,E.G., Cao,J., Cheng,J.F., Clark,R.M., Fahlgren,N., Fawcett,J.A., Grimwood,J., Gundlach,H. *et al.* (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.*, **43**, 476–481.
14. Initiative,I.B. (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**, 763–768.
15. Ming,R., Hou,S., Feng,Y., Yu,Q., Dionne-Laporte,A., Saw,J.H., Senin,P., Wang,W., Ly,B.V., Lewis,K.L. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, **452**, 991–996.
16. Merchant,S.S., Prochnik,S.E., Vallon,O., Harris,E.H., Karpowicz,S.J., Witman,G.B., Terry,A., Salamov,A., Fritz-Laylin,L.K., Marechal-Drouard,L. *et al.* (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–250.
17. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
18. Ouyang,S., Zhu,W., Hamilton,J., Lin,H., Campbell,M., Childs,K., Thibaud-Nissen,F., Malek,R.L., Lee,Y., Zheng,L. *et al.* (2007) The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.
19. Rensing,S.A., Lang,D., Zimmer,A.D., Terry,A., Salamov,A., Shapiro,H., Nishiyama,T., Perroud,P.F., Lindquist,E.A.,



- Kamisugi, Y. *et al.* (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.
20. Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A. *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.
21. Banks, J.A., Nishiyama, T., Hasebe, M., Bowman, J.L., Gribskov, M., dePamphilis, C., Albert, V.A., Aono, N., Aoyama, T., Ambrose, B.A. *et al.* (2011) The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science*, **332**, 960–963.
22. Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.
23. Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.
24. Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., Miller, S.M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L.K. *et al.* (2010) Genomic analysis of organismal complexity in the multicellular green alga *Volvox carterii*. *Science*, **329**, 223–226.
25. Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
26. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
27. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
28. Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S. *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.*, **5**, R7.
29. Mi, H., Guo, N., Kejariwal, A. and Thomas, P.D. (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res.*, **35**, D247–D252.
30. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
31. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
32. Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. and Dubchak, I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.
33. Hartmann, S., Lu, D., Phillips, J. and Vision, T.J. (2006) Phytome: a platform for plant comparative genomics. *Nucleic Acids Res.*, **34**, D724–D730.
34. Wall, P.K., Leebens-Mack, J., Muller, K.F., Field, D., Altman, N.S. and dePamphilis, C.W. (2008) PlantTribes: a gene and gene family resource for comparative genomics in plants. *Nucleic Acids Res.*, **36**, D970–D976.
35. Remm, M., Storm, C.E. and Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
36. Li, L., Stoeckert, C.J. Jr and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
37. Chiu, J.C., Lee, E.K., Egan, M.G., Sarkar, I.N., Coruzzi, G.M. and DeSalle, R. (2006) OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics*, **22**, 699–707.
38. Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L.J., Guo, Y., Heriche, J.K., Hu, Y., Kristiansen, K., Li, R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.
39. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
40. Rasmussen, M.D. and Kellis, M. (2007) Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res.*, **17**, 1932–1942.
41. Lynch, M. (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl Acad. Sci. USA*, **104**(Suppl. 1), 8597–8604.
42. Ohno, S., Wolf, U. and Atkin, N.B. (1968) Evolution from fish to mammals by gene duplication. *Hereditas*, **59**, 169–187.
43. Dehal, P.S. and Boore, J.L. (2006) A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics*, **7**, 201.
44. Butler, G., Rasmussen, M.D., Lin, M.F., Santos, M.A., Sakthikumar, S., Munro, C.A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J.L. *et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, **459**, 657–662.
45. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
46. Eddy, S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.
47. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
48. Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigan, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
49. Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggo, M., Pruss, D. *et al.* (2010) The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nat. Genet.*, **42**, 833–839.
50. Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S.P. *et al.* (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.*, **43**, 109–116.
51. Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J. *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.
52. Smit, A.H., R; Green, P. (1996–2010).
53. Smit, A.H., R; Green, P. (2008–2010).
54. Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K. Jr, Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D. *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.
55. Slater, G.S. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
56. 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.
57. Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.*, **10**, 516–522.