Original Paper Open Access

# Comparative genomics with maize and other grasses: from genes to genomes!

**James C. Schnable[1] and Eric Lyons[2]***

[1]Department of Plant and Microbial Biology, University of California-Berkeley, Berkeley, CA, USA
[2]iPlant Collaborative, Bio5 Institute, University of Arizona, Tucson, AZ, USA
*Corresponding author: E-mail: ericlyons@email.arizona.edu

## Abstract

Of all the major plant groups, the grasses, with the complete genomes of five species, are the best positioned to take advantage of comparative genomics to obtain insight into functional genetic elements. Of all the grasses, maize is the best characterized in terms of genetics, development, and evolution. We provide several examples of how the web-based comparative genomics system CoGe may be used to aid in the interpretation of the maize genome sequence. These examples include verifying gene models, identifying differences between genome assemblies, identifying conserved non-coding sequences, identifying syntenic regions between species and polyploidies, and identifying homeologs within maize and orthologs between maize and other grass genomes. In addition, a comprehensive list of orthologous gene sets is provided between maize and Sorghum, foxtail millet, rice, and Brachypodium.

**Keywords**: comparative genomics, CoGe, synteny

## Introduction

On February 28th, 2008 at the Smithsonian National Museum of Natural History, while dining on sushi under the watchful gaze of Henry, the 8-ton 14-foot-tall African elephant on display in the Museum's entrance rotunda, attendees to the 50th annual Maize Meeting welcomed the completion of the maize B73 genome. The production and assembly of these 2 billion nucleotides represented a significant accomplishment in and of itself - the genome of maize remains the largest sequenced to date (Schnable et al, 2009) - yet the focus of the community was already on the interpretation of the maize genome to gain insight into its function, evolution, and possible improvement (as reviewed in Walbot, 2009).

The grasses are a particularly attractive system for comparative genomics for two reasons. The genomes tend to retain high levels collinearity (Moore et al,1995), enabling the identification of orthologous genomic regions in diverse species, and the grasses are the plant family most heavily sampled with complete genome sequences (Figure 1). Research into the evolution of genomes is based upon the comparison of equivalent genomic regions in multiple species, and relies on computational tools that enable researchers to intuitively compare these regions and accurately identity the evolutionary events responsible for changes among genes, genomic regions, and whole genomes between multiple species. However, these same comparative genomics tools may be leveraged to increase the accessibility and usefulness of the maize genome to researchers engaged in a broad range of research activities.

Here, we demonstrate how CoGe, a web-based comparative genomics platform (Lyons et al, 2008a, 2008b), may be used to navigate the maize genome and draw useful conclusions by comparing portions of the maize genome to the genomes of related species, existing sequences deposited with GenBank (http://www.ncbi.nlm.nih.gov/genbank/) (Benson et al, 2011), or data generated by individual researchers. We present a number of examples, ranging from the identification of conserved regulatory sequences associated with individual genes to comparisons of chromosome level rearrangements, each of which should be of interest to a segment of maize research community. Each example begins with data from the community, walks through analyses step-by-step, and provides links to CoGe which allow readers to regenerate each stage of the analytical process.

The use-cases presented are:

- Proofing computationally generated gene models presently assigned to the maize genome using experimentally verified data from GenBank
- Tracking down genes which appear to vanish between releases of the maize genome
- Visually comparing gene promoter regions using sequence data from multiple inbreds
- Identifying conserved non-coding sequences (putative regulatory elements) associated with orthologous genes in multiple grass species
- Identifying and visualizing large-scale changes in genome structure due to evolution, polyploidy, and assembly updates.
- Identifying syntenic orthologs of maize genes in other grass species

 RECEIVED 18/07/2011

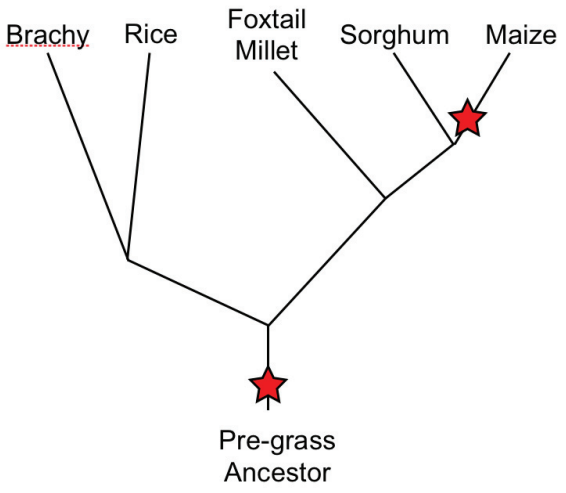


Figure 1 - Phylogeny of grasses with complete genome sequences. Red stars indicated whole genome duplication events. Branch lengths are not to scale.

## Methods

### *Overview of CoGe*

CoGe is comprised of three major systems:

1. A suite of interconnected web-based tools for analyzing and comparing genomic data;
2. A core data system to manage any number of genomes from any set of organisms in any state of assembly and annotation (currently 12,000 genomes from 10,000 organisms);
3. A genomic visualization system for creating intuitive and interactive graphics.

Together, these systems permit the open-ended exploration, analysis, and comparison of any genome within the system. CoGe currently contains the partial or complete genomes of 26 flowering plant species (http://genomevolution.org/wiki/index.php/Sequenced_plant_genomes) as well as thousands of genomes from other branches of the tree of life. Links between different tools allow a researcher to move seamlessly from one to the next. In addition, individual tools are designed to easily import user-generated data or data from NCBI, export results and data, and provide links to quickly regenerate an analysis. CoGe's homepage is located at http://genomevolution.org, and several tutorials for getting started with the system are available: http://genomevolution.org/r/4a3.

## How to verify and compare maize gene models

video: www.youtube.com/watch?v=y_AMfpnuwII

Gene models for the B73 reference genome are currently produced by automated computational pipelines (Liang et al, 2009). While these computational methods have high rates of success, often finding coding sequences, they do make errors, especially with regards to assigning introns and exons to the appropriate gene. In addition, local assembly mistakes in the maize genome may split genes (Schnable and Freeling, 2011). By comparing the region of the maize genome containing a computationally generated gene model to manually generated data such as cDNA clones, researchers can verify that the structure of their gene within the sequenced maize genome reflects the underlying biological reality.

CoGe has a tool, GEvo, that facilitates the comparison of genomic regions and sequences by providing the means to easily retrieve sequence and annotation data from a variety of sources, providing several analytical tools to compare sequences, and providing an interactive system for visualizing the results of the comparison.

For this example, the genomic gene model corresponding to *sugary1*, originally cloned in 1995 (James et al, 1995), will be validated against the published mRNA sequence as archived in GenBank.

Analysis steps:

(quick link: http://genomevolution.org/r/3uub)

1. Launch GEvo: http://genomevolution.org/CoGe/GEvo.pl.
2. Make sure the "Sequence Submission" tab is selected. Next to the "Sequence 1" submission box, make sure that "CoGe Database Name" is selected. This will search by name for genomic features stored in CoGe's database. Search for "GRMZM2G138060" the maizesequence.org gene ID identified as corresponding to *sugary1* (Schnable and Freeling, 2011). When the mouse is clicked on another text-box on the page, GEvo will automatically start searching for genomic features by that name. If multiple genomes and multiple genomic features match the given name, the appropriate one may be selected from a drop-down menu. By default, the highest version number of a genome is presented first. Note: a major difficulty with genes are tracking all of the various names that may be associated with a gene. Since CoGe contains thousands of genomes from all domains of life, this problem is compounded. Usually, the gene names assigned by maizesequence.org – those starting with "GRMZM" or "AC" are the best for searching within CoGe.
3. Next to the "Sequence 2" submission, select "NCBI GenBank". This will search GenBank for an accession, and retrieve the sequence and its associated annotations automatically. Enter "U18908.1" to retrieve Su1p's (*sugary1*) annotated mRNA sequence.
4. Select the "Algorithm" tab and choose "BLASTN: Small Regions". BLASTN is ideal for identifying small blocks of similar sequence while the default (Altschul et al, 1990), BLASTZ, is ideal for identifying large blocks of similar sequence (Harris, 2007). Identifying small blocks of similar sequence is appropriate for this analysis where individual exons are to be proofed.
5. The analysis is now configured. Press "Run GEvo Analysis!" to proceed.

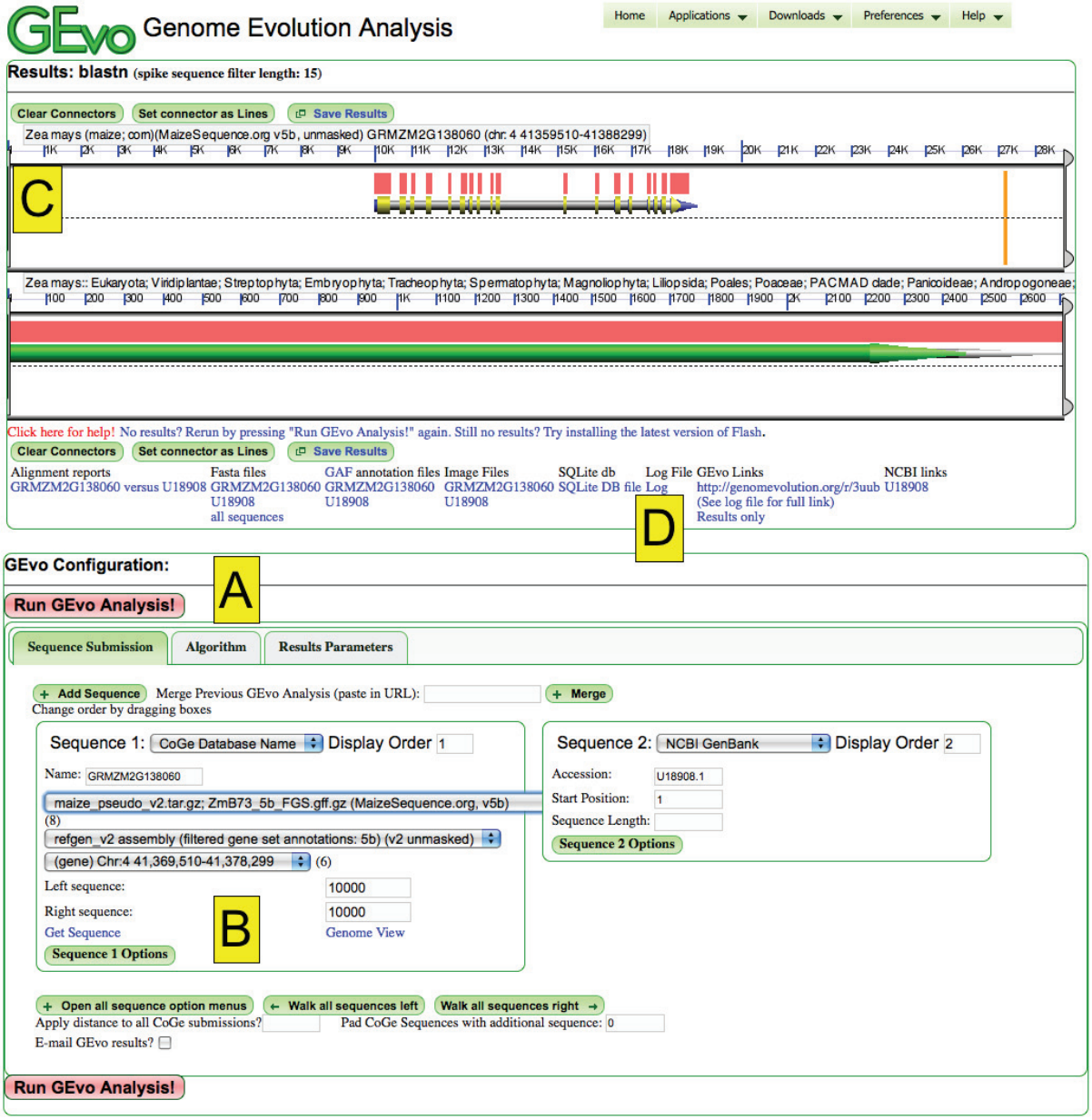


Figure 2 - GEvo analysis comparing a maize gene and surrounding genomic sequence extracted from CoGe's genome database to the cloned cDNA's sequence deposited at GenBank. **A)** The button for running the analysis. **B)** One sequence submission box. **C)** Interactive graphical display of results. Arrows are gene models, pink boxes are regions of sequence similarity. **D)** Links to input files, output files, the analysis log file, and a tiny-url that links back to GEvo pre-configure to regenerate the same results.

6. When the analysis is complete, the results will appear above the analysis configuration options (Figure 2). These results contain two panels, each representing one of the compared sequences. The dashed line in each panel separates the top and bottom strands of DNA, and the composite colored arrows represent annotated genes, with each color representing a different part of the gene's structure: grey is the full extent of the gene, blue is annotated exons, and green/yellow is protein coding sequence also referred to as CDS. These regions are drawn in ascending order so gray is only visible in non-exon regions of the gene (introns), and blue is only visible in those exons which do not code for protein (UTRs). The pink rectangles above the gene model are regions of sequence similarity as identified by BLASTN. Clicking on the genes or BLAST hits will cause a box to appear with information about the feature. Also, a transparent wedge will appear connecting regions of sequence similarity. By highlighting all regions of sequence similarity (by clicking on each, or holding shift and clicking), the entire CDS sequence from NCBI is seen to be present and annotated correctly in the genome, validating the gene model.

7. To save this analysis in order to return to it in the future, there is a link located below the results graphics under the heading "GEvo Links". When loaded, this link will load GEvo and configure your analysis as it was last run.
8. As an example of a gene model that was split across two contigs and mis-assembled in B73 RefGen_v1, see http://genomevolution.org/r/3uug for *pericarp color1*. Note, in this example a subset of the sequence retrieved from NCBI is used, and the sequence was reverse complemented in order to place it in the same orientation as the genomic sequence. The extent of a region analyzed may be changed by clicking and dragging the slider bars located at the ends of the images; various permutations may be done to a sequence, including reverse complementing it, by clicking on the "Sequence X Options" located below each sequence submission box.

## How to visualize assembly and annotation changes between versions of the maize genome

video: www.youtube.com/watch?v=URKq537JYhE

Due to the highly repetitive nature of the maize genome, the Maize Genome Sequencing Consortium decided to sequence the B73 maize genome using a BAC-by-BAC approach, stitching together chromosomes using a combination of genetic and physical maps as well as overlapping BAC contigs (Schnable et al, 2009). Individual BACs were sequenced using a shotgun approach, which often resulted in multiple

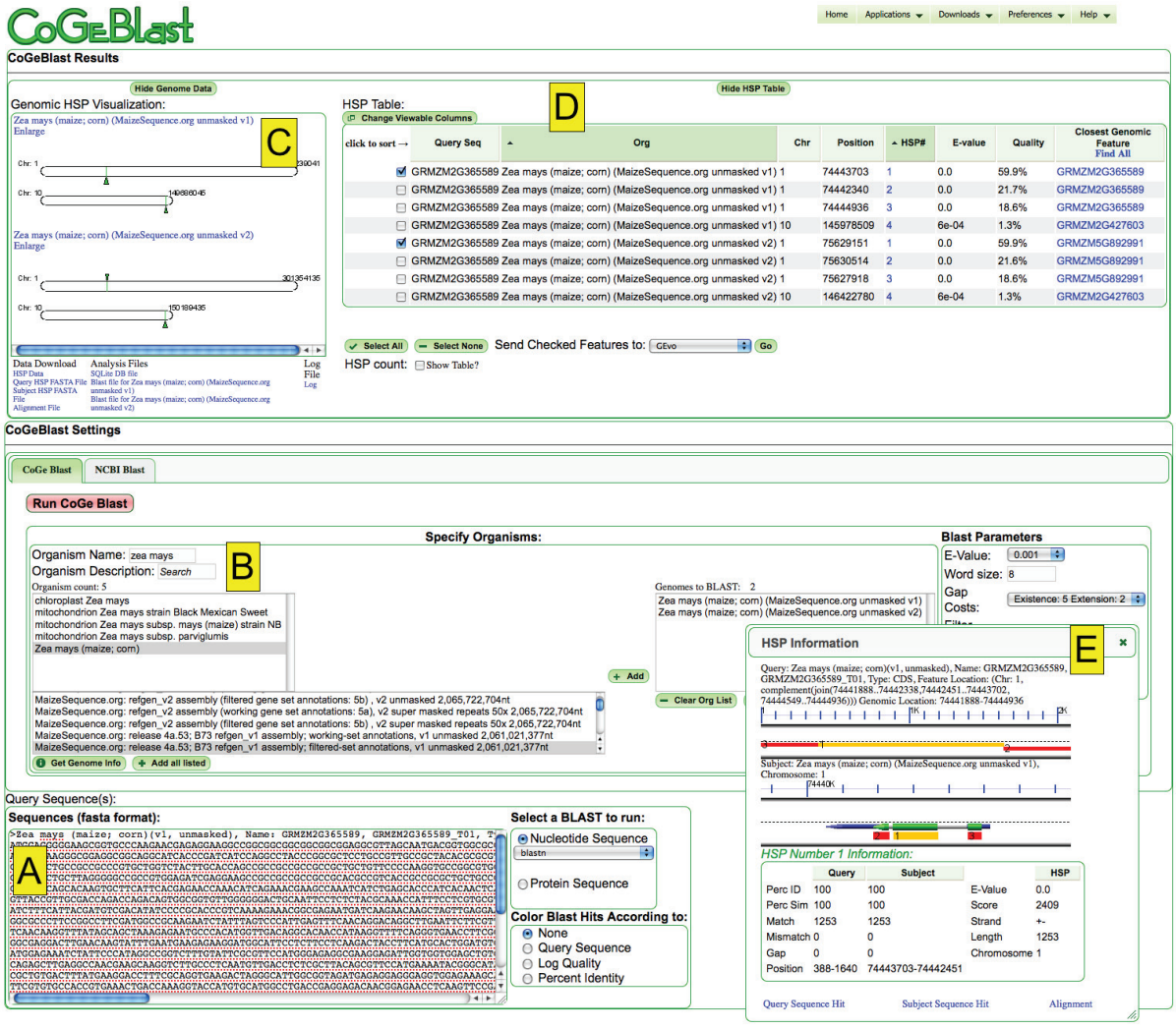


Figure 3 - CoGe Blast analysis. **A)** Input sequences. **B)** Search and select any set of genomes from the thousands available in CoGe. **C**) Graphical overview of blast hits to chromosomes in selected genomes. **D)** Interactive table of blast hits. The check boxes permit the selecting of overlapping genomic features (e.g. gene) that were hit. Those genomic features may be sent to other tools in CoGe for additional analysis or data extraction. **E)** Information box showing an overview of a selected blast hit (yellow bar) on the query sequence (top graphic) and in the matching genomic region (bottom graphic). Additional blast hits to the query sequences and neighboring genomic sequences are shown as red boxes.

assembled contigs separated by unassembled gaps. These gaps often contain transposons and other repetitive sequences.

Thus, while the location of each BAC within a chromosome has been reliably determined, the order of sequenced contigs within each BAC can be much less accurate. B73 RefGen_v2 significantly improved the accuracy of contigs ordered and orientated within BACs by using evidence rather than at random. However, one effect of this reordering of contigs in the new assembly is that gene models may add exons, loose exons, fuse together, split apart, or change strands as well as move to new locations in the genome.

In this example, we follow a gene from the first complete release of the maize genome, GRMZM2G365589, which "disappeared" in the update to version 2. The gene was the focus of a recent publication reporting the cloning of ragged seedling2 a gene required for mediolateral expansion of maize leaves (Douglas et al. 2010), so it may confidently be said that the B73 RefGen_v1 ID corresponded to a real and biologically relevant gene.

Analysis steps:

1. Launch FeatView: http://genomevolution.org/CoGe/FeatView.pl. FeatView is CoGe's tool for searching for genomic features by name.
2. Search for "GRMZM2G365589" by typing that name in the text-box next to "Name: " and clicking "Search".
3. The first identified result is a CDS from B73 RefGen_v1. At the bottom of the page is the information stored in CoGe about the gene. This area also includes links to various tools in CoGe.
4. Send the sequence to CoGeBlast by clicking on the button labeled "CoGeBlast" (quick link http://genomevolution.org/CoGe/CoGeBlast.pl?featid=56210971;gstid=1).
5. When CoGeBlast loads, the nucleotide sequence for GRMZM2G365589 will be pre-loaded in the sequence submission box (Figure 3A).
6. CoGeBlast is CoGe's interface for using BLAST to search against any set of genomes in its system. Search for maize genomes be typing "Zea mays" in the text-box next to "Organism Name" (Figure 3B).
7. From the identified organisms, select "Zea mays (maize; corn)".
8. Select "B73 RefGen_v2 assembly (filtered gene set annotations: 5b), v2 unmasked" and add it to the list of genomes to BLAST by clicking the "+ Add" button.
9. Select "release 4a.53; B73 RefGen_v1 assembly; filtered-set annotations, v1 unmasked" and add it to the list of genomes to blast by clicking the "+ Add" button.
10. Run the analysis by pressing "Run CoGeBlast".
11. When the results return, they will be presented above the configuration area. These consist of two views of the results: a graphical view of the chromosomes with locations of BLAST hits (aka HSP or High Scoring Pair) denoted by triangles (Figure 3C) and a

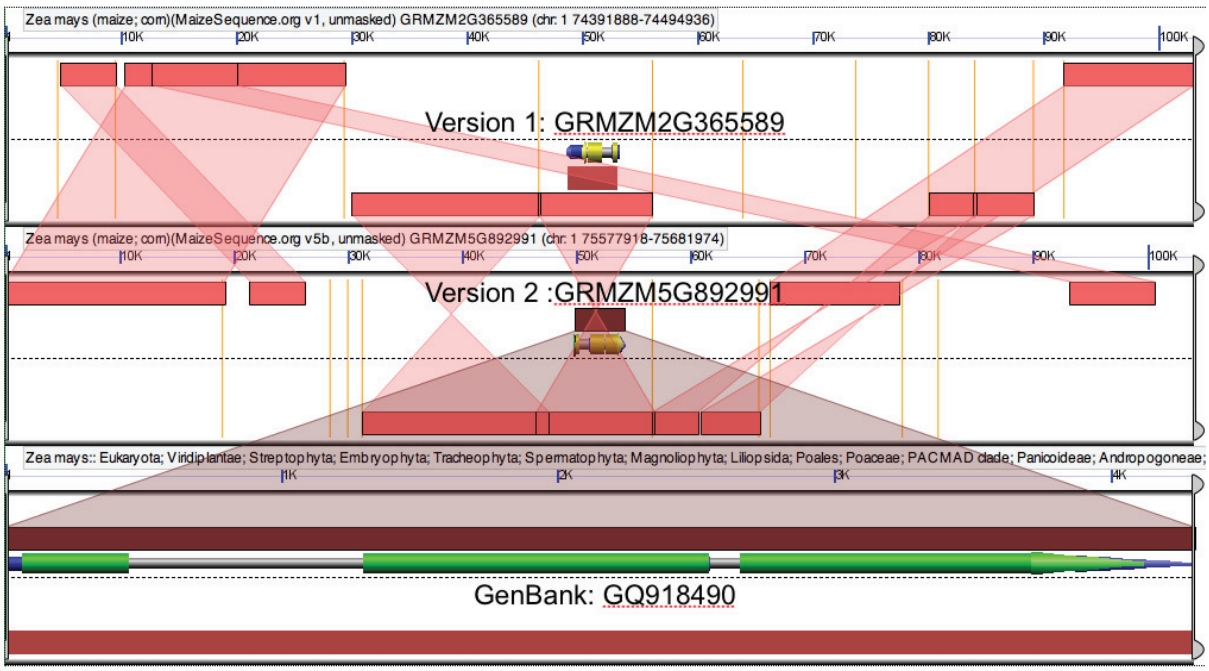


Figure 4 - GEvo analysis comparing the assemblies of two versions of the maize genome and the gene sequence extracted from GenBank. The assembly at a local level changed significantly between versions one and two of the maize genome. This resulted in the gene name being changed between versions of maize. Regions of sequence similarity are identified by colored boxes with transparent wedges connecting them. The vertical orange bars in the first two panels mark gaps between assembled sequence contigs.

table of hits (Figure 3D).
12. BLAST hits may be evaluated in their genomic context by clicking on a triangle in the genomic visualization or on the HSP number in the BLAST hits list. This will bring up a box with two graphical panels showing the query sequence and the genomic region (Figure 3E top and bottom respectively). The selected BLAST hit will be drawn in yellow and any other hits between that query sequence and neighboring sequence in the genomic region will be drawn in red. Using this visualization it is possible to quickly assess how much of a query sequence is covered by multiple BLAST hits from a single genomic region.
13. Evaluate all BLAST hits by clicking on their HSP number in the table of BLAST hits (Figure 3D, E). Note that there are two identical genomic features hit: GRMZM2G365589 in maize version 1 (the original query sequence) and GRMZM5G892991 in B73 RefGen_v2. It appears the genome name has changed and the next steps will determine why.
14. Select the check-box next to HSP 1 from each genome (Figure 3D). This selects the genomic feature that overlaps the BLAST hit and permits it to be sent to another tool in CoGe.
15. Next to "Send Checked Features to:", select "GEvo" and press Go. This option is located below the BLAST hit table and will send the identified genomic features to GEvo to compare those genomic regions.
16. Note: Since we are comparing two regions of maize to examine assembly structure, and maize's genome contains many repetitive sequences, the GEvo analysis will require different settings than the previous example. To expand the window of sequence displayed within GEvo to include 50 kb of up and downstream sequence type "50000" in the text box next to "Apply distance to all CoGe submissions" located beneath the sequence submission boxes.
17. Select the "Algorithm" tab and make sure BLASTZ is the selected algorithm. Since we are examining identical - if rearranged - sequences change the "Score threshold" to "100000" to filter out smaller and lower percent similarity hits.
18. Run the analysis (quick link: http://genomevolution.org/r/3uuj).
19. Analyze the results (Figure 4). Note: there are orange vertical lines in background of the genomic regions. These represent gaps between sequenced regions of the maize genome and are represented in the raw maize genome sequence as stretches of "N". The length of these gaps is predetermined by the type of gap (gaps between sequence contigs within a single BAC, gaps between BACs, etc) and do not represent the total length of missing sequence. For a description of all the glyphs used in GEvo's images, please see: http://genomevolution.org/r/3uum.
20. Highlight regions of sequence similarity. Note how the size, placement, and orientation of contigs have changed between the versions. Of importance, gene GRMZM2G365589 in B73 RefGen_v1 is now in a larger contig in B73 RefGen_v2 which has been flipped relative to the overall arrangement of the chromosome. During the annotation process for maize version two, possibly because of the inversion of sequence, this gene was not recognized as GRMZM2G365589 and was assigned a new name.
21. The GenBank sequence published in the study of this gene (GQ918490) may be added to the analysis to validate the structure of both gene models, as covered in example #1. This is accomplished by clicking the button "+ Add Sequence" to create a new sequence submission box under the "Sequence Submission" tab, selecting "NCBI GenBank" for the sequence, and pasting in the sequence name (quick link: http://genomevolution.org/r/3uuo).

## Promoter variation between inbreds

example: http://genomevolution.org/r/3v7g

Much natural variation is the result, not of differences in the structure or activity of the proteins encoded by different alleles of the same gene, but rather by changes in the pattern and level of gene expression as a result of polymorphisms in noncoding sequences surrounding a gene. To investigate the polymorphisms responsible for this variation, many research groups re-sequence the region surrounding a gene in a wide range of accessions.

In this example, we compare promoter sequences for *opaque-2*, a classical mutant of maize which regulates the abundance of different proteins within the endosperm (Schmidt et al, 1990), isolated from a variety of maize inbred lines to the sequence of the B73 reference genome (Manicacci et al, 2009). Using GEvo, these promoter sequences are retrieved from GenBank, aligned to the B73 genome sequence, and their polymorphisms visualized in reference to B73 sequence.

Analysis steps:
1. Launch GEvo: http://genomevolution.org/CoGe/GEvo.pl.
2. The gene id which corresponds to opaque2 in B73 RefGen_v2 was found to be "GRMZM2G015534" (Schnable and Freeling, 2011). Enter this name in the the Sequence 1 submission box.
3. The promoter sequences for *opaque-2* (Manicacci et al, 2009) are deposited in GenBank with the accessions FJ935730 to FJ935747. To add some of these sequence to the GEvo analysis, first add however many sequence submission boxes are needed by pressing the "+Add Sequence" located above the Sequence 1 submission box. For each of these new sequence submission boxes, select "NCBI GenBank" and type in a GenBank accession next to the text box labeled "Accession".
4. By default, GEvo runs pairwise sequence comparisons against all the submitted sequences. For this example, only the B73 sequence will be considered a "reference sequence" to which all other sequences

will be compared. To set up this configuration, press the "Open all sequence option menus" located below the sequence submission boxes. This will open all the sequence options for each sequence submission box. For all sequence submission except the one for B73 (retrieved from the CoGe database), select "No" for the line labeled "Reference Sequence".
5. To visualize polymorphisms, GEvo can be configured to generate graphics for sequence similarity that contain gaps of color to represent mismatches, insertions, and deletions. To turn on this option, select the "Results Parameters" tab and select "Yes" for "Color matches in HSPs". This option is second from the top in the second column of options.
6. Change the sequence comparison algorithm to BLASTN by selecting the "Algorithm tab" and selecting "BLASTN: Small Regions" from the list of options next to "Alignment Algorithm" (quick link: http://genomevolution.org/r/3v9s).
7. Press "Run GEvo Analysis".
8. By default, CoGe adds 10,000 nucleotides upstream and downstream of any feature used to anchor a genomic region. Since the sequences extracted from GenBank contain the 5' end of the gene and the upstream promoter region, only a small portion of the displayed maize B73 region is matched. To adjust this display, there are slider-bars located at the ends of each displayed genomic region (though the one on the left may be partially hidden in the viewer). For the B73 region, slide these bars so they boarder the matched region and rerun the analysis (quick link: http://genomevolution/orgr/3v7g).
9. Examining these results show the approximate location of polymorphisms within each sequence in the

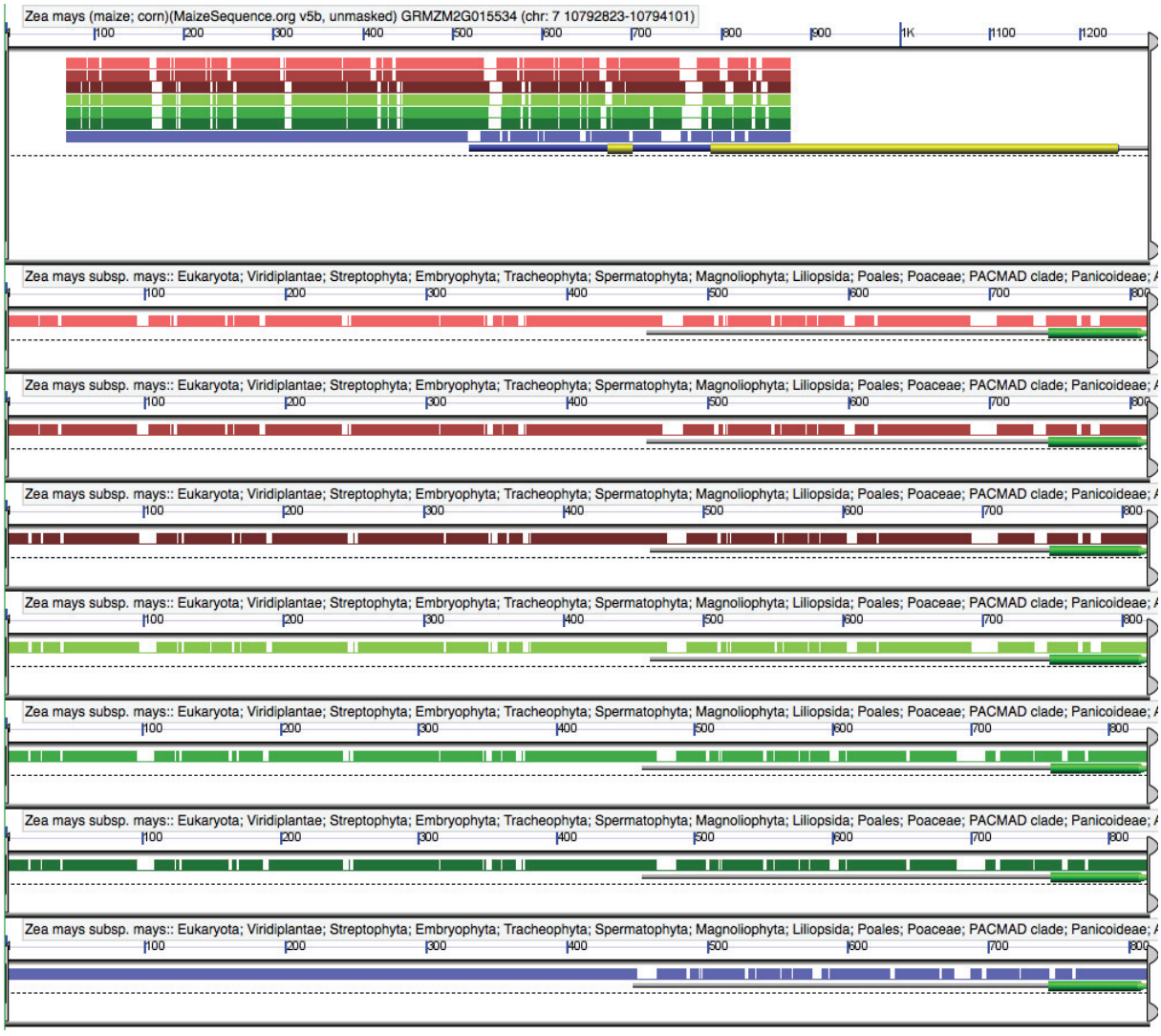


Figure 5 - GEvo analysis of the 5' region of the B73 maize gene *opaque-2* (top panel) aligned to seven sequenced promoter regions from different maize inbred lines. Regions of sequence similarity are shown as colored boxes. These boxes contain non-colored regions indicating portions of the alignments without matches. These are due to sequence variations and indels, and indicate polymorphisms between the B73 sequence and each of the other inbred lines.

colored boxes indicating regions of sequence similarity. Areas of the colored box that are not colored indicate sequences that did not matched due to sequence variation or indels (Figure 5).
10. In order to get the actual sequence alignments, click on a colored box representing a region of sequence similarity. Besides from drawing a transparent wedge connecting the regions of similar sequence, an box entitiled "GEvo Results Info" appears containing information about the match. At the top of this box is a link called "full annotation". Click on this link to launch HSPView.
11. HSPView provides a detailed view of regions of sequence similarity, including statistics of the match, the sequences participating in the match, and an alignment of the match.

## Every maize gene and its syntenic orthologs in other grasses

BLAST is likely the only computational tool almost every biologist can name off the top of their head. For identifying homologous genes (genes that share similar sequence as a result of descent from a common ancestor) BLAST does a great job. However, for the trickier task of distinguishing orthology and paralogy BLAST leaves much to be desired, especially within large gene families. In the grasses, which diverged from a common ancestor only fifty million years ago, synteny - the conservation of gene order between genomic regions related by common descent - provides a robust method of identifying true orthologous genes.

CoGe makes identifying syntenic orthologs between two species easy using a utility called SynMap (Lyons et al, 2008b). For most model organisms, pairwise identification of syntenic orthologs is enough. However, as previously mentioned, the grasses are extraordinarily well represented among plant genome sequences. Using SynMap, we have identified syntenic orthologs for maize genes in each of the other four grass species with sequenced genomes as well as homeologous genes within the maize genome itself. These gene lists, as well as GEvo links permitting anyone to quickly compare syntenic orthologs between all grass species are provided as supplemental dataset S1. This dataset is a resource to allow anyone to quickly look up a maize gene by name, and immediately find its homeolog in maize and orthologs in Sorghum, foxtail millet, rice, and Brachypodium. Generating these datasets for a pair of grass genomes is described later in this document as an advanced topic.

## How to identify putative regulatory sequence through comparison to related genomes

Detailed characterization of sequences involved in regulating the expression of a single gene often requires lengthy promoter-bashing and/or immunoprecipitation enrichment experiments. However, a first pass to identify candidate functional regulatory sequences can take advantage of these fact that these sequences are under functional constraints and are therefore constrained by purifying selection. Therefore, such regulatory sequences should have lower substitution rates than functionless DNA. For a thorough review of the study of these conserved noncoding sequences in plants (Freeling and Subramaniam, 2009). The excellent conservation of gene order and wide evolutionary range of sequenced species within the grasses makes them ideal for the identification of functionally conserved noncoding sequences (CNSs).

This example with involve a homeologous pair of maize genes, each with homology to the Arabidopsis circadian clock mutant GIGANTEA (Fowler et al, 1999; Park et al, 1999). The two maize genes are compared to their shared single orthologs in the Sorghum, foxtail millet, rice and Brachypodium genomes. The conserved sequences neighboring all these genes are potential regulators of GIGANTEA expression in the grasses. The rice homolog of GIGANTEA, (gene model id: LOC_Os01g08700), has also been shown to be involved with photoperiodic control of flowering (Hayama et al, 2002; Izawa et al, 2011).

As discussed above, there are multiple methods which can be used to identify and define orthologs. If you would like to define orthology using "best BLAST hit" begin at step #1 below. If you would like to use synteny to filter out well conserved paralogs, you would normally have to extract syntenic data for each pairwise species comparison using SynMap or validate each best BLAST match gene's genomic region. However, since we have already done these comparisons between maize and all the other grass species with sequenced genomes, you can simply search supplemental dataset S1 for "Os01g08700", copy the pregenerated GEvo link from the right-most column the spreadsheet into your browser and proceed to step #4.

Analysis steps:
1. Launch FeatView and search for the previously identified rice GIGANTEA gene, Os01g08700, and send it to CoGeBlast (quick link: http://genomevolution.org/CoGe/CoGeBlast.pl?featid=57688315).
2. Find and BLAST against the following genomes: rice, Brachypodium, maize, Sorghum, and foxtail millet.) The genome of rice is searched against in order to rapidly identify the underlying gene for use downstream analyses.
3. Evaluate the BLAST hits and select the best overlapping features to send to GEvo. For this example, the top BLAST hit for each organism is the one to select except for maize, where the two top hits should be selected (GRMZM2G107101; GRMZM5G844173). Maize has two copies of this gene due to its lineage-specific whole genome duplication event.
4. Next, type "1000" in the "Apply distance to all

CoGe submissions" text-box. This will limit the genomic sequence searched to 1000nt upstream and downstream from each gene. Run the analysis (quick link: http://genomevolution.org/r/3uuy).

5. When the results are returned note that there are many colored bars representing regions of sequence similarity between each pair of sequences. Some of these bars are located below the gene models indicating that the similar sequences are on opposite strands in this comparison. Select "reverse complement" in the "Sequence Options" menu under the appropriate genes so that all sequences are running in the same direction. Rerun the analysis (quick link: http://genomevolution.org/r/3uv0). Note: you can

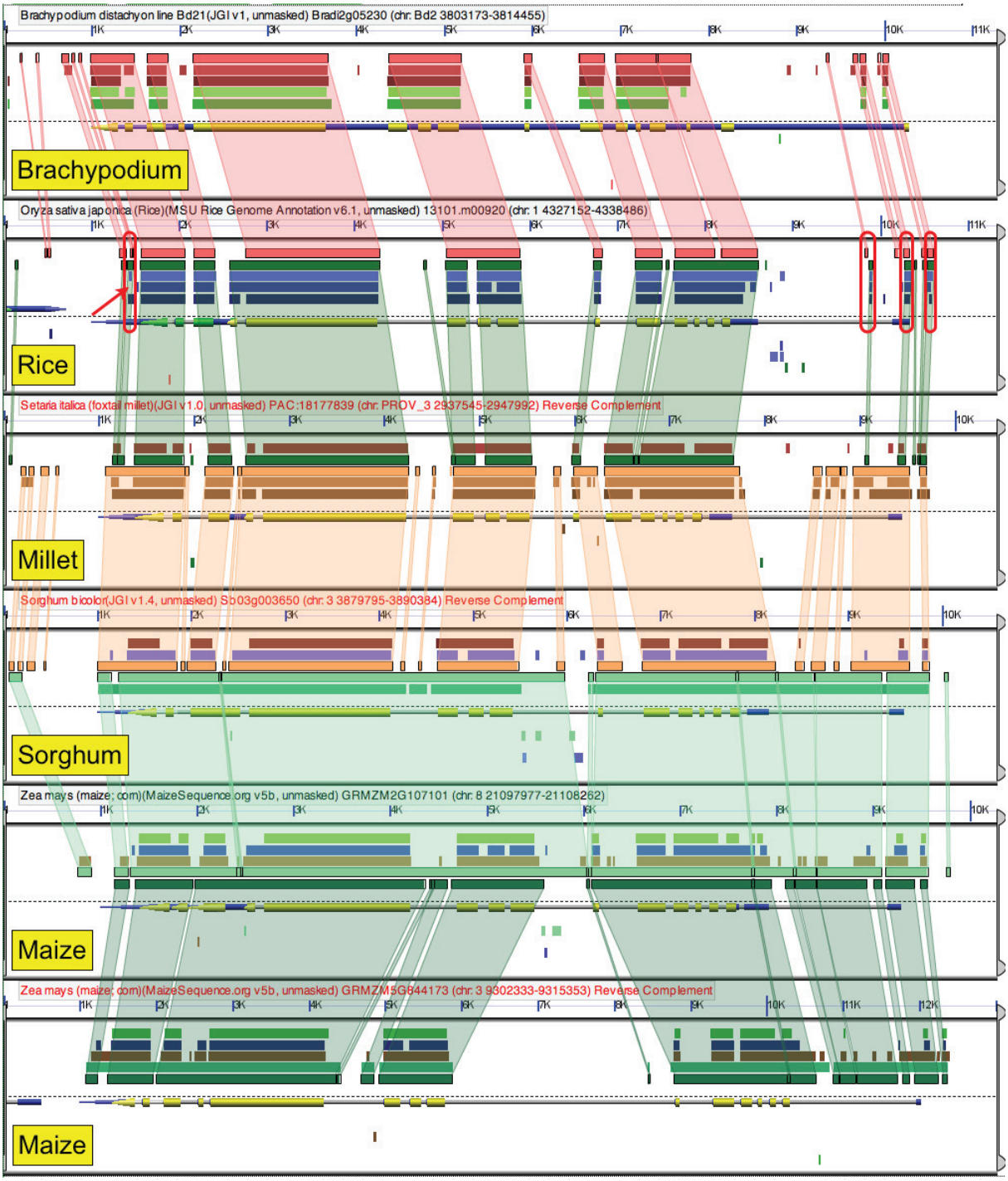


Figure 6 - GEvo analysis of GIGANTEA orthologs in five grass species to identify conserved non-coding sequences (CNS). Maize has two homeologies copies. CNSs are identified in rice by red ovals. The red arrow points to a conserved sequence found in all orthologous grass genes examined except one of the two homeologous genes in maize.

open all of the sequence submission boxes' options menu by clicking the "Open all sequence option menus" button located below the sequence submission boxes and above the "Apply distance to all CoGe submissions" text-box.
6. Plant CNSs are relatively short sequences and are usually merged or missed by BLASTZ. To visualize CNSs, change the algorithm to BLASTN (quick link: http://genomevolution.org/r/3uv2). Note, visualizing six genomic regions with GEvo often requires a significant amount of vertical computer monitor resolution. The size of the panels may be adjusted under the "Results Parameters" tab. The height of the panels may be adjust by changing the value for "Feature Height (Pixels)". This refers to the height of each track of gene models and regions of sequence similarity.
7. Interpreting the results: Identifying CNSs requires visually inspecting GEvo's results (Figure 6). By chance, the sequences submitted to GEvo by CoGeBlast are arranged by their phylogenetic relationships (Figure 1). However, their relative order may be changed by clicking and dragging the sequence submission boxes around relative to one another. In Figure 6, the two maize homeologs are located in the bottom two panels, Sorghum above it, preceded by millet, rice, and Brachypodium.

The evolutionary relationship of these organisms is reflected in the degree of sequence similarity they share with one another. The two maize homeologs and Sorghum are very similar in sequence over most of the compared sequences including introns and upstream and downstream regions, reflective of the shared ancestry of Sorghum and the two subgenomes of modern maize within the last 12 million years (Swigoňová et al, 2004). The common ancestor of both Brachypodium and rice branched off from the maize lineage ~50 million years ago during the radiation of the major grass tribes (The International Brachypodium Initiative 2010), while foxtail millet shares a more recent common ancestor with maize and Sorghum than these two species.

The rice panel is most informative for discerning CNSs in maize and among the other grasses. Note that there are sets of conserved sequences that do not overlap protein coding sequences (Figure 6, red ovals). There is one set in the 3' UTR, one set in the proximal 5' intron, one set in the 5' UTR, and one set 5' of the gene. These sequences are conserved among these genes in nearly all cases despite a maximal divergence of ~50 million years. In addition, of the two maize homeologs, there is one case where a CNSs is missing from one copy (Figure 6, red arrow, 3' UTR) potentially producing a change in the regulation of this homeolog's expression.

## How to identify changes between different versions of the maize assembly at a whole genome level

video: www.youtube.com/watch?v=d8I1XWLGEcM

Over time, new versions of existing genomes are generated. These updates include corrections to the original assembly and gene annotations, and addition of more sequences and gene models. While essential, these updates can also be frustrating, especially when the details of changes are undocumented. CoGe has a tool, SynMap, designed to aid in the comparison of whole genomes by the identification of syntenic regions where homologous genes are arranged in the same order in multiple species (Lyons et al, 2008b). SynMap can also be used to visualize changes from one version of a genome assembly to the next. The program generates interactive dot plots which allow researchers to visualize rearrangements at a genome-wide, chromosome by chromosome, or – through links to GEvo – gene by gene level. In this example, we will use SynMap to compare the structural differences in assembly between B73 RefGen_v1 and B73 RefGen_v2 of the maize genome.

Analysis steps:
1. Launch SynMap (quick link: http://genomevolution.org/CoGe/SynMap).
2. Search for version 2 of maize for "Organism 1 Search" by typing its name in the text-box next to "Name". Note that there are several versions of the maize genome assembly, and several sets of gene models for each assembly (http://www.maizesequence.org/info/website/help/index.html). In this case search for "refgen_v2 assembly (filtered gene set annotations: 5b)". Since SynMap begins with a whole genome comparison, and genomes contain many repeated sequences, it is usually best to used masked genomes when available. This is especially true when comparing grass genomes, which usually contain many transposons. In this example, select the "super masked repeats 50x" sequence. If gene annotations are available for a given genome SynMap will select "CDS" by default and only use annotated coding sequences in its genome comparisons. If no gene models are available, SynMap will select "genomic sequence" by default and it will compare all sequence included in the genome assembly. Comparing genomic sequences will take much longer then only comparing CDS sequences because of the larger amount of sequence to compare and the time it takes to process highly repetitive sequences.
3. Search for version 1 of maize in "Organism 2 Search". This will be "release 4a.53; B73 RefGen_v1 assembly; filtered-set annotations: masked repeats 50x" (quick link: http://genomevolution.org/r/3va8).
4. By default, SynMap uses BLASTZ for the whole genome comparison, and predefined parameters for filtering repetitive sequences and identifying syntenic regions. These settings are usually sufficient for most genome comparisons (plants, animals, fungi, bacteria, archaea, etc.). If you wish to change these options, they are found by clicking on the tab labeled "Analysis Options" in SynMap's configuration menu.
5. Depending on the size of the genomes under anal-

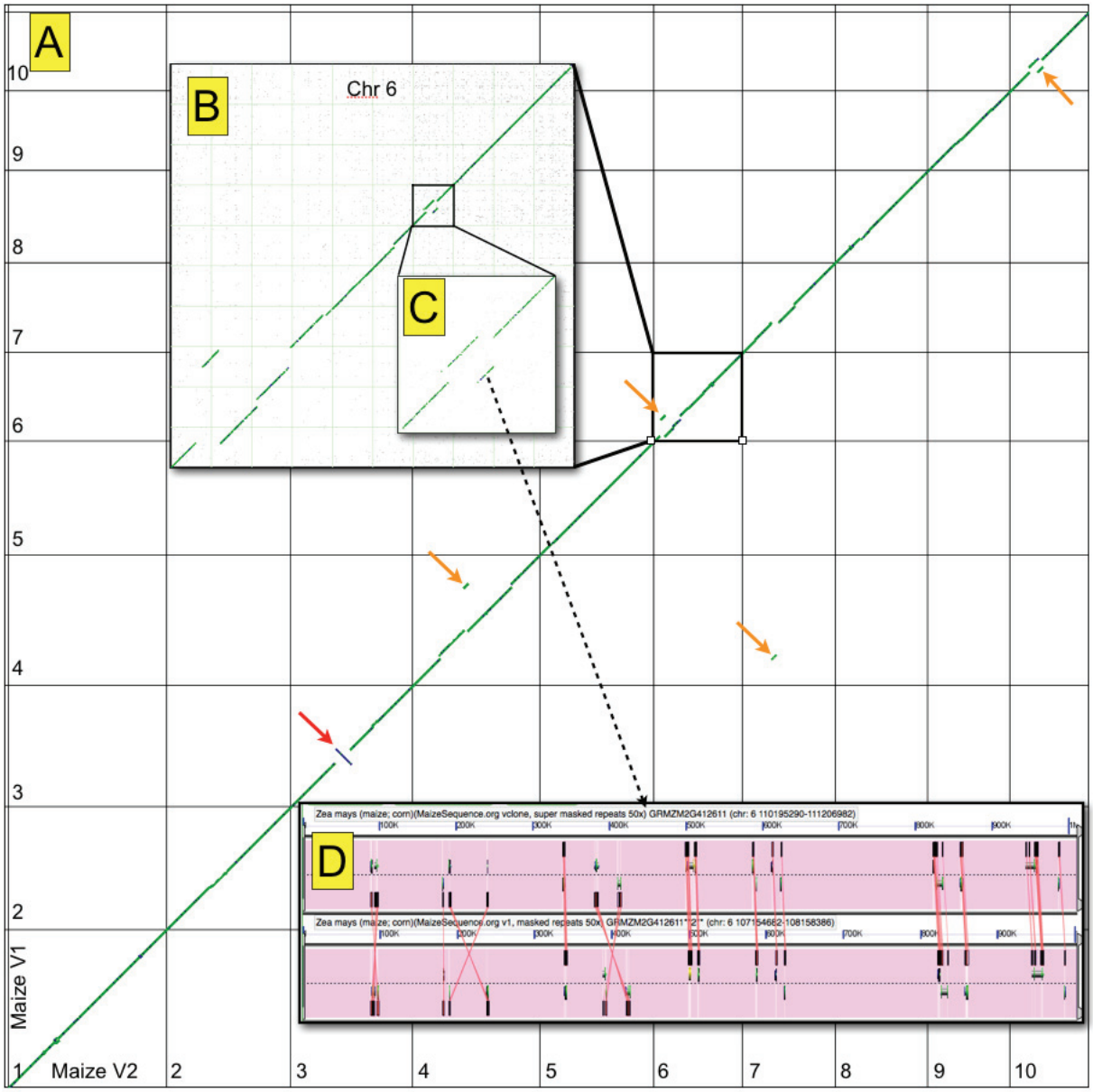


Figure 7 - Whole genome to high-resolution analysis of differences in genome structure between version 1 and version 2 of the maize genome assembly. **A)** Whole genome sytnenic dotplot generated by SynMap of two versions of the B73 maize genome assembly. X-axis: version 1; y-axis: version 2. Syntenic gene pairs are colored green if they are in the same orientation between the two genomes and blue of they are inverted with respect to one another. Red arrow highlights a large-scale inversion; orange arrows highlight "translocations" of genomic segments between genome versions. **B)** Close-up of the syntenic dotplot between chromosome 6 of the two maize versions. Note several blue dots representing inverted genes and translocation of segments between genome versions. **C)** Close up of a the syntenic dotplot between a portion of chromosome 6 of the two maize versions. **D)** GEvo's high-resolution sequence analysis of 1MB of chromosome 6. All non-protein coding sequences are masked in the analysis (purple background); and regions of sequence similarity which overlap genes are connected by lines. Note that while there is an overall pattern of collinear gene arrangement between the genomic regions, there are several inverted genes. Inverted genes are discernible by the location of the sequence similarity box below the gene models, and the "X" patterns formed by lines connecting regions of sequence similarity.

ysis and the types of sequences being compared, the computation may take some time to complete. While the analysis is running, its status will periodically be updated and displayed on a background of a spinning DNA double helix. Fortunately, the results from individual steps of the analysis are cached and automatically reused. If a particular comparison between genomes has been run previously, regenerating the results takes much less time.

6. Run the analysis by clicking the "Run SynMap"

button (quick link: http://genomevolution.org/r/3va9).
7. When completed, the results are displayed above the SynMap's configuration menu. Near the top of the results will be a dotplot identifying syntenic genes. Each genome is arranged along one of the two axes (version 2 on the x-axis and version 1 on the y-axis). Each horizontal and vertical black line separate chromosomes along the two genomes. By default, the chromosomes are ordered by size. To order chromosomes by their name, select the "Display Options" tab in the configuration menu and select "Name" for the line labeled "Sort Chromosomes By".
8. Rerun the analysis by click the "Run SynMap" button (quick link: http://genomevolution.org/r/3vaa).
9. Dots in the dotplot that are shown in a color (green by default) identify syntenic genes. These often form what appear to be lines and highlight syntenic genomic regions. A set of cross-hairs will follow the cursor as the mouse is moved over the syntenic dotplot. The horizontal and vertical lines of the cross-hairs permit the rapid identification of multiple regions in one genome that are syntenic to a region in the other genome. This is accomplished by centering the cross-hairs on a syntenic region and identifying additional syntenic regions that traverse the cross-hairs. By examining several regions of the dotplot, a syntenic relationship will emerge where there is a strong syntenic line running diagonally though the dotplot, and many additional regions of synteny to other regions in the maize genome. The major diagonal line is the self-self comparison of the two genome versions, and the others are due to polyploidies in the maize lineage. Since this analysis is focused on determining changes in genome structure between versions of the genome, only the self-self comparison is wanted.
10. SynMap has an option to screen syntenic regions such that the highest scoring regions are selected in order to enforce a specific syntenic relationship between two genomes. This accomplished by an algorithm called "Quota Align" (Tang et al, 2011), and may be turned on in SynMap by selecting the "Analysis Options" tab in the configuration menu, and selecting "Quota Align" from the row labeled "Syntenic Depth". Next, specify a 1 maize : 1 maize syntenic relationship in the text-boxes labeled "Ratio of coverage depth". The option called "overlap distance" sets the number of genes two syntenic regions may overlap without mutually excluding one another from the final set of syntenic regions, and the default value of 40 is fine for this example. Press "Generate SynMap" to rerun the analysis (quick link: http://genomevolution.org/r/3uzg). Notice that only one set of syntenic regions remain in the resulting dotplot.
11. SynMap has an option to color inversions with a different color. To turn this on, select the "Display Options" tab under the configuration menu and select "Inversions" in the row labeled "Color diagonals by". Rerun the analysis (quick link: http://genomevolution.org/r/3uzh).
12. Overall, the structure between the two versions of the maize genome are quite similar (Figure 7A). This is due to the fantastic physical map of BACs tiled across its genome. There is one large-scale inversion in chromosome three (inversions are colored blue and have a negatively sloped line) between B73_refgen1 and B73_refgen2 and a few chromosome pieces have changed their location (Figure 7A; red and orange arrows, respectively).
13. To examine a region in more detail, SynMap permits zooming in on a chromosome/chromosome comparison by clicking on it in the dotplot (Figure 7B). The size of these images may be adjusted using the "Image Width" in the "Zoomed SynMap" box located above the dotplot. By zooming in on a chromosome/chromosome comparison, finer details of the structural rearrangements may be seen (Figure 7C). Besides the translocation of large contigs, there are also several small blue regions visible in the line, representing small inversions.
14. Finer-scale analyses of genomic regions may be performed by clicking on a gene pair in the zoomed-in dotplot. When the cross-hairs turn red, the mouse is on top of a gene pair. This will update information in a box located near the top of the zoomed dotplot, and signify that you may click on the gene-pair. When the gene pair is clicked, GEvo will be launched and pre-configured with the selected pair of genes.
15. Running GEvo against two maize regions will often produce many regions of sequence similarity due to repeat sequences. These may be removed from the analysis by masking all sequence from the analysis except CDS sequence. This is accomplished by opening the sequence options in the sequence submission box and selecting "Non-CDS" for "Mask Sequence". Select a gene-pair from the zoomed dotplot, launch GEvo, mask non-CDS sequence, and examine 500,000 nucleotides up and downstream from each gene. (quick link: http://genomevolution.org/r/3uzp).
16. For many regions of the maize genome, there will be many rearrangements of contigs between assembly versions (Figure 7D). Small inversion are easy to visualize by regions of sequence similarity displayed below gene models, and "X" like patterns of lines connecting regions of sequence similarity for multiple genes. A similar pattern was seen for a previous example examining a gene model that was removed in version two of the maize assembly.
17. Note that like GEvo, SynMap contains links to all the files used and generated by its analysis, and a link to regenerate the analysis.

## Advanced Topic: How to generate orthology gene lists between maize and another grass (e.g. Sorghum)

video: www.youtube.com/watch?v=-fejg_O1aRs

While a maize-specific homeolog and ortholog dataset is distributed with this paper (dataset S1),

the future promises a rich harvest of additional grass spaces with sequenced genomes. This section details how to generate ortholog gene lists between maize and Sorghum, however, the same techniques could be used to identify syntenic orthologs between any two grass species.

Analyses of grass genomes have shown that their overall structure is highly conserved (Moore et al, 1995). Large regions of their genomes maintain synteny with one another permitting the identification of homologous genes derived from the same ancestral genomic region. However, the grass lineage has an ancient whole genome duplication that predated the radiation of major grass lineages (Paterson et al, 2004; Figure 1). This paleotetraploidy may obfuscate the identification of orthologous genes among grass genomes. Out-paralogous sequences (those derived from the pre-grass tetraploidy) may be erroneously assessed as being orthologous, especially when a true ortholog is no longer present in the genome. Given the relatively recent maize-specific tetraploidy (Swigoňová et al, 2004; Bomblies and Doebley, 2005) and resulting fractionation of duplicate genes (Woodhouse et al, 2010), this problem is magnified within maize. However, SynMap is able to filter out these more ancient syntenic regions using the Quota Align algorithm (Tang et al, 2011).

For this example, we will use CoGe's tool SynMap to identify syntenic regions between maize and Sorghum, determine which regions are derived from the divergence of their lineages or the pre-grass whole genome duplication, and create sets of orthologous genes between the two subgenomes of maize and Sorghum. In addition, SynMap has methods for measuring the evolutionary distance between syntenic gene pairs and options for removing ancient polyploidies from an analysis.

Analysis steps:

1. Launch SynMap (quick link: http://genomevolution.org/CoGe/SynMap).
2. In the organism search boxes, search for:
   a. Maize: refgen_v2 assembly (filtered gene set annotations: 5b): super maksed repeats 50x;
   b. Sorghum: masked repeats 50x;
   c. Make sure CDS sequences are selected for both genomes;
   d. quick link: http://genomevolution.org/r/3uvp.
3. To run the analysis, press "Generate SynMap" (quick link: http://genomevolution.org/r/3uvo).
4. To order chromosomes by their name, select the "Display Options" tab in the configuration menu and select "Name" for the line labeled "Sort Chromosomes By". Rerun the analysis by pressing "Generate SynMap" (quick link: http://genomevolution.org/r/3uvr).
5. Examine the resulting dotplot and use the crosshairs to determine the syntenic relationship between the maize and Sorghum genomes: many regions of the maize genome are syntenic to two regions of Sorghum; many regions of Sorghum are syntenic to four regions in maize. The 2 Sorghum (S) : 4 maize (M) syntenic relationship is expected as these two lineages share a whole genome duplication event (creating a 2S:2M syntenic relationship) and maize has had an additional independent whole genome duplication following the divergence of their lineages (creating a 2S:4M syntenic relationship; see Figure 1). This relationship does not apply to the entire genomic sequence of both organisms. Regions lacking one or more syntenic regions in either species can be

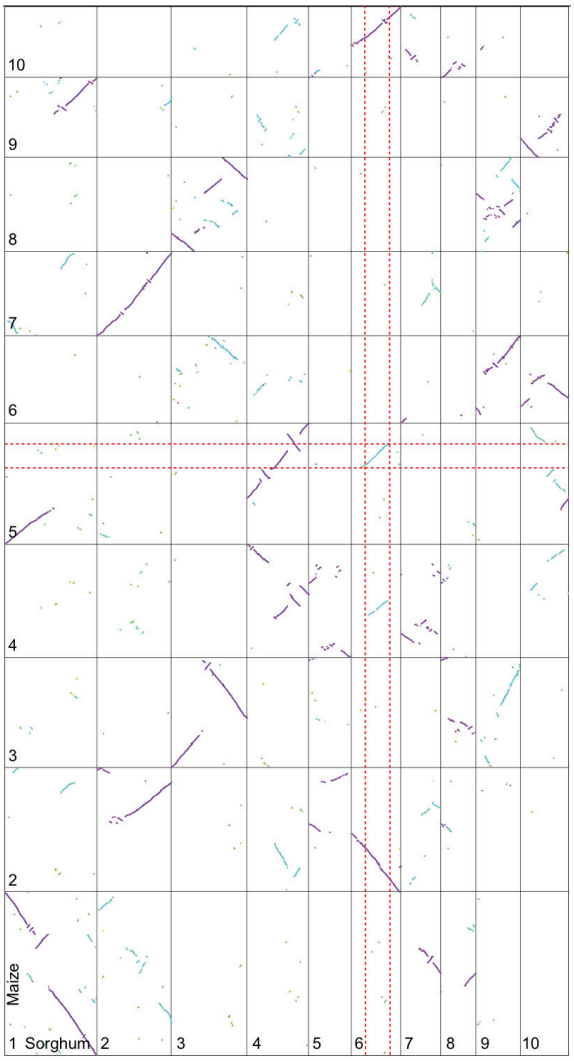


Figure 8 - Syntenic dotplot generated by SynMap. Sorghum is on the x-axis and maize is on the y-axis. Black lines separate chromosomes. Colored dots/lines are syntenic gene pairs: purple pairs are derived from the divergence of maize and sorghum's lineages; cyan from their shared whole genome duplication event. Red dashed lines show syntenic relationship between these genomes. For each region of sorghum, there are four syntenic maize regions; for each region of maize, there are two syntenic sorghum regions. The additional set in maize is due to the maize specific whole genome duplication event (Figure 1).

the result of high levels of genomic rearrangement, degrading the syntenic signal below SynMap's detection threshold. SynMap can be set to correct for apparent gaps caused by gene poor regions by selecting the "Display Options" tab in the configuration menu and selecting "Genes" in the row labeled "Dotplot axis metric". This will use the number of genes to determine the size of the chromosomes in the dotplot instead of the number of nucleotides. This option will also tend to correct for large differences in total genome size between related species which do not increase gene content, such as the transposon blooms seen in maize. Select this option and rerun the analysis (quick link: http://genomevolution.org/r/3uvw).

6. Each pair of identified syntenic genes has a relationship rooted in their evolution history. Pairs of genes derived from the pre-grass whole genome duplication event have been diverging for a longer time and are less similar to each other than pairs of genes which shared a common ancestral sequence as recently as the divergence of the the maize and Sorghum lineages (Figure 1). The maize-specific whole genome duplication event duplication created extra copies of both types of genes within maize. SynMap contains a method for determining the relative ages of syntenic gene pairs by calculating the rate of synonymous mutations (Ks) between a pair of sequences (Yang, 2007). To turn on this option, select the "Analysis Options" tab in SynMap's configuration menu and select "Synonymous (Ks)" for the row entitled "CodeML Calculate syntenic CDS pairs and color dots". When this option is selected, SynMap will calculate the synonymous substitution rate for each pair of identified syntenic genes and color them according to the derived value. Press "Generate SynMap" to rerun the analysis (quick link: http://genomevolution.org/r/3uvv).

7. When the dotplot is returned, notice that the syntenic gene pairs are generally colored purple or cyan, and lines are consistently composed of only one color (Figure 8). Examine the histogram of these values located below the dotplot (Figure 9). This histogram contains the log10 transformed Ks values with small values on the left (representing more closely related sequence). Its color scheme is used to color the dots in the syntenic dotplot. Note that while the dots in the dotplot fall into two major color types, there are a range of colors in the histogram. SynMap draws more diverged gene pairs first and more similar gene pairs will cover them up. This results in the clear demarcation of syntenic regions. Also note the large orange-red peak on the right of the histogram with non-log10 transformed Ks values between 50-150 substitutions per synonymous site. These values are in excess of what CodeML (Yang, 2007) can reliably estimate for Ks values and may be considered noise. Apparently extremely diverged gene pairs can be the result - among other things - of errors in sequence alignment, gene pairs where one or both genes are actually psuedogenes, and the attempted comparison of gene pairs which are not truly homologous. Combining the information in the histogram with the colored dots in the dotplot, it is clear that the purple lines are evolutionarily younger and represent orthologous regions in maize and Sorghum while than the cyan lines represent genomic regions which diverged much longer ago, in the pre-grass whole genome duplication.

8. The older homeologous regions can be filtered out - leaving only orthologous gene pairs - using "Quota Align". Turn on this feature in SynMap and specify a 2 maize : 1 Sorghum syntenic relationship in the text-boxes labeled "Ratio of coverage depth". Press "Generate SynMap" to rerun the analysis (quick link: http://genomevolution.org/r/3uvx).

9. The dotplot in the results now contains a subset of the syntenic genes identified in the previous analyses. Also, nearly all of the syntenic genes are colored purple (there are a few small cyan regions that were selected by Quota Align, but represent a small fraction of the total data). Note that the values in the Ks histogram are significantly biased toward less diverged gene pairs.

10. Below the dotplot and histogram (if a CodeML analysis is selected) is a log file containing all the steps of the analysis and a section entitled "Links and Downloads". The latter area contains all files used in each step of the analysis. Of note are the files under "Results" and specifically the ones labeled: "Results with synonymous/nonsynonymous rate values", "Final syntenic gene-set output with GEvo links", and "Condensed syntelog file with GEvo links". The first file (synonymous/nonsynonmous rate values) will be generated if those calculations were selected. It contains all the identified syntenic gene pairs organized by the syntenic block in which they reside. The format of the information in each line is complex, The description of this format and the information necessary to parse it is available here: http://genomevolution.org/r/3uvy. The next file (final syntenic gene-set with GEvo links) follows the same format as the prior file, but also has pre-generated GEvo links for analyzing each identified pair of syntenic genes in their genomic context. The last file (condensed syntelog) groups together syntenic gene sets across all overlapping syntenic regions and is the file of interest for this example. Since Quota Align was used to screen all non-orthologous syntenic regions, the condensed syntelog file contains all orthologous gene-sets between maize and Sorghum. Each line of the file contains one or two maize genes (the latter case if both homeologs are still present in the maize genome) and the orthologous Sorghum gene. Many maize genes do not have a homeolog because many genes have been lost following the maize-specific whole genome duplication event. In addition to the list of genes, this file contains several links to CoGe to compare genomic regions with GEvo, retrieve FASTA sequences, or create a list of the genomic features that may be

sent to other tools in CoGe for data extraction, querying other genomes, phylogenetics, etc.
11. Below the “Links and Downloads” section is an important link: “Regenerate this analysis”. This link will load SynMap with the exact configuration used to generate the results and automatically start running the analysis.

## Conclusion

Comparing multiple related genomes creates many opportunities to understand genomes at various levels. However, to take advantage of these opportunities appropriate tools must be available to the broader genetics community, the data that fuels these tools should be easily imported, and the results should be intuitive to interpret. Here we use the web-based comparative genomics package CoGe to analyze and compare the maize genome at a variety of levels using specific use-cases that many researchers may encounter. Maize, with its rich genetic history and in-depth functional characterization of many of its gene products, provides one of the most promising systems for plant comparative genomics, particularly given the number of additional sequenced grass

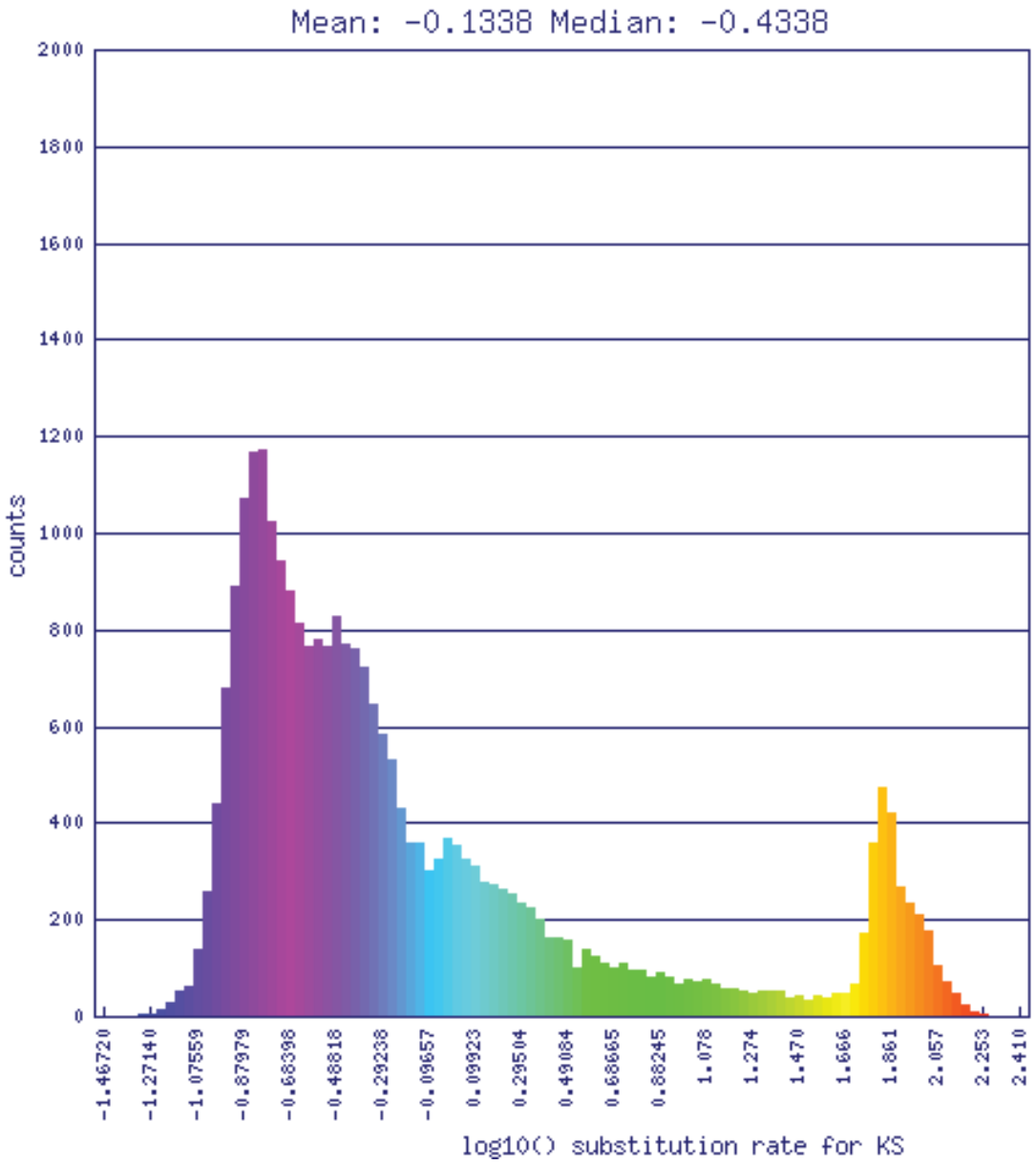


Figure 9 - Synonymous rate value histogram of syntenic gene-pairs generated by SynMap for the comparison of the maize and sorghum genomes. Values have been log10 transformed and colors on histogram match the syntenic dot colors shown in figure 6. Large peak on right of graph represents noise in the analysis due to erroneous gene models, pseudogenes, non-homologous gene-pairs, etc.

genomes available for comparison. As DNA sequencing costs continue to decrease, maize research, and the study of grasses in general, will continue to be a cornerstone for understanding the molecular function of plant genes and their products, the basic processes of plant development, and the evolution of plant genomes. Grasses as a whole will continue to be a fantastic experimental system. Their importance as one of the most dominant plant groups is undisputed, and from an agronomic standpoint, they will continue to feed most of the world and play an increasing role in fueling the planet. However, maize would not be so well positioned if it were not for the hundreds of researchers involved with bringing the maize genome to fruition, and the thousands of maize geneticists who have unraveled so many mysteries of plant function, regulation, response, development, and breeding over the past century.

## Acknowledgements

This work was supported by the iPlant Collaboration [grant number DBI 0735191] to EL and a Chang-Lin Tien Graduate Fellowship to JCS. Additional thanks to the ancient Americans who domesticated maize, without whom none of this work would have been possible.

## References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. J Mol Biol 215: 403-410

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW, 2011. GenBank. Nucleic Acids Res 39: D32-37

Bomblies K, Doebley JF, 2005. Molecular Evolution of FLORICAULA/LEAFY Orthologs in the Andropogoneae (Poaceae). Mol Biol Evol 22: 1082-1094

Douglas RN, Wiley D, Sarkar A, Springer N, Timmermans MCP, Scanlon MJ, 2010. *ragged seedling2* Encodes an ARGONAUTE7-like protein required for mediolateral expansion, but not dorsiventrality, of maize leaves. Plant Cell 22: 1441-1451

Fowler S, Lee K, Onouchi H, Samach A, Richardson K, Morris B, Coupland G Putterill J, 1999. GIGANTEA: a circadian clock-controlled gene that regulates photoperiodic flowering in Arabidopsis and encodes a protein with several possible membrane-spanning domains. EMBO J 18: 4679-4688

Freeling M, Subramaniam S, 2009. Conserved noncoding sequences (CNSs) in higher plants. Curr Opin Plant Biol 12: 126-132

Harris RS, 2007. Improved Pairwise Alignment of Genomic Data. http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf

Hayama R, Izawa T, Shimamoto K, 2002. Isolation of rice genes possibly involved in the photoperiodic control of flowering by a fluorescent differential display method. Plant Cell Physiol 43: 494-504

Izawa T, Miharab M, Suzukic Y, Guptaa M, Itoha H, Naganoa AJ, Motoyamad R, Sawadae Y, Yanog M, Yokota Hiraie M, Makinoc A, Nagamurad Y, 2011. Os-GIGANTEA Confers Robust Diurnal Rhythms on the Global Transcriptome of Rice in the Field. Plant Cell 23: 1741-1755

James MG, Robertson DS, Myers AM, 1995. Characterization of the maize gene *sugary1*, a determinant of starch composition in kernels. Plant Cell 7: 417-429

Liang C, Mao L, Ware D, Stein L, 2009. Evidence-based gene predictions in plant genomes. Genome Res 19: 1912-1923

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M, 2008a. Finding and Comparing Syntenic Regions among Arabidopsis and the Outgroups Papaya, Poplar, and Grape: CoGe with Rosids. Plant Physiol 148: 1772-1781

Lyons E, Pedersen B, Kane J, Freeling M, 2008b. The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. Tropical Plant Biol 1: 181-190

Manicacci D, Camus-Kulandaivelu L, Fourmann M, Arar C, Barrault S, Rousselet A, Feminias N, Consoli L, Francès L, Méchin V, Murigneux A, Prioul JL, Charcosset A, Damerval C, 2009. Epistatic Interactions between *Opaque2* Transcriptional Activator and Its Target Gene CyPPDK1 Control Kernel Trait Variation in Maize. Plant Physiology 150: 506 -520

Moore G, Devos KM, Wang Z, Gale MD, 1995. Cereal genome evolution. Grasses, line up and form a circle. Curr.Biol 5: 737-739

Park DH, Somers DE, Kim YS, Choy YH, Lim HK, Soh MS, Kim HJ, Kay SA, Nam HG, 1999. Control of circadian rhythms and photoperiodic flowering by the Arabidopsis GIGANTEA gene. Science 285: 1579-1582

Paterson AH, Bowers JE, Chapman BA, 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA 101: 9903-9908

Schmidt RJ, Burr FA, Aukerman MJ, Burr B, 1990. Maize regulatory gene *opaque-2* encodes a protein with a “leucine-zipper” motif that binds to zein DNA. Proc Natl Acad Sci USA 87: 46-50

Schnable JC, Freeling M, 2011. Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. PLoS One 6:e17855. doi: 10.1371/journal.pone.0017855

Schnable PS et al, 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326:1112-1115

Swigoňová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J, 2004. Close Split of Sorghum and Maize Genome Progenitors. Genome Res 14: 1916-1923

Tang H, Lyons E, Pedersen B, Schnable JC, Paterson AH, Freeling M, 2011. Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics 12: 102 doi: 10.1186/1471-2105-12-102

The International Brachypodium Initiative, 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463: 763-768

Walbot V, 2009. 10 reasons to be tantalized by the B73 maize genome. PLoS Genet. 5:e1000723. doi: 10.1371/journal.pgen.1000723

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M, 2010. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. PLoS Biol. 8:e1000409. doi: 10.1371/journal.pbio.1000409

Yang Z, 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591