- JSMC practical course script -

Inferring phylogeny based on sequence information

Collection of homologous sequences using synteny

Search for the homologous sequences of the following genes from *Arabidopsis thaliana*. The sequence search will be conducted on the Brassica Database Webseite (http://brassicadb.org/brad/searchSyntenytPCK.php).

Use the following Locus IDs for your search to find the homologous genes in the species *Brassica rapa*, *Schrenkiella parvula*, *Leavenworthia alabamica* and *Capsella rubella*.

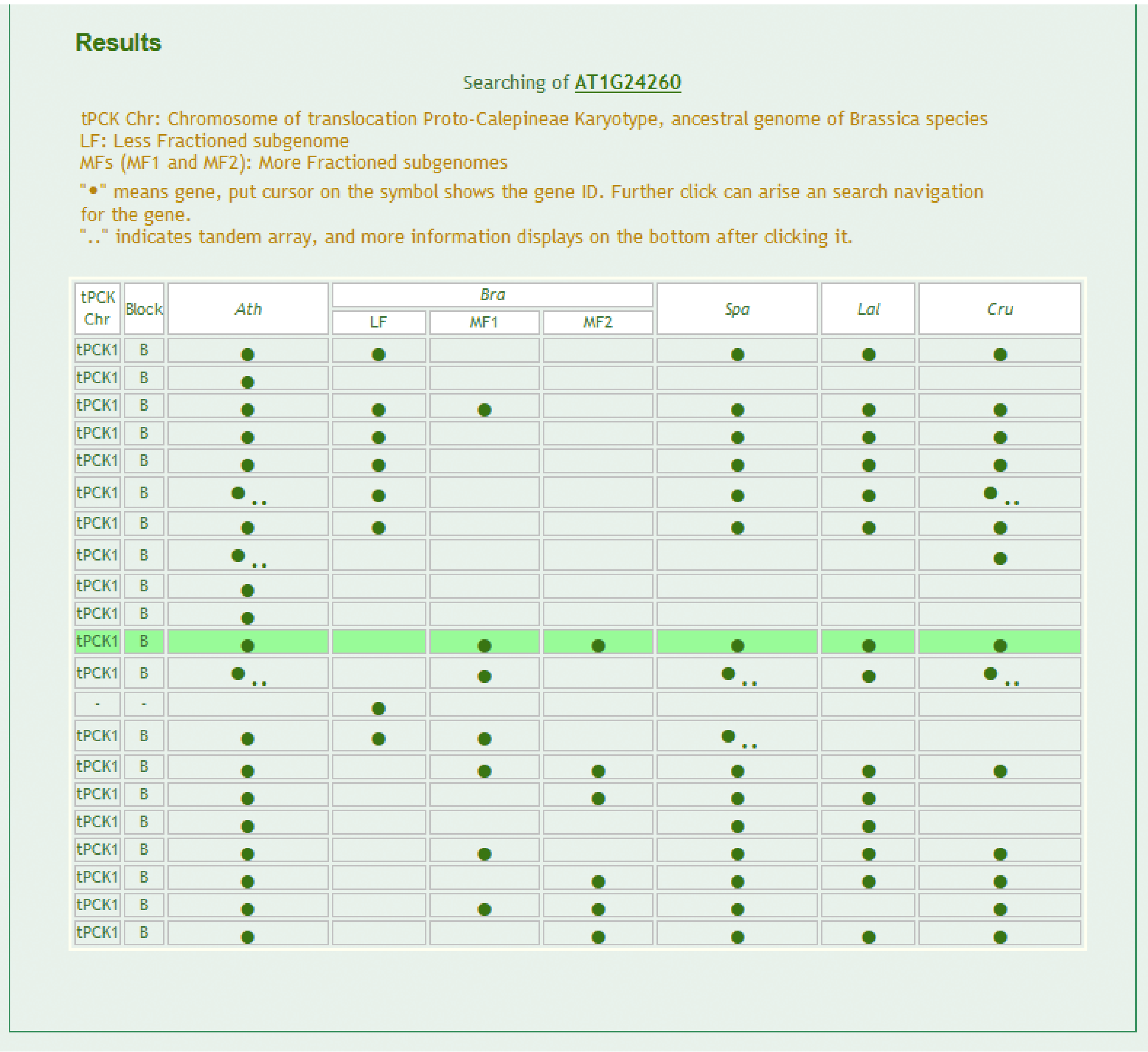
Locus-IDs:

APETALA1	AT1G69120
APETALA3	AT3G54340
PISTILLATA	AT5G20240
AGAMOUS	AT4G18960
SEPALLATA3	AT1G24260
AGAMOUS-LIKE6	AT2G45650

The result will look as in Figure 1. The row highlighted in green marks the sought-after locus. If there is a homologous gene in the corresponding species, this will be indicated by a green dot. Click on the green dot to display and save the "gene sequence" (which is actually the coding sequence of the gene). Copy the coding sequences of all five species for all six genes to a fasta-file. To simplify recognition in the phylogeny that will be reconstructed, make sure to assign informative names to all of the sequences (including species and gene names and using underscores instead of blanks, e.g. Arabidopsis_thaliana_SEP3).

Multiple sequence alignment (see above)

Translate the collected coding sequences into the corresponding amino acid sequences using 'Transeq' (http://www.ebi.ac.uk/Tools/st/emboss_transeq/).
Ensure that all coding sequences have been translated correctly by loading the amino acid sequences into Jalview (http://www.jalview.org/).
If all sequences have been translated correctly create a multiple sequence alignment using MAFFT (implemented in Jalview, web service > alignment > run MAFFT with preset).



Results

Searching of AT1G24260

tPCK Chr: Chromosome of translocation Proto-Calepineae Karyotype, ancestral genome of Brassica species
LF: Less Fractioned subgenome
MFs (MF1 and MF2): More Fractioned subgenomes
"•" means gene, put cursor on the symbol shows the gene ID. Further click can arise an search navigation for the gene.
".." indicates tandem array, and more information displays on the bottom after clicking it.

tPCK Chr	Block	*Ath*	*Bra*			*Spa*	*Lal*	*Cru*
			LF	MF1	MF2			
tPCK1	B	●	●			●	●	●
tPCK1	B	●						
tPCK1	B	●	●	●		●	●	●
tPCK1	B	●	●			●	●	●
tPCK1	B	●	●			●	●	●
tPCK1	B	●..	●			●	●	●..
tPCK1	B	●	●			●	●	●
tPCK1	B	●..						●
tPCK1	B	●						
tPCK1	B	●						
tPCK1	B	●		●	●	●	●	●
tPCK1	B	●..		●		●..	●	●..
-	-		●					
tPCK1	B	●	●	●		●..		
tPCK1	B	●		●	●	●	●	●
tPCK1	B	●			●	●	●	
tPCK1	B	●				●	●	
tPCK1	B	●		●		●	●	●
tPCK1	B	●			●	●	●	●
tPCK1	B	●		●	●	●		●
tPCK1	B	●			●	●	●	●

Figure 1: Result of synteny search

Improving gene predictions using FGENESH+

Check the MAFFT alignment in Jalview and identify sequences that do not fit nicely into the alignment. Figure 2 shows two examples.

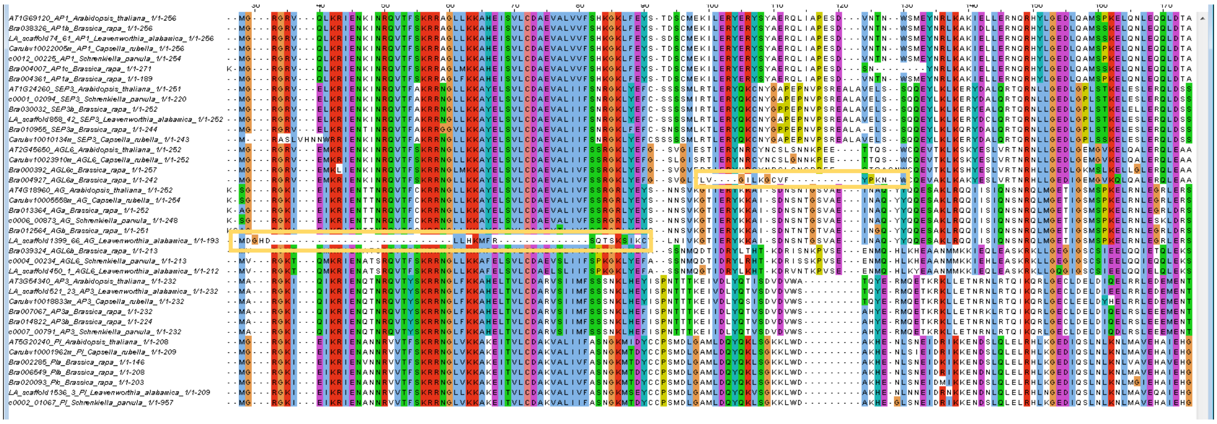
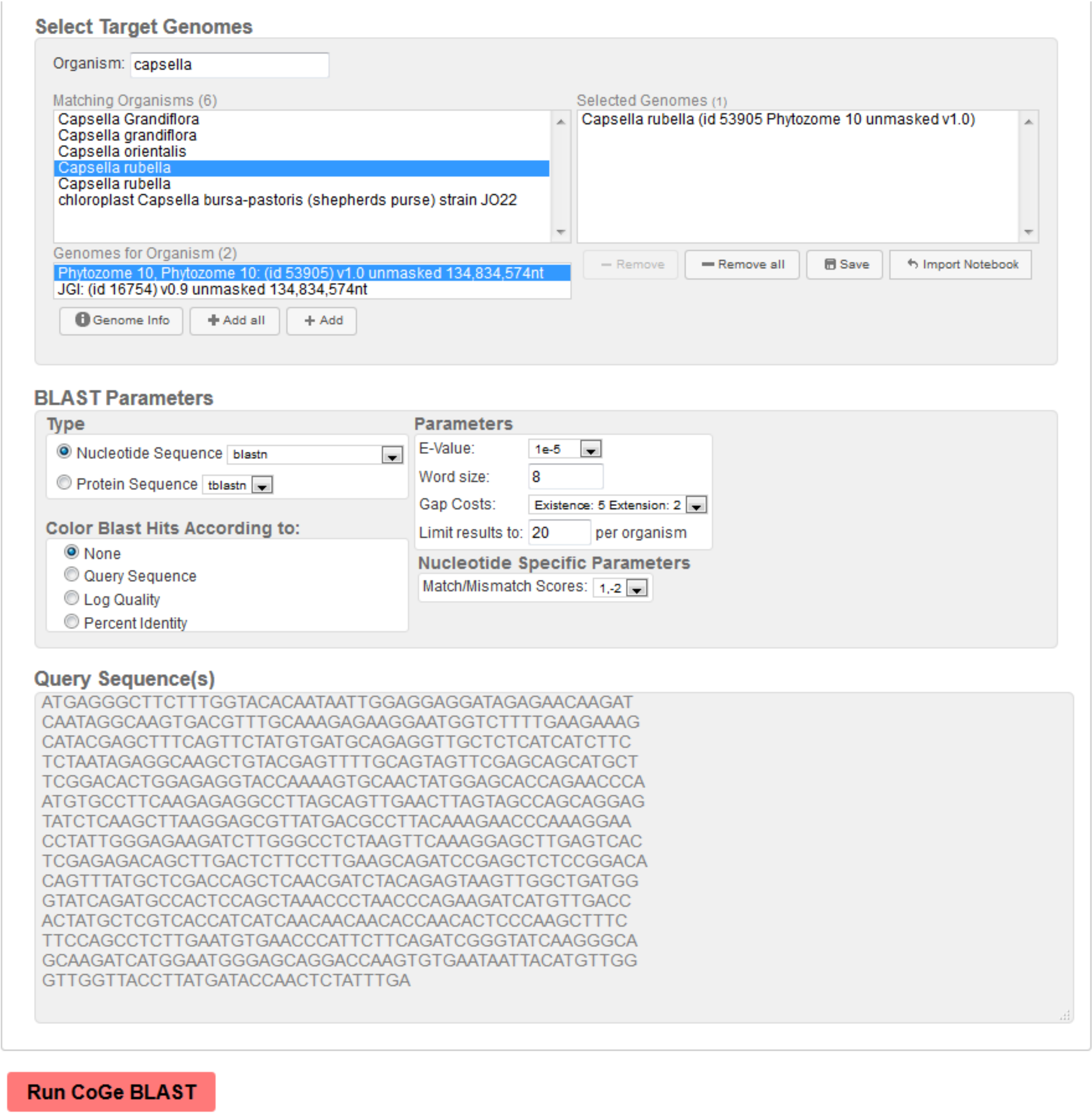


Figure 2: MAFFT alignment where sequences that do not fit nicely into the alignment are boxed.

Get the genomic locus of the suspiciously looking sequences using CoGe BLAST (https://genomevolution.org/coge/CoGeBlast.pl) as follows. Choose the corresponding species and click "+ Add" (Figure 3). Enter the coding sequence in the field "Query Sequence(s)" and click "Run CoGe BLAST.



Select Target Genomes

Organism: capsella

Matching Organisms (6)
Capsella Grandiflora
Capsella grandiflora
Capsella orientalis
Capsella rubella
Capsella rubella
chloroplast Capsella bursa-pastoris (shepherds purse) strain JO22

Selected Genomes (1)
Capsella rubella (id 53905 Phytozome 10 unmasked v1.0)

Genomes for Organism (2)
Phytozome 10, Phytozome 10: (id 53905) v1.0 unmasked 134,834,574nt
JGI: (id 16754) v0.9 unmasked 134,834,574nt

BLAST Parameters

Type
● Nucleotide Sequence: blastn
○ Protein Sequence: tblastn

Color Blast Hits According to:
● None
○ Query Sequence
○ Log Quality
○ Percent Identity

Parameters
E-Value: 1e-5
Word size: 8
Gap Costs: Existence: 5 Extension: 2
Limit results to: 20 per organism

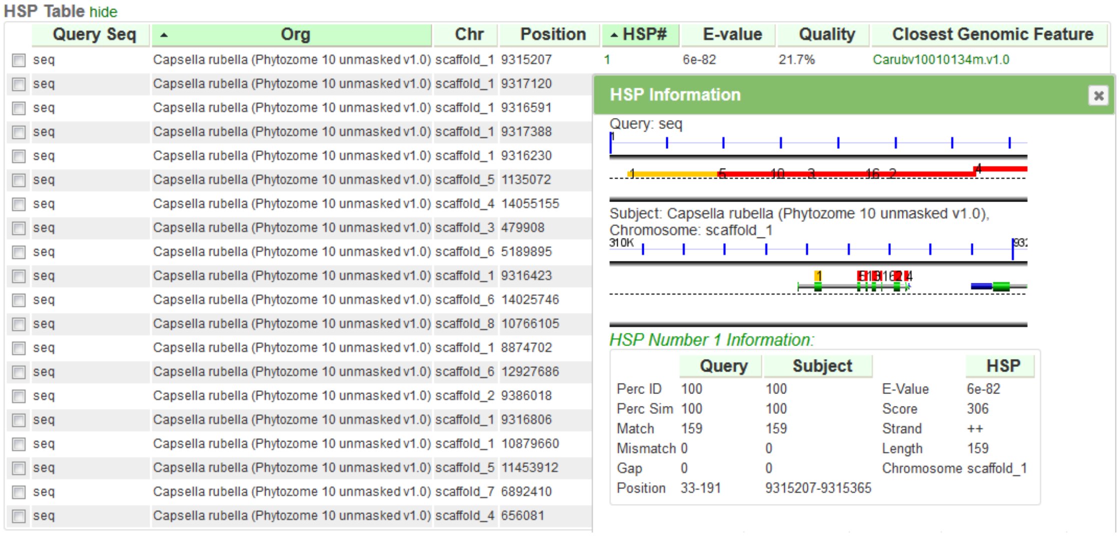
Nucleotide Specific Parameters
Match/Mismatch Scores: 1,-2

Query Sequence(s)

```
ATGAGGGCTTCTTTGGTACACAATAATTGGAGGAGGATAGAGAACAAGAT
CAATAGGCAAGTGACGTTTGCAAAGAGAAGGAATGGTCTTTTGAAGAAAG
CATACGAGCTTTCAGTTCTATGTGATGCAGAGGTTGCTCTCATCATCTTC
TCTAATAGAGGCAAGCTGTACGAGTTTTGCAGTAGTTCGAGCAGCATGCT
TCGGACACTGGAGAGGTACCAAAAGTGCAACTATGGAGCACCAGAACCCA
ATGTGCCTTCAAGAGAGGCCTTAGCAGTTGAACTTAGTAGCCAGCAGGAG
TATCTCAAGCTTAAGGAGCGTTATGACGCCTTACAAAGAACCCAAAGGAA
CCTATTGGGAGAAGATCTTGGGCCTCTAAGTTCAAAGGAGCTTGAGTCAC
TCGAGAGACAGCTTGACTCTTCCTTGAAGCAGATCCGAGCTCTCCGGACA
CAGTTTATGCTCGACCAGCTCAACGATCTACAGAGTAAGTTGGCTGATGG
GTATCAGATGCCACTCCAGCTAAACCCTAACCCAGAAGATCATGTTGACC
ACTATGCTCGTCACCATCATCAACAACAACACCAACACTCCCAAGCTTTC
TTCCAGCCTCTTGAATGTGAACCCATTCTTCAGATCGGGTATCAAGGGCA
GCAAGATCATGGAATGGGAGCAGGACCAAGTGTGAATAATTACATGTTGG
GTTGGTTACCTTATGATACCAACTCTATTTGA
```

Figure 3: BLAST search on CoGe.

On the table with the BLAST results, click on the HSP# 1 and on the then appearing pop-up window, click on the display of the subject locus (Figure 4). A new window will open with the genome browser of the corresponding species at the desired locus. Zoom out of the locus such that at least 20 Kb are shown and view the genomic sequence by choosing “Sequence” > “Save track data” > “View”.



HSP Table hide

	Query Seq	Org	Chr	Position	HSP#	E-value	Quality	Closest Genomic Feature
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_1	9315207	1	6e-82	21.7%	Carubv10010134m.v1.0
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_1	9317120				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_1	9316591				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_1	9317388				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_1	9316230				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_5	1135072				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_4	14055155				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_3	479908				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_6	5189895				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_1	9316423				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_6	14025746				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_8	10766105				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_1	8874702				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_6	12927686				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_2	9386018				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_1	9316806				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_1	10879660				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_5	11453912				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_7	6892410				
☐	seq	Capsella rubella (Phytozome 10 unmasked v1.0)	scaffold_4	656081				

HSP Information

Query: seq

Subject: Capsella rubella (Phytozome 10 unmasked v1.0), Chromosome: scaffold_1

HSP Number 1 Information:

	Query	Subject		HSP
Perc ID	100	100	E-Value	6e-82
Perc Sim	100	100	Score	306
Match	159	159	Strand	++
Mismatch	0	0	Length	159
Gap	0	0	Chromosome	scaffold_1
Position	33-191	9315207-9315365		

Figure 4: Result table of BLAST search on CoGe.

Copy the genomic sequence to the gene prediction tool FGENESH+ (http://www.softberry.com/cgi-bin/programs/gfs/fgenes_plus.pl). Paste the genomic sequence into the text area “Paste nucleotide sequence here”. Into the other text area “Paste protein sequence here“ enter the protein sequence of the most closely related gene from *Arabidopsis thaliana*. Under “Select organism specific gene-finding parameters“ choose „Dicot plants, Arabidopsis (generic)“ and press “search”.

With the resulting gene prediction, first conduct a BLAST search with the protein sequence on NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome). If the BLAST search returns the expected results, replace the corresponding sequence in your fasta-file with the nucleotide sequence of the gene prediction.

Repeat this step until you predicted new genes for all sequences that did not fit nicely into the alignment.

New phylogeny reconstruction

Add the sequences in your fasta-file to the sequence collection of 24 coding sequences from yesterday. Delete duplicate sequences from your new fasta-file (i.e. delete AP1, AP3, PI, AG and SEP3 from *Arabidopsis thaliana* **once** from your dataset [not both copies!]).
Translate the nucleotide sequences into amino acid sequences (see above). Align the sequences using MAFFT (see above). Remove alignment parts with low conservation using TrimAl (http://phylemon2.bioinfo.cipf.es/utilities.html > choose “start as anonymous user” > Utilities > Alignment Utilities > TrimAl (v. 1.3)). Upload your protein alignment and choose method “strict”. Save the trimmed alignment and reconstruct a RAxML phylogeny via the online tool CIPRES (https://www.phylo.org/portal2/login!input.action). Upload the trimmed alignment, select the input file and the calculation tool (RAxML) and adjust the following parameter:

Maximum Hours to Run (click here for help setting this correctly): 2

Please select the Data Type: Protein

Outgroup (one or more comma-separated outgroups, see comment for syntax): CgMADS1

Advanced Parameters

Conduct a rapid Bootstrap analysis and search for the best-scoring ML tree in one single program run. (-f a): Ticked

Bootstrap iterations (-#|-N): 1000

Save the parameters and start the tree calculation.

After the tree calculation is completed display the output data via the ‘output’ button.

First of all open the error report file and check for any error messages.

The maximum likelihood tree is located in the ‘RAxML_bipartitions.result’ file. Download the file and visualize the tree with the FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).