Gramene Comparative Genome Views: How Can Gramene Leverage Rice For The Other Grasses?

Ken Youens-Clark[1], Kiran Kumar[1], Immanuel Yap[2], Pankaj Jaiswal[2], JunJian Ni[2], Steven Schmidt[1], William Spooner[1], Wei Zhao[1], Liya Ren[1], Susan McCouch[2], Edward Buckler[3,4], Lincoln Stein[1], Doreen Ware[1,3]

[1]Cold Spring Harbor Labs, Cold Spring Harbor, NY 11724, [2]Department of Plant Breeding, Cornell University, Ithaca, NY 14853
[3]USDA-ARS NAA Plant, Soil & Nutrition Laboratory Research Unit US Plant, Soil & Nutrition Laboratory, Tower Road, Ithaca, NY 14853-2901,
[4]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853

Due to the large size and complexity of many of the cereals genomes, finished genomic assemblies are unlikely to be available in the next few years. Many of these genomes will be represented by genomic sequences, ESTs, genetic and finger print contig physical maps. While these resources themselves are useful, it is often not possible to anchor many of the sequences to a physical location in their respective genomes. By leveraging the rice genome assembly, it is possible to order and orient many unanchored cereal sequenced based upon synteny in rice. These alignments will accelerate identification of genes using traditional mapped based cloning and the development of genetic and physical marker resources. Gramene (http://www.gramene.org/), with support of funds from NSF and USDA, contributes to the development and implementation of bioinformatic resources for the plant community. One such resource, CMap, is a web-based comparative genetic and physical map tool that allows a user to dynamically generate comparative map views between the cereal genomes. Gramene implements the EnsEMBL genome browser, to display the rice assembly with anchored cereal annotations and a BLAST view. Both resources provide the users with interactive displays to link within the Gramene database and to act as a web-based portal to other genome resources. Over the next year we will be improving the CMap tool (available at http://www.gmod.org/) and supporting additional plant genome views in the EnsEMBL framework.

Software

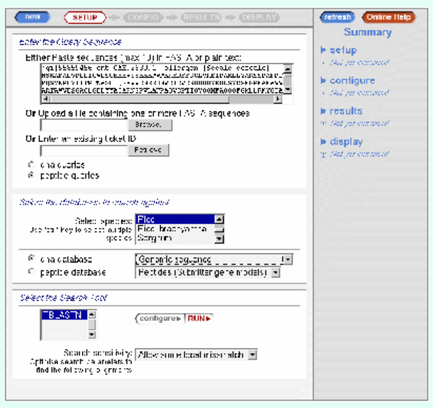
CMap and the EnsEMBL genome browser are written in Perl and rely only on open-source tools such as the Apache web server, the MySQL database, the libgd image library, and other Perl modules available on the Comprehensive Perl Archive Network (CPAN). The tools are customizable through configuration files and include tools for loading and interacting with the database.

The tools are available from the following locations:

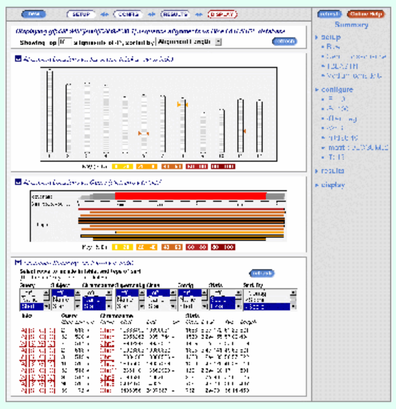
CMap	http://www.gmod.org/cmap/
Genome Browser	http://www.ensembl.org/

How do users use the Gramene website for "Comparative Genomic" analyses?

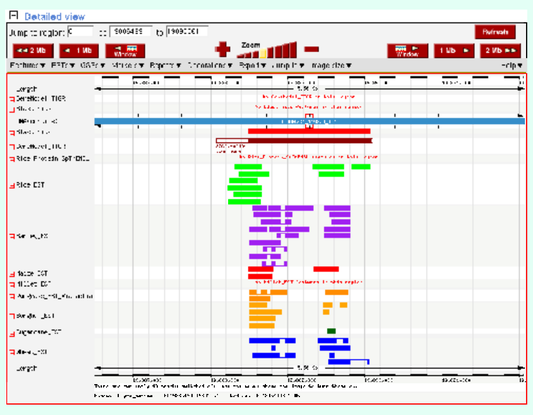
1. A TBLASTN search of a rye protein against the rice genomic sequence was carried out using BlastView. A sensitivity of "allow some local mismatch" was selected, which is a reasonable setting for the detection of hologous coding regions in related species.



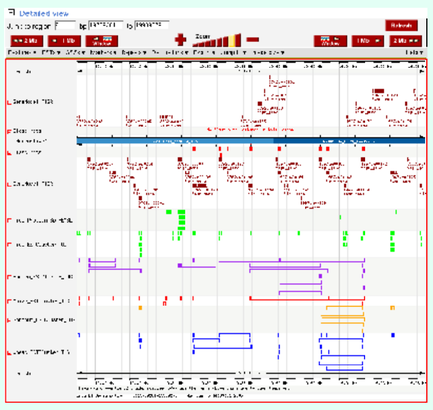
2. The top 10 alignments sorted by length are summarized in both graphical and tabular formats.



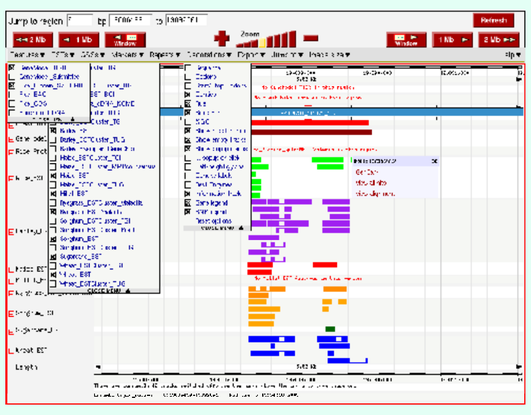
3. The link to ContigView was followed for the highest scoring alignment. This displays the region of the alignment extended by 2000 bp flanks.The ContigView display was configured to show rice gene models, proteins, and ESTs for several species. The rye protein corresponds very closely with an annotated rice gene model, and there are corresponding ESTs for all species except millet. This suggests that the protein has widespread homology across grass species. There is, however, no SpTrEMBL protein mapped to this region.

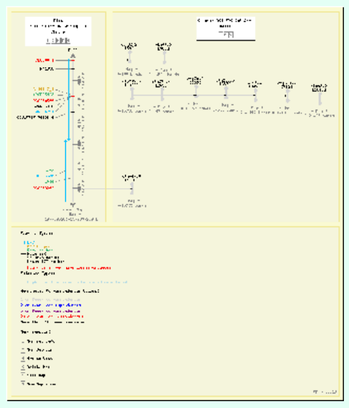


4. By using ContigView's zoom controls, a larger genomic area can be viewed - in this case 200,000bp. As the depth of overlapping EST features can be large at such resolutions, the EST tracks have been replaced by EST cluster (TUG) tracks. The rye protein sequence has seven significant alignments in this region, each spanning the greater part of the length of an annotated Rice Gene Model. This pattern suggests gene duplication in rice. The EST clusters for non-rice species are elongated as they match against multiple genomic locations probably due to the duplications in rice.



5. The CMAP TIGR assembly was displayed by following the link from ContigView's 'jump-to' menu. The Maize Curated AGI FPC map has been opened for comparison.





What 'Comparative Genome' data is there in the database?

Mapping of 3rd party sequence databases to the TIGR2 Oryza sativa (ssp. japonica) genome assembly. Results for each mapping contribute a track to the Gramene ContigView display.

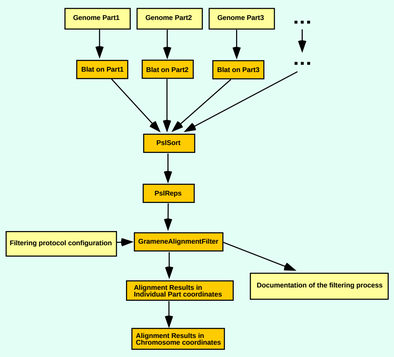
Database/Track name (SequenceType_Project)	Sequences Total Aligned Alignments
Rice	
GeneModel_TIGR:	58561 58561
Protein_SpTrEMBL:	23881 36620
BAC:	2172 12617
CDS:	1954 7864
EST:	193240 455341
ESTCluster_TGI:	51863 182863
ESTCluster_TUG:	48554 118585
Marker_RFLP:	6753 6753
Marker_SSR:	2625 3083
Marker_TOS17:	16328 16793
RiceBrachyantha	
BACend_OMAP:	34930 57825
RiceIndica	
EST_BGI:	67893 152320
ESTCluster_BGI:	19775 52443
RiceJaponica	
BACend_IRGSP:	71673 76867
cDNA_KOME:	23485 103433
RiceNivara	
BACend_OMAP:	76621 86263
RiceRufipogan	
BACend_OMAP:	53049 59178
Barley	
EST:	226577 616030
ESTCluster_TGI:	23221 80407
ESTCluster_TUG:	25867 80421
Exemplar_GeneChip:	10987 46316
Maize	
BACend:	73397 150596
EST:	175532 467840
ESTCluster_MMPcornsensus:	6778 28126
ESTCluster_TGI:	23760 83180
ESTCluster_TUG:	24175 73938
HiCot_Bennetzen:	73686 145231
HiCotCluster_TIGR:	34572 83099
HiCotMethylFilterCluster_TIGR:	52421 143021
MethylFilter_CSHL:	12559 22750
MethylFilter_Orion:	127294 255765
Mu_insert:	29544 49040
Marker:	2750 11529
Millet	
EST:	694 1836
Ryegrass	
EST_Vialactia:	10210 27825
ESTCluster_Vialactia:	4893 14790
MethylFilter_Orion:	36216 68577
MethylFilterCluster_Orion:	21859 50487
Sorghum	
CDNA:	1855 5078
EST:	121745 315686
ESTCluster_Pratt:	11683 30012
ESTCluster_TGI:	19949 64899
ESTCluster_TUG:	12522 36274
GSS_Klein:	117 235
Marker:	743 1835
MethylFilter_Orion:	35346 83526
Sugarcane	
EST:	141702 407248
Wheat	
EST:	296094 814457
ESTCluster_TGI:	47206 146388
ESTCluster_TUG:	33160 94361
Marker:	3442 9217

CMap Maps by Type	
QTL	93
Genetic	35
Physical	5
Sequence	2
Deletion	1

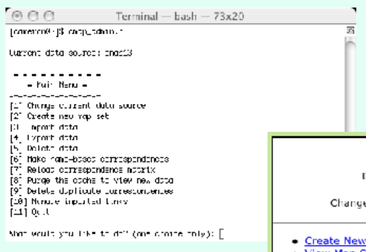
How does 'Comparative Genomic' data get into Gramene?

Alignment Pipeline of the Comparative Datasets to Rice Genome

- BLAT is used to generate alignments.
- These alignments are passed through various filtering steps which are configurable
- Documentation for the whole alignment process is generated automatically based on the configuration of alignment parameters and filtering parameters
- Different alignment and filtering protocols are used for different types of datasets.
- The filtered alignments are loaded into the browser.

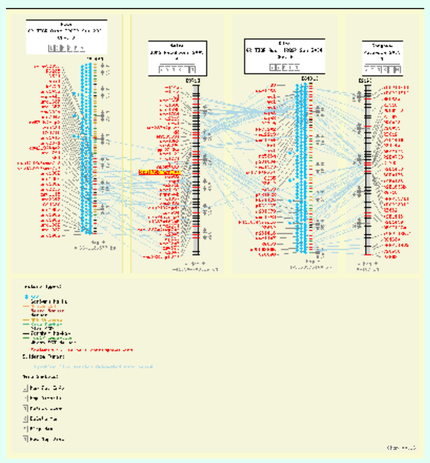


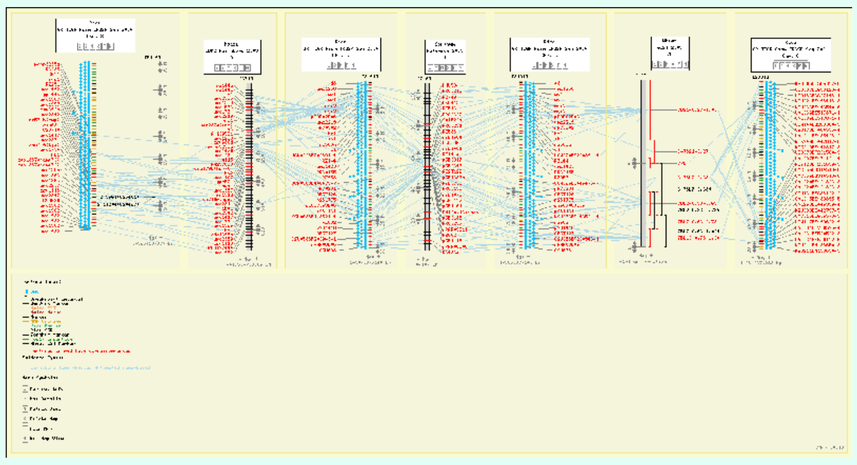
CMap has two curation main tools: a command-line tool for batch uploads and a web-based administration interface.



Other Comparative Views

Here are additional comparative views from CMap showing synetry among many species (rice, maize, sorghum and wheat)





Future Plans

- Adding a GUI front end to configure and run the analysis pipeline
- CMap 0.14 will offer more controls over how maps are selected, aligned, oriented, ordered, and ornamented
- The new 'comparative' functionality for the Genome Browser will be the addition of the Arabidopsis assembly and completed Maize contigs into Ensembl Compara

Funded By

Gramene is funded by a Plant Genome Initiative grant from the National Science Foundation, an IFAFS grant from the USDA Cooperative State Research and Education Service (CSREES), and was previously supported by a Specific Cooperative Agreement through the USDA Agricultural Research Service.