

Gramene

A Resource for Comparative Grass Genomics

Doreen Ware

Summary

Grasses are one of the largest agricultural crops, providing food, industrial materials and renewable energy sources. Due to their large genome size and the number of the species in the taxa, many of the genomes are not targeted for complete sequencing. Gramene seeks to provide basic researchers, industry and educators with a resource that can be used as a tool for knowledge discovery across grass species. This chapter briefly outlines system requirements for end users and database hosting, outlines data types and basic navigation within Gramene and provides an example of how a maize researcher would use Gramene to leverage rice genome organization and phenotypic information to support targeted experimental research in maize.

Key Words: Rice; maize; wheat; barley; genome sequence; comparative genomics.

1. Introduction

Gramene seeks to provide the plant community with a resource for the comparative genomics of grass species where reference genomes and their respective annotations, including sequence and phenotypes, serve as a scaffold species where the complete genome sequence is unavailable.

Because of the high degree of conservation of gene order in grass species and conserved gene function between closely related species, the information available for a reference grass genome is useful to make informed decisions about genome organization, gene function and potential phenotypes in related species. In comparison, the genome information for *Arabidopsis* may provide

researchers with detailed information on potential gene function, though not necessarily genome organization.

The Gramene Genome Browser (*1,2*) hosts an implementation of the Ensembl Genome Browser (*3*) containing collated information on rice genes. This information includes predicted function as assigned by GO annotations (*4,5*) and Pfam domain similarity (*6*), single-nucleotide polymorphisms (SNPs) and rice phenotypes [as quantitative trait loci (QTL)]. Comparative genome information is provided by the alignment of plant expressed sequence tags (ESTs) with the best rice ortholog, and maize synteny information is based on the maize physical map. The front Web page acts as portal to internal pages in Gramene as well as external links to other databases (*see Table 1*). Internal links within Gramene include Mapviewer (an implementation of the CMap comparative genetic viewer developed as part of the GMOD consortium), markers, genes, QTL and an ontology browser. Contextual links to external sources include reference resources such as GenBank (*7*) (*see Chapters 2 and 3*), model organism databases such as MaizeGDB (*8*) (*see Chapter 16*) or GrainGenes (*9,10*) (*see Chapter 14*) and project specific databases such as Barleybase (*11*) (*see Chapter 17*) and Panzea (*12,13*) (*see Table 1*).

2. Materials

2.1. Hardware and Software Requirements for Users

A computer with internet access and a standard web browser such as Mozilla/Firefox, Netscape 6 and above, or Safari.

2.2. Gramene System Components

Gramene is a web-based application that allows users to search and view biological data, making use where appropriate of graphics viewers such as the Ensembl genome browser or the CMap genetic and comparative map package. Data are maintained in distinct relational databases (MySQL), and users connect to the site using a standard web browser. User queries for static (HTML) and dynamic content are negotiated by the Apache web server and a middleware layer written in Perl. Bulk downloads of data are provided through an FTP site.

2.3. Local Installation of Gramene

Local installation of Gramene requires a computer running a Unix operating system (such as Linux, FreeBSD and Solaris), Perl (5.6.1 or greater), Apache 1.x (1.3.26 or greater, *see Note 1*), MySQL (4.1 or greater), Perl modules (Comprehensive Perl Archive Network, CPAN), CMap (0.15) and the Ensembl

Table 1
Links to External Websites, Databases

Resource	URL	Data type
EBI (InterPro)	http://www.ebi.ac.uk/interpro	Protein families
EMBL (<i>see</i> Chapter 1) (SMART)	http://smart.embl-heidelberg.de	Protein domains
ExPASy (PROSITE)	http://www.expasy.ch/prosite	Protein domains
GenBank (<i>see</i> Chapters 2 and 3)	http://www.ncbi.nlm.nih.gov/	Protein, gene
Gene Ontology Consortium (<i>see</i> Chapter 24)	http://www.geneontology.org	Protein, gene, Ensembl genes
Georgetown University (PIR)	http://pir.georgetown.edu/pirwww/index.shtml	Protein domains
GrainGenes (<i>see</i> Chapter 14)	http://www.graingenes.org	ESTs
GRIN	http://www.ars-grin.gov/	Protein
INRA (ProDom)	http://protein.toulouse.inra.fr/prodom/ current/html/home.php	Protein domains
IRRI (<i>see</i> Chapter 22)	http://www.iris.irri.org/	Protein
KEGG (<i>see</i> Chapter 21)	http://www.genome.ad.jp/	Protein
KOME	http://cdna01.dna.affrc.go.jp/cDNA	cDNA sequences
Laboratory for Genomics & Bioinformatics	http://www.funagen.org/Sorghum.htm	ESTs, clustered ESTs

(Continued)

Table 1 (Continued)

Resource	URL	Data type
MaizeGDB (<i>see</i> Chapter 16)	http://www.maizegdb.org	ESTs, clustered ESTs, gene
Maize Oligonucleotide Array Project	http://www.maizearray.org	Microarray target sequences
Manchester University (SPRINT)	http://lumber.sbs.man.ac.uk/dbbrowser/sprint	Protein domains
MGOS	http://www.mgosdb.org	SAGE tag sequences
MRC-LMB (Superfamily)	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY	Protein domains
NASC (<i>Arabidopsis</i>) (<i>see</i> Chapter 9)	http://atensembl.arabidopsis.info	Gene models
NCBI (GenBank, dbEST)	http://www.ncbi.nih.gov/dbEST/index.html	ESTs
NCBI (GenBank, dbGSS) (<i>see</i> Chapters 1 and 2)	http://www.ncbi.nih.gov/dbGSS/index.html	Genome Survey Sequences,
		Bacterial Artificial Chromosome (BACs), Methyl-filtered/Hi-cot selected reads, FSTs
NCBI (GenBank, dbPLN)	http://www.ncbi.nih.gov/Genbank/index.html	CDS sequences, cDNA sequences, Genomic clone sequences
NSF Rice Oligo Array Project	http://www.ricearray.org	Microarray oligonucleotide sequences

Oryzabase	http://www.shigen.nig.ac.jp/rice/oryzabase/top/top.jsp	Gene
PlantGDB (<i>see</i> Chapter 25)	http://www.plantgdb.org	Clustered ESTs, FST sequences
PLEXdb (<i>see</i> Chapter 17)	http://www.plexdb.org	Microarray oligonucleotide sequences, Microarray target sequences
Sanger (Pfam)	http://www.sanger.ac.uk/Software/Pfam	Protein domains
TIGR (Gene Indices) (<i>see</i> Chapter 5)	http://www.tigr.org/tdb/tgi/plant.shtml	Clustered ESTs
TIGR (Rice Genome Annotation)	http://www.tigr.org/tdb/e2k1/osa1	Gene models
TIGR (TIGRFAM)	http://www.tigr.org/TIGRFAMs	Protein domains
TIGR (TIGR Maize Database)	http://maize.tigr.org	Clustered Genome Survey Sequences
Uniprot (<i>see</i> Chapter 4)	http://us.expasy.org	Protein
University of Delaware Massively Parallel Signature Sequencing (MPSS) (<i>see</i> Chapter 19)	http://mpss.udel.edu/	MPSS tag sequences

EST, expressed sequence tag; FST, flanking sequence tag.

genome browser. The current Gramene server (<http://www.gramene.org/>) is running on a Fedora Core 1 Linux machine with dual CPUs, 6 GB memory and 600 GB of disk space available.

3. Methods

3.1. Basic Navigation of the Gramene Website

Gramene contains many different data types (*see Table 2*) that are maintained as multiple modules within the database. The main entry point for the system is through the front Web page (<http://www.gramene.org/>) (*see Note 2*). Every Gramene page contains the main navigation bar as well as module specific navigation bars, a general search, a link to the site map and a feedback page. The main navigation bar is found at the top of each Gramene page and is the

Table 2
Major Data Types in Gramene

Data type	Description
Sequence	Genome assembly such as rice pseudomolecules, maize BACs, Arabidopsis assembly, Genomic survey sequences such as BAC ends, mRNAs such as full-length cDNA, ESTs
Map	Sequence, genetic, physical, Maize Bin, Deletion, QTL
Marker	Gramene has a broad definition for Marker data type, which include all features located on chromosomes, eg, SSR, RFLP, AFLP, RAPD markers, BAC end sequences, BAC Clones, cDNAs, ESTs, Gene Primers, STS, Tos17 FSTs
Gene	Genetically defined phenotype genes, sequenced genes with or without experimental evidence, tRNAs, rRNAs, pseudogenes
Protein	Swiss-Prot/TREMBL, from predicted gene models, families and domains
Diversity	Variations with species on the same allele, e.g., SNPs, allelic data of SSR, genes
Pathway	Biochemical or genetic pathways based EcoCyc
Germplasm	Related to the species and other data in Gramene
Controlled Vocabularies	GO, TO, EO, PO, GRO, cross-reference to other data in Gramene
Reference	Publications associated with the data types and methods associated with data integration

BAC, bacterial artificial chromosome; EST, expressed sequence tag; QTL, quantitative trait loci.

main entry point to the search modules (Search), genome browsers (Genomes), software and data sets (Downloads), general information (Resources), help documents (Help), and information about the project and participants (About). In addition to the navigation bar, there is a general search which can be refined to interrogate the different data modules, as well as the Feedback link. The Feedback link is set up to provide the user with a comment page where the URL from the page the user was viewing at the time of the response is automatically included in the message. From the main navigation bar, a dropdown menu is available from each of the main headings containing refined subjects. For example, from the Help menu, entry points are available to the site map, help documents, release notes, tutorials, frequently asked questions (FAQs), workshop exercises, mailing list, rice gene nomenclature and ask us feedback links. By selecting the site map, a user is provided with detailed set of options available within each one of the search modules, as well as information available from reference pages, data analysis, help documents and downloads within Gramene. The interfaces within Gramene are interactive, providing the user with links to external reference databases (*see Table 1*) as well as links to internal modules with in Gramene.

3.2. Example Use of Gramene

Within the constraints of this chapter it is will not be possible to go through all of the Gramene interfaces. Instead, this example examines a sample question and walks through how to use Gramene to obtain information to facilitate genomic research. In this example, a maize line expressing a mutant phenotype has identified candidate genes for the mutation based on microarray gene expression data. We would like to identify candidate rice genes orthologous to the differentially expressed maize genes, as well as related genes from other cereals, ascertain the potential function of these rice genes based on annotation of the region in rice, and use synteny to identify candidate genomic position in maize.

In this example, we will use the BLAST search module as an entry point to the Rice Genome Browser.

From the Gramene home Page select Search from the main navigation bar and select “Sequences-BLAST”. This will take the user to the BLAST home page where a sequence can be entered either pasting it directly into the sequence box or uploading a local file (*see Fig. 1*). Once the sequence has been entered we must identify the sequence type by selecting “DNA queries” or “peptide queries”. In this example, we are using a DNA sequence and check “DNA queries”. Next we must determine which sequence database to query. The

The screenshot shows the Gramene BLAST home page. At the top, there is a navigation bar with 'GRAMENE Multi' and a search box. Below this are tabs for 'new', 'SETUP', 'CONFIG', 'RESULTS', and 'DISPLAY'. The main content area is divided into three sections: 'Enter the Query Sequence', 'Select the databases to search against', and 'Select the Search Tool'. The 'Enter the Query Sequence' section has a text input field containing a DNA sequence, a 'Browse...' button, and a 'Retrieve' button. The 'Select the databases to search against' section has a 'Select species' dropdown menu with 'Rice' selected, a 'dna database' dropdown menu with 'Genomic sequence' selected, and a 'peptide database' dropdown menu with 'Peptides (Fgenesh gene models)' selected. The 'Select the Search Tool' section has a dropdown menu with 'BLASTN' selected, a 'configure' button, and a 'RUN' button. The search sensitivity is set to 'Near-exact matches'. On the right side, there is a 'Summary' section with links to 'setup', 'configure', 'results', and 'display', each with a 'Not yet initialised' status.

Fig. 1. The Gramene BLAST home page.

Gramene sequence databases are composed of reference genomes available in the Genome Browser module (*Arabidopsis*, rice and maize) as well as the sequence of features aligned to these reference genomes. Sequences are organized first by species and then by sequence type. In this example we select the species “Rice” and then “Genomic sequence”. The search tool used will be “BLASTN” and the search sensitivity “Near-exact matches”.

The Gramene BLAST result page contains three views, the alignment locations vs. karyotype, the alignment locations vs. query sequence and the alignment summary. All three views are interactive and allow the user to view details of the alignment and link to other pages within Gramene.

In the alignment location vs. karyotype view (see **Fig. 2**), the top scoring hit is highlighted by mousing over the glyph, we are presented with a drop down box containing links to the BLAST alignment with the genome (align), the alignment in context to the submitted sequence (query sequence) or genomic sequence (genomic sequence), and a link to the rice genome browser (ContigView).

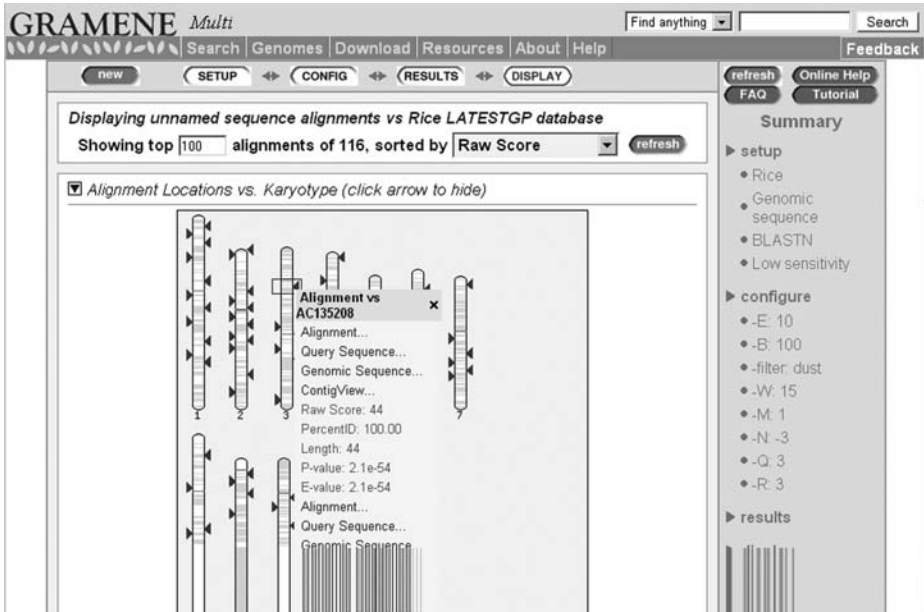


Fig. 2. The alignment location vs. karyotype view.

By selecting ContigView we leave the BLAST search module and enter the rice genome browser (*see Note 3*).

The ContigView of the EnSEMBL genome browser contains four views, the Chromosome or Clone view, Overview, Detailed view and Basepair view. All views contain a small red box that positions the user for each successive view. Views can be turned off or on by right clicking on the “+” or “-” symbol. In nearly complete genomes such as rice and *Arabidopsis*, the highest level view is of a chromosome, whereas in emerging genome sequences that use a bacterial artificial chromosome (BAC) based sequencing approach (such as maize), the highest level sequence view is of a clone or BAC.

The Chromosome or Clone view provides users with a representation of the complete pseudomolecule (a chromosome in the case of rice) and the position (in a red box) represented by the Detailed view.

The Overview provides the user with information on long range-derived features such as QTL, and synteny in the context of smaller sequence features such as genes and markers (*see Fig. 3*).

The Detailed view (*see Fig. 4*) provides semantic zooming, customizable tracks and links to internal pages, as well as contextual links to data sources

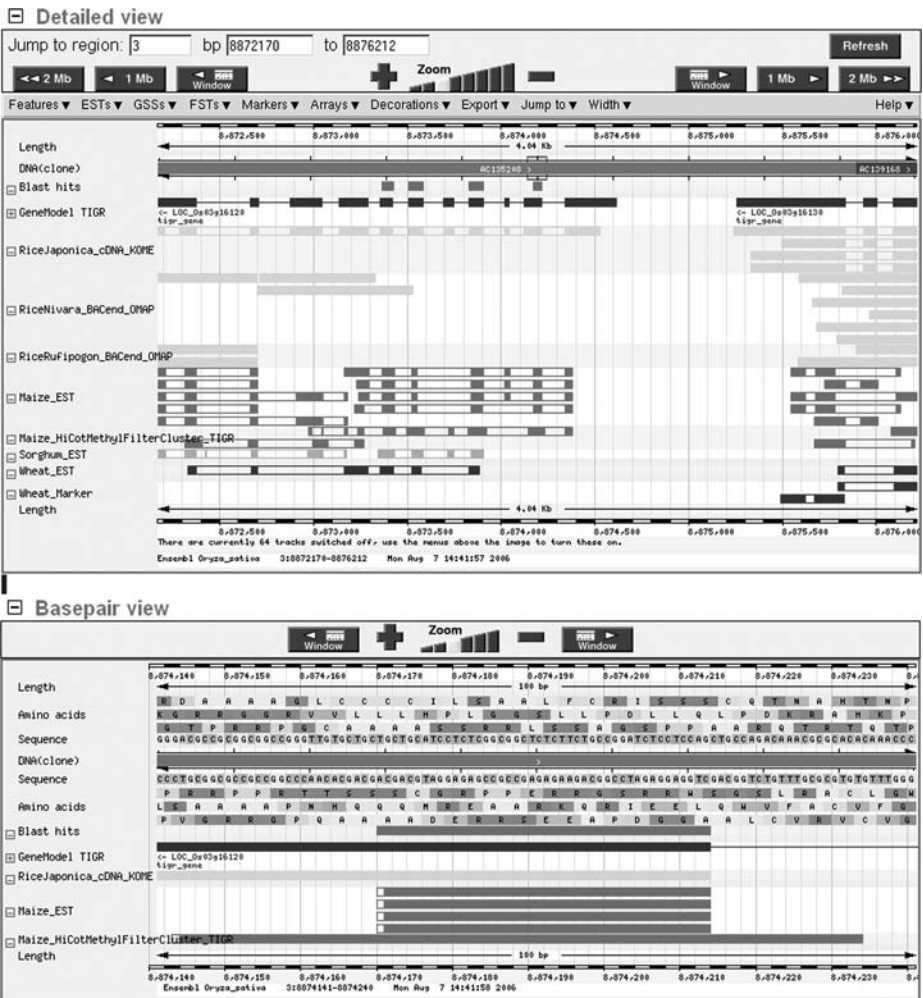


Fig. 4. The Detailed view provides semantic zooming, customizable tracks, links to internal pages as well as links to external data. The Basepair view displays the rice genome sequence for both the forward and reverse strands as well as predicted amino acid sequences representing the three potential reading frames for each strand.

displayed as the first feature track in the Detailed view (see Fig. 4, see Note 4). This region is annotated with rice QTL and is syntenic with maize chromosomes 1 and 9. The region also contains 21 different QTL reflecting 11 traits (grain number, soluble protein content, 1000 seed weight, seed number, spikelet

fertility, seed dormancy, chalky endosperm, spikelet number, root number, and days to heading panicle length) in five trait categories (yield, biochemical, vigor, quality, development, and anatomy) (see Fig. 3). For more detailed information we can link directly to the QTL database for details of the QTL and information on the studies and trait descriptions.

In this example, the BLAST hit overlaps the rice gene model LOC_os03g16120. The cereal alignments show that the rice gene model has sequence similarity to ESTs from wheat, sorghum as well as maize (see Fig. 4, see Note 5).

The cereal features are a useful validation for the rice gene model structure and for the development of species-specific molecular genetic markers.

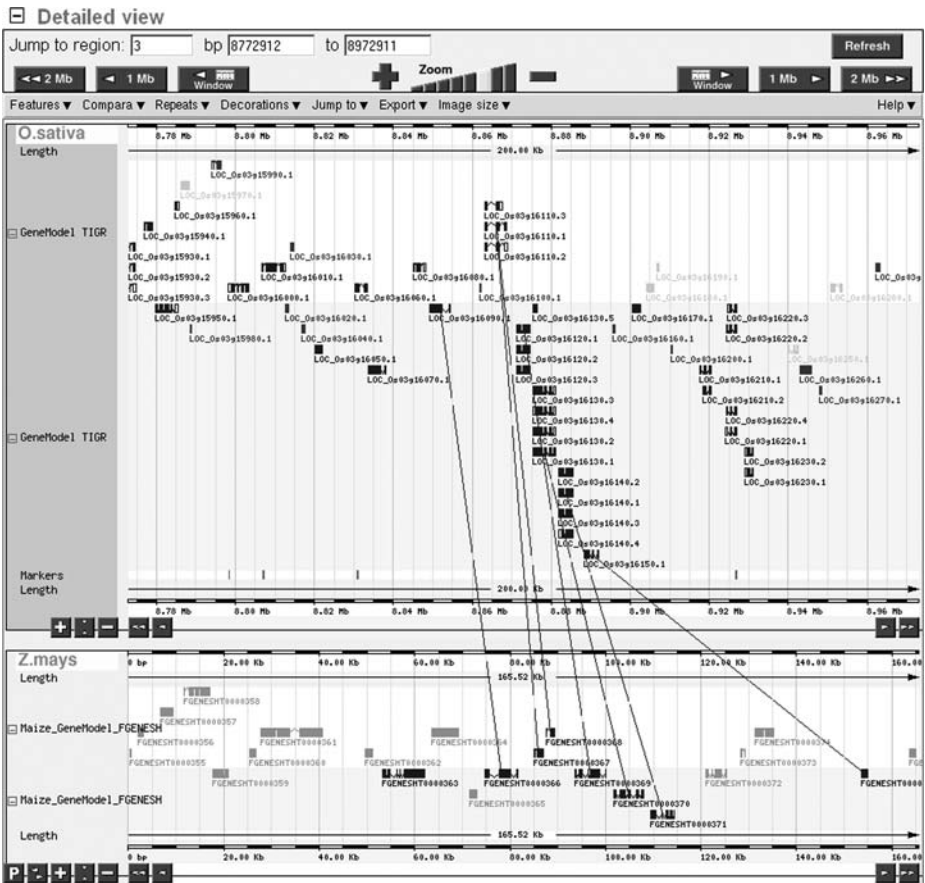
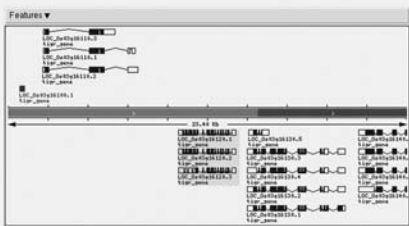


Fig. 5. MulticontigView between rice and maize.

Gene	LOC_Os03g16120 (TIGR_LOCUS ID)												
Ensembl Gene ID	LOC_Os03g16120												
Genomic Location	View gene in genomic location: 8871212 - 8874611 bp (8.9 Mb) on chromosome 3 This gene is located in sequence: AC135208												
Description	expressed protein												
Prediction Method	This gene was annotated by TIGR through a process of automatic and manual curation.												
Sequence Markup	View genomic sequence for this gene with exons highlighted												
Export Data	Export gene data in EMBL, GenBank or FASTA												
SNP information	View information about variations on this gene.												
Transcript Structure	<p>1: LOC_Os03g16120.1 (LOC_Os03g16120.1) [Transcript information] [Exon information] [Protein information] 2: LOC_Os03g16120.2 (LOC_Os03g16120.2) [Transcript information] [Exon information] [Protein information] 3: LOC_Os03g16120.3 (LOC_Os03g16120.3) [Transcript information] [Exon information] [Protein information]</p> 												
Orthologue Prediction	<p>The following gene(s) have been identified as putative orthologues by reciprocal BLAST analysis:</p> <table border="1"> <thead> <tr> <th>Species</th> <th>Type</th> <th>dN/dS</th> <th>Gene Identifier</th> </tr> </thead> <tbody> <tr> <td><i>Arabidopsis thaliana</i></td> <td>UBRH</td> <td>0.07876</td> <td>At5g59210-TIGR-G (At5g59210.2) [MultiContigView] [Align]</td> </tr> <tr> <td><i>Zea mays2</i></td> <td>UBRH</td> <td>0.21109</td> <td>PCGENESHG0000369 (Novel Ensembl prediction) [MultiContigView] [Align]</td> </tr> </tbody> </table> <p>View alignments of homologies.</p> <p>UBRH - (U)nique (B)est (R)eciprocal (H)its MBRH - one of (M)any (B)est (R)eciprocal (H)its RWS = Reciprocal Hit based on Synteny around BRH DWGA = Derived from Whole Genome Alignment</p>	Species	Type	dN/dS	Gene Identifier	<i>Arabidopsis thaliana</i>	UBRH	0.07876	At5g59210-TIGR-G (At5g59210.2) [MultiContigView] [Align]	<i>Zea mays2</i>	UBRH	0.21109	PCGENESHG0000369 (Novel Ensembl prediction) [MultiContigView] [Align]
Species	Type	dN/dS	Gene Identifier										
<i>Arabidopsis thaliana</i>	UBRH	0.07876	At5g59210-TIGR-G (At5g59210.2) [MultiContigView] [Align]										
<i>Zea mays2</i>	UBRH	0.21109	PCGENESHG0000369 (Novel Ensembl prediction) [MultiContigView] [Align]										

Transcript/Translation Summary


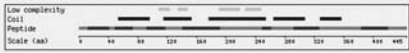
LOC_Os03g16120.1	<p>Stable ID: LOC_Os03g16120.1 Exons: 15 Transcript length: 1814 bp Translation length: 445 residues [Transcript information] [Exon information] [Protein information]</p>
Similarity Matches	<p>This Tigr_gene entry corresponds to the following database identifiers:</p> <p>EntrezGene: 3005950 RefSeq peptide: NP_012366.1 [Target %id: 85, Query %id: 92] SpTrEMBL: OS4_I00 [Target %id: 85, Query %id: 92] OS4_I00 [Target %id: 60, Query %id: 98] TIGR FN: 11973.m07049 TIGR GeneIndex: TC265437 [Target %id: 95, Query %id: 99] TIGR LOCUS: LOC_Os03g16120.1</p>
Transcript Structure	
Protein Features	

Fig. 6. Gene model page displaying gene annotation.

Sequence features also provide links to the alignment, reference databases such as GenBank, or Model Organism databases such as MaizeGDB, or project specific databases such as BarleyBase or Panzea. A graphical representation of this analysis is available from a Multicontig view (see Fig. 5). A link to the Multicontig view is available from the gene model page (see Fig. 6).

In this example the user is able to view several rice and maize genes that have their respective best reciprocal hits within the same genomic region, as

represented by the maize BAC and the rice assembly, suggesting this region is syntenic (*see Fig. 5*). In addition to the ortholog information, the user may also obtain information or links to related genomic location, transcript structure, SNPs and protein. This information can be used to suggest potential function and develop molecular reagents for marker assisted selection or genome walking.

In this example, we entered Gramene through the BLAST module and traversed to the genome browser view to view information on the related genomic region in rice. Based on the position in rice we were able to identify syntenic locations in maize on chromosomes 1 and 9. We assessed the function of the predicted orthologous rice gene and were able to ascertain that this region of the genome in rice is associated with yield, vigor, protein content, and flowering time phenotypes.

Notes

1. Apache 2.x is not supported yet due to significant differences in the persistent Perl interpreter module (mod_perl).
2. As Gramene matures in the next 18 months it is anticipated that much of the steps described here will become part of GrameneMart and will reduce the independent steps that a user would need to answer this question within Gramene.
3. Alternatively, we could enter the ContigView by selecting “[c]” in the alignment summary.
4. The tracks displayed can be adjusted by the user. For the purpose of this example, we have selected features: GeneModel_TIGR, Rice_CDS and Rice_QTL, EST; RiceJaponica_cDNA_Kome, maize, wheat, rice and sorghum and GSS; MaizeBacEnd Maize_WGS_JGI and Sorghum_Methylfiltered_Orion.
5. For genomes that are not completely sequenced, ESTs may provides the best representation of orthologous genes. In the case of a genome that is sequenced or has sequencing in progress, ESTs provide potential orthologous gene sets based on best reciprocal hit, many best reciprocal hits, and reciprocal hit based on synteny around a best reciprocal hit.

References

1. Jaiswal, P., Ni, J., Yap, I., Ware, D., Spooner, W., Youens-Clark, K., Ren, L., Liang, C., Zhao, W., Ratnapu, K., Faga, B., Canaran, P., Fogleman, M., Hebbard, C., Avraham, S., Schmidt, S., Castelevens, T.M., Buckler, E.S., Stein, L., and McCouch S. (2006) Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res.* 34, D717–D723.
2. Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., Teytelman, L., Schmidt, S., Zhao, W., Cartinhour, S., et al. (2002) Gramene: a resource for comparative grass genomics. *Nucleic Acids Res.* 30 (1), 103–105.

3. Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyra, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H.R., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, K. C., Cameron, G., Durbin, R., Cox, A., Hubbard, T., and Clamp, M. (2004) An overview of Ensembl. *Genome Res.* 14, 925–928.
4. Lomax, J. (2005) Get ready to GO! A biologist's guide to the Gene Ontology. *Brief. Bioinform.* 6, 298–304.
5. Clark, J.I., Brooksbank, C. and Lomax, J. (2005) It's all GO for plant scientists. *Plant Physiol.* 138, 1268–1278.
6. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C., Eddy, S.R. (2004) The Pfam protein families database. *Nucleic Acids Res.* 32, D138–D141.
7. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2006) Genbank. *Nucleic Acids Res.* 34, 16–20.
8. Lawrence, C.J., Dong, Q., Polacco, M.L., Seigfried, T.E., and Brendel, V. (2004) MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.* 32, D393–D397.
9. Matthews, D.E., Carollo, V.L., Lazo, G.R., and Anderson, O.D. (2003) GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res.* 31, 183–186.
10. Carollo, V., Matthews, D.E., Lazo, G.R., Blake, T.K., Hummel, D.D., Lui, N., Hane, D.L., and Anderson, O.D. (2005) GrainGenes 2.0. An improved resource for the small-grains community. *Plant Physiol.* 139, 643–651.
11. Shen, L., Gong, J., Caldo, R.A., Nettleton, D., Cook, D., Wise, R.P., and Dickerson, J.A. (2005) BarleyBase—an expression profiling database for plant genomics. *Nucleic Acids Res.* 33 (1), D614–D618.
12. Du, C.G., Buckler, E., and Muse, S. (2003) Development of a maize molecular evolutionary genomic database. *Comp Funct Genomics* 4 (2), 246–249.
13. Zhao, W., Canaran, P., Jurkuta, R., Fulton, T., Glaubitz, J., Buckler, E., Doebley, J., Gaut, B., Goodman, M., Holland, J., Kresovich, S., McMullen, M., Stein, L., and Ware, D. (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Res.* 34 (1), D752.