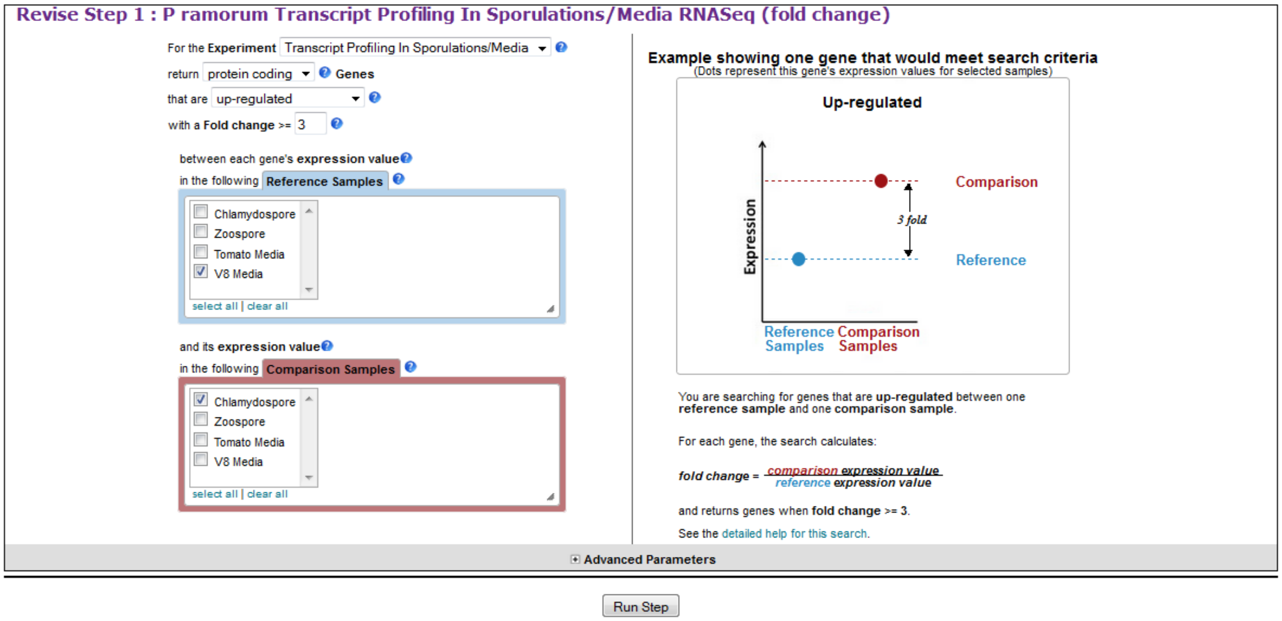
Complex strategies with Genomic Colocation
Exercise 14

14.1 Divergent genes with similar expression profiles.

Identify Phytophthora ramorum genes that meet these four criteria:

1. are located within 1000 bp of each other
2. are divergently transcribed (on opposite strands),
3. are up-regulated in either zoospore or chlamydospore compared to either media,
4. show at least a 3-fold increase in expression.

- Hint: first use the “Genes bases on RNAseq expression” -> “Transcript Profiling In Sporulations/Media” -> “P.r. Sporulations/Media RNASeq (fc)” search.



Revise Step 1 : P ramorum Transcript Profiling In Sporulations/Media RNASeq (fold change)

For the **Experiment** Transcript Profiling In Sporulations/Media

return protein coding **Genes**

that are up-regulated

with a **Fold change** >= 3

between each gene's **expression value**

in the following **Reference Samples**

□ Chlamydospore
□ Zoospore
□ Tomato Media
☑ V8 Media

select all | clear all

and its **expression value**

in the following **Comparison Samples**

☑ Chlamydospore
□ Zoospore
□ Tomato Media
□ V8 Media

select all | clear all

Example showing one gene that would meet search criteria
(Dots represent this gene's expression values for selected samples)

You are searching for genes that are **up-regulated** between one **reference sample** and one **comparison sample**.

For each gene, the search calculates:

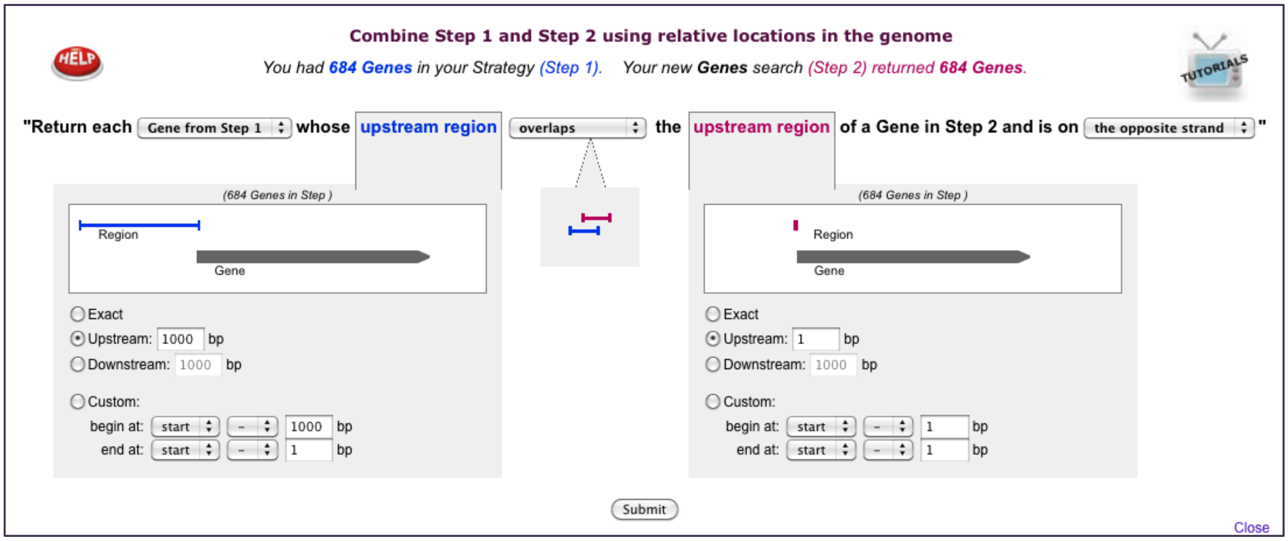
$$\textit{fold change} = \frac{\textit{comparison expression value}}{\textit{reference expression value}}$$

and returns genes when **fold change** >= 3.

See the detailed help for this search.

- Add a step that is the same as the first step and select the genomic colocation (1 relative to 2) operation.
- Set up the form to identify those genes that are transcribed on the **opposite strand** that have their starts located within 1000 bp of another genes start.
- Turn on the “Pr Sporulations/Media RNAseq – rpkm Graph” and “Pr Sporulations/Media RNAseq – percentile graph“ columns to assess how well the pairs of genes compare in terms of expression. The pairs of genes are located one above the other in the result table if sorted by location.
- Identify paired genes that have similar expression profiles based on the graphs.
- Note that you could do similar types of experiments to look at potential co-regulation / shared enhancers / divergent promoters with other sorts of data such as:

- ○ DNA motifs for transcription factor binding sites.
  - ○ Of course other expression queries.
  - ○ Etc …
- The screenshot below shows one way (there are MANY) to configure the genome colocation form to identify genes that are divergently transcribed located with their start within 1000 bp of each other.



Combine Step 1 and Step 2 using relative locations in the genome

*You had **684 Genes** in your Strategy (Step 1).   Your new **Genes** search (Step 2) returned **684 Genes**.*

"Return each [Gene from Step 1] **whose upstream region** [overlaps] **the upstream region of a Gene in Step 2 and is on** [the opposite strand] **"**

(684 Genes in Step)

○ Exact
● Upstream: 1000 bp
○ Downstream: 1000 bp

○ Custom:
begin at: start - 1000 bp
end at: start - 1 bp

(684 Genes in Step)

○ Exact
● Upstream: 1 bp
○ Downstream: 1000 bp

○ Custom:
begin at: start - 1 bp
end at: start - 1 bp

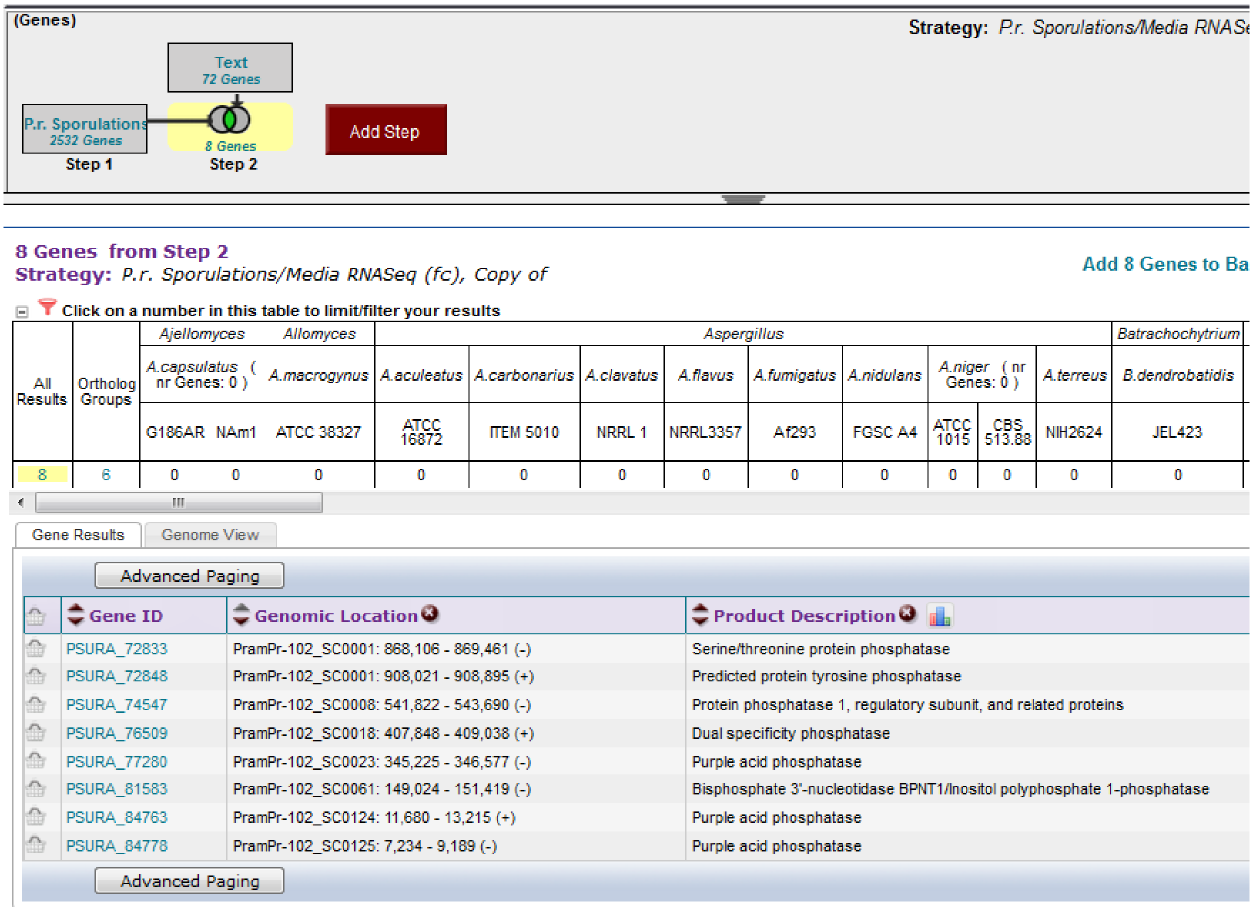
14.2 Identify potential transcription factor binding motifs

The goal of this exercise is to identify DNA motifs in the promoter regions of similarly expressed phosphatase genes, and then search for these motifs in un-annotated genes that also show similar expression. Maybe these un-annotated genes have related functions or are in the same pathways.

a. Use the same RNAseq dataset from the previous example, up-regulated 3-fold increase P.ramorum Chlamydospore vs V8 media reference.

b. Restrict this set to genes that have “phosphatase” entered in the gene product. This should give you 8 genes as shown below.

.



(Genes)

Strategy: *P.r. Sporulations/Media RNASe*

8 Genes from Step 2

Strategy: *P.r. Sporulations/Media RNASeq (fc), Copy of*

Add 8 Genes to Ba

Click on a number in this table to limit/filter your results

All Results	Ortholog Groups	*Ajellomyces*		*Allomyces*	*Aspergillus*									*Batrachochytrium*
		A.capsulatus (nr Genes: 0)		*A.macrogynus*	*A.aculeatus*	*A.carbonarius*	*A.clavatus*	*A.flavus*	*A.fumigatus*	*A.nidulans*	*A.niger* (nr Genes: 0)		*A.terreus*	*B.dendrobatidis*
		G186AR	NAm1	ATCC 38327	ATCC 16872	ITEM 5010	NRRL 1	NRRL3357	Af293	FGSC A4	ATCC 1015	CBS 513.88	NIH2624	JEL423
8	6	0	0	0	0	0	0	0	0	0	0	0	0	0

Gene Results | Genome View

Advanced Paging

Gene ID	Genomic Location	Product Description
PSURA_72833	PramPr-102_SC0001: 868,106 - 869,461 (-)	Serine/threonine protein phosphatase
PSURA_72848	PramPr-102_SC0001: 908,021 - 908,895 (+)	Predicted protein tyrosine phosphatase
PSURA_74547	PramPr-102_SC0008: 541,822 - 543,690 (-)	Protein phosphatase 1, regulatory subunit, and related proteins
PSURA_76509	PramPr-102_SC0018: 407,848 - 409,038 (+)	Dual specificity phosphatase
PSURA_77280	PramPr-102_SC0023: 345,225 - 346,577 (-)	Purple acid phosphatase
PSURA_81583	PramPr-102_SC0061: 149,024 - 151,419 (-)	Bisphosphate 3'-nucleotidase BPNT1/Inositol polyphosphate 1-phosphatase
PSURA_84763	PramPr-102_SC0124: 11,680 - 13,215 (+)	Purple acid phosphatase
PSURA_84778	PramPr-102_SC0125: 7,234 - 9,189 (-)	Purple acid phosphatase

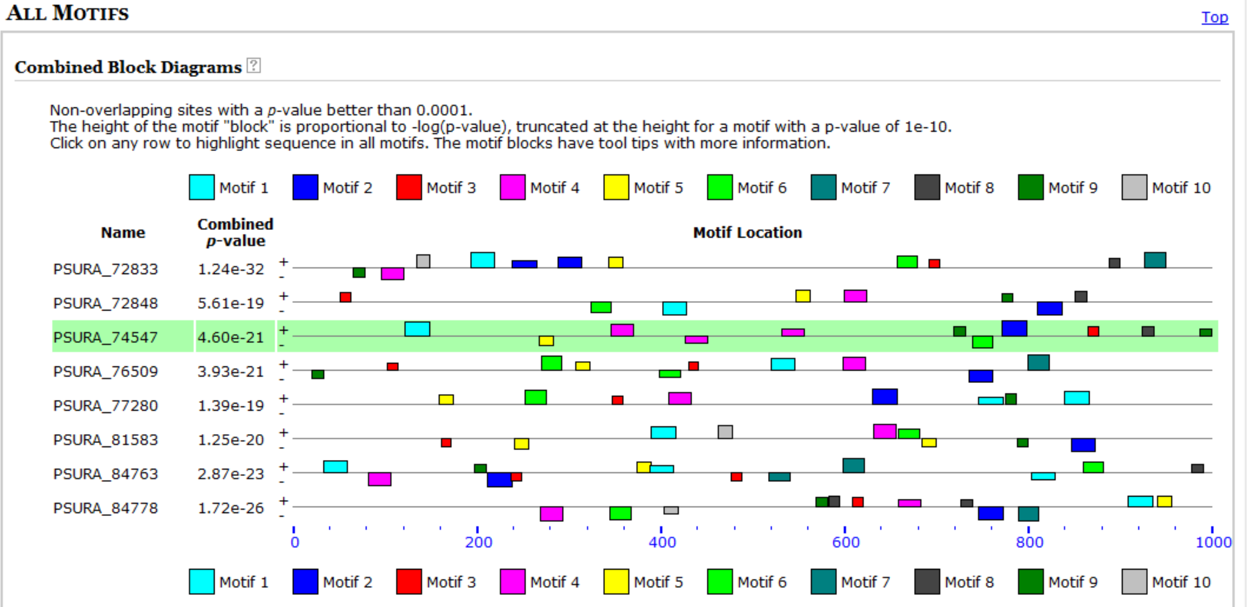
Advanced Paging

c. Now download the promoter region sequence for these 8 genes. Most oomycete genes do not have UTR regions identified in the annotations, so we will take a large region upstream from the translation start site. Take 1Kb upstream from the ATG. Hint: use the download # genes link shown in exercise 10. Select FASTA sequence, and change the options to get the upstream region.

d. You should now have 8 1Kb long sequences. We now want to identify the over-represented DNA patterns found in these sequences. Run these sequences in the

DNA motif finder MEME (http://meme.sdsc.edu/meme/intro.html). However, it will take a while to return these results, especially if we all submit jobs. Pre-run results can be accessed here (http://nbcr-222.ucsd.edu/opal-jobs/appMEME_4.9.11401058916868-1781741983/meme.html). (This link should be active during the workshop but will not last forever. If you are following this example outside of the workshop you will need to actually run and wait for the MEME results.)

e. Take a look at the DNA motif results. Several interesting motifs are found. Motif 2 (CCAAAT) is very similar to a CAAT-box. Scroll all the way down to the bottom where the motif placement map is seen. Motif 5 is often found ~500-600bp upstream of Motif 2.



ALL MOTIFS

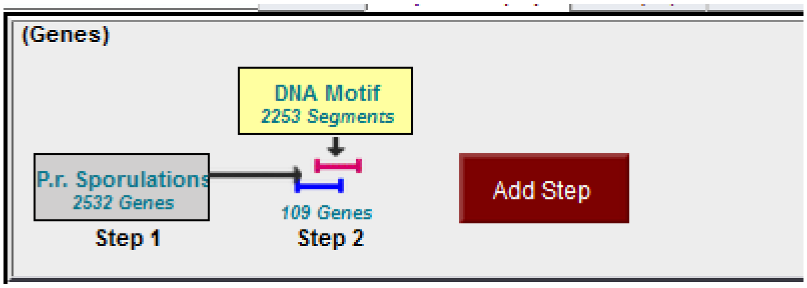
Top

Combined Block Diagrams

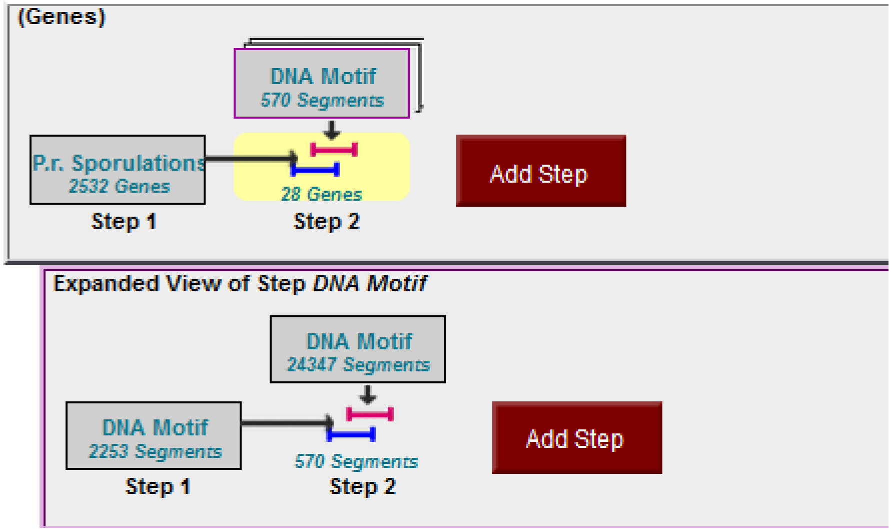
Non-overlapping sites with a *p*-value better than 0.0001.
The height of the motif "block" is proportional to -log(p-value), truncated at the height for a motif with a p-value of 1e-10.
Click on any row to highlight sequence in all motifs. The motif blocks have tool tips with more information.

Name	Combined *p*-value
PSURA_72833	1.24e-32
PSURA_72848	5.61e-19
PSURA_74547	4.60e-21
PSURA_76509	3.93e-21
PSURA_77280	1.39e-19
PSURA_81583	1.25e-20
PSURA_84763	2.87e-23
PSURA_84778	1.72e-26

f. Search for all occurences of Motif 5 in the 1Kb upstream regions of the expressed gene set originally used in this example (up-regulated 3-fold increase P.ramorum Chlamydospore vs V8 media reference). Meme gives the RegEx for Motif 5 as G[CA][AT][AG]TTGC[CG]TG[CT]A[AC]. Using this will only return back 3 of the 8 phosphatase genes, so instead use a more general RegEx: G[CA][AT].[TC]T[GA][CA].[TG]G[CT]A[AC], this returns 7 of the 8 phosphatase genes.



g. See how many of Motif 5 are found in close proximity to the CAAT-box like Motif 2. Make a nested strategy for the motif identification, search for the exact motif ‘CCAAAT’ found within 750bp downstream of Motif 5.



How many previously un-annotated genes did you find?

Optional:

h. Motifs 1 and 3 also have interesting sections. See if you can make a RegEx to represent these DNA motifs. You can search for these individually or try to identify any co-location patterns.