



Gramene Exercises
Cereal Genomics Course
Cold Spring Harbor, New York, USA
October 19th, 2016

These exercises will illustrate the power of comparative plant genomics in research using the resources in Gramene.

Exercise 1. View a phylogenetic tree for a family of transcription factors

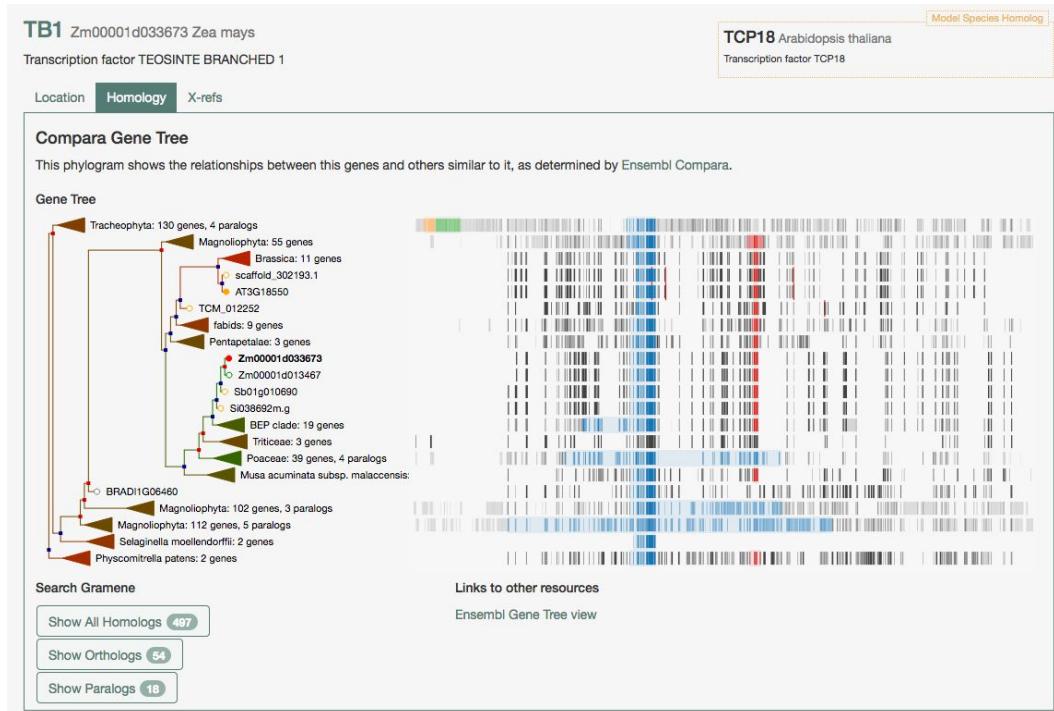
In this exercise, we will navigate a phylogenetic tree for plant genes in the TCP family of transcription factors (named after the first characterized protein members: maize TB1, snapdragon CYC, and rice PCF), highlight species-specific orthologs/paralogs with particular GO annotations in the tree. We will then proceed to generate lists of orthologs/paralogs and download both, images and tables with our results.

- a. How many orthologs can you identify for maize TB1?

Hint: You may find the answer for this through different approaches. Gramene's new search will give you the quickest answer through a snapshot of the *tb1* (Zm00001d033673) gene tree. Other approaches are described in subsequent exercises.

1. Go to www.gramene.org. This is Gramene's homepage.
2. Enter TB1 in the search box. This will redirect you to search.gramene.org.
3. Find the maize *tb1*(Zm00001d033673) gene. Click on the "Homology" tab.

Answer: There are 54 orthologs of maize TB1 in the current Gramene build #51.



- b. What is the most prominent TCP domain among members of the gene tree? How many maize genes have a TCP domain?

Note: By looking at the maize TB1 gene tree in Gramene's genome browser, 3 InterPro domains with TCP features appear to be shared among family members. IPR005333 is considered a "family" of protein domains as it encompasses TCP domains: IPR017887 and IPR017888.

Answer: Again, there are multiple ways to answer a question.

- 1) Via search.gramene.org:
 - a) Go to the "Homology" tab of the Search results for maize *tb1* (see above).
 - b) Click on the most prominent blue colored domain (IPR017887).
 - c) Simple answer: IPR017887 - Transcription factor TCP subgroup.
Note: From a closer look, wheat genes in the tree appear to be lacking protein domain annotation even though their protein sequence lines up with TCP domains.
- 2) Via the genome browser:
 - a) From the "Homology" tab in your search results (see above), click on "Ensembl Gene Tree view" OR go to ensembl.gramene.org, search for maize TB1 and click on (Plant Compara) Gene Tree ([EPIGT00820000103607](#)).

- b) Select InterPro domains in the annotations table. By selecting an individual domain, all members that share it will be highlighted in the tree.
- c) Detailed answer:
- 479 members have IPR005333 Transcription factor, TCP.
 - 477 members IPR017887 Transcription factor TCP subgroup
 - 141 members IPR017888 CYC/TB1, R domain

highlight	Accession	Description
<input type="radio"/> 479 members	IPR005333	Transcription factor, TCP
<input checked="" type="radio"/> 477 members	IPR017887	Transcription factor TCP subgroup
<input type="radio"/> 141 members	IPR017888	CYC/TB1, R domain
<input type="radio"/> 4 members	IPR020467	Potassium channel, voltage dependent, Kv1.4
<input type="radio"/> 1 member	IPR001932	PPM-type phosphatase domain
<input type="radio"/> 1 member	IPR032675	Leucine-rich repeat domain, L domain-like

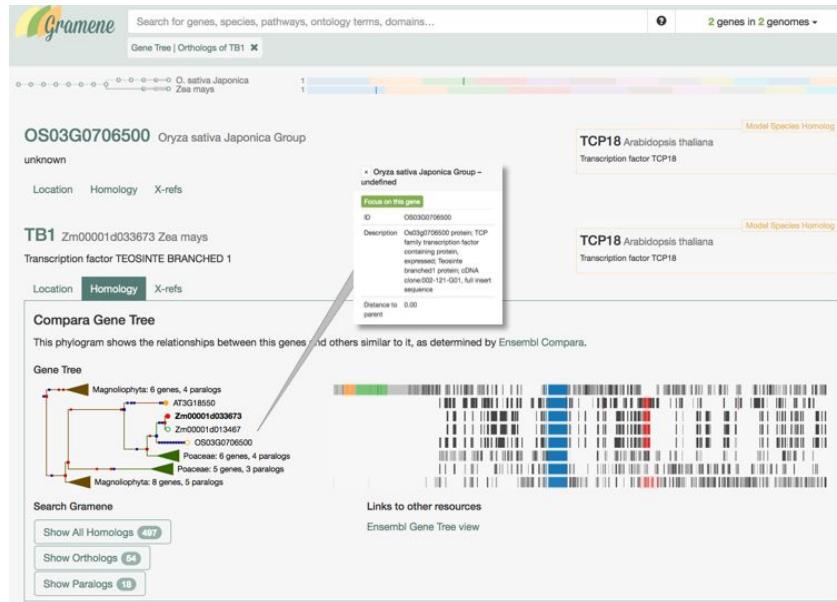
- 3) Customized data dump: Using the BioMart utility.
- Go to <http://ensembl.gramene.org/biomart/martview>.
 - Select Database: “Plant Genes” and Dataset: “Zea mays genes”.
 - Under “Protein Domains”, select “Limit to genes with these family or (InterPro) domain IDs” and enter “IPR005333”, “IPR017887” or “IPR017888”.
 - Click on “Count”. Alternatively, under “Attributes” select the associated data (e.g., gene or transcript ID, position, sequence, variants, GO terms, etc.) that you would like to download for these genes.
 - Answer: There are 46 maize genes with IPR005333, 45 with IPR017887 and 4 with IPR017888.

★ URL

Dataset 46 / 44300 Genes Zea mays genes (AGPv4 (CampbellMaker2015Dec))	Please restrict your query using criteria below
Filters InterPro ID(s): [ID-list specified]	
Attributes Gene stable ID Transcript stable ID	
Dataset [None Selected]	
REGION: <input type="checkbox"/>	
GENE: <input type="checkbox"/>	
GENE ONTOLOGY: <input type="checkbox"/>	
PLANT ONTOLOGY: <input type="checkbox"/>	
ENVIRONMENT ONTOLOGY: <input type="checkbox"/>	
GRAMENE TAXONOMIC ONTOLOGY: <input type="checkbox"/>	
GROWTH STAGE ONTOLOGY: <input type="checkbox"/>	
TRAIT ONTOLOGY: <input type="checkbox"/>	
MULTI-SPECIES COMPARISONS: <input type="checkbox"/>	
PROTEIN DOMAINS:	
<input type="checkbox"/> Limit to genes ... <input type="checkbox"/> with coiled coils (Ncoils) <input checked="" type="radio"/> Only <input type="radio"/> Excluded	
<input checked="" type="checkbox"/> Limit to genes with these family or domain IDs: <input type="text" value="IPR005333"/>	
<input type="checkbox"/> Transmembrane domains <input checked="" type="radio"/> Only <input type="radio"/> Excluded	
<input type="checkbox"/> Signal domains <input checked="" type="radio"/> Only <input type="radio"/> Excluded	
VARIATION: <input type="checkbox"/>	

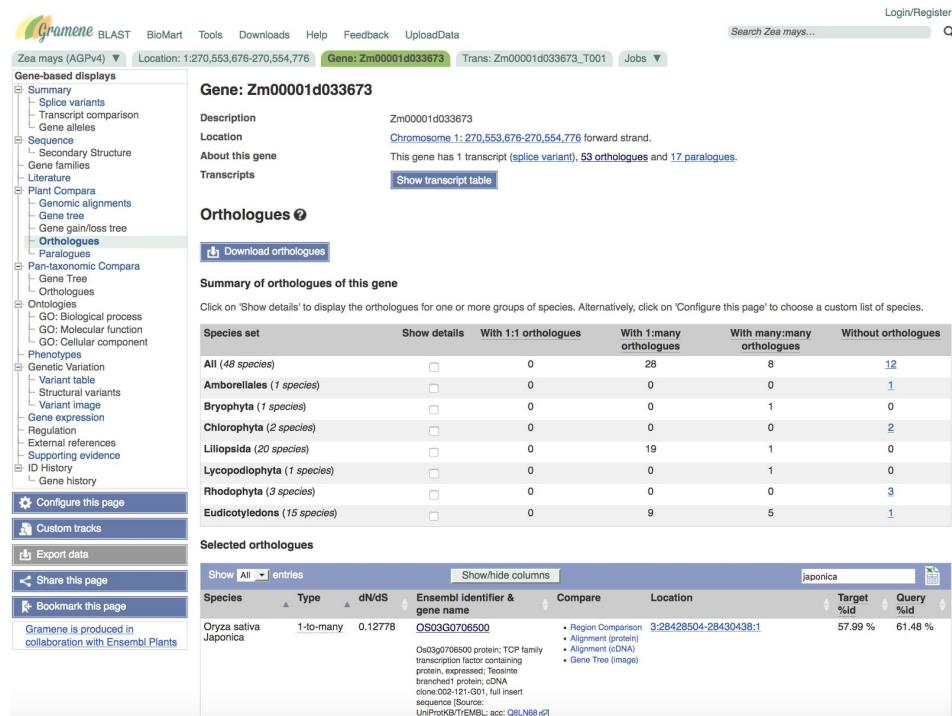
- c. You have learned 3 ways to find orthologs for a given gene (via Search, Genome Browser and BioMart). Can you identify the (*Japonica*) rice ortholog of the maize *tb1* gene and highlight both genes in the TCP gene family tree?

- 1) Via Search:
 - a) From the “Homology” tab in Search results (see above), select *Zea mays* and *Oryza sativa japonica* from the drop-down menu on the top right of the Search page.
 - b) Click on “Show Orthologs”.

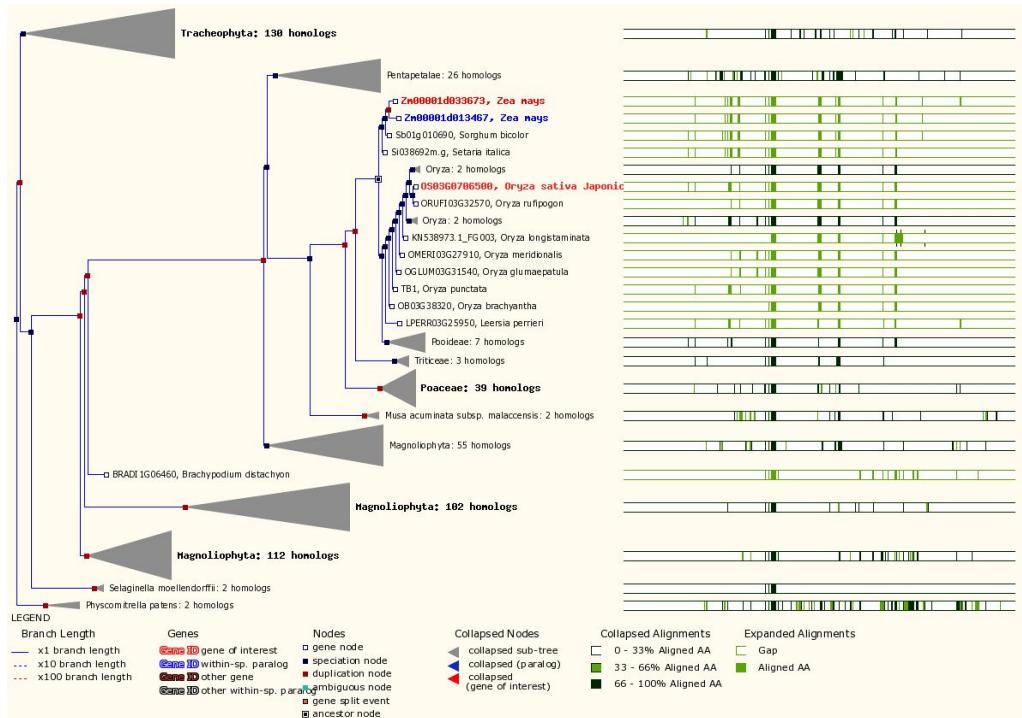


2) Via Genome Browser:

- From the left side menu of the Gene Summary page or the Plant Compara Gene Tree view (see above), select the (Plant Compara) “Orthologues” option
- Type “japonica” on the “Filter” box to select to show only rice orthologues in the results table.



- Click on the “Gene Tree (image) link” for the rice orthologue.



3) Via BioMart:

- From the “Zea Mays genes” data set in BioMart (see above), under the “Gene” filter, select “ID list limit”.
- Enter “Zm00001d033673” as the “Gene stable ID” for maize *tb1*.
- Under “Attributes”, select “Homologs”.
- From the “Homologs” attributes form, under “Gene Attributes” select “Gene stable ID”, and under “Orthologs” select “Oryza sativa Japonica gene stable ID” and any additional data desired (e.g., % identity).
- Click on “Results”. Customize how to view and export your results.

New **Count** **Results** **URL** **XML** **Perf** **Help**

Please select columns to be included in the output and hit 'Results' when ready

Features Homologs
 Structures Sequences
 Variation

GENE:
Gene Attributes
 Gene stable ID
 Transcript stable ID
 Protein stable ID
 Chromosome/scaffold name
 Gene start (bp)
 Gene end (bp)
 Strand
 Band
 Gene name
 Source of gene name
 Gene description
 Gene biotype
 % GC content
 Transcript count

ORTHOLOGS:
Aegilops tauschii Orthologs
 Aegilops tauschii gene stable ID
 Aegilops tauschii protein stable ID
 Aegilops tauschii chromosome/scaffold
 Aegilops tauschii start (bp)
 Aegilops tauschii end (bp)
 Representative protein or transcript ID
 Ancestor
 Homology type
 % identity
 Aegilops tauschii % identity
 dN
 dS
 Orthology confidence [0 low, 1 high]

Amborella trichopoda Orthologs
 Amborella trichopoda gene stable ID
 Amborella trichopoda protein stable ID
 Amborella trichopoda chromosome/scaffold
 Amborella trichopoda start (bp)
 Amborella trichopoda end (bp)
 Representative protein or transcript ID
 Ancestor
 Homology type
 % identity
 Amborella trichopoda % identity
 dN
 dS
 Orthology confidence [0 low, 1 high]

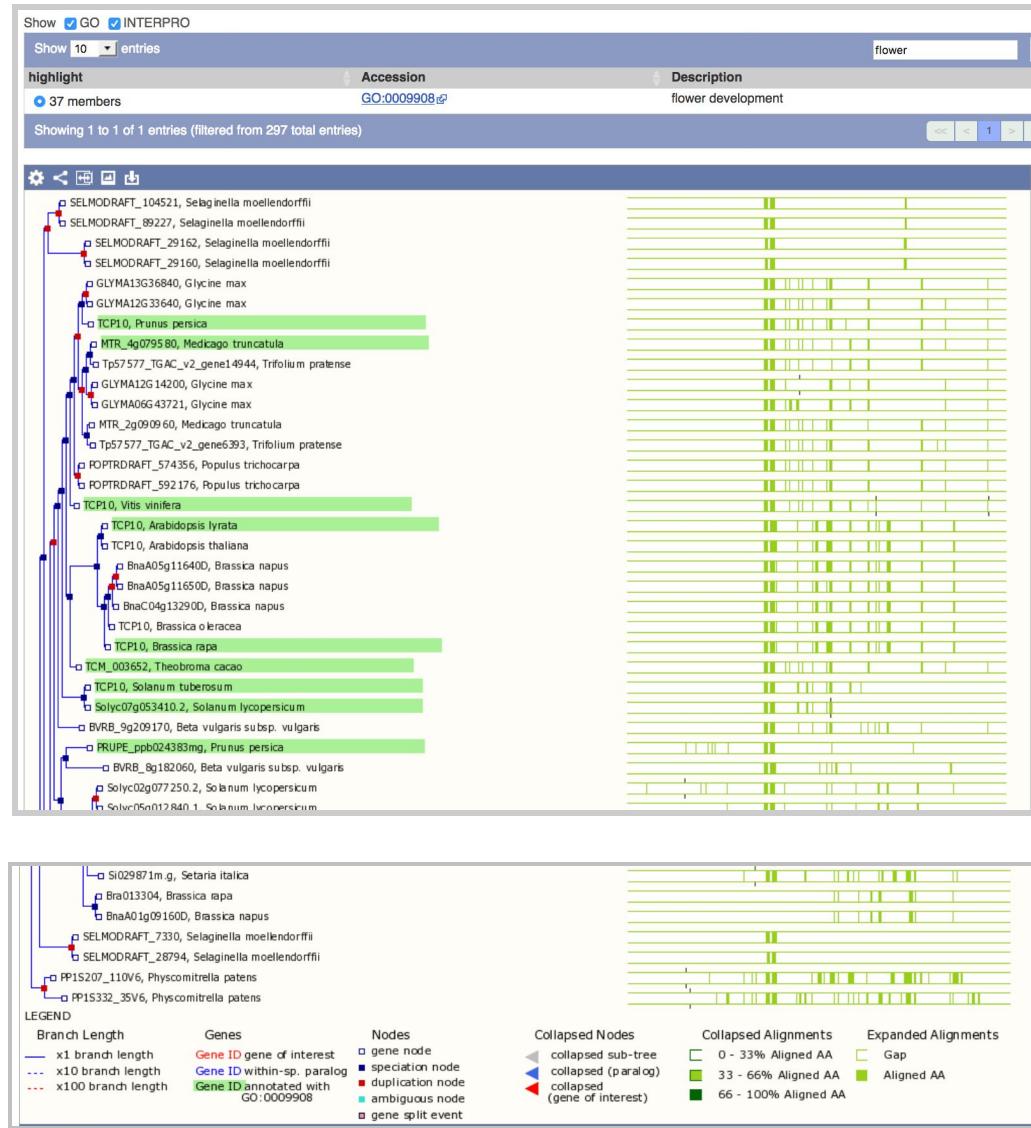
Arabidopsis lyrata Orthologs
 Arabidopsis lyrata gene stable ID
 Homoloav type

Answer: [OS03G0706500 \(IRGSP1\)](#) or [LOC_Os03g49880 \(MSU6\)](#).

- d. Identify genes in the tree that have been associated with flower development.

Hint: [GO:0009908](#) is the GO term identifier for “flower development”.

1. From the Plant Compara Gene Tree view (see above), enter the term “flower” in the Filter box to identify GO or InterPro term(s) for flower development.
2. Select GO:0009908



Exercise 2. Identify tomato transcription factors within the TCP gene family with a SNP that results in a truncated peptide.

Note: The SNP will introduce a stop codon(*) resulting in a truncated protein product.

Hints:

- 1) Via Search and/or Genome Browser:
 - a) From the Transcript page of Solyc06g069240.1, the tomato ortholog of maize *tb1* (proceed as above for the rice ortholog of TB1)
 - b) Select “Domains & features” from the left side menu.

- c) Find the IRP017887 domain and click on “Display all genes with this domain”.
- d) Copy the resulting gene list and use it as input to mine for tomato variants with a “stop_gained” as functional consequence.

The screenshot shows the Solanum lycopersicum (SL2.50) transcript-based displays page for Solyc06g069240.1. The transcript is located on Chromosome 6: 43,006,781-43,008,127, forward strand. It has 2 exons, 8 domains/features, and 116 variations. The Domains & features table lists PANTHER, PROSITE profiles, and Pfam domains. The Other features table lists a low-complexity segment from position 185 to 204.

Domain type	Start	End	Description	Accession	InterPro
PANTHER	17	173	-	PTHR31072_0	
PANTHER	17	173	-	PTHR31072_SF6_0	
PANTHER	191	297	-	PTHR31072_0	
PANTHER	191	297	-	PTHR31072_SF6_0	
PROSITE profiles	255	272	CYC/TB1, R domain	PS51370_0	IPR017887_0 [Display all genes with this domain]
PROSITE profiles	115	173	Transcription factor TCP subgroup	PS51369_0	IPR017887_0 [Display all genes with this domain]
Pfam	114	275	Transcription factor, TCP	PF03634_0	IPR005333_0 [Display all genes with this domain]

- 2) Via BioMart: First use IPR017887 (TCP domain) to Filter the tomato genes data set. Then select tomato variations as a second data set and under Filters, use “stop_gained” as “Consequence type”.

The screenshot shows the BioMart results page for Dataset 34 / 38735 Genes. A filter for IPR017887 (TCP domain) has been applied. The results table includes columns for InterPro ID, short description, gene stable ID, variation ID, chromosome name, position on chromosome (bp), and consequence to transcript. The consequences listed are mostly stop_gained.

InterPro ID	InterPro short description	Gene stable ID	Variation ID	Chromosome name	Position on chromosome (bp)	Consequence to transcript
IPR017887	TF_TCP_subgr	Solyc02g0094290.1	vcZ11HNA9	2	54869495	stop_gained
IPR017887	TF_TCP_subgr	Solyc01g103780.2	vcZ11HNA9	1	92305148	stop_gained
IPR017887	TF_TCP_subgr	Solyc02g007259.2	vcZ13UIURK	2	42244608	stop_gained
IPR017887	TF_TCP_subgr	Solyc02g007259.2	vcZ13UIURK	2	42244773	stop_gained
IPR017887	TF_TCP_subgr	Solyc02g006200.1	vcZ13LECR	2	38016698	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g009030	vcZ17BLBRI	9	1505249	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g009030	vcZ17BLBSE	9	1505460	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g048390	vcZ17INTMO	8	13139533	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g048390	vcZ17INTNO	8	13139560	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g048390	vcZ17NTIOE	8	13139634	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g048390	vcZ17NTIOL	8	13139867	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g048390	vcZ17NTIP7	8	13140047	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g048390	vcZ17NTIP9	8	13140068	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g048390	vcZ17NTIP9	8	13140107	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g048390	vcZ17NTIP9	8	13140160	stop_gained
IPR017887	TF_TCP_subgr	Solyc09g048390	vcZ17NTIP9	8	13140221	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g089240.1	vcZ17HQBHB	8	43008047	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g018940.1	vcZ17CCDAV	5	5987540	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g018940	vcZ17CCDAY	5	5987573	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g018940	vcZ17CCDBU	5	5987629	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g018940	vcZ17CCDCB	5	5987604	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g018940	vcZ17CCDCG	5	5988050	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vcZ17NT61M	8	13379406	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vcZ17NT61P	8	13379473	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vcZ17NT62P	9	13379447	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vcZ17NT63B	8	13380111	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vcZ17NT63M	8	13380265	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vcZ17NT63T	8	13380326	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g033780	vcZ17DEEDS	5	44839431	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g033780	vcZ17DEEPQ	5	44839867	stop_gained

Exercise 3. Explore the genetic variation associated with a gene

We will now explore genetic variants along the *Arabidopsis c/e18* gene to find 2 SNPs reported to have drastic functional consequences for the CLE18 peptide. CLE18 is a CLAVATA3/ESR-related (CLE) peptide with diverse roles in plant growth and development. Two functional consequences (Cao *et al*, 2011)[Nature Genetics].

- a. Visualize the genetic variants for this gene
- b. Are there any new stop codons introduced (nonsense variants) in this gene? Compare your findings with Supplementary Table 3
- c. Are there any transcript-specific variants for this gene?
- d. Download a subset of the variants (e.g., those that introduce an amino acid change in the protein)

Note: In addition to the Ensembl “Tools” for genomic analysis, other genetic analysis (e.g., Simple Sequence Repeat Identification Tool or SRIT) tools can be accessed through Gramene’s archival Diversity pages at
<http://archive.gramene.org/diversity/tools.html>

Exercise 4. Explore a metabolic pathway and compare it in two species

Let’s now look at the carotenoid biosynthesis pathway in Gramene. You will notice that we currently offer two pathway platforms: Plant Reactome (rice reference pathways & orthology-based projections to 64 species) and BioCyc-based (10 plant species now served via Cyverse @
<http://pathway.iplantcollaborative.org>).

- a. Search for “carotenoid biosynthesis”
- b. Browse through results categories
- c. Select annotations in the Plant Reactome (rice)
- d. Download a list of proteins associated with this rice pathway in Reactome
- e. Go back to the search results and look up for what species have annotated pathways in the BioCyc platform
- f. Select rice and compare with maize
- g. Download a list of all the genes associated with a carotenoid biosynthesis pathway in each, rice and maize
- h. Check out other resources for maize pathways in MaizeGDB.Org

Exercise 5. Upload, visualize and share your own data into a new genome browser track

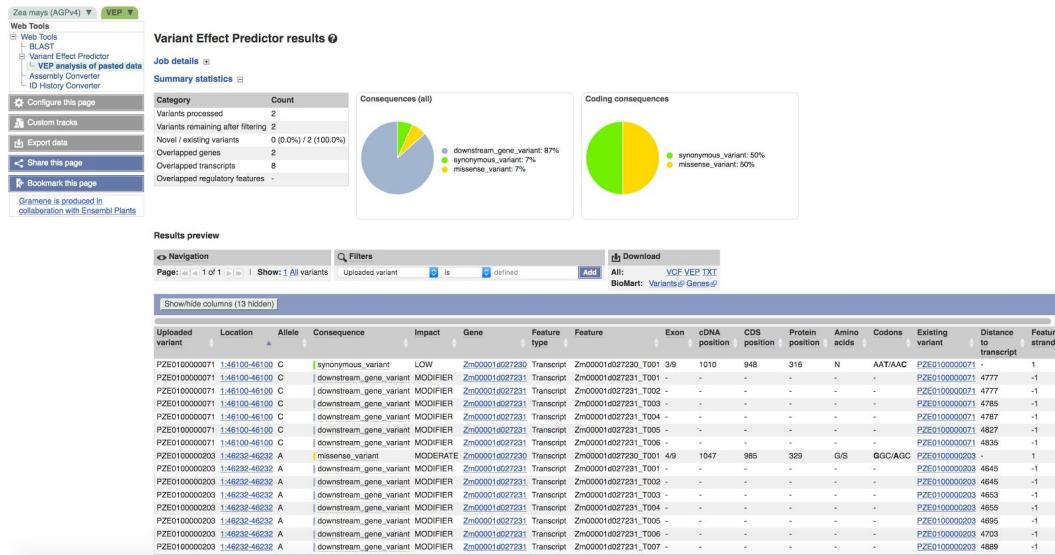
The Ensembl genome browser allows users to upload their own data and visualize it on a custom track. Data may be formatted in various file formats including GFF, GTF, BED, BAM, VCF, bedGraph, gbrowse, PSL, WIG, BigBed, BigWig, and TrackHub. Some data like GFF annotations may be directly uploaded from a local machine. Large data files like BED/BAM alignments or BigWig graphic display configurations need to be uploaded onto a local server that is accessible to the browser via an URL. Another way to share third-party data is via a DAS (Distributed Annotation System) registry, which would need to be set up by a software engineer.

The test data sets that we will upload and visualize for this exercise have been preloaded onto a local server that is publicly accessible:

http://data.gramene.org/public/Zea_mays4m/methylome/. The data consists of BAM alignments and CpG methylation for B73 & Mo17 maize lines used in the study by Regulski *et al* (2013) [Genome Research 23:1651] and were used to create expression tracks in Gramene build 45.

Copy/paste the following sample VCF:

```
##fileformat=VCFv4.0
##fileDate=20161018
##source=MaizeHapMapMockUp
##reference=RefGenv4
##phasing=no
##INFO=<ID=MQ,Number=.,Type=Float,Description="RMS mapping quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT
B73:MZ M97:MZ MKN009:MZ MKN010:MZ MKN011:MZ
1 46100 PZE0100000071 T C . PASS MQ=92 GT 0/0
0/0 0/0 0/0 0/0
1 46232 PZE0100000203 G A . PASS MQ=91 GT 0/0
0/0 0/0 0/0 0/0
```



Exercise 6. BLAST a sequence. Determine synteny for a genomic region. Convert coordinates between different genome assemblies.

In this exercise, we will identify orthologues of a species whose reference genome is not available in Gramene via BLASTX and find corresponding synteny blocks in other species.

- a. Use the nucleotide sequence of the *Sorghum virgatum* Sh1 gene taken from Lin *et al* (2012) [Nature Genetics 44:720] to identify orthologous genes in maize, rice, and *Arabidopsis thaliana*.

Note: The corresponding gene in *S. bicolor* appears to be missing two exons.

- b. Highlight the orthologs in two of those species in the tree as you learned in Exercise 1.
- c. Download the genetic variation for each of the maize Sh1 orthologs as you learned in Exercise 2. How many nonsense substitutions can you find in each of these genes?
- d. Lin *et al* (2012) also provide RefGen_v2 coordinates for maize shattering QTLs in Supplementary Table 5. Identify synteny blocks for the intervals at maize chromosomal regions (RefGen_v2) chr1: 259,223,260 - 261,622,457 and chr5: 15,806,322 - 16,428,681 in rice and sorghum. Download the synteny images that you generate.

Note: You will need to first use the Assembly converter tool to map the QTL intervals to RefGen_v3 and subsequently to RefGen_v4 coordinates. This will be publicly available in the upcoming Gramene build 52.

- e. Download all the genes for a given synteny block. Can you identify a *Sh1* orthologous (YABBY-like) gene in it?
- f. Compare your results with those in Lin *et al* (2012) [Nature Genetics 44:720]

>S. variegatum Sh1 CDS

```
ATGTCGGCCCAGCAGATCGCGCCGGTGCCGGAGCATGTGTGCTACGTGCACTGCAACTT  
CTGCAATAACAATTCTCGCGGTCACTGTCCCCGAGTCACAGCATGCTGAACATCGTGACAG  
TCCGTTGTGGGCACTGCACTAGCCTGCTGTCAGTGAACTTGAGAGGACTCCTCCAATCA  
CTCCCTGTCCAGAACATCACTACTCGCAGGAGAATAATTCAAGGTCCAAAATTTCAGCTT  
TACTGAAAACCTACCCCTGAGTATGCACCTTCGTCTCGAAATACCGCATGCCAACGATGT  
TGTCAGCAAAAGGTGATCTGGATCATATGCTGCACGTGCGTGGTAAGCTCCAGAGAAGA  
GGCAACGTGTTCTTCAGCATATAACAGATTATTAAAGGAAGAGATACGAAGGGATTA  
GCAAGCAACCCAGACATAAGCCACAGGGAAGCCTTCAGCACTGCAGCAAAAAATTGGGC  
ACATTTCACAAACATTCACTTTGGACTAGGGCCCTATGAAAGTAGCAACAAAGCTTGATG  
AGGCCATTGGTGCAACGGGCCATCCCCAGAAAGTCCAAGATCTCTACTAA
```