This document provides an overview of features and capabilities of version 5 of Phytozome, the Joint Genome Institute's Plant Genomics portal. Before getting started, please make sure you have the following:

1. The Firefox browser installed on your laptop (http://www.mozilla.com/en-US/firefox/personal.html). Phytozome features and appearance are optimizedfor Firefox.
2. The Java plug-in for Firefox (https://addons.mozilla.org/en-US/firefox/browse/type:7). This is required to use the Jalview analysis tool.
3. Cookies and javascript enabled.

And here are a few quick pointers:

1. The latest version of Phytozome is always available at http://www.phytozome.net
2. If you ever get "lost" navigating the system, you can always return to the home page by clicking on the "Phytozome" banner at the top of every page
3. All (or almost all) text in green is a hyperlink, so can be clicked on to perform an action or go to a page with more information.
4. Green question marks (❓) are linked to context-appropriate help sections.

So, let's get started!

Overview

The Phytozome website and database provides you with access to gene families and individual genes for 18 plant genomes (in version 5). Sequence similarity (BLAST/BLAT) and keyword-based searching can help you pinpoint genes and gene families of interest. You can navigate the evolutionary history of each gene family, and identify closely related families (via shared functional annotation or consensus sequence similarity). The site also provides direct and uniform access to much of the information associated with the individual genomes (e.g, annotations, deflines, functional assignments, sequences). A uniform browsing environment (using GMOD's Gbrowse browser) is also provided for each genome in Phytozome, which includes gene calls, supporting homologies and EST alignments, and VISTA tracks showing the degree of nucleotide conservation between related plant genomes. Finally, all data on gene families and individual genomes in the Phytozome database are downloadable via BioMart, a popular open source data retrieval tool.

Let's take a look at the home page, at http://www.phytozome.net (Figure 1). Here we see the eighteen Phytozome genomes arranged phylogenetically (though the branch lengths are drawn purely by aesthetic criteria; they do not represent evolutionary distance!). Organisms can be selected for
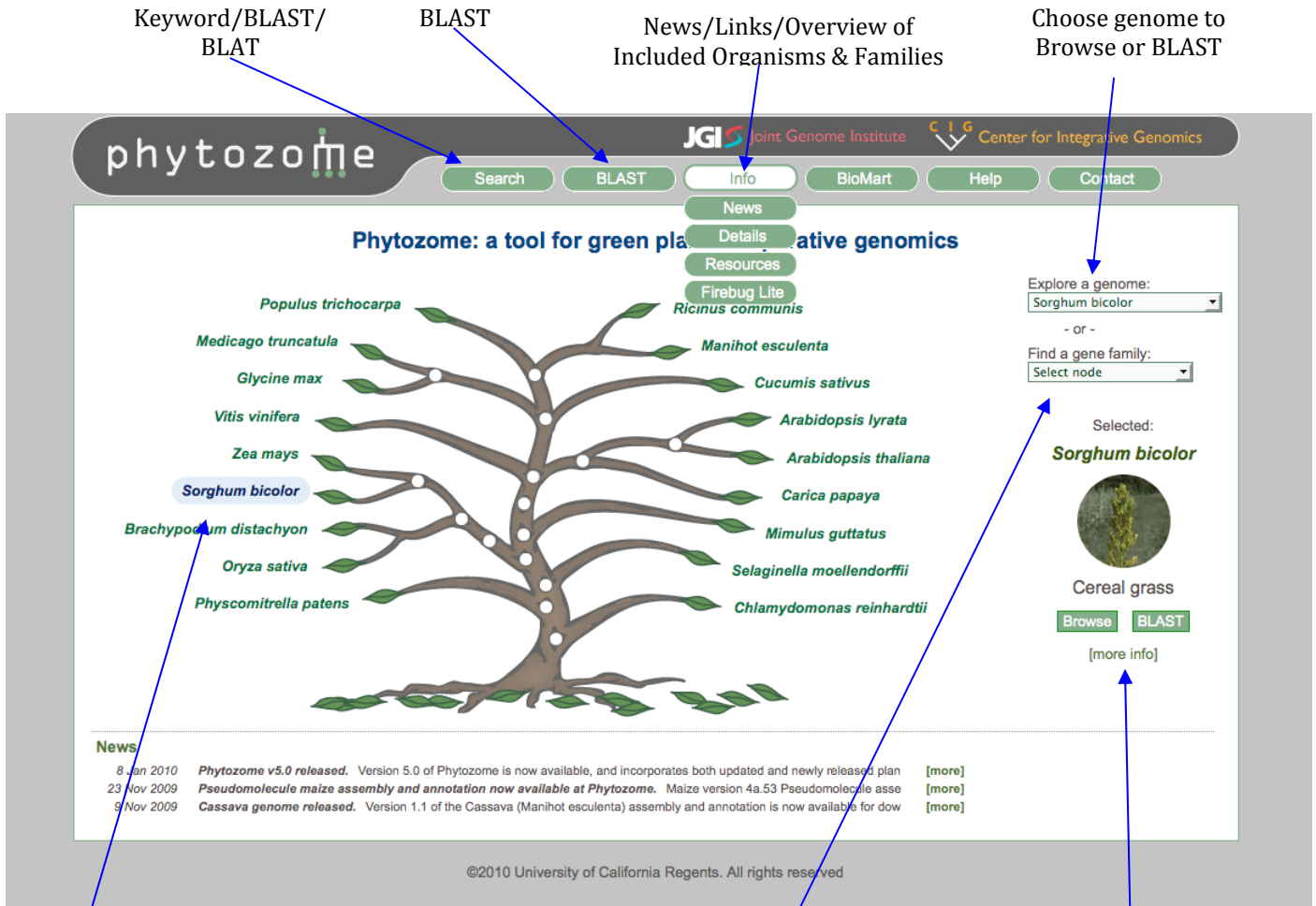
Keyword/BLAST/BLAT

BLAST

News/Links/Overview of Included Organisms & Families

Choose genome to Browse or BLAST

Sorghum is selected

Choose a node to search for families via keyword/BLAST

Go to Sorghum information page

Figure 1. Phytozome home page with Sorghum genome selected.

browsing or BLAST searching either by clicking on a crown node (the organism names next to the tree leaves) or using the "Explore a genome" pulldown. This will bring up a thumbnail image of the plant selected, as well as buttons to launch Gbrowse or go to the BLAST interface for that genome. The "more info" link below these buttons will take you directly to the selected organism's information page (Figure 3).

If instead, you're interested in gene families, you can select a node directly on the tree (click
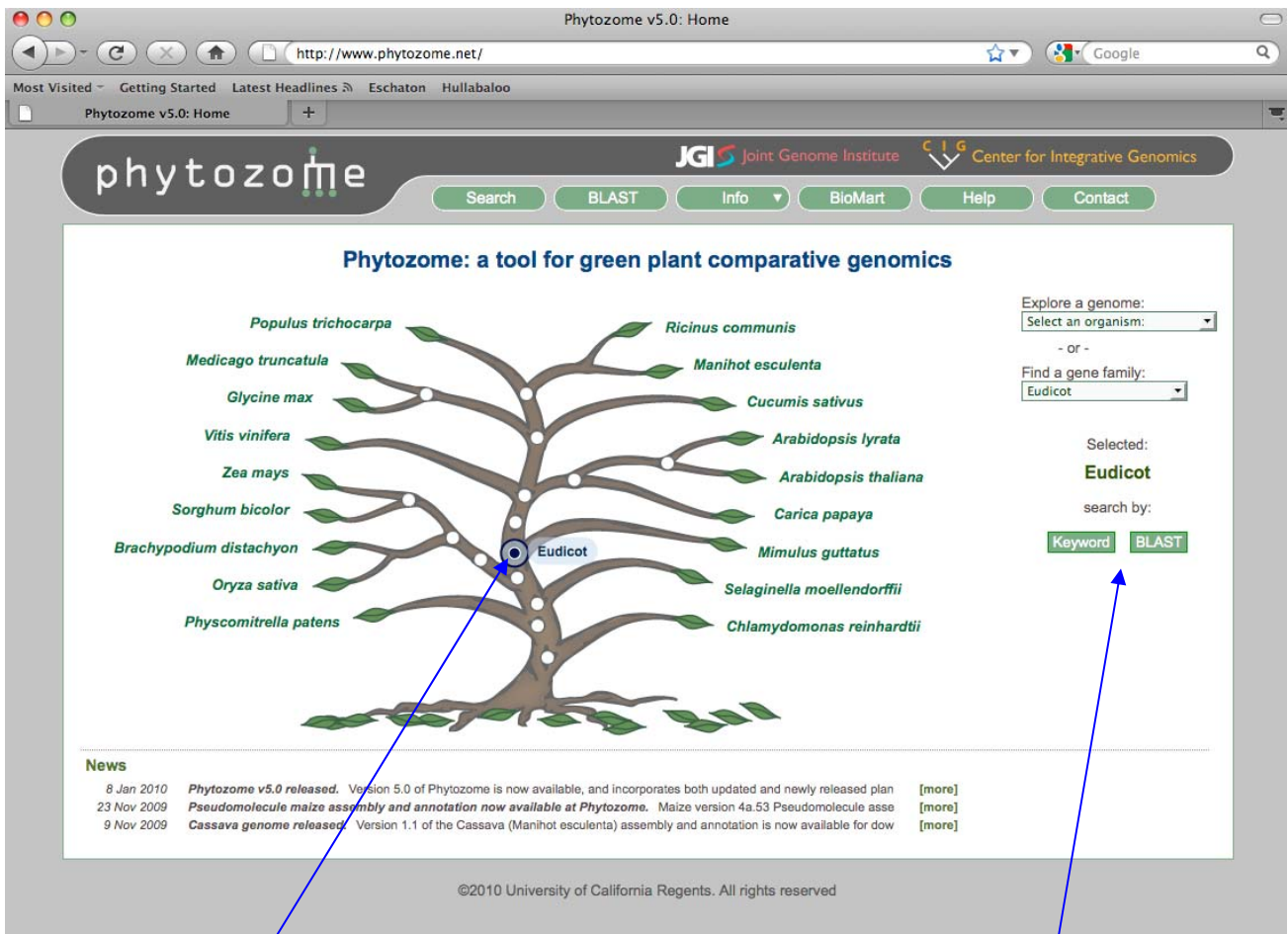
Figure 2. Phytozome 5 home page with the Eudicot node selected.

Eudicot node is selected

Find Eudicot gene families by keyword or BLAST

on one of the open white circles) or use the "Find a gene family" pulldown (Figure 2). By selecting a node, you have access to gene families that contain genes from all the species descended from that node. For example, eudicot gene families can have genes from mimulus, papaya, both arabidopsis species, grape, soybean, medicago, poplar, castor bean, cassava, and cucumber). There's no equivalent to a genomic browser for gene families, so the choices are instead to find families via keyword search, or to BLAST against a database of eudicot gene family consensus sequences..

Returning momentarily to Figure 1, click on the "[more info]" link, and load the Sorghum information page (Figure 3). Organism information pages provide convenient links to

FTP site for Sorghum data

Click to go to a different organism page

Figure 3. Sorghum information page.

browsing, BLASTing, and bulk data download (via the Phytozome FTP site). For organisms that were sequenced and/or annotated at JGI, there's also a summary of the genome sequencing project and more detailed information about the assembly and annotation. To the right is a simple horizontal cladogram of the Phytozome organisms. Click on any one of the names to be taken to that organism's information page.

Let's now turn to using the search system to find gene families of interest.

Keyword Search Interface

Figure 4. Keyword search interface. In this example, the user is searching for eudicot gene families whose members are annotated as "celluose*", with no restrictions placed on family membership (Restrict Organisms is checked just to show the included orgs. Each is set to "Allow").

Keyword search finds gene families by searching for member genes that match the search terms. You can search using gene symbols (e.g., *AHP6*), identifiers (e.g, *AT3G46060*), words or phrases that might be part of a defline (e.g., *cellulose*), or using any of the ontology ids that are automatically assigned by Phytozome (e.g., *go:7264*, *pfam:PF08477*, *panther:PTHR11708*, etc.). By default, the following types of data are searched:

- Definition lines (deflines)
- symbols, and aliases (which we call "synonyms")
- transcript, peptide, and locus names
- external database identifiers: SwissProt/Tremlb, Unigene, EntrezGene, HGNC
- Ontology ids: GeneOntology, PFAM, Enzyme Classification, KOG, Panther, KEGG Orthology

You can also explicitly specify which fields to search using the convention "fieldname:value" in your search term (e.g, "synonym:ahp6"), which will only search against symbols and aliases for a match. Fieldnames recognized by Phytozome are:

- **General**: defline transcriptname locusname peptidename synonym
- **External identifiers**: sptremblid kogclassdesc keggclassdesc entrezgeneid hgncid unigeneid
- **Ontologies**: goid pfamid ec kogid cogid pantherid keggorth
- **Classifiers**: kogclassdesc keggclssdesc clusterkogletter clusterkeggorth
- **Phytozome internal identifiers** (you probably won't use this directly): transcriptid primary_transcriptid clustered proteomes

Note that you can also specify the species membership of the gene families you'd like to find. Select "Require" ("Exclude") for any species you want (don't want) present in each family returned.

Another type of keyword search is based on ontology descriptions (see Figure 5). Here the search is done in two steps. All the ontologies whose descriptions match your keywords are first found, and then those ontology ids are used to search the gene families. This is useful when you suspect the gene or gene family of interest is poorly curated , so is unlikely to have biologically meaningful terms in its defline.
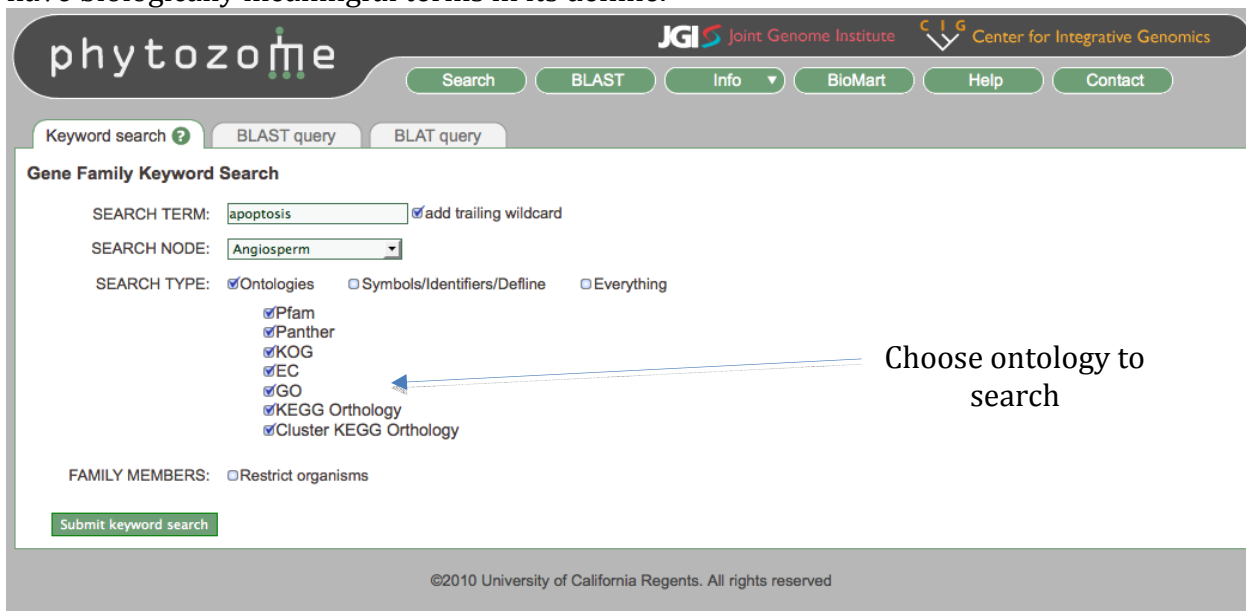

Figure 5. Ontology search interface

By searching for the ontology annotations corresponding to your keywords, you can take advantage of the fact that all genes and gene families in Phytozome are automatically assigned PFAM, Panther, KEGG-EC, KEGG-Orthology, KOG, and GO (via pfam2go) ids, which effectively annotates them with the corresponding ontology descriptions.

A successful ontology search will also show you the ontology annotations that were found using your keywords (Figure 6). Those are then used to find corresponding gene families.

Figure 6. Search of all ontologies using term "beta-amylase". Three matching annotations were found, which are then automatically used to find matching gene families (the set of found families is not shown in this image).

Some last pointers on search syntax:

- Search is case-insensitive
- By default, all terms are appended with a wildcard
- You can't prepend a wildcard
- Terms separated by whitespace are treated as an "OR" search (find something matching this term OR this term OR this term….). Place a "+" in front of terms to do an AND search (e.g., *+cell +wall* will only find gene families in which the words cell AND wall both appear). Prepend with "-" to only find gene families that DON'T match the term (but you have to have at least one term they do match…pure NOT searches aren't supported).
- Use parentheses to group search terms, e.g, *+(cell membrane) +wall* will find gene families that match the terms wall AND match either "cell" OR "membrane"
- If you want to search for a phrase rather than a word (e.g, *catalytic component*), enclose the phrase in quotes (*"catalytic component"*). This will only return families where the complete phrase appears

Search Results Page

A successful search will return one, or possibly thousands of gene families (try searching for *kinase* or *repeat* !), so the results page presents an overview of the gene families that have been found (Figure 7). Results are ordered in decreasing family size. Each row shows the family size, which node it was found at (if you performed an "all nodes" search; otherwise the node is simply the one shown in the "Search Node" field up above), an automatically assigned defline for the family (or "Hypothetical gene (#XXX) if no name could be assigned), and a "fingerprint" showing the number of members from each species present in the family. These are color-coded to aid the eye: darker blocks correspond to larger member counts. If you're not particularly interested in the family membership of one or more species, click the "Display" tab up top and uncheck that species…it won't appear in the fingerprint.

At this point you can revise your search (typically trying to make it more specific) by clicking the "revise your search" link next to SEARCH CRITERIA on the Summary tab. You can also click on any of the rows and be taken to that family's Gene Family page. Before we consider the Gene Family page, there are a few other features of the Search Results page worth knowing about.

Figure 7. Search results page for a standard symbols/identifiers/deflines search of eurosid I families using "degradation" as the search term.

If you select one or more Gene families by clicking their checkboxes, the "Analyze Results / Get Data" tab will open (Figure 8). At this point, depending which of the options show that you choose, you can

- Send the members of the selected families to Jalview, for alignment, and phylogenetic analysis
- Open the data retrieval tool BioMart, with your selected families already chosen for data download.
- Create a Composite Gene Family from the selected families, which can viewed in a Gene Family pageFigure 4. Analyze Resuts / Get Data tab on Search Results page. The Align sequences in JalView option has been selected, allowing the user to send off family member sequences (peptide or CDS), or cluster (family) consensus sequences, to the Jalview analysis tool.

8

Figure 8. Analyze Results / Get Data tab on Search Results page. The Align sequences in JalView option has been selected, allowing the user to send family member sequences (peptide or CDS), or cluster (family) consensus sequences, to the Jalview analysis tool

But at this point, let's look at a gene family in detail. Click on the 2nd family, Hypothetical gene (#22238032), to open its Gene Family page.

Gene Family Page

Figure 9. Default Gene Family view of a 13 member Eurosid I family.

The Gene Family page (Figure 9) contains detailed information on each family member, as well as various views and tools to help assess the family as a whole. Different views are organized in two rows of tabs, with the lower rows corresponding to different "information" views, and the upper row corresponding (with the exception of "Classification") to different "actions." The default is to show the family Classification and the "Genes in this family" tab. The classification shows any KOG or KEGG-Orthology annotations that are assigned to more than 50% of family members.

Each row in this tab corresponds to a single member gene. The gene's organism, symbol(s) and defline are shown (longer deflines can be viewed by clicking the "+" icon for that row,
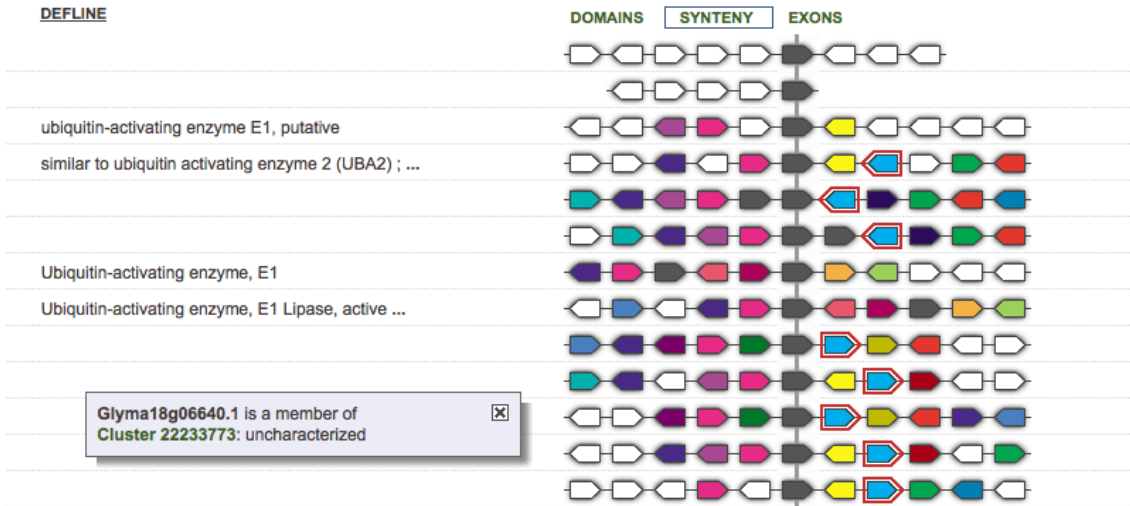
Figure 10. Synteny view, showing the 5 upstream and 5 downstream neighboring genes of each family member. All family member genes and shown in the same orientation. Genes that are rendered in the same color are members of the same family at this node. Clicking on a gene icon pops up the rectangle shown in the lower left, which shows the gene name and its family membership.

or the upper "+" icon to expand all rows). Buttons are provided that open the appropriate genome browser centered on this gene ("Genomic View"), or a more detailed Gene Page (see Figure 19). The DbXref column displays hyperlinked identifiers in other databases (e.g, TAIR, MSU Rice project, MaizeSequence.org) corresponding to the gene in question.

By default, the graphical display to the right (Figure 10) shows a schematic of the genic environment of each member gene. This view provides a quick check of the local syntenic support for the gene family. Five upstream and 5 downstream neighbor genes are shown (if less than 5 are shown, this indicates a chromosome or scaffold end). Members of the currently family always appear in the center, colored charcoal. Other genes that share a family are rendered in a common color. White is reserved for those genes that don't have another family member appearing in the synteny view. Mousing over any gene brings up an info box showing the gene's transcript name, and the identifier and defline of the its family (the identifier is hyperlinked to that family's page).

Clicking on the "Exons" heading switches the display to a cartoon of the gene structure of the family members (Figure 11). This allows for a quick check of whether exon count and intron/exon size are consistent for this family.

Clicking on the "Domains" heading (Figure 12) brings up a view of the PFAM domains called on member peptides. The peptide lengths are scaled to appear the same in the graphic (actual peptide length is indicated to the right of each row). Each domain is assigned a unique color; mousing over the domain reveals its identifier, description, and locations within the peptide.
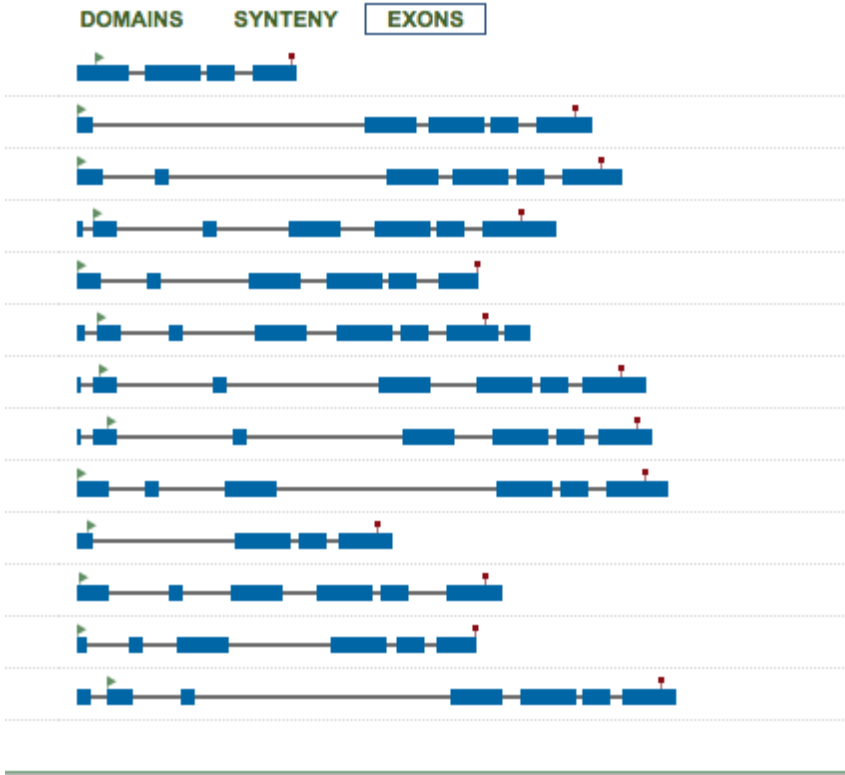
Figure 11. Exon/Intron structure of family members. Translation starts and stops are indicated with blue flags and red blocks, respectively.
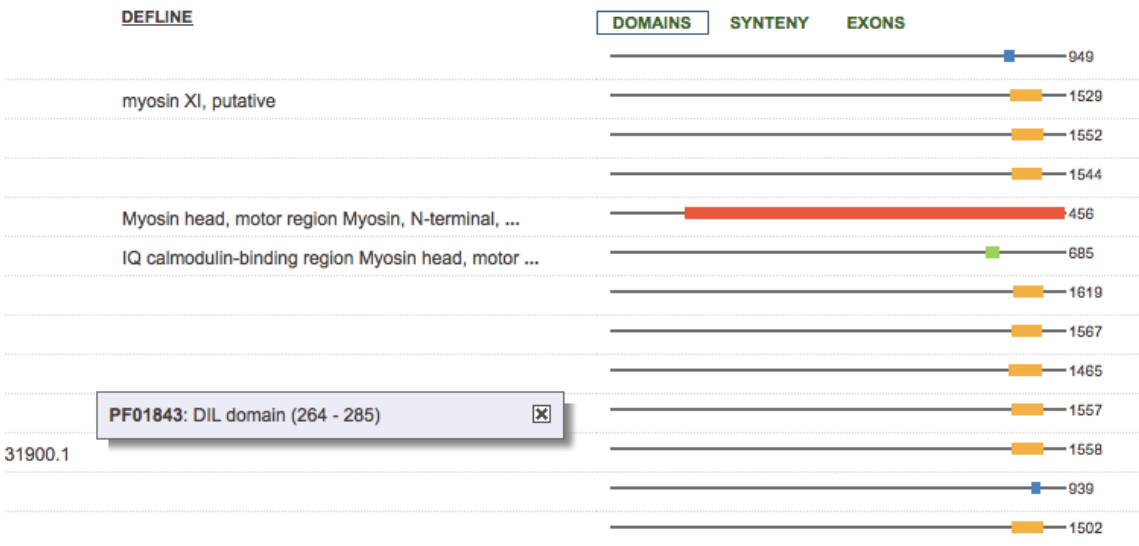

Figure 12. Domain view of a gene family (not the same family as shown above).

Switching to the Functional Annotation tab (Figure 13) shows all the domain and functional annotations assigned to members of this family, broken down by species. This provides a quick summary of which annotations are shared by all family members. Identifiers that can be looked up in external databases (e.g, GO at EBI, PFAM at Sanger) are hyperlinked.

Clicking any of the checkboxes next to the ontology identifiers will open up the Find related families tab in the upper row. The two selected annotations appear now appear under the heading "Find EurosidI clusters with selected annotation(s)". Clicking on that heading will find gene families that contain both these annotations. This capability supports echking whether a given families particular combination of annotations is unique at that node.

The Multiple Sequence Alignment (MSA) tab shows the precomputed MUSCLE alignment of the family member peptides. mView (http://www.ebi.ac.uk/Tools/mview/FAQ.html) color-coding is used (Figure 14). You can directly download an HTML file to insert this multiple sequence alignment into a web page or figure. Downloading the MSA in other formats (CLUSTALW, FASTA, etc.) is best done thru Jalview, BioMart of the Get Data tab.

Figure 13. Functional Annotation tab of Gene Family page, showing the distribution across species of the domain and functional annotations assigned to members of this family. One Panther and one KOG annotation have been selected, which activates the "Find related families" tab. Clicking on "Find EurosidI clusters with selected annotations" will search for other EurosidI families containing both these annotations.

Figure 14. Multiple Sequence Alignment of this family's member peptides.

The last tab in the lower row, Family History (Figure 15), shows all the ancestors and descendants of this family, in the style of the Search Results page. Click on any row to be taken to the corresponding Gene Family page.

Moving now to the upper row, we've already encountered the Find related families tab. In Figure 16, however, I've selected three member genes, whose identifiers now appear under the "Find all families with selected gene(s)" heading. Clicking this heading will bring back all gene families at all nodes that contain these 3. This is similarly to a Family History, but focused solely on the origins and fate of a subset of our current family's genes.

The third capability on the Find related families tab is BLAST for similar clusters. This link brings up the BLAST interface with the query sequence prepopulated with this family's consensus peptide sequence, and the BLAST target set to a database of all consensus sequences from families at this node. This lets you find gene families with similar consensus sequences.

phytozome   JGI Joint Genome Institute   C I G Center for Integrative Genomics

Search | BLAST | Info ▾ | BioMart | Help | Contact

Hypothetical EurosidI gene
Cluster 22238032. 13 members:
Gma Rco Csa Mtr Mes Ptr
4 1 1 2 2 3

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):
01200 : Genetic Information Processing
01230 : Folding, Sorting and Degradation
04120 : Ubiquitin mediated proteolysis [PATH:ot04120 tab04120]
K03178 : UBE1, UBA1; ubiquitin-activating enzyme E1 [EC:6.3.2.19] [COG:COG0476] [GO:0004839]
01500 : Human Diseases
01510 : Neurodegenerative Disorders
05020 : Parkinson's disease [PATH:hsa05020]
K03178 : UBE1, UBA1; ubiquitin-activating enzyme E1 [EC:6.3.2.19] [COG:COG0476] [GO:0004839]

Genes in this family | Functional Annotation | Multiple Sequence Alignment | Family History ❓

These clusters include genes from this family:

	SIZE	NODE	DESCRIPTION	M. esculenta	R. comunis	P. trichocarpa	M. truncatula	G. max	C. sativus	A. thaliana	A. lyrata	C. papaya	V. vinifera	M. guttatus	S. bicolor	Z. mays	O. sativa	B. distachyon	S. moellendorffi	P. patens	C. reinhardtii
🔍	44	Viridiplantae	Hypothetical gene (#23255333)	3	2	3	2	4	1	2	2	2	2	2	4	4	4	3	1	2	1
🔍	43	Embryophyte	Hypothetical gene (#23109299)	3	2	3	2	4	1	2	2	2	2	2	4	4	4	3	1	2	-
🔍	41	Tracheophyte	Hypothetical gene (#22968518)	3	2	3	2	4	1	2	2	2	2	2	4	4	4	3	1	-	-
🔍	40	Angiosperm	Hypothetical gene (#22826673)	3	2	3	2	4	1	2	2	2	2	2	4	4	4	3	-	-	-
🔍	25	Eudicot	Hypothetical gene (#22627997)	3	2	3	2	4	1	2	2	2	2	2	-	-	-	-	-	-	-
🔍	23	Rosid Pre-Duplication	Hypothetical gene (#22491066)	3	2	3	2	4	1	2	2	2	2	-	-	-	-	-	-	-	-
🔍	21	Rosid Post-Duplication	Hypothetical gene (#22364995)	3	2	3	2	4	1	1	1	2	2	-	-	-	-	-	-	-	-
🔍	13	EurosidI	Hypothetical gene (#22238032)	2	1	3	2	4	1	-	-	-	-	-	-	-	-	-	-	-	-
🔍	6	Malpighiales	Hypothetical gene (#22102622)	2	1	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
🔍	6	Legume	Ubiquitin-activating enzyme, E1	-	-	-	2	4	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure 15. Family (evolutionary) history of this gene family. Only internal node families are shown (the list does not include crown node, singleton families). The current family is always highlighted in blue.

The Align family members tab (not shown selected) enables you to send the entire family's coding or peptide sequence set to the Jalview analysis tool, discussed later. The Get Data tab (Figure 17), provides a quick interface (Get Sequences (basic)) for grabbing family member sequences, raw MSA alignment files, or consensus sequence. Data can be sent to a file or shown in the current browser. If more involved data sets (e.g, more annotations, flanking sequence, etc.) is needed, choose Get family data (advanced) to launch BioMart.

The rightmost tab on the upper row, Display Options (Figure 18), helps you customize the contents of the Gene Family page. You can disable the display of various columns, and change the gene coloring rules for the synteny display (the default is to color genes from families with at least 2 non-tandem members visible in the synteny display. You can increase the requirement so that even more members have to be present for a gene to receive a color. Try it!). You can also choose not to display information about particular species, even if they have members in the family. This is useful, for example, if you're interested in poplar-arabidopsis relationships, which first appear in the Rosid node, which also includes papaya, grape, the legumes, castor bean, cassava and cucumber, which might not be of interest. This feature lets you hide those "uninteresting" species.

Figure 16. Selection of three family members opens up the "Find related families" tab. Click on the "Find all clusters with selected gene(s)" to search for gene families at all nodes that contain at least these three genes.

Figure 17. Get Data tab allows you to retrieve sequence data for the gene family (for viewing in your browser or download), as well as MUSCLE multiple sequence alignments of the family members. To retrieve more information (e.g, annotations, flanking sequences, etc.), select "Get family data (advanced)."

Figure 18. Display option tab on Gene Family page. Settings apply to this view of this family only. To apply the settings to all subsequent family views, be sure to click the appropriate "Save settings" button.

Gene Page

Having completed our tour of the main Gene Family page, let's focus on the Gene Page, which is devoted to information about a single gene (Figure 19). To get there from the gene family page, click on the Gene page link next to the second poplar gene (POPTR_0009s07950.1). This loads up a default view (About this gene) that includes name and defline information, assigned domains and functional annotations, a PFAM domain view, and a simplified Gbrowse view of this gene in its genomic context, showing alternative transcripts (if any), peptide homology, and sometimes EST support. To get to the full Gbrowse view of this gene, you can click the "View in Gbrowse environment" link. But let's stay on the gene page for now. As for the family page, additional information is available in a row of tabs at the top. Select Sequences (Figure 20), and a panel showing genomic, transcript, coding (CDS) and peptide sequences for this gene appears. You can expand the sequence views one at a time, or all at once. It's possible to add upstream and downstream flanking sequence to the genomic sequence using the "update sequence" button.To the right of each sequence heading are buttons to send that sequence off to either the Phytozome BLAST interface or NCBI's BLAST utility. The shading indicates exons, introns, UTR and flanking sequences. Note that even the peptide has exon shading to help you keep track of where splices occur.

Figure 19. Gene Page default view.

Figure 20. Sequences tab on Gene Page.

The Peptide Homologs tab (Figure 21) shows the highest scoring peptide global alignments (dual-affine Smith-Waterman) of other Phytozome genes to this gene. Both gaps and insertions are indicated schematically in the rightmost column. Each target gene is identified by its species, name and defline (if available), and the score and percent similarity to the query sequence. The MRSF column (Most Recent Shared Family) lists the most recent node (going backwards from the crown nodes) at which the query and target gene are in the same family. For example, our gene is in the same family as its best papaya (Cpa) hit at the RPO (Rosid Post-hexaploidy) node. Click on the MRSF entry to be taken to the page for that shared family. Click on any target gene's Gene Page button to go to its Gene Page. Click on the "+" expand icon to see the textual version of any alignment.

You can also filter the display of homologs to only include certain species or clades, using the Homolog filter.

Figure 21. Peptide Homologs tab on Gene Page, with Homolog filter opened.

The Gene Ancestry tab (not shown selected) is completely analogous to the Family History tab on the Gene Family page. It lists all families that are ancestors of the current gene (which, by definiton, has no descendants).

Let's now turn away from the gene family and gene-centric view and look at genomic browsing in Phytozome. Go back to the About this gene tab, and click on "View in Gbrowse environment." This will take you to the page shown in Figure 22.

Gbrowse genomic browsing

Phytozome uses v1.69 of Gbrowse, GMOD's Generic Genome Browser. There is extensive help and tutorials for this browser at http://gmod.org/wiki/GBrowse#Documentation , including interactive tutorials at OpenHelix. This is just an overview to get you working comfortably with the main features.

Figure 22. Gbrowse view centered on gene from previous Gene Page.

The Gbrowsers at Phytozome always include at least the following tracks:

- Transcript and Alternate Transcripts
- Peptide homologs (pulled from diverse plant proteomes), aligned by BLASTX and/or BLATX
- ESTs or EST assemblies (TA = transcript assembly) for both same-species and related species expressed sequence tabs
- VISTA genomic synteny tracks, both self-synteny and pairwise synteny, between this and other Phytozome plant genomes.

Gbrowsers for some of our genomes also contain repeat-masking, syntenic block, transposable elements, map markers, and short read depth density, and other data.

Only those tracks that are selected (by checkbox) in the lower Tracks panel are shown. If you change the track selections, make sure to click any of the Update Image buttons appearing at the page. Note that the page reloading can be slow, depending on how many features are included in the tracks you've enabled. This is a limitation of v1.69, and is much improved in the just released version 2.0 Gbrowse, which we hope to have deployed well before summer.

You can change the order of tracks in the Details display by click-dragging them with your mouse (place the mouse pointer anywhere over the track label).

Navigation in Gbrowse is accomplished in one of three ways:

1. Click the "<" or ">" arrows next to "Scroll/Zoom" . The "<<" and ">>" are just larger size left and right moves
2. Click somewhere within the current Details view, to recenter the image, or within the Overview panel to move large distances across the scaffold/chromosome
3. Enter a landmark (in the form of scaffold_name:START..END) in the Search box

Zooming in and out is accomplished either using the "Show" pulldown (Figure 23), or the large yellow "-" and "+" icons. **IMPORTANT NOTE**: A Details view is only available for regions of 100kbp or smaller. This is another v1.69 performance limitation that we will relax when moving to v2.

Most features in the Detail view are hyperlinked to informative pages, and provide additional information (deflines, aliases, symbols) when moused over. The transcripts (primary and alternative) link out the our Gene Pages. Aligned features (ESTs and Peptides) link out to a standard Gbrowse details page, showing the "query" sequence and its alignment to the genome in detail. The VISTA tracks link out to the VISTA browser, which also has extensive online help available from OpenHelix (http://www.openhelix.com/downloads/vista/vista_tutorialhome.shtml).
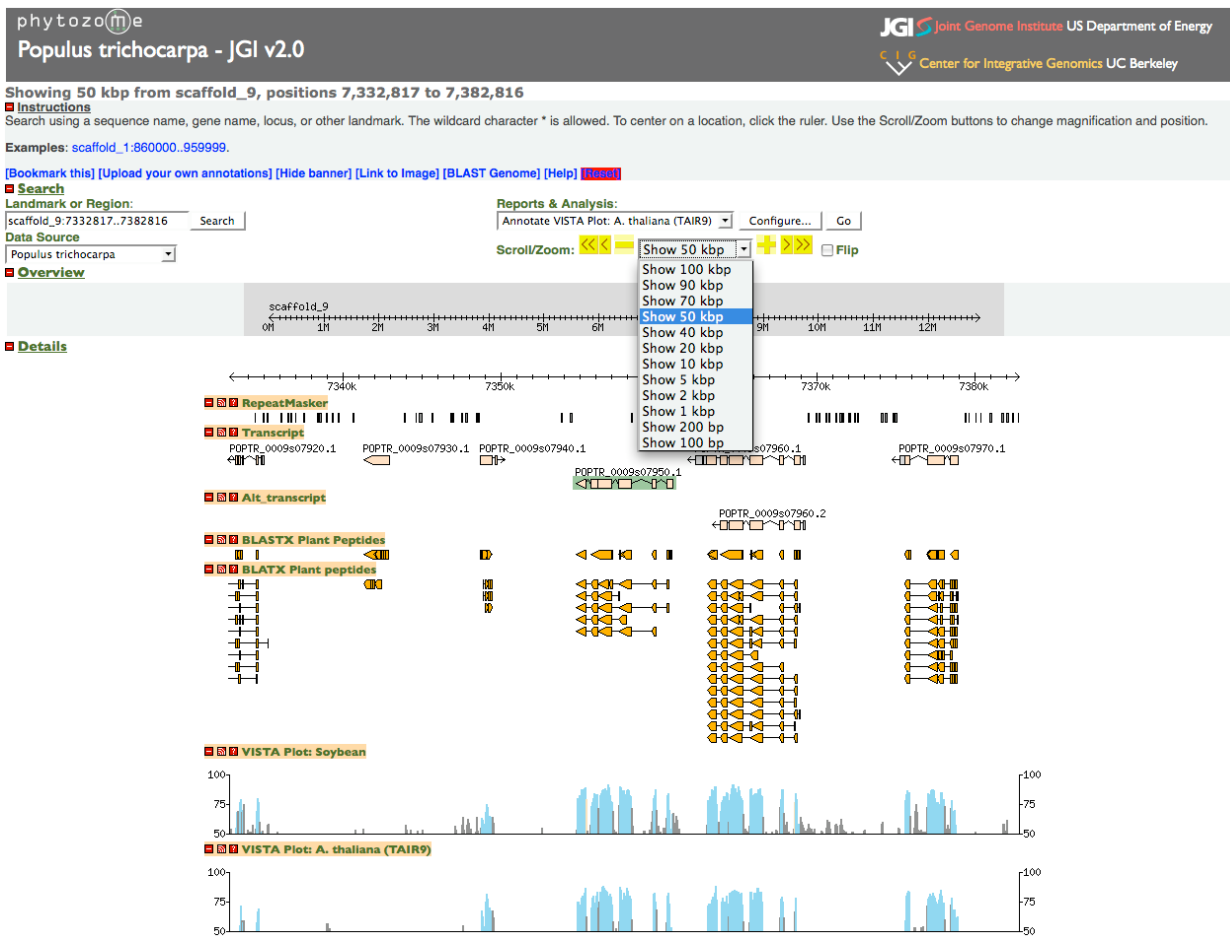
Figure 23. Expanded region view (50kbp) centered on the region in Figure 22.

The series of links at the top (e.g, "[Bookmark this]", etc.) are quite useful. If you want to share a particular browser view with someone, including all your track selections and their order, click on [Bookmark this], which will load your browser's URL field with a complete URL that can be emailed to anyone and will guarantee them the exact view you're currently seeing. If you just want the Overview and Details image, click [Link to Image]. [BLAST Genome] links directly to the Phytozome BLAST interface for this genome.

You can also search directly in Gbrowse for gene names, aliases, symbols, using the Search box mentioned earlier. Note that this is much, much slower than the gene family search implemented in Phytozome, but has the advantage that you can search for names of aligned features (ESTs, related non-phytozome peptides) that are not available through Phytozome's search.

Let's take a look at the BLAST interface now. Go back to one of Gene Pages (click thru on a transcript in Gbrowse), select the sequences tab, and then click "BLAST this sequence PHYTOZOME" to load the BLAST interface.

BLAST/BLAT Search Interface
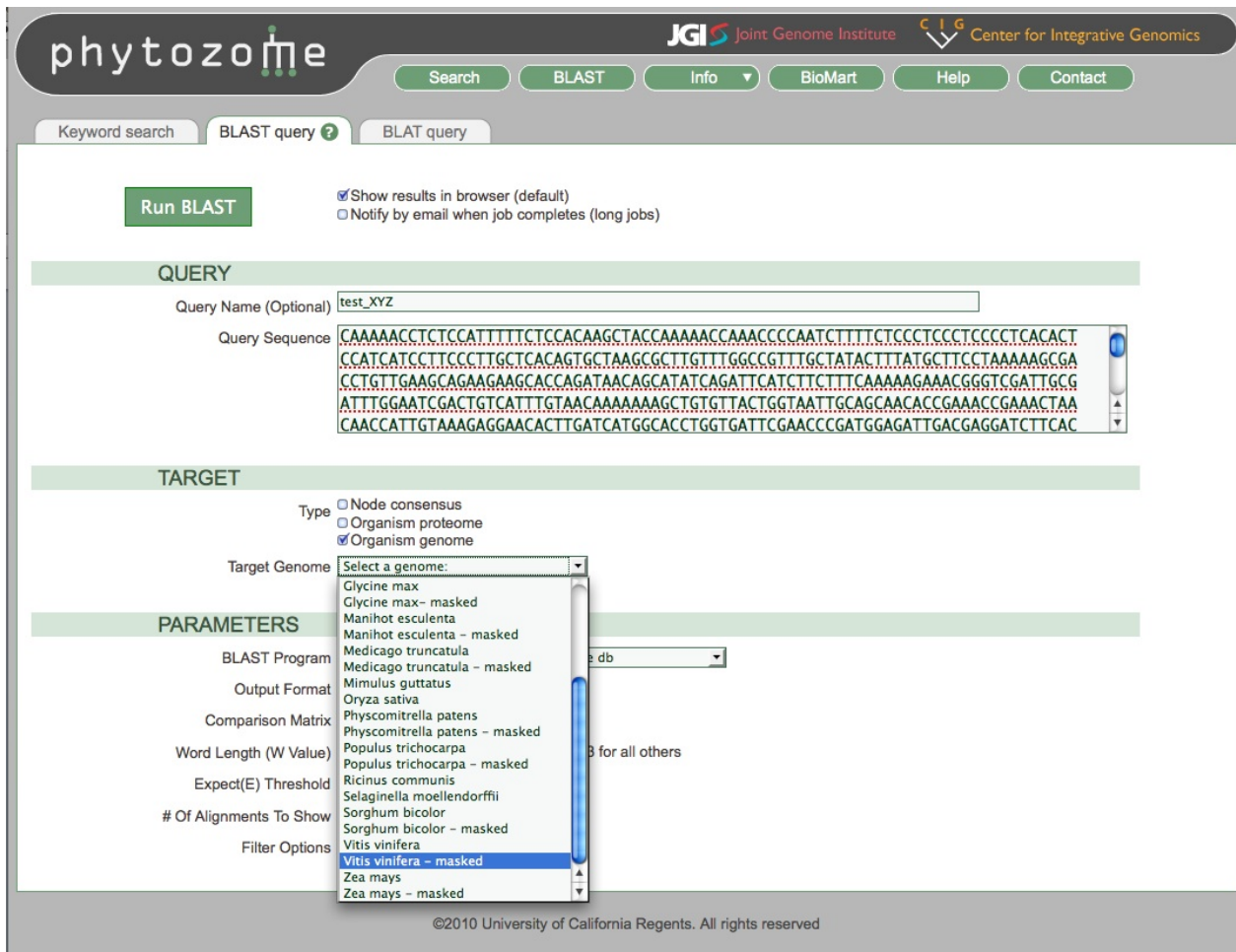
Figure 24. BLAST interface. Genome targets (both unmasked and, where available, masked) can be selected from the set of Phytozome organisms.

Phytozome's BLAST/BLAT interface (Figure 24) should look familiar to anyone who uses NCBI's BLAST utility. The most obvious difference is the set of target choices. BLAST databases exist for all Phytozome genomes (with repeat-masked versions available for Arabidopsis lyrata, brachypodium, cassava, medicago, physcomitrella, poplar, sorghum, grape, and maize), all proteomes (with peptides from both primary and alternatively spliced transcripts), as well as (peptide) family consensus sequences for all 15 internal nodes where gene families have been constructed.

The results page for BLAST against peptide databases (proteomes and node consensus sequences) are self-explanatory, with each hit containing a link out to the appropriate Gene page (proteome database) or Gene Family page (node consensus database). Results for BLAST against a genomic target (Figure 25) are a bit more interesting. The user can switch the results view between query coordinates (the same as for proteome and node consensus BLASTs) and target coordinates (the default for genomic BLAST), using the "switch views"

button.   The target view includes a feature-scale graphic as well as a target scale view to provide some sense of where along the chromosome or scaffold a hit is found. Click on a
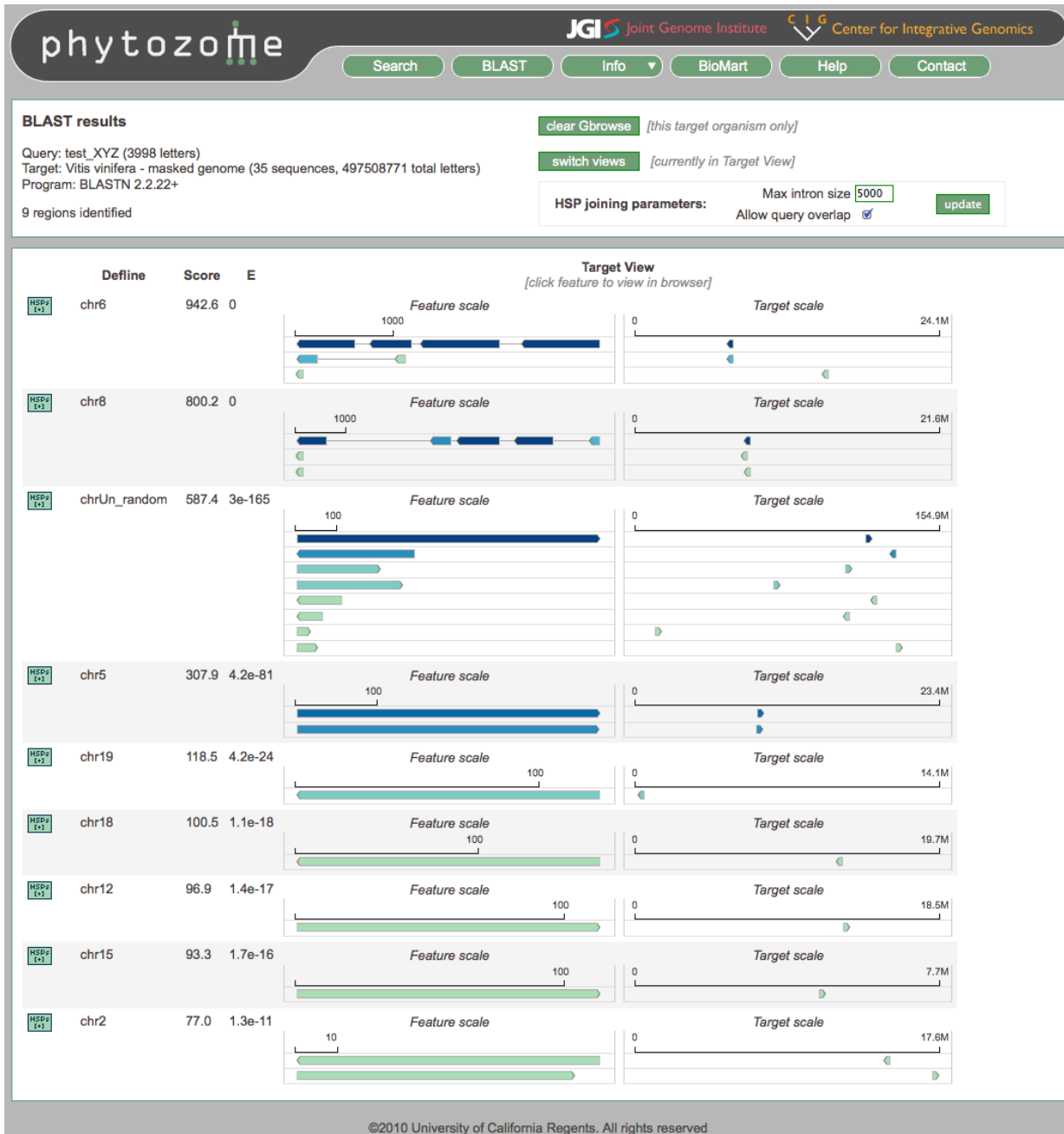
Figure 25. Genomic BLAST results, showing the default Target view.
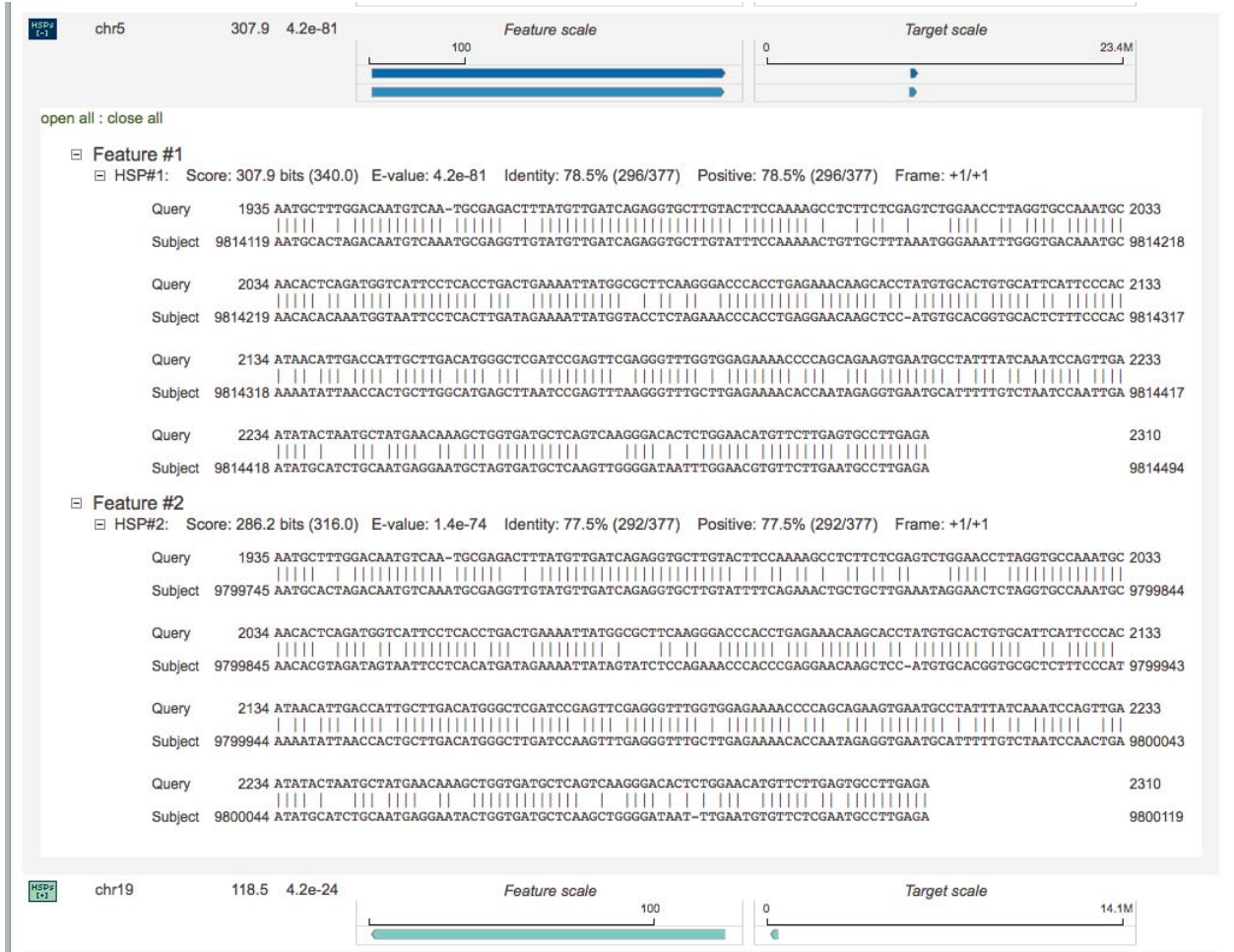
Figure 26. **Expanded text view of Feature HSPS, from clicking HSP button and then expanding features.**

feature in either column, and you're taken to the appropriate region in Gbrowse with your BLAST hit(s) displayed in a track called "UserBlast" (Figure 27)

Phytozome attempts to stitch together HSPs into features in a sensible (but sadly not splice-aware) manner. You can control the maximum allowed intron size in the HSP joining parameters up top. Here you can also allow or forbid query overlap (meaning the same part of query sequence might be used in two independent HSPs).

If you'd like a detailed, text-style view of the alignment for a particular target sequence, click on the green "HSP [+]" button to the left. Then expand one or all of the Feature lists (Figure 26). Click "close all" to collapse the list.

By design, the BLAST results you view in Gbrowse are private (no one else can see them), and will persist there until you remove or overwrite them. They'll be overwritten if you redo the HSP joining parameters on the results page, and they can be removed by clicking the "clear gbrowse" button.

Figure 27. Gbrowse view of BLAST feature, after zooming out and recentering. The query and target names, the feature label, and the evalue are all shown.

Data Retrieval with BioMart

The BioMart tool will help you perform bulk downloads of genome or gene family associated data, with control over exactly what information is included and how it's formatted. BioMart is deployed for a number of different model organism and comparative genomics websites; take a look at http://www.biomart.org/ . A list of help docs and presentation on BioMart is available here: http://www.biomart.org/install.html .

You can get to BioMart interface directly (by clicking on the BioMart menu button at the top of any Phytozome page), by choosing "Get Data -> Get Family data (advanced)" on a Gene Family page, or by selecting gene families on a Search Results or BLAST results page and choosing "Analyze Results / Get Data". All but the direct method will bring up BioMart with its Filter pre-set to restrict the results to only the families you're interested in. But let's take a look at the interface when it's loaded directly (no filter preset) as in Figure 28:
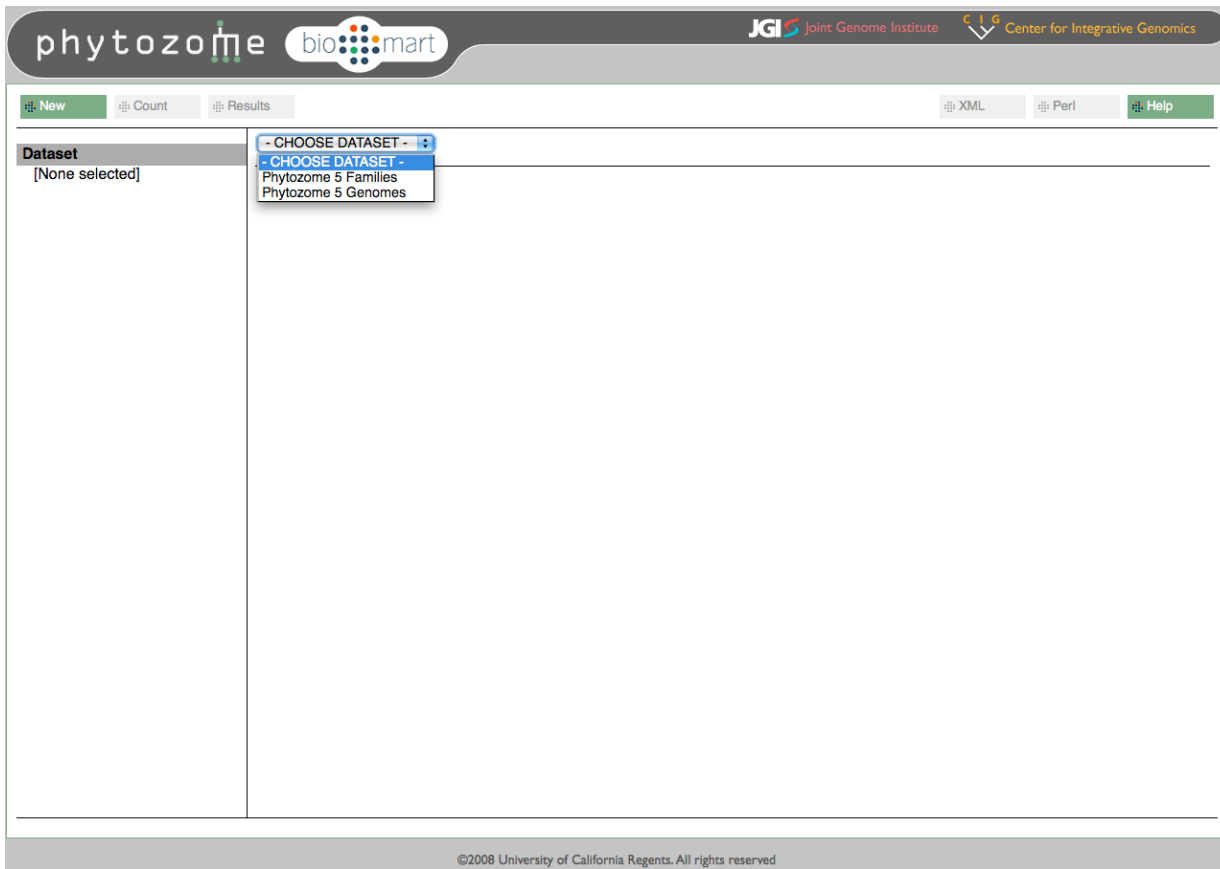
Figure 28. BioMart startup page, showing the option to choose Genome or Family datasets.

The first step in extracting data is deciding whether you want family-centric or genome-centric data. Use the CHOOSE DATASET pulldown to pick one or the other. We'll choose Genomes for now. Our next step is to click on the Filters heading in the left column, to open up the filters (no, that isn't obvious), as seen in Figure 29. This allows you to specify any restrictions on the data you want to download (only for a particular organism(s), certain genomic regions, certain genes with certain properties, etc.). Set the organism filter to only include Arabidopsis thaliana. Now click on the Attributes heading, and specify what information you'd like included in your data set. As shown in Figure 30, we've asked for a simple data set consisting gene names, chromosome, description (defline), strand, start and end coordinates, and any PFAM identifiers (with *their* descriptions) assigned to the genes. Click on the "Count" button in the upper left, and you'll see there are 27379 entries (corresponding to the number of protein-coding loci in Athaliana TAIR9). Now click on "Results" to get the view shown in Figure 31. The default it to show 10 rows of results, formatted for viewing on the web. You can change data format and send the set off to a file rather than the browser. For large data sets (which can take a while to assemble), it best to choose "Compressed web file (notify by email)" in the Export pulldown, providing an email address where you can be contacted once the file is ready.
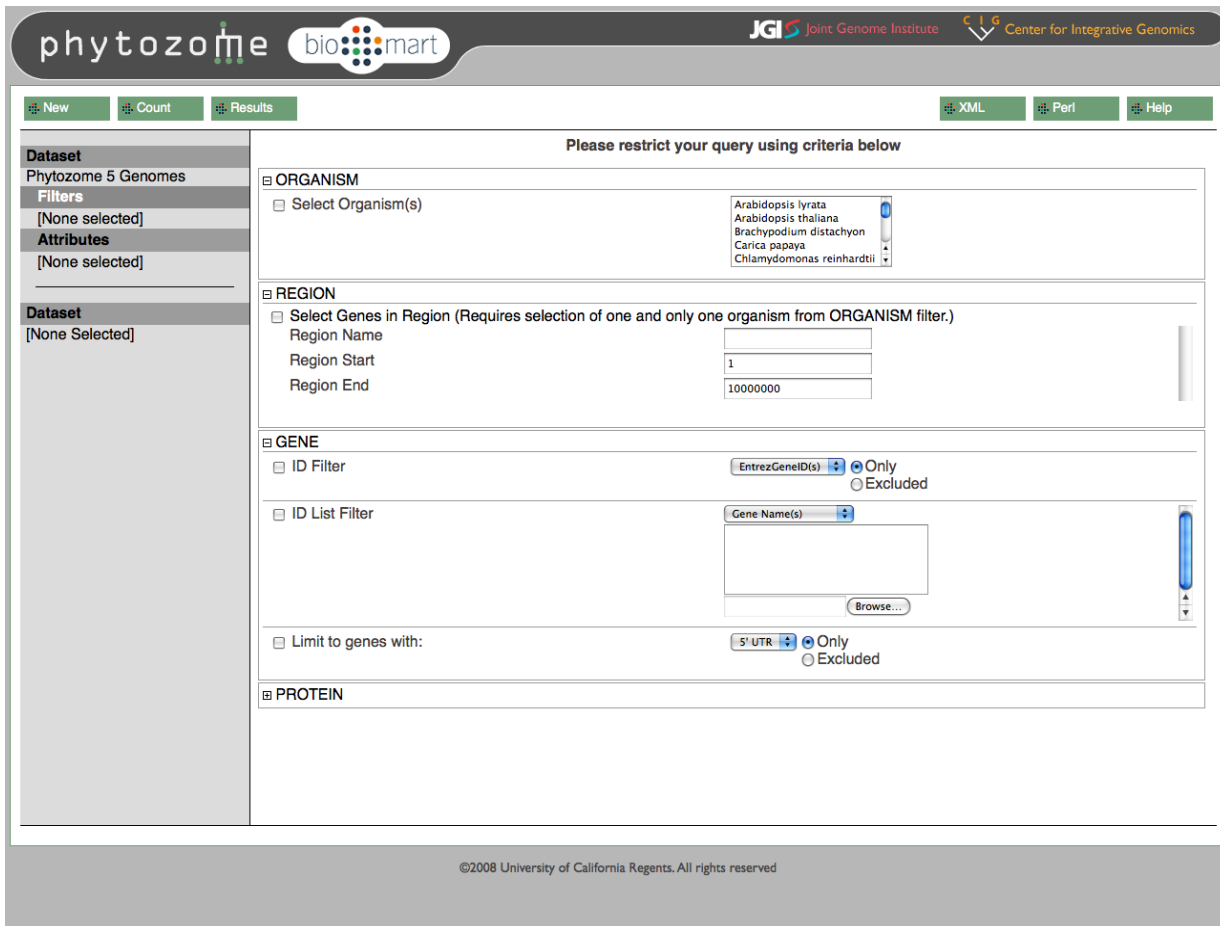
Figure 29. After selection of "Phytozome 5 Genomes" as the dataset, clicking on "Filters" in the left column displays the various filters that restrict the data that's retrieved. The "PROTEIN" set of filters has not been expanded in this image.

You can also download detailed gene structure information (exon coordinates, cds starts and stops) and sequence data (gene, transcript, CDS, UTR, etc.) with or without flanking data. More complicated queries are also possible, including intersections between 2 data sets, which you'll try in the exercises. If you're interested, do check out the BioMart tutorial docs mentioned above.

We're now ready to view the final Phytozome component, which is Jalview. Please return to one of the previous Gene Family pages, click the "Align member sequences" tab, and click "Load member proteine sequences". This should launch Jalview.

Figure 30. Attributes selection page, which appears after clicking on "Attributes" in the left column. The requested attributes are gene name, descripton, chromosome, strand, start and end, as well as any assigned PFAM domains (ids and descriptions). Clicking the Count button at left reports that 27379 genes will be included (this corresponds to the number of protein-coding loci in Arabidopsis thaliana TAIR9).

Figure 31. BioMart results preview, showing first 10 rows of the result set. The results can be exported in various formats, and, for larger result sets that make take time to assemble, the user can choose email notification.

Jalview Analysis applet

Jalview (Figure 32) provides an environment for editing and recalculation of MSAs (multiple sequence alignments), calculation of Neighbor Joining and Maximum likelihood trees, and downloading calculated results in various text and image formats. More information and tutorials on Jalview are available at http://www.jalview.org/. The version used at Phytozome was branched off several years ago, so all features described at the jalview website might not be available at Phytozome (for now).

Note that when you launched Jalview from a Gene Family page, two panels came up. The first, "Unaligned Sequences", always appears. The second "ClustalW Multiple Sequence Alignment" will appear if either a) the data you're loading is a gene family with a precomputed MSA (as is the case for most Phytozome families), or b) the data you're loading has 25 or fewer sequences, in which case a de novo MSA will be calculated on the fly.
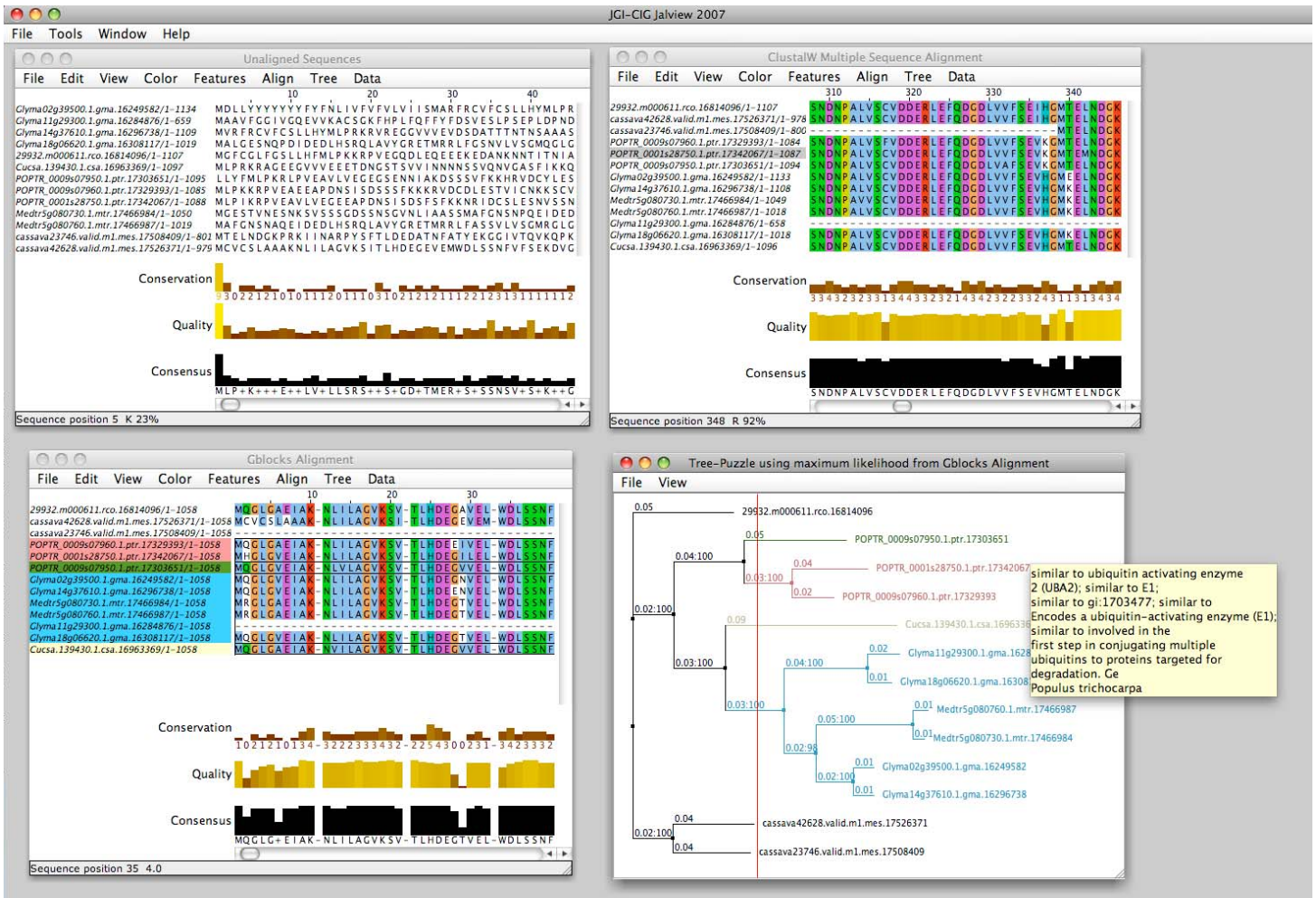
Figure 32. Jalview analysis tool. The 4 panels shown are the initial loading panel (Unaligned Sequences), the Multiple Sequence Alignment Panel (which loads automatically if the family's MSA has been precomputed, or if the sequences set has no more than 25 members), the Gblocks Alignment panel (generated automatically when the user attempts to build a phylogenetic tree), and the Tree panel ("Tree-Puzzle using maximum likelihood from Gblocks Alignment). Clicking on the Tree panel "cuts" the tree at that point, and distinct subfamilies are indicated by different colors (reflected in the Gblocks panel simultaneously). Mousing over a crown node brings up a defline, if available.

On the MSA panel, select the Tree menu item, and then "Tree Puzzle using Maximum Likelihood" Because "Run Gblocks…" was preselected, Jalview first performs a Gblocks operation on the MSA, removing columns that are unlikely to correspond to meaningful parts of the alignment, before performing the tree calculation.

Mouse over any of the crown nodes (sequence names) in the tree and a popup will appear with any defline information that's available for that sequence.

By clicking anywhere in the Tree panel, you instruct Jalview to perform a "slice"of the tree, with the different subclades assigned different colors in both the tree and the associated

Gblocks panel. You can save the tree in either Newick format, or as an image, by clicking the File menu in the Tree panel. You can do the same for the Multiple Sequence Alignment, using File->Export and File->Output to Textbox, in the MSA panel (try it!).

Jalview's true utility is revealed on the editing front. You can select sequences in the alignment (by clicking), and then remove them (Edit->Delete). You can also remove columns or blocks of aligned residues, or add additional sequences (Edit->Add Sequence(s)), at which point you can recomputed the MSAs and associated trees. This is very useful for refining alignments, removing outlier sequences, recomputing trees with outgroup sequences present, etc.

Conclusion

That ends our tour of Phytozome. Please help make Phytozome more useful to plant biologists, by

1. Completing the Phytozome user survey at
    http://www.zoomerang.com/Survey/WEB229RWN4493P
2. Sending us an email at phytozome@jgi-psf.org, with bugs, new feature and new data requests.
3. Joining our mailing list (extremely low volume – a few messages a month). Sign up at
    http://www.phytozome.net/contact.php .

Thank you,

The Phytozome Development Team: Richard D. Hayes, Joni Fazo, Russel Howson, David M. Goodstein, Rochak Neupane, Shengqiang Shu

The Plant Genomics Team at JGI: all the above plus Kerrie Barry, Uffe Hellsten, Therese Mitros, Simon Prochnik, Megan Hall, Albert Wu, and Daniel S. Rokhsar