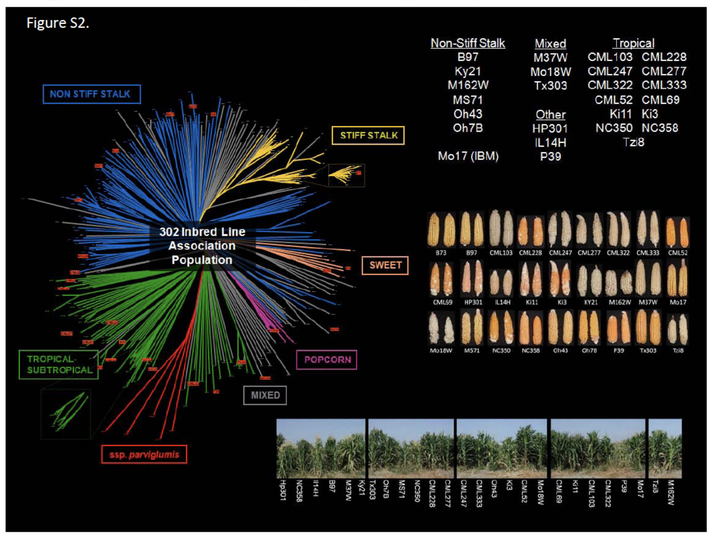
THE GRAMENE MAIZE PANGENOME BROWSER

Sharon Wei1, Joshua C Stein1, Andrew Olson1, Yinping Jiao1, Bo Wang1, Michael Campbell1, Marcela K. Tello-Ruiz1, Doreen Ware 1,2
1 Cold Spring Harbor Laboratory; Cold Spring Harbor, NY, USA 11724
2 USDA ARS NEA Plant, Soil & Nutrition Laboratory Research Unit; Ithaca, NY, USA 14853

ABSTRACT: Maize is the most genetically diverse crop in the world, with differences in gene content estimated between 5-20% among lines. Capturing the pan-Zea gene space and structural variation requires additional reference genomes, and the infrastructure to store, analyze and make accessible. To support this effort, Gramene has developed a dedicated genus-level browser resource: maize-pangenome.gramene.org, built upon the Ensembl infrastructure and guided by FAIR practices. Our first pass of this resource includes B73, W22 and PH207 complete reference genomes, along with 4 monocots, 3 dicots, 3 lower plants and 2 non-plant outgroup species. These served as input to generate phylogenetic resources based on protein and whole-genome DNA alignments. Insights into ancestrally conserved regions and structural rearrangements are defined by pairwise whole-genome alignments and displayed in a number of informative ways, including a multi-species view that allows graphical stacking of browsers and interspecies navigation. The gene trees can be used to programmatically identify gene expansions and losses between different maize accessions, which may explain evolutionary adaptations, inaccuracies in the gene models, or errors in the underlying reference genome assemblies. We anticipate maize accessions like the NAM populations being added to this resource. To test the utility of these resource and to assess quality of the gene structure predictions, Gramene outreach efforts include the first maize annotation jamboree co-organized with the MaizeCODE project. This work constitutes an initial prototype to support the infrastructure to identify misannotated gene structures and a process to correct these guided by the gene trees. In addition to providing resources to support quality assessment, as well as insights into many outstanding questions in the evolutionary history of the *Zea* genus, this resource will provide a basis for functional characterization of genes and the identification of targets for agronomic improvement of maize. This project is funded by NSF (IOS-1127112) and partially from USDA-ARS (1907-21000-030-00D).

Maize NAM Parents

A genetic resource for crop improvement

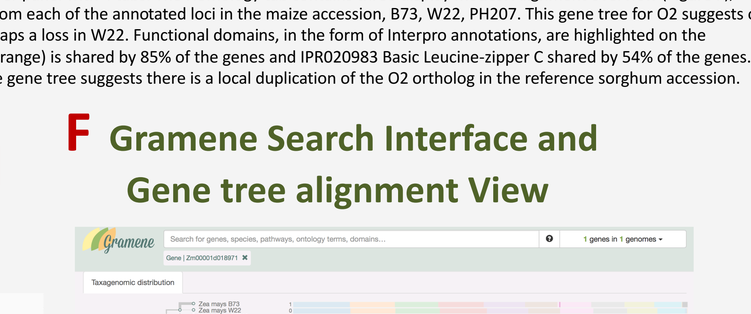


McMullen, Michael D., Stephen Kresovich, Hector Sanchez Villeda, Peter Bradbury, Huihui Li, Qi Sun, Sherry Flint-Garcia, et al. 2009. "Genetic Properties of the Maize Nested Association Mapping Population." *Science* 325 (5941): 737–40.

Maize genomes at panMaize

- Zea_mays (B73): Zm-B73-REFERENCE-GRAMENE-4.0
- Zea_maysw22: Zm-W22-REFERENCE-NRGENE-2.0
- Zea_maysph207: Zm-PH207-REFERENCE_NS-UIUC_UMN-1.0

Reference genomes for gene tree analysis:
3 maize lines: B73, W22 and PH207
4 grass outgroups
3 dicots
3 lower plants
2 non-plant outgroup species

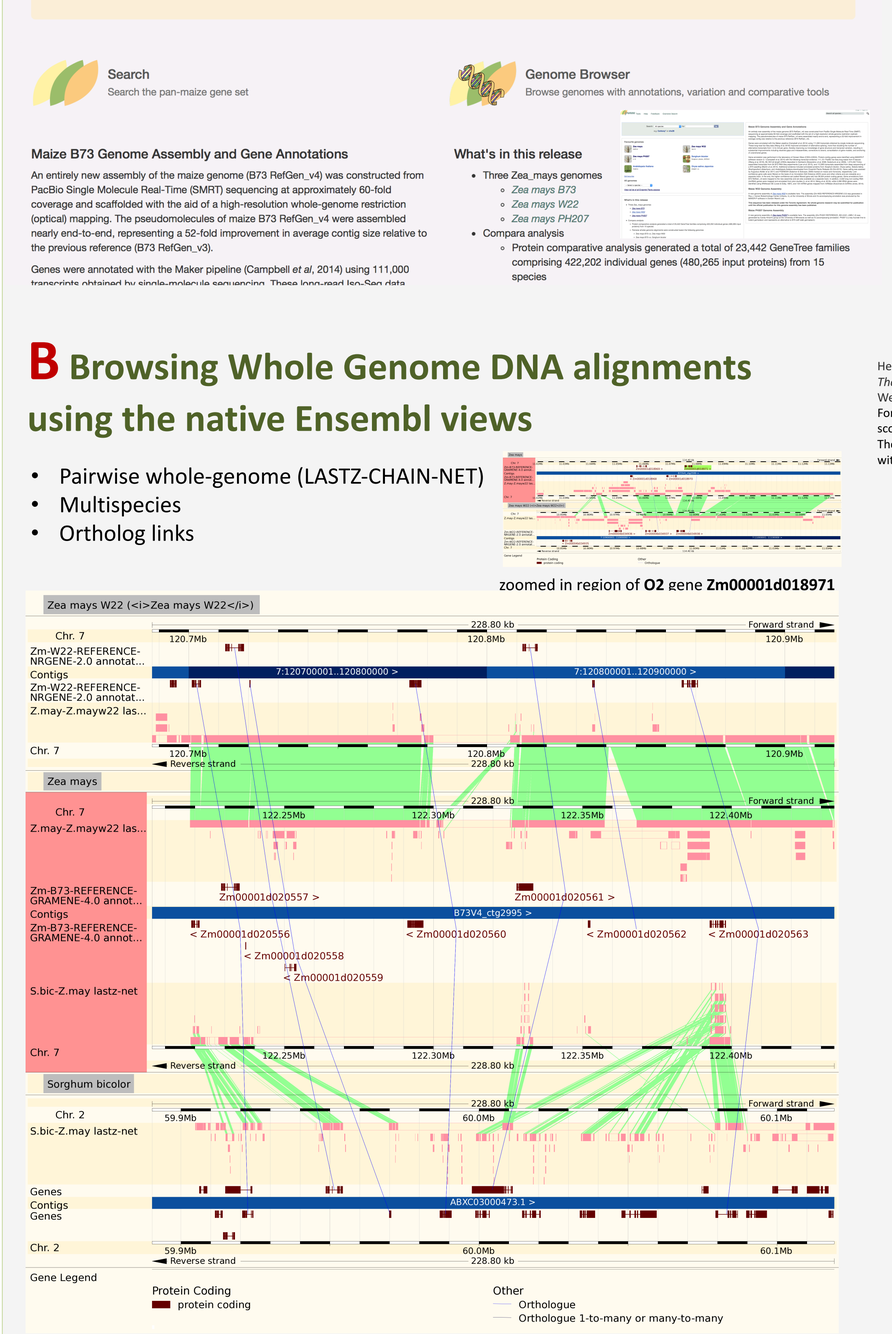


This is the beta release of the website (maize-pangenome.gramene.org). The platform will be used to house the reference assemblies from recently funded NAM reference genomes project (Dawe PI), as well as the encode data sets from Maize Code project (Gingeras PI). The Maize Pan Genome website (Figure A) is built using the Ensembl & Gramene infrastructure as a proof of concept deployment of pan-genome resource. The Ensembl infrastructure is used to store and access the public reference genome assemblies, gene annotations, pre-computed whole genome DNA alignments (Figure B) and gene trees (Figure D) utilizing the native Ensembl views. The global search capabilities, and novel comparative views are based on the Gramene interface (Figure F, G, H). The data are also available through the Gramene API.

The foundation of the resource is the ability to host the reference assemblies and pre-computed phylogenomic data in the form of whole genome DNA alignments and gene trees based upon the Ensembl Compara pipeline (Figure C). Pre-computed whole genome alignments provide insights on conservation of coding and non-coding regions. For the maize Pan Genome browser, we included whole DNA alignments between maize accessions and sorghum (Figure B). The gene trees provide information on paralogs and orthologs. For the maize pan genome browser we have included 4 monocots, 3 dicots, 3 lower plants, 2 non-plants for phylogenetic reference and functional annotation (Top panel). Using the gene tree information and genomic context, we have developed a workflow which characterizes candidate gene structural miss-annotations. In this example, using the native Ensembl gene tree view, we see a W22 example of a split gene (Figure D). The genes, Zm00004b008634 and Zm00004b008635, each align to the C-terminal and N-terminal halves of the consensus protein alignments, indicating a potential miss-annotation, where a gene is split into two separate gene models. The synteny maps, available through the Ensembl views, provide high-level views of conservation (E). In this example, we are looking at synteny between B73 chr7, PH207, W22, and Sorghum. These views are generated based on the orthologs derived from the gene trees, and provide interactive links to ortholog tables. An alternative view for micro level synteny is provided through the Gramene search interface Homology Gene Neighborhood view. An example is shown on the right.

To highlight the functionality of the Gramene search and views, maize gene opaque endosperm2 (O2) is used as an example. In this case we searched with the gene ID and the results include the top-level view of the genomes and a lower panel with 4 tabs. The Homology tab contains three display modes: Alignment overview (Figure F), Multiple sequence alignment (Figure G), and Neighborhood conservation (Figure H). In each display mode you can interactively expand or collapse nodes in the tree. The gene trees are built using the longest protein coding transcript from each of the annotated loci in the maize accession, B73, W22, PH207. This gene tree for O2 suggests one orthologs/allele for each of the maize accessions. Stepping down the leaves of the gene tree, the next set of paralogs, there appears to be an expansion in the B73 accessions, genes (Zm00001d034455 and Zm00001d034457), and perhaps a loss in W22. Functional domains, in the form of Interpro annotations, are highlighted on the alignment overview. For O2 there are two annotated functional domains. Clicking on a highlighted domain displays its name and information on how many of the genes in the tree share the domain. In the case of O2, IPR004827 bZIP (orange) is shared by 85% of the genes and IPR020983 Basic Leucine-zipper C shared by 54% of the genes. The O2 PH207 gene Zm0008a06946 is missing amino acids that are directly impacting the bZIP domain. This could be an artifact of the reference assembly, a miss-annotation or represent the biological model in PH207. Interestingly, the gene tree suggests there is a local duplication of the O2 ortholog in the reference sorghum accession.

A *BETA* maize-pangenome.gramene.org



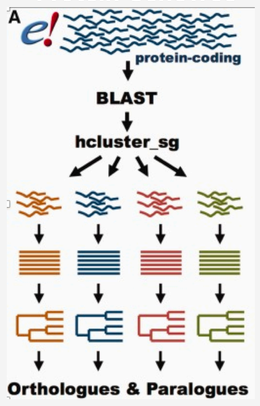
B Browsing Whole Genome DNA alignments using the native Ensembl views

- Pairwise whole-genome (LASTZ-CHAIN-NET)
- Multispecies
- Ortholog links

An example of WGA between Sorghum, B73 and W22, at B73 **Chromosome 7:122,210,000-122,430,000**. Green shaded areas indicate alignments, blue lines connect orthologs. This region shows a potential insertion in B73 compared to W22 and Sorghum.

C Protein and DNA Genome Comparisons

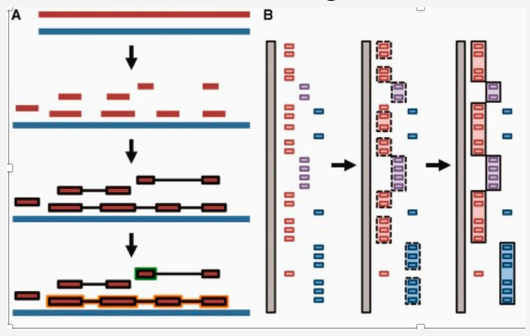
Protein GeneTree



Herrero, Javier et al. "Ensembl Comparative Genomics Resources." *Database: The Journal of Biological Databases and Curation* 2016 (2016): bav096. *PMC*. Web. 5 Mar. 2018.

For each protein-coding gene in Ensembl, a representative protein is used. BLAST scores are provided to hcluster_sg for grouping the sequences into gene families. The proteins are aligned with MCoffee or MAFFT and a phylogenetic tree is built with TreeBeST. Finally, orthologues and paralogues are inferred from the tree.

Whole Genome Alignment



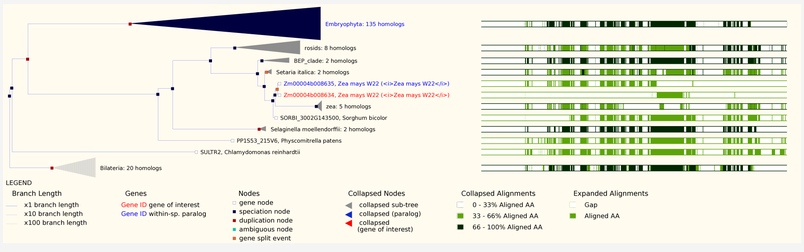
Whole genome analysis pipeline. (**A**) Pairwise alignments. A reference genome (blue) is aligned to another genome (red) with LASTZ. The raw alignments that are in the same order and orientation are grouped in chains (highlighted in black). On each region of the reference genome, the best chain is selected to single out the set of nets. A top-level net (orange) can include a nested net (green) in regions it does not cover. (**B**) Large-scale syntenies. LASTZ-net alignments are sorted on a reference genome (grey). The red, magenta and blue boxes represent alignments to different chromosomes in the other genome. For simplicity, we assume that they are in the same order and orientation. Contiguous collinear alignments are joined in a first-pass, forming a nascent syntenic block. In the second pass, the nascent blocks are joined and extended further to build macro-synteny blocks.

D Browsing Gene trees using the native Ensembl views

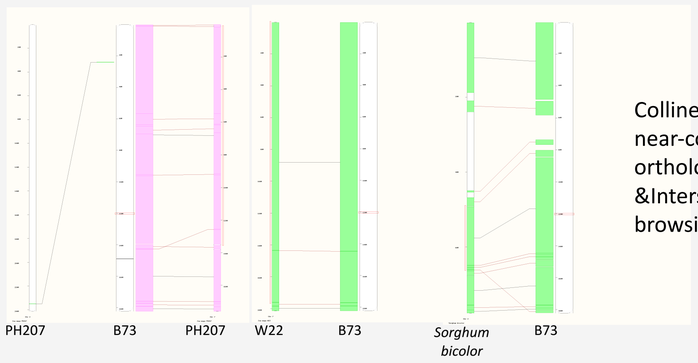
- 3 maize species & 12 outgroups
- 23,442 family clusters
- Ortholog & paralog prediction
- Candidate split genes

Candidate Split Genes

Genome	Split Gene Count
oryza_sativa	46
sorghum_bicolor	41
zea_mays B73	40
zea_maysw22	36
zea_maysph207	50

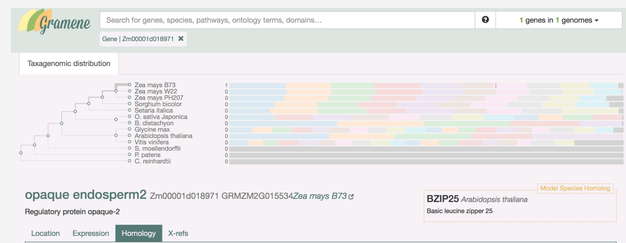


E Browsing Synteny Maps

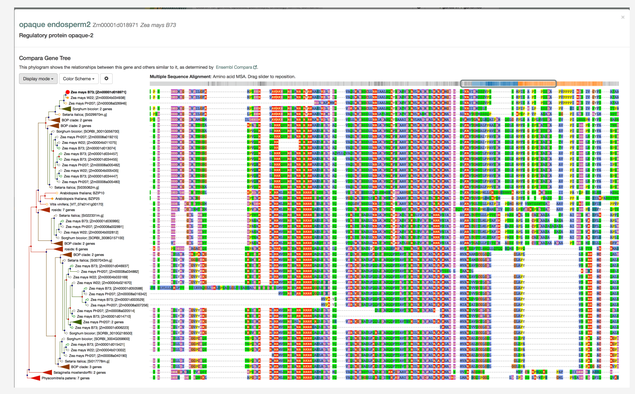


Collinear and near-collinear ortholog mapping &Interspecies browsing

F Gramene Search Interface and Gene tree alignment View

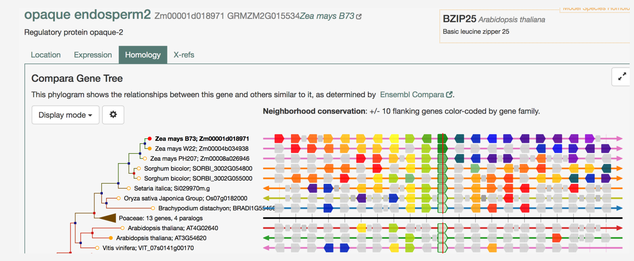


G Gramene Search: Multiple alignment view zoomed in



The multiple protein alignments view allows a user to zoom into the protein level alignments directly. Users can see overall transcript structure conservation, as well as specific amino acid conservation. Scanning down toward the bottom of the tree, there are several genes that do not contain Basic Leucine-zipper C and are conserved across all 3 accessions suggesting these are likely to be true models and not artifacts.

H Gramene Search: Neighborhood view



The Neighborhood Conservation mode provides information on local micro-synteny. The colors of genes are defined based on the top gene in the tree and the neighborhood of each homolog is displayed on an arrow to indicate strand with a distinct color for each chromosome. In this example 5' (left) of O2 there is a high level of conservation of gene order in the W22 & B73, 3' (right) of the gene there are 4 genes that are intervening before micro-synteny is observed. This view provides the users with information on local structural rearrangements, miss-annotations, or errors in the underlying assembly.

This project is funded by
USDA ARS 8062-21000-041-00D Enhancing Plant Genome Function Maps Through Genomic, Genetic, Computational and Collaborative Research
NSF IOS-1127112 IPGA: Gramene - Exploring Function through Comparative Genomics and Network Analysis
NSF IOS-1445025 MaizeCODE - An Initial Analysis of Functional Elements in the Maize Genome
NSF IOS1744001 Whole Genome assembly of Maize NAM Founders