

**Mapping Relationships**  
**Among Polyploid Genomic Features found in Soybean (*Glycine max*)**

By Chet Langin\*, Jeffry Shultz, Nagajyothi Lavu, M. Javed Iqbal, Kay Shopinski,  
and David Lightfoot

Plant Biotechnology and Genomics Core-facility, Southern Illinois University, Carbondale, Illinois

**Contact:** [clangin@siu.edu](mailto:clangin@siu.edu).

**Supplementary Information:** A summary of online supplementary material is at  
<http://soybeangenome.siu.edu/soybean/summary.html>.

\*To whom correspondence should be addressed.

## **Abstract**

In bacterial artificial chromosome (BAC) based physical maps of polyploid and paleo-polyploid genomes duplicated features form a complex web of associations. BAC clones can have multiple relationships with loci, contigs, Minimum Tiling Paths (MTP's), sequences, related genes, Expressed Sequence Tags (EST's), Quantitative Trait Loci (QTL's), BAC end sequences, and other features. Each feature, can have multiple relationships with each additional features. The aim of this study was to develop an ontology and methods to describe, store, manipulate, retrieve, and display relationships found in polyploid genomes. An ontology to describe genomic relationships was developed. A SQL multi-table database was created to store and query the information. Extropy, object-oriented Perl software, was programmed with heuristic algorithms to load the database, facilitate the manipulation of the data, and create genomic flat file (GFF) output. The GFF output represented tracks suitable for displaying polyploid genomic regions in the Genome Browser (GBrowse) and other tools. This visual output greatly enables education, discovery, the generation of hypotheses, and the motivation of new methods for data analysis concerning the soybean genome. All data, source code, and the GBrowse display are available at <http://soybeangenome.siu.edu>.

Keywords Gbrowse / polyploid / physical map / Perl database / ontology

## Introduction

Soybean (*Glycine max*) is among the world's most important commercial crops, accounting for 48 percent of the world market in oil crops (Zhang, *et al*, 2004). The soybean genome is duplicated having undergone polyploidization events at least twice 3.5-5.0 MYA and 15.5-16.5 MYA (Blanc and Wolfe 2004). Southern hybridization with cDNA or hypomethylated genomic DNA probes shows 4-6 homoeologous loci per probe (Shoemaker *et al.*, 1996). There is discontinuous variation among EST paralogs (Shoemaker *et al.*, 2002; Tian *et al.*, 2004). The duplicated regions have been segmented and reshuffled (Grant *et al.*, 2000; Yan *et al.*, 2003). Sequence divergence among some duplicated regions causes PCR based markers to be able to amplify single loci from genomic DNA (Song *et al.*, 2004) but not BAC pools (Shultz *et al.*, 2005). However, about 35% of EST probes, genetic markers and BES do identify BACs from single loci (Shopinski *et al.*, 2003; Shultz *et al.*, 2005). Together these problems have delayed large scale sequencing projects.

Relationships between genomic features are indicators for placing duplicated features onto a physical map. In recently polyploid or duplicated genomes many of the same types of features fall into the same place, causing conflicts (Cheung, *et al*, 2003; Wu *et al.*, 2004; Shultz *et al.*, 2005). The result is a complex web of associations between features.

The relationships between a locus (DNA marker) and a BAC (also called a *clone*, herein) include the presence or absence of hybridization, where *hybridization* means the similarity of sequences that exceed 75 percent of nucleotide identity over more than 20 base pairs (Shultz *et al.*, 2005). If a locus and BAC hybridize, then the locus and the BAC *have a relationship* and the potential approximate location of the BAC can be determined directly by the location locus (Cone, *et al*, 2002). If the locus and the BAC do not hybridize then these features *do not have a relationship* and the location of the BAC cannot be determined directly by the locus. In a polyploid genomes a BAC can possibly have a relationship with more than one locus (Shoemaker, *et al*, 1996).

A locus does not have a direct relationship with a contig. However, the relationship between a locus BAC and a contig is inferred via fingerprinting and FPC (Arizona Genomics

Institute, 2003): A BAC might have a starting band and an ending band in a contig (Soderlund, *et al*, 2000; and Soderlund, *et al*, 1997). A BAC may only have a relationship with one contig. But, a contig has relationships with many BAC's. A BAC starts at a specified band in a contig and ends at another specified band in a contig but orders in FPC are approximate particularly at contig ends.

An ideal situation is a ternion that occurs with a locus related to a BAC which is related to a contig (Soderlund, *et al*, 2000). Composite features can all be given potential locations in such a case. Additionally, any other related features, such as EST's, related genes, sequences, and BAC clone minimum tiling paths (MTP's), can also be given potential locations. However, centiMorgan distances compress and expand compared to genetic distances. Therefore, incompatible features may appear to improperly overlap or be widely spaced, and the order of some features is not certain. In addition, biological data does not all fit neatly into organized ternions. Experiments are rarely error free (Bhandarkar, *et al*, 2001). False-positive and false-negative hybridization errors, the presence of chimeric clones, and gaps in library coverage lead to ambiguity and error in the clone order (Hall, *et al*, 2001; and Shultz, *et al*, 2003). There are several types of conflict: Probes detect only a single BAC in a contig; a probe that maps to one genetic location also maps BAC's that have been assembled into more than one contig; and, two probes corresponding to distinct genetic locations associate with a single contig (Cone, *et al*, 2002).

The Genome Browser GBrowse (Stein, *et al*, 2002) has been used to visually display polyploid soybean genomic features and their relationships in potential estimated locations. A *Glycine max* GBrowse physical map involved the coordinated placement of loci, BAC's, contigs and other features. Three views of contigs and their BAC's are provided; forward, reverse, and spread.

## **Materials and Methods**

### **Systems and Methods**

A LAMP system was used comprised from Linux, Apache, MySQL, and Perl. We created a novel database and software. Programs were written in text mode so that the software can be readily accessed over the Internet via the Secure Shell (ssh).

### **Ontology**

Types of relationships were labeled so that they could be processed in the database and the programming code. These labels are alpha, beta, and gamma (Figure 1). An alpha feature does not have any relationships with any other features. A beta feature has a relationship with one other feature. A gamma feature is part of, or the basis of, a ternion. To further clarify BAC's, a beta<sub>1</sub> BAC has a relationship with a locus, but not with a contig. A beta<sub>2</sub> BAC is part of a contig, but does not have a relationship with a locus. In the database and programming code, alpha, beta, and gamma are spelled out as special characters are not supported.

A gamma or beta locus is also called an *anchor*, because it anchors a BAC. A gamma BAC is also called an *anchor* because it indirectly anchors a contig. The genetic map locations of the loci are used to determine the locations of the other features that are directly or indirectly related to the loci.

### **Genomic Framework**

The loci locations were obtained from the USDA (Cregan, *et al*, 1999; and Cregan, *et al*, 2003) and SoyBase (Marek, *et al*, 1996; and Iowa State University SoyBase, 2003). The BAC's were obtained in laboratories at Southern Illinois University at Carbondale, Texas A&M University, the University of Minnesota (Danesh, *et al*, 1998), and Iowa State University (Marek, *et al*, 1996). The contigs were obtained via FPC (Soderlund, *et al*, 2000; and Soderlund, *et al*, 1997) from builds at Southern Illinois University at Carbondale (Shultz, *et al*, 2005) and Texas A&M University (Wu, *et al*, 2004).

### **Database Components**

The Soybean GBrowse Database consists of 27 tables containing the various relationships between the genomic features (Figure 2). Our custom Extropy program loads these tables from 10 various sources of information. The Extropy program then appropriately manipulates this data and writes a GFF file suitable for loading into GBrowse. Mouse clicking on many features in GBrowse accesses a custom CGI Perl script, named *detail*, that re-accesses the database in order to produce a web page with more detailed information about that feature. Mouse clicking on other features accesses additional information either in SoyBase or GenBank.

Typical tables contain relationships between clones and loci, clones and contigs, EST's and clones, MTP's and clones, related genes and clones, related genes and loci, sequences and clones, sequences and loci, molecular linkage groups (MLG's) and their lengths, loci and their locations, clones and their anchored locations, clones and their placed locations, contigs and their anchors, contigs and their placed locations, contig end matches and associated information, and QTL's and their locations. Additional tables contain other necessary information, such as audit trails. Figure 1 illustrates just some of the complex relationships between table fields in our database (*EID* in each case stands for *Entry ID*.)

### **Database Access and Algorithms**

All data, source code, and the GBrowse display are available at <http://soybeangenome.siu.edu>. A summary of the available online supplementary material is at <http://soybeangenome.siu.edu/soybean/summary.html>. A link to a MySQL dump file of the database is on the summary web page. This dump can be used to reinstall the database at one's own site. The algorithms create and manipulate the database are explained at that site. Briefly, the Extropy directory structure begins with a directory we named gbrowse plus subdirectories organized as follows: The main file, extropy, in the gbrowse/round4 subdirectory, merely calls the MenuMain.pm module, in the gbrowse/round4/Extropy/Extropy subdirectory, which displays the Main Menu and obtains selections from the user. Then, MenuMain.pm calls whichever other module is appropriate to process the user's selection. All of these other modules are in the gbrowse/round4/Extropy/Extropy subdirectory. Two general purpose modules were created: ExtropyConstants.pm for constants, and ExtropyUtils.pm, both in the gbrowse/round4/Extropy subdirectory, for common subroutines. All of the beginning data is in the gbrowse/given\_data

subdirectory and all of the intermediate data is placed in the `gbrowse/working_data` subdirectory. The configuration files and the final GFF file are written to the `gbrowse/round4` subdirectory.

Extropy organizes the genomic data with four broad categories; activating a project; reading the data; processing the data; and creating the GFF file. Briefly, activation creates 25 empty database tables for later use: `bad_clone2locus`, `bad_qtl`, `clone2locus`, `clone2locus2`, `clone2locus3`, `clone_anchors`, `clone_locations`, `clone_loci`, `clone_loci2`, `clones`, `confirmed`, `contig2clone`, `contig_anchors`, `contigs`, `contigs_onQ`, `end_counts`, `end_matches`, `est`, `loci`, `mlg`, `mtp`, `qtl`, `qtl2locus`, `related`, and `sequence`. Every single table has an EID field, for *Entry ID*, which uniquely identifies that entry in the table.

Reading the data: Extropy currently reads 10 types of input data files: MLG, loci, FPC, QTL, end matches, MTP, EST, sequence, related genes, and confirmeds. Data files come from a variety of sources, including Excel spreadsheets, and other databases. Excel spreadsheets are first saved as tab-delimited text files. A Perl module was written for each type of data input.

Processing the Data: Extropy has six Perl modules to prepare the data for further use: `CrosscheckLocusNames.pm`, `UpdateLocusNames.pm`, `CountCloneAnchors.pm`, `UpdateCloneAnchors.pm`, `CountContigAnchors.pm`, and `UpdateContigAnchors.pm`. Each one of these modules corresponds to a Main Menu item and they must be selected in order. The user corrects as many locus names as possible and puts the corrections into a `locus_name_changes.txt` file

Creating the GFF File: A single module, `Gff.pm` writes the GFF file. MLG data must be used. Other types of data can be turned on or off allowing some customization of the GFF file.

## Results and Discussion

Figure 3 shows a resulting abbreviated GBrowse display for MLG G from bases 720,000 to 1,220,000. Sample relationships are highlighted in red. The relationships between features becomes apparent in this view. For example, following the red lines, Sequence 38286516 is placed according to the location of Beta 2 Clone B36J21, which is placed according to the location of Contig 3325, which is placed according to the location of Gamma Clone H20J22, which is placed according to the location of Locus Satt275. If the location of Satt275 is changed in the database, the Perl Extropy program also adjusts all of these other feature locations accordingly.

All of the known relationships between all of the features have been stored in our database. As we obtain more accurate data, such as a different relationship, a new locus, or a different location for a locus, Extropy adjusts all of the locations accordingly, simply by the user running the Main Menu from top to bottom. Further, the modular construction of Extropy is scalable. New types of features can be added by programming new modules and making relatively simple adjustments to existing modules. Figure 4 shows that the GFF file output can be used to support other formats and databases.

The features of a physical map, in a GFF (Sanger Institute, 2003) file, are entered into a database which GBrowse can access to display views of up to a million bases at a time, or down to individual bases. GBrowse necessarily displays data more accurately than empirical data error would justify. Often singular base pair accurate data is not available (Engler, *et al*, 2003). We use a custom database for processing the data as well as the GBrowse database. The custom database uses Boyce-Codd Normal Form (Codd, 1974) adapted for speed. Some tables are duplicated before manipulation in order to maintain an audit trail. Our centiMorgan to base conversion method was obtained from Mouse Genetics (Silver, 1995).

Much more can be accomplished in this area of research. The algorithms should be expanded to include more indirect types of relationships, which would assist in determining specific duplicated areas of the genome. The Perl script could be rewritten with a graphical interface and generalized for use by other localities. The Perl script should be rewritten to take

better advantage of subroutines and objects. One could see if there is a way to place the features using simultaneous equations. A scoring system should be devised to determine on which MLG's to place a feature with multiple anchors.

### **Acknowledgements**

This research was funded in part by a grant from the NSF 9872635, NSF 0405819, ISPOB 02-127-03 and USB 2228-4228. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, Illinois Soybean Board or United Soybean Board. The continued support of SIUC, College of Agriculture and Office of the Vice Chancellor for Research to MJJ and the Genomics Core Facility is highly appreciated. The authors thank Dr. Bill Beavis for sharing unpublished results. J. Shultz is thanked for technical assistance. Bill Beavis and Andrew Farmer are thanked for instituting CMAP at LIS.

## References

- Arizona Genomics Institute, BioFPC, <http://www.genome.arizona.edu/software/fpc>, 11/6/03.
- Bhandarkar, S.M., Machaka, S.A., Shete, S.S., and Kota, R.N. (2001) Parallel Computation of a Maximum-Likelihood Estimator of a Physical Map, *Genetics*, **157**, 1021-1043.
- Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*. 16:1667-1678.
- Cheung, J., Wilson, M.D., Zhang, J., Khaja, R., MacDonald, J.R., Heng, H.H.Q., Koop B.F., and Scherer, S.W. (2003) Recent Segmental and Gene Duplications in the Mouse Genome, *Genome Biology*, **4**, R47.
- Codd, E.F. (1974) Recent Investigations into Relations Data Base Systems, *Proceedings of the International Federation for Information Processing (IFIP) Congress*, 1974.
- Cone, K.C., McMullen, M.D., Bi, I.V., Davis, G.L., Yim, Y.-S., Gardiner, J.M., Polacco, M.L., Sanchez-Villeda, H., Fang, Z., Schroeder, S.G., Havermann, S.A., Bowers, J.E., Paterson, A.H., Soderlund, C.A., Engler, R.W., Wing, R.A., and Coe, Jr., E.H. (2002) Genetic, Physical, and Informatics Resources for Maize, on the Road to an Integrated Map, *Plant Physiology*, **130**, 1598-1605.
- Cregan, P.B., Jarvik, T., Bush, A.L., Shoemaker, R.C., Lark, K.G., Kahler, A.L., Kaya, N., VanToai, T.T., Lones, D.G., Chung, J., Specht, R.E. (1999) An Integrated Linkage Map of the Soybean Genome, *Crop Science*, **39**, 1464-1490.
- Grant, D., Cregan, P., Shoemaker, R.C. (2000) Genome Organization in Dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*, *Proc. Natl. Acad. Sci. USA*, **94**, 4168-4173.
- Cregan, P.B., Jarvik, T., Bush, A.L., Shoemaker, R.C., Lark, K.G., Kahler, A.L., Kaya, N., VanToai, T.T., Lones, D.G., Chung, J., Specht, J.E. (2003) An Integrated Genetic Linkage Map of the Soybean, [http://bldg6.arsusda.gov/~pooley/soy/cregan/soy\\_map1.html](http://bldg6.arsusda.gov/~pooley/soy/cregan/soy_map1.html), 11/6/03.
- Danesh, D., Penuela, S., Mudge, J., Denny, R., Nordstrom, H., Martinez, J. and Young, N. (1998) A bacterial artificial chromosome library for soybean and identification of clones near a major cyst nematode resistance gene, *Theor Appl Genet*, **96**, 196-202.
- Engler, F.W., Hatfield, J., Nelson, W., and Soderlund, C.A. (2003) Locating Sequences on FPC Maps and Selecting a Minimum Tiling Path, *Genome Research*, **13**, 2152-2163.
- Hall, D., Bhandarkar, S.M., and Wang, J. (2001) ODS2: A Multiplatform Software Application for Creating Integrated Physical and Genetic Maps, *Genetics*, **157**, 1045-1056.
- Iowa State University Soybase, <http://129.186.26.94>, 11/6/03.
- Keim, P., Diers, B.W., Olson, T.C., and Shoemaker, R.C. (1990) RFLP Mapping in Soybean: Association Between Marker Loci and Variation in Quantitative Traits, *Genetics*, **126**, 735-742.
- Marek, L.F., and Marek, R.C. (1996) Construction and size characterization of a bacterial artificial chromosome (BAC) library from soybean, *Soybean Genet. Newslet.*, **23**, 126-129.
- Meyers BC, Scalabrin S, Morgante M. 2004. Mapping and sequencing complex genomes: let's get physical! *Nat Rev Genet*.5(8):578-588.
- Sanger Institute, <http://www.sanger.ac.uk/Software/GFF>, 11/4/03.
- Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J.P., Kochert, G., and Boerma, H.R. (1996) Genome Duplication in Soybean (*Glycine* subgenus *soja*), *Genetics*, **144**, 329-338.
- Shopinski, K., J. Iqbal, J. Yuan, A. Afzal, R. Ahsan, J. Shultz, K. Meksem, and D. Lightfoot. 2003. EST integration with the soybean physical map. *Agronomy Abstracts* 102:388.

- Shultz, J., Meksem, K., and Lightfoot, D. (2003) Evaluating Physical Maps by Clone Location Comparisons, *Genome Letters*, **2**, 98-105.
- Shultz, J., Meksem, K., Langin, C., Zobrist, K., Lavu, N., Iqbal, J., Potter, J., Yesudas, C., Watson, D., Wu, C., Zhang, H.-B., Town, C., and D.A. (2005) Browsing the soybean genome: physical map builds from recently duplicated genomes, *Mol. Gen. Genome*, submitted for publication.
- Silver, L.M., (1995) *Mouse Genetics: Concepts and Applications*, Sec. 5.1.3., Oxford University Press, 1995, adapted for the web, <http://www.informatics.jax.org/silver/frame5.1.shtml>, 11/6/03.
- Shoemaker, R.C., K. Polzin, J. Labate, J. Specht, E.C. Brummer, T. Olson, N. Young, V. Concibido, J. Wilcox, J.P. Tamulonis, G. Kochert, and H.R. Boerma. 1996. Genome duplication in soybean (*Glycine subgenus soja*). *Genetics* 144:329-38.
- Soderlund, C., Humphrey, S., Dunham, A., French, L. (2000) Contigs built with fingerprints, markers and FPC v4.7, *Genome Research*, **10**, 1772-1787.
- Soderlund, C., Longden, I., Mott, R. (1997) FPC: a system for building contigs from restriction fingerprinted clones, *Computational Applied Bioscience*, **13**, 101-106.
- Stein, L.D., Mungall, C., Shu, S.Q., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., and Lewis, S. (2002) The Generic Genome Browser: A Building Block for a Model Organism System Database, *Genome Research*, **12**, 1599-1610.
- Tian, A.G., J. Wang, P. Cui, Y.J. Han, H. Xu, L.J. Cong, X.G. Huang, X.L. Wang, Y.Z. Jiao, B.J. Wang, Y.J. Wang, J.S. Zhang, and S.Y. Chen. 2004. Characterization of soybean genomic features by analysis of its expressed sequence tags. *Theor Appl Genet* 108:903-13.
- Wu, C., Sun, S., Nimmakayala, P., Santos, F.A., Springman, R., Ding, K., Meksem, K., Lightfoot, D.A., and Zhang, H.B. (2004) A BAC and BIBAC-based Physical Map of the Soybean Genome, *Genome Research*.
- Yan, H.H., J. Mudge, D.J. Kim, D. Larsen, R.C. Shoemaker, D.R. Cook, and N.D. Young. 2003. Estimates of conserved microsynteny among the genomes of *Glycine max*, *Medicago truncatula* and *Arabidopsis thaliana*. *Theor Appl Genet* 106:1256-65
- Zhang, W.-K., Wang, Y.-J, Luo, G.-A., Zhang, J.-S, He, C.-Y, Wu, X.-L., Gai, J.-Y., and Chen, S.-Y. (2004) QTL Mapping of Ten Agronomic Traits on the Soybean (*Glycine max* L. Merr.) Genetic Map and their Association with EST Markers, *Theor Appl Genet*, **108**, 1131-1139.

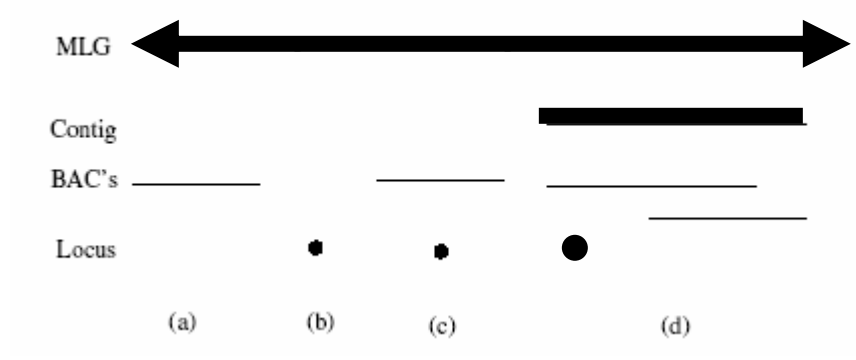
Figure 1: Ontology for genomic features. Panel A: Feature (a) shows a alpha clone (thin black line). Feature (b) shows a gamma marker (black dot). Feature (c) shows a gamma marker and clone. Feature (d) shows a gamma marker, a gamma clone, a gamma contig and a beta clone. Panel B shows two ternions from conserved duplicated genomic regions anchored to one gamma marker. Panel C shows the arrangement of the ternions equally spaced from the anchor. Panel D shows an example of boats on a lake where marker Satt382 identifies 3 regions of the genome by hybridizing to 3 BACs assigned to different contigs. Relationships of this type are common and are catalogued in Shultz et al., 2005 supplementary information. All three contigs are shown at each anchor location aligned and spread.

Figure 2: Illustrates just some of the complex relationships between table fields in our database. *EID* in each case stands for *Entry ID*. Typical tables contain relationships between clones and loci, clones and contigs, EST's and clones, MTP's and clones, related genes and clones, related genes and loci, sequences and clones, sequences and loci, molecular linkage groups (MLG's) and their lengths, loci and their locations, clones and their anchored locations, clones and their placed locations, contigs and their anchors, contigs and their placed locations, contig end matches and associated information, and QTL's and their locations. Additional tables contain other necessary information, such as audit trails.

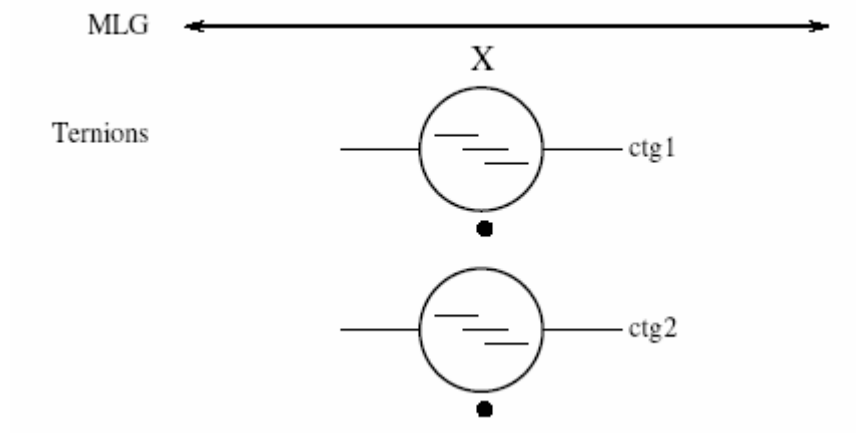
Figure 3: An abbreviated GBrowse display for MLG G from bases 720,000 to 1,220,000. Sample relationships are highlighted in red: Sequence 38286516 is a BAC end sequence labeled as belonging to Clone B36J21. This clone is similarly labeled as having Sequence 38236519, as well as another sequence. Clone B36J21 is also labeled as having a beta relationship, determined by FPC, with Contig 3325. The only gamma clone showing in this area is Clone H20J22. Clicking on Contig 3325 would also indicate Clone H20J22 as being the gamma clone. Clone H20J22 is labeled as having hybridized with Locus Satt275, whose approximate location is known. All of these mentioned features have been placed based on the location of Satt275.

Figure 4: Demonstration of the use of the Gbrowse GFF file for the institution of a CMap representation of the soybean genome at NCGR in LIS.

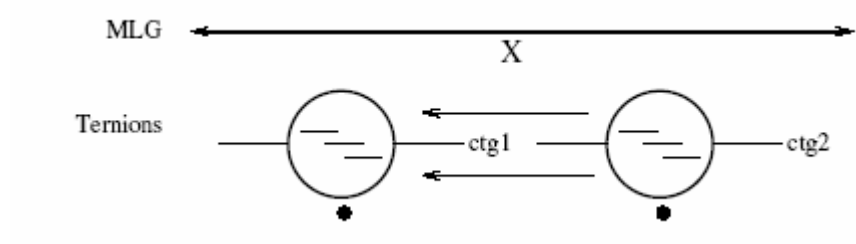
Figure 1  
A.



B.



C.



D.

### G. max, Version 4, SIU 2003

Showing 10 Mbp from A1, positions 4,950,000 to 14,949,999

**Instructions [Hide]:** Search using a sequence name, gene name, locus, or other landmark. The wildcard character \* is allowed. Example: A1:9500000..10400000, A2:1..10000000, B1:1..10000000, B2:1..10000000, C1:1..10000000, C2:49000000..51000000, D1AQ:19000000..23000000, D1BV:1..10000000, D2:1..10000000, E:1..10000000, F:1..10000000, G:1..10000000, H:1..10000000, I:1..10000000, J:1..10000000, K:1..10000000, L:1..10000000, M:1..10000000, N:1..10000000, O:1..10000000, Queue:1..10000000. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position. To save this view, bookmark link.



*This dataset is Version 4. Methodology and further information.*

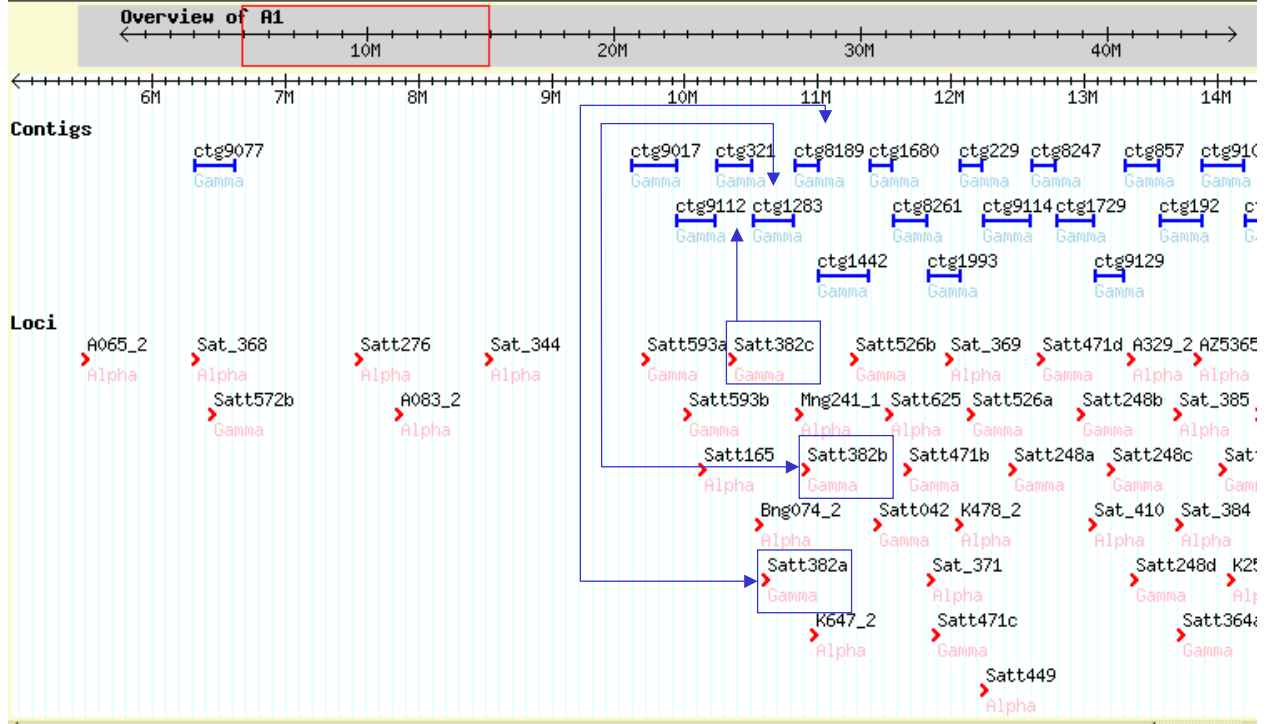


Figure 2

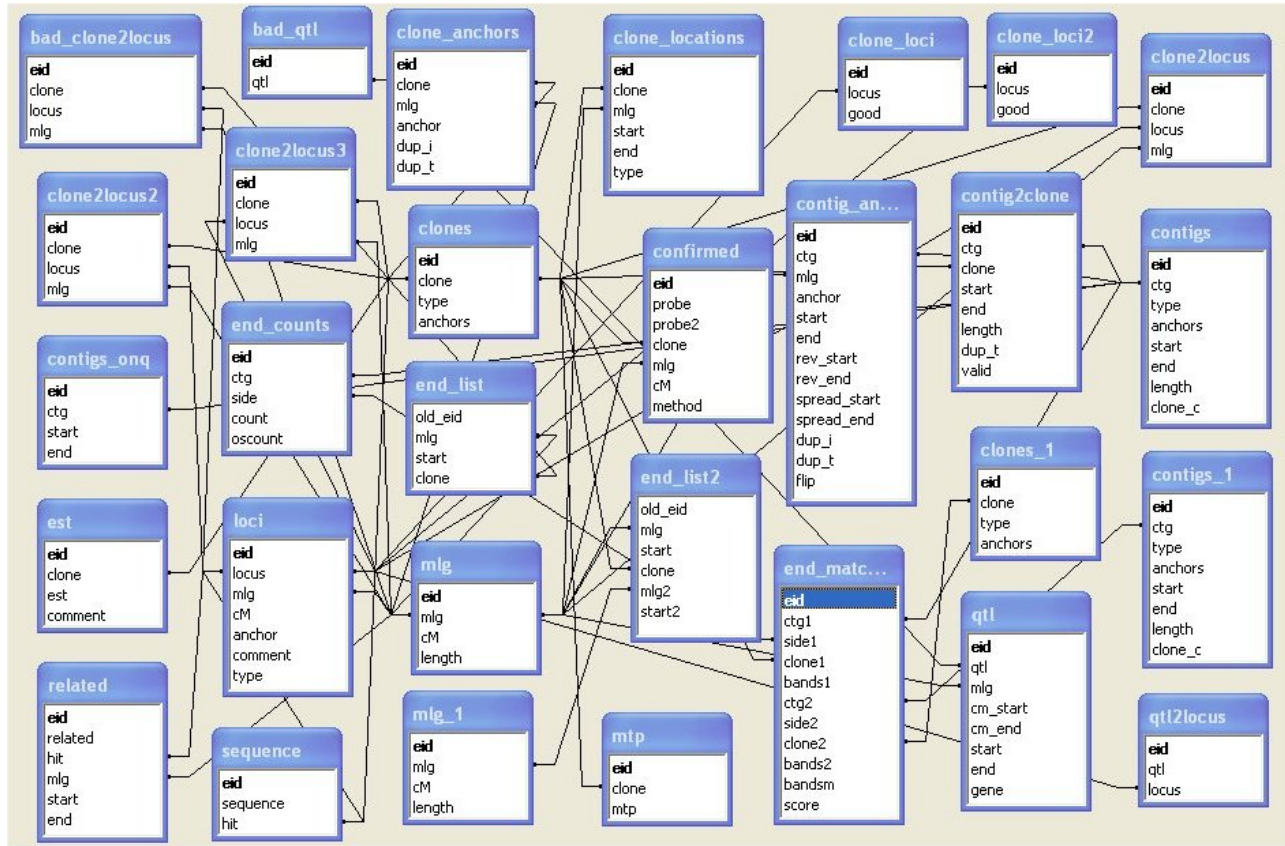


Figure 3

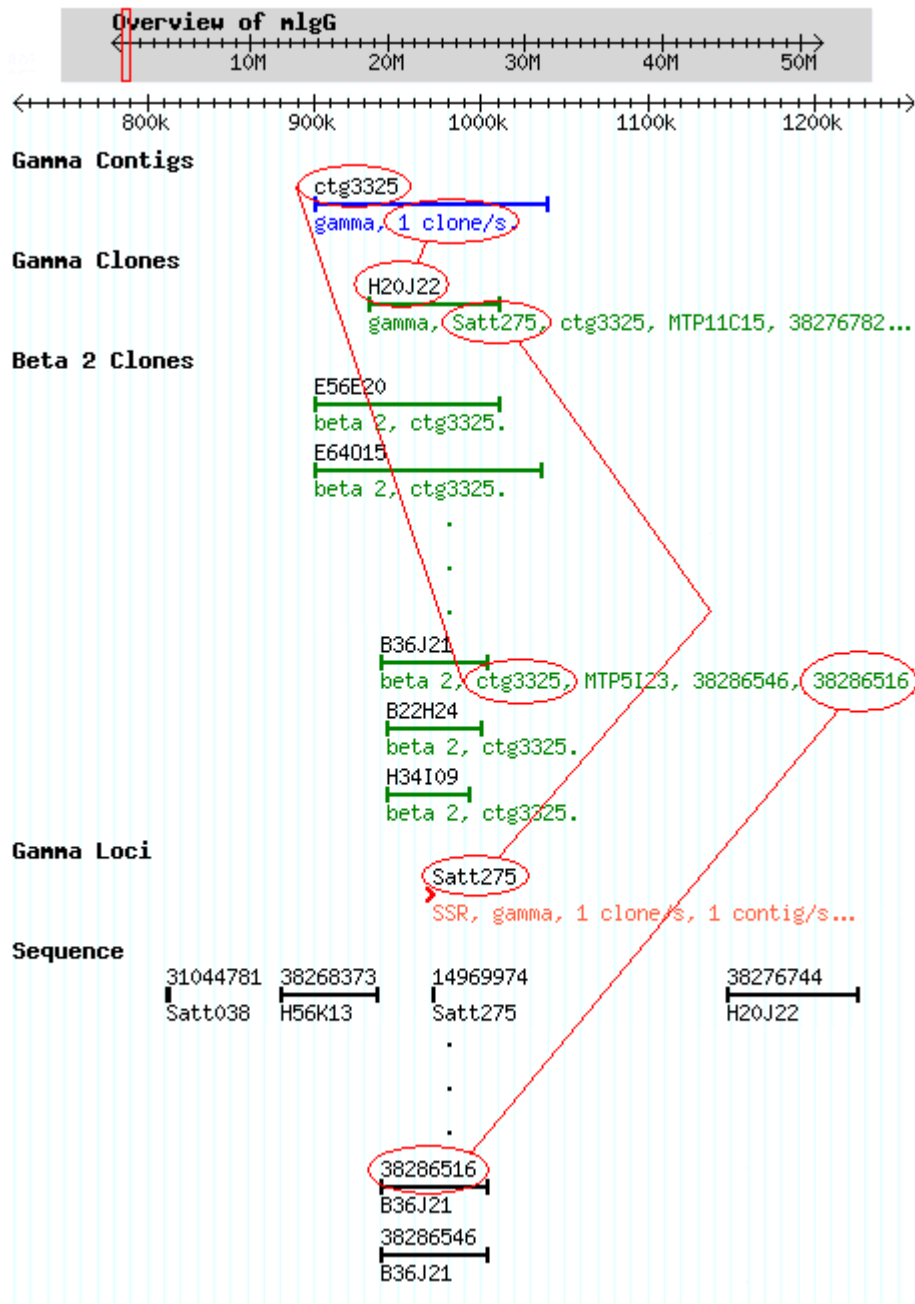
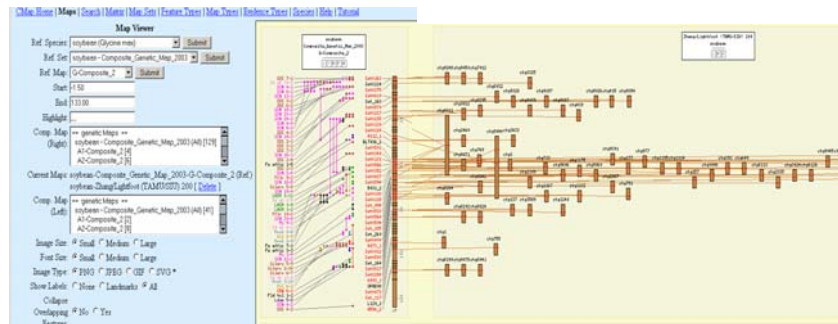


Figure 4.

**Figure 4: C-map representation of the soybean physical map found at NCGR**

**A. Linkage group A1**



**B. Linkage group G.**

