



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

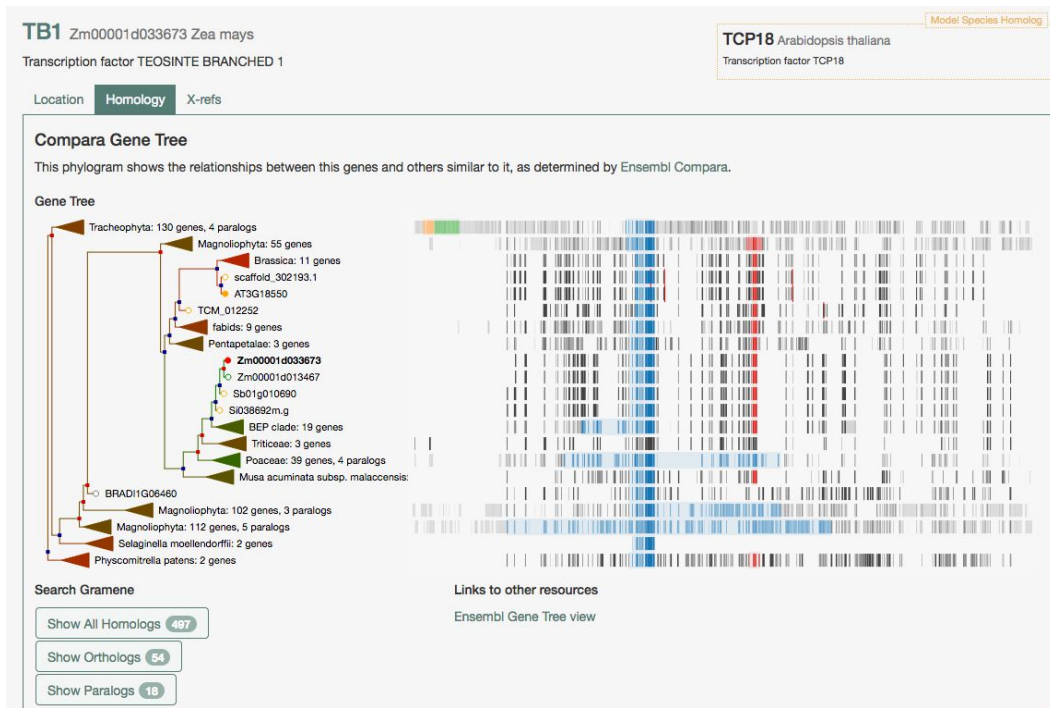
In this exercise, we will navigate a phylogenetic tree for plant genes in the TCP family of transcription factors (named after the first characterized protein members: maize TB1, snapdragon CYC, and rice PCF), highlight species-specific orthologs/paralogs with particular GO annotations in the tree. We 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:
 - i) 479 members have [IPR005333](#) Transcription factor, TCP.
 - ii) 477 members [IPR017887](#) Transcription factor TCP subgroup
 - iii) 141 members [IPR017888](#) CYC/TB1, R domain

Gene: Zm00001d033673

Gene tree

Number of genes: 497
 Number of speciation nodes: 361
 Number of duplication nodes: 128
 Number of ambiguous nodes: 6
 Number of gene split events: 1

Annotations table

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

Showing 1 to 6 of 6 entries (filtered from 297 total entries)

- 3) Customized data dump: Using the BioMart utility.
 - a) Go to <http://ensembl.gramene.org/biomart/martview>.
 - b) Select Database: “Plant Genes” and Dataset: “Zea mays genes”.
 - c) Under “Protein Domains”, select “Limit to genes with these family or (InterPro) domain IDs” and enter “IPR005333”, “IPR017887” or “IPR017888”.
 - d) 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.
 - e) Answer: There are 46 maize genes with IPR005333, 45 with IPR017887 and 4 with IPR017888.

Dataset 46 / 44300 Genes
Zea mays genes (AGPv4 (CampbellMaker2015Dec))

Filters
InterPro ID(s): [ID-list specified]

Attributes
Gene stable ID
Transcript stable ID

Dataset
[None Selected]

Please restrict your query using criteria below

REGION:

GENE:

GENE ONTOLOGY:

PLANT ONTOLOGY:

ENVIRONMENT ONTOLOGY:

GRAMENE TAXONOMIC ONTOLOGY:

GROWTH STAGE ONTOLOGY:

TRAIT ONTOLOGY:

MULTI-SPECIES COMPARISONS:

PROTEIN DOMAINS:

Limit to genes ... with coiled coils (Ncoils) Only Excluded

Limit to genes with these family or domain IDs:

InterPro ID(s)

Browse... No file selected.

Transmembrane domains Only Excluded

Signal domains Only Excluded

VARIATION:

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”.

Gramene Search for genes, species, pathways, ontology terms, domains... 2 genes in 2 genomes

Gene Tree | Orthologs of TB1

OS03G0706500 *Oryza sativa* Japonica Group
unknown

Location Homology X-refs

TB1 Zm00001d033673 *Zea mays*
Transcription factor TEOSINTE BRANCHED 1

Location Homology X-refs

Compara Gene Tree
This phylogram shows the relationships between this genes and others similar to it, as determined by Ensembl Compara.

Gene Tree
Magnoliophyta: 6 genes, 4 paralogs
Zm00001d033673
Zm00001d013467
OS03G0706500
Poaceae: 6 genes, 4 paralogs
Poaceae: 5 genes, 3 paralogs
Magnoliophyta: 8 genes, 9 paralogs

Search Gramene
Show All Homologs (497)
Show Orthologs (54)
Show Paralogs (18)

Links to other resources
Ensembl Gene Tree view

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.

Gramene BLAST BioMart Tools Downloads Help Feedback UploadData Login/Register

Zea mays (AGPv4) Location: 1:270,553,676-270,554,776 Gene: Zm00001d033673 Trans: Zm00001d033673_T001 Jobs

Gene-based displays
Summary
Splice variants
Transcript comparison
Gene alleles
Sequence
Secondary Structure
Gene families
Literature
Plant Compara
Genomic alignments
Gene tree
Gene gain/loss tree
Orthologues
Paralogues
Pan-taxonomic Compara
Gene Tree
Orthologues
Ontologies
GO: Biological process
GO: Molecular function
GO: Cellular component
Phenotypes
Genetic Variation
Variant table
Structural variants
Variant image
Gene expression
Regulation
External references
Supporting evidence
ID History
Gene history
Configure this page
Custom tracks
Export data
Share this page
Bookmark this page

Gramene is produced in collaboration with Ensembl Plants

Gene: Zm00001d033673
Description Zm00001d033673
Location Chromosome 1: 270,553,676-270,554,776 forward strand.
About this gene This gene has 1 transcript (splice variant), 53 orthologues and 17 paralouges.
Transcripts Show transcript table

Orthologues
Download orthologues

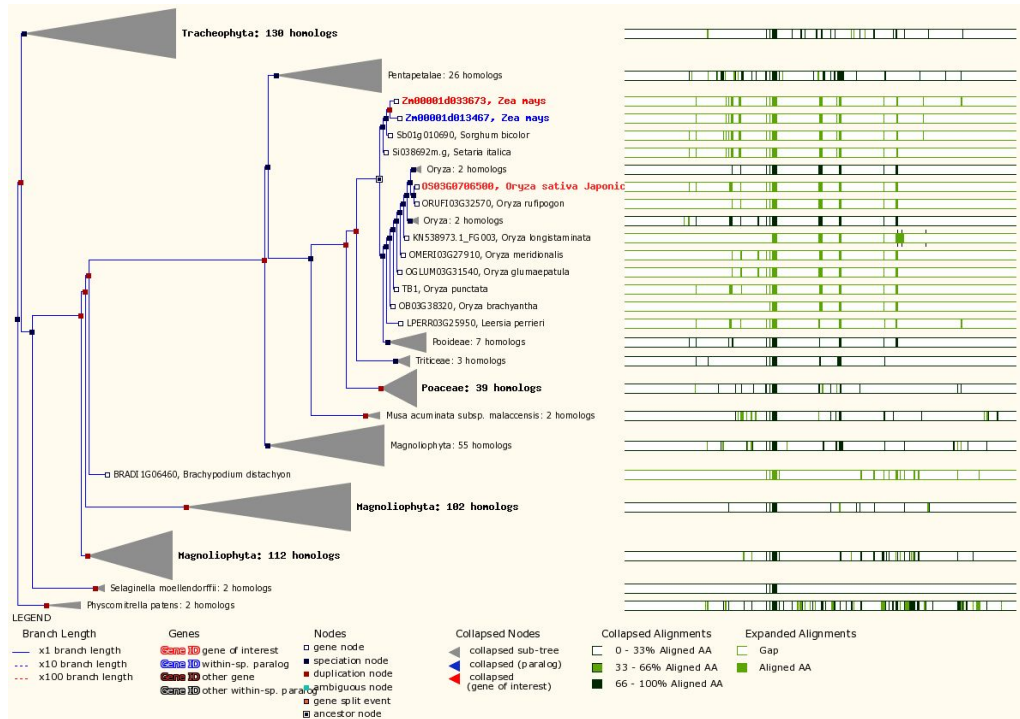
Summary of orthologues of this gene
Click on 'Show details' to display the orthologues for one or more groups of species. Alternatively, click on 'Configure this page' to choose a custom list of species.

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (48 species)	<input type="checkbox"/>	0	28	8	12
Amboresiales (1 species)	<input type="checkbox"/>	0	0	0	1
Bryophyta (1 species)	<input type="checkbox"/>	0	0	1	0
Chlorophyta (2 species)	<input type="checkbox"/>	0	0	0	2
Liliopsida (20 species)	<input type="checkbox"/>	0	19	1	0
Lycopodiophyta (1 species)	<input type="checkbox"/>	0	0	1	0
Rhodophyta (3 species)	<input type="checkbox"/>	0	0	0	3
Eudicotyledons (15 species)	<input type="checkbox"/>	0	9	5	1

Selected orthologues

Species	Type	dN/dS	Ensembl Identifier & gene name	Compare	Location	Target %id	Query %id
Oryza sativa Japonica	1-to-many	0.12778	OS03G0706500	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) Gene Tree (image) 	3:28428504-28430438:1	57.99 %	61.48 %

- 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 Port Help

Dataset 1 / 44300 Genes
Zea mays genes (AGPv4 (CampbellMaker2015Dec))

Filters
Gene stable ID(s): [ID-list specified]

Attributes
Gene stable ID
Oryza sativa Japonica gene stable ID
% identity

Dataset
[None Selected]

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

Features Homologs
 Structures Sequences
 Variation

GENE:

Gene Attributes

<input type="checkbox"/> Gene stable ID	<input type="checkbox"/> Band
<input type="checkbox"/> Transcript stable ID	<input type="checkbox"/> Gene name
<input type="checkbox"/> Protein stable ID	<input type="checkbox"/> Source of gene name
<input type="checkbox"/> Chromosome/scaffold name	<input type="checkbox"/> Gene description
<input type="checkbox"/> Gene start (bp)	<input type="checkbox"/> Gene biotype
<input type="checkbox"/> Gene end (bp)	<input type="checkbox"/> % GC content
<input type="checkbox"/> Strand	<input type="checkbox"/> Transcript count

ORTHOLOGS:

Aegilops tauschii Orthologs

<input type="checkbox"/> Aegilops tauschii gene stable ID	<input type="checkbox"/> Homology type
<input type="checkbox"/> Aegilops tauschii protein stable ID	<input type="checkbox"/> % identity
<input type="checkbox"/> Aegilops tauschii chromosome/scaffold	<input type="checkbox"/> Aegilops tauschii % identity
<input type="checkbox"/> Aegilops tauschii start (bp)	<input type="checkbox"/> dN
<input type="checkbox"/> Aegilops tauschii end (bp)	<input type="checkbox"/> dS
<input type="checkbox"/> Representative protein or transcript ID	<input type="checkbox"/> Orthology confidence [0 low, 1 high]
<input type="checkbox"/> Ancestor	

Amborella trichopoda Orthologs

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

Arabidopsis lyrata Orthologs

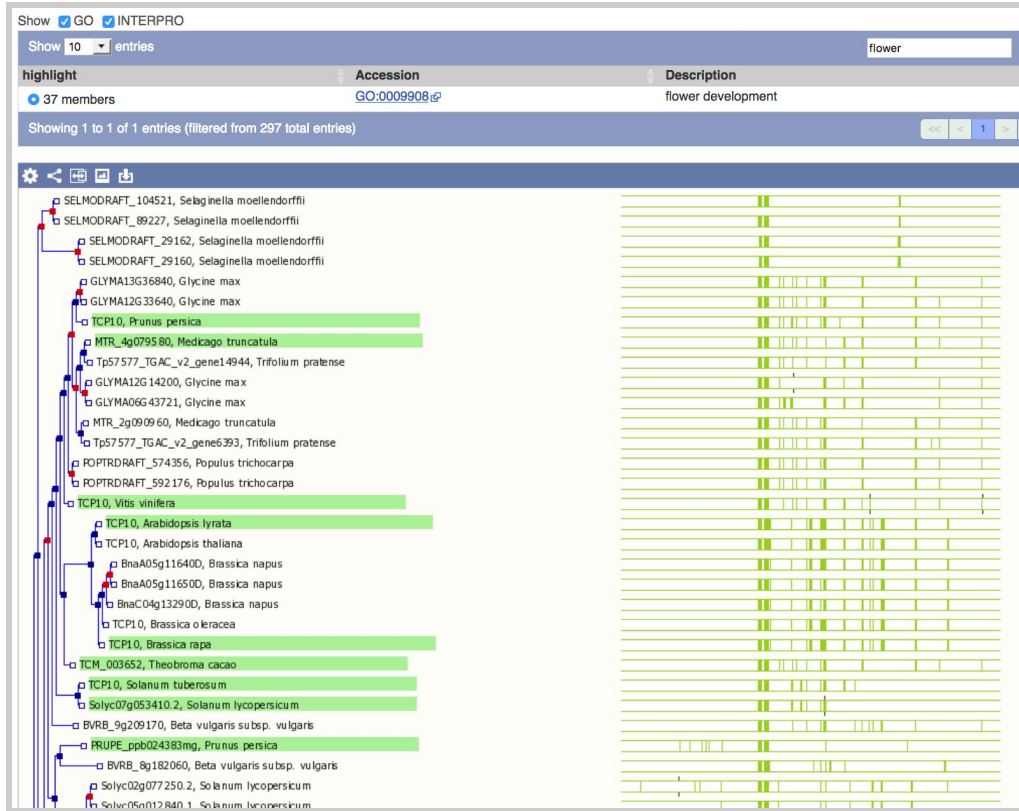
<input type="checkbox"/> Arabidopsis lyrata gene stable ID	<input type="checkbox"/> Homology type
--	--

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

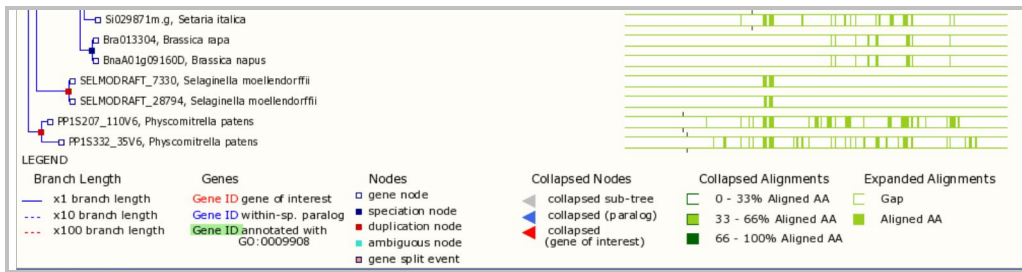
- d. Identify genes in the tree that have been associated with flower development.

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.

Transcript: Solyc06g069240.1.1

Location: Chromosome 6: 43,006,781-43,008,127 forward strand.

About this transcript: This transcript has 2 exons, is annotated with 8 domains and features and is associated with 116 variations.

Gene: This transcript is a product of gene Solyc06g069240.1

Domains & features

Domain type	Start	End	Description	Accession	InterPro
PANTHER	17	173	-	PTHR31072	-
PANTHER	17	173	-	PTHR31072:SF6	-
PANTHER	191	297	-	PTHR31072	-
PANTHER	191	297	-	PTHR31072:SF6	-
PROSITE profiles	255	272	CYC/TB1, R domain	PSS1370	IPR017888 [Display all genes with this domain]
PROSITE profiles	115	173	Transcription factor TCP subgroup	PSS1369	IPR017887 [Display all genes with this domain]
Pfam	114	275	Transcription factor, TCP	PF03634	IPR005333 [Display all genes with this domain]

Other features

Feature type	Start	End
Low complexity (Seg)	185	204

- 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”.

Export all results to: File TSV Unique results only

View: 50 rows as HTML Unique results only

InterPro ID	InterPro short description	Gene stable ID	Variation ID	Chromosome name	Position on chromosome (bp)	Consequence to transcript
IPR017887	TF_TCP_subgr	Solyc02g084290.1	vc214MABR	2	54869495	stop_gained
IPR017887	TF_TCP_subgr	Solyc01g103780.2	vc211HNA9	1	82305148	stop_gained
IPR017887	TF_TCP_subgr	Solyc02g077250.2	vc213UIBK	2	42244608	stop_gained
IPR017887	TF_TCP_subgr	Solyc02g077250.2	vc213UIBW	2	42244773	stop_gained
IPR017887	TF_TCP_subgr	Solyc02g069200.1	vc213LECC	2	35219628	stop_gained
IPR017887	TF_TCP_subgr	Solyc02g069200.1	vc213LECR	2	35219656	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g006030.1	vc21PBLHI	6	15052249	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g006030.1	vc21PBLSE	6	15054640	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048390.1	vc21NTIMO	6	13139333	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048390.1	vc21NTIRJ	6	13139968	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048390.1	vc21NTICL	6	13139867	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048390.1	vc21NTIP7	6	13140047	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048390.1	vc21NTIP9	6	13140068	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048390.1	vc21NTIPJ	6	13140137	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048390.1	vc21NTIPL	6	13140180	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048390.1	vc21NTIP5	6	13140221	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g069240.1	vc21HOBH8	6	43008047	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g012840.1	vc21CCDAV	5	5987540	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g012840.1	vc21CCDAW	5	5987573	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g012840.1	vc21CCDCE	5	5987705	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g012840.1	vc21CCDBU	5	5987804	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g012840.1	vc21CCDC2	5	5987870	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g012840.1	vc21CCDC3	5	5988050	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vc21NT61M	6	13379406	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vc21NT62A	6	13379783	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vc21NT62P	6	13379948	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vc21NT638	6	13380111	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vc21NT63M	6	13380265	stop_gained
IPR017887	TF_TCP_subgr	Solyc06g048370.2	vc21NT63T	6	13380326	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g032780.1	vc21EEDCS	5	44839431	stop_gained
IPR017887	TF_TCP_subgr	Solyc05g032780.1	vc21EEDPQ	5	44839867	stop_gained

Exercise 3. Explore the genetic variation associated with a gene

We will now explore genetic variants along the Arabidopsis *cle18* 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
```

Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	2
Variants remaining after filtering	2
Novel / existing variants	0 (0.0%) / 2 (100.0%)
Overlapped genes	2
Overlapped transcripts	8
Overlapped regulatory features	-

Consequences (all)

- downstream_gene_variant: 87%
- synonymous_variant: 7%
- missense_variant: 7%

Coding consequences

- synonymous_variant: 50%
- missense_variant: 50%

Results preview

Uploaded variant	Location	Allele	Consequence	Impact	Gene	Feature type	Feature	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Distance to transcript	Feature strand
PZE0100000071	1:46100-46100	C	synonymous_variant	LOW	Zm00001.0027230	Transcript	Zm00001.0027230_T001	3/9	1010	948	316	N	AATAAC	PZE0100000071	-	1
PZE0100000071	1:46100-46100	C	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T001	-	-	-	-	-	-	PZE0100000071	4777	-1
PZE0100000071	1:46100-46100	C	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T002	-	-	-	-	-	-	PZE0100000071	4777	-1
PZE0100000071	1:46100-46100	C	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T003	-	-	-	-	-	-	PZE0100000071	4785	-1
PZE0100000071	1:46100-46100	C	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T004	-	-	-	-	-	-	PZE0100000071	4787	-1
PZE0100000071	1:46100-46100	C	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T005	-	-	-	-	-	-	PZE0100000071	4827	-1
PZE0100000071	1:46100-46100	C	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T006	-	-	-	-	-	-	PZE0100000071	4836	-1
PZE0100000203	1:46232-46232	A	missense_variant	MODERATE	Zm00001.0027230	Transcript	Zm00001.0027230_T001	4/9	1047	985	329	G/S	GGCAGC	PZE0100000203	-	1
PZE0100000203	1:46232-46232	A	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T001	-	-	-	-	-	-	PZE0100000203	4645	-1
PZE0100000203	1:46232-46232	A	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T002	-	-	-	-	-	-	PZE0100000203	4645	-1
PZE0100000203	1:46232-46232	A	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T003	-	-	-	-	-	-	PZE0100000203	4653	-1
PZE0100000203	1:46232-46232	A	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T004	-	-	-	-	-	-	PZE0100000203	4655	-1
PZE0100000203	1:46232-46232	A	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T005	-	-	-	-	-	-	PZE0100000203	4655	-1
PZE0100000203	1:46232-46232	A	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T006	-	-	-	-	-	-	PZE0100000203	4703	-1
PZE0100000203	1:46232-46232	A	downstream_gene_variant	MODIFIER	Zm00001.0027231	Transcript	Zm00001.0027231_T007	-	-	-	-	-	-	PZE0100000203	4889	-1

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
CTGCAATACAATTCTCGCGGTCAGTGTCCCGAGTCACAGCATGCTGAACATCGTGACAG
TCCGTTGTGGGCACTGCACTAGCCTGCTGTCAGTGAAC TTGAGAGGACTCCTCCAATCA
CTCCCTGTCCAGAATCACTACTCGCAGGAGAATAATTTCAAGGTCCAAAATTTAGCTT
TACTGAAAAC TACCCTGAGTATGCACCTTCGTCTTTCGAAATACCGCATGCCAACGATGT
TGTCAGCAAAAGGTGATCTGGATCATATGCTGCACGTGCGTGGTAAGCTCCAGAGAAGA
GGCAACGTGTTCCCTCAGCATATAACAGATTTATTAAGGAAGAGATACGAAGGATTAAA
GCAAGCAACCCAGACATAAGCCACAGGGAAGCCTTCAGCACTGCAGCAAAAAATTGGGC
ACATTTTCCAAACATTCATTTTGGACTAGGGCCCTATGAAAGTAGCAACAAGCTTGATG
AGGCCATTGGTGCAACGGGCCATCCCCAGAAAGTCCAAGATCTCTACTAA
```