Genetic Circuit Evolution

Steven Kuntz

Worms and Rodents

- Sternberg rotation project update
- Dicer update
- Circuit evolution approaches
  - Expression data
  - Genomic data
- Anchor cell project update
- Muscle cell project update
- CompClust update

Part I

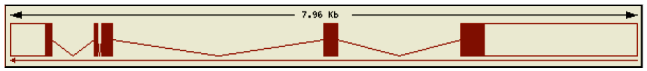
Sternberg Lab rotation project update

Homeobox Genes in Nematodes

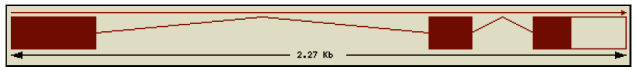
- Six hox genes in three pairs
  - Lin-39 and ceh-13; egl-5 and mab-5; php-3 and nob-1
- Role in development
  - Lin-39: vulval development
  - Ceh-13: anterior patterning
  - Egl-5: posterior patterning (anus to tail)
  - Mab-5: posterior patterning (gonad to anus)
  - Php-3: posterior patterning
  - Nob-1: posterior patterning
- Promoter dissection
  - Mussa analysis

C. elegans vs. C. briggsae vs. CB5161: lin-39 and ceh-13

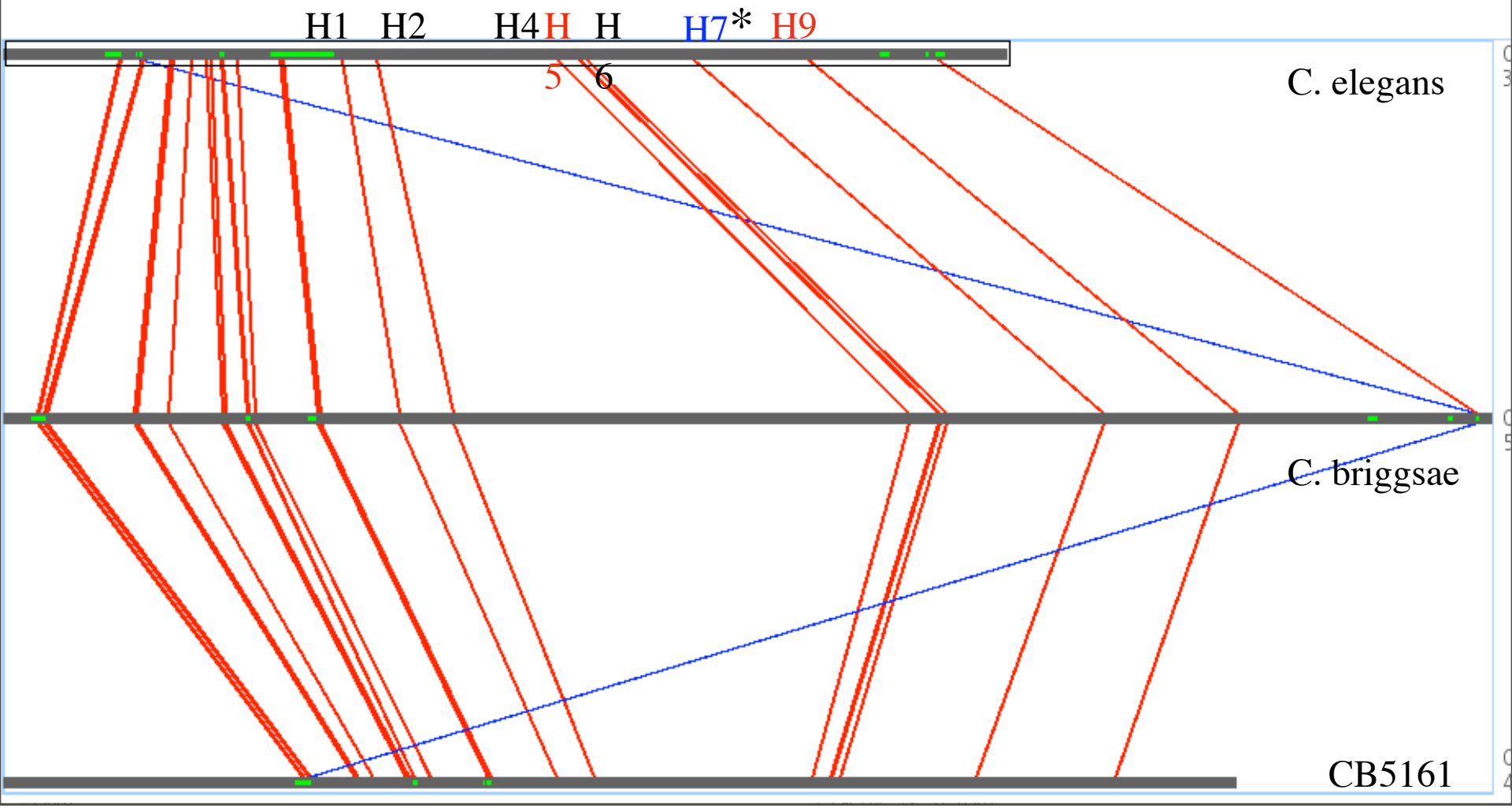
- Window: 30 Threshold: 25



Lin-39 (- strand)

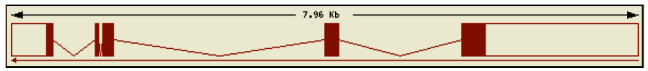


Ceh-13 (+ strand)

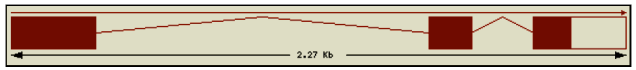


C. briggsae vs. C. elegans vs. PS1010: lin-39 and ceh-13

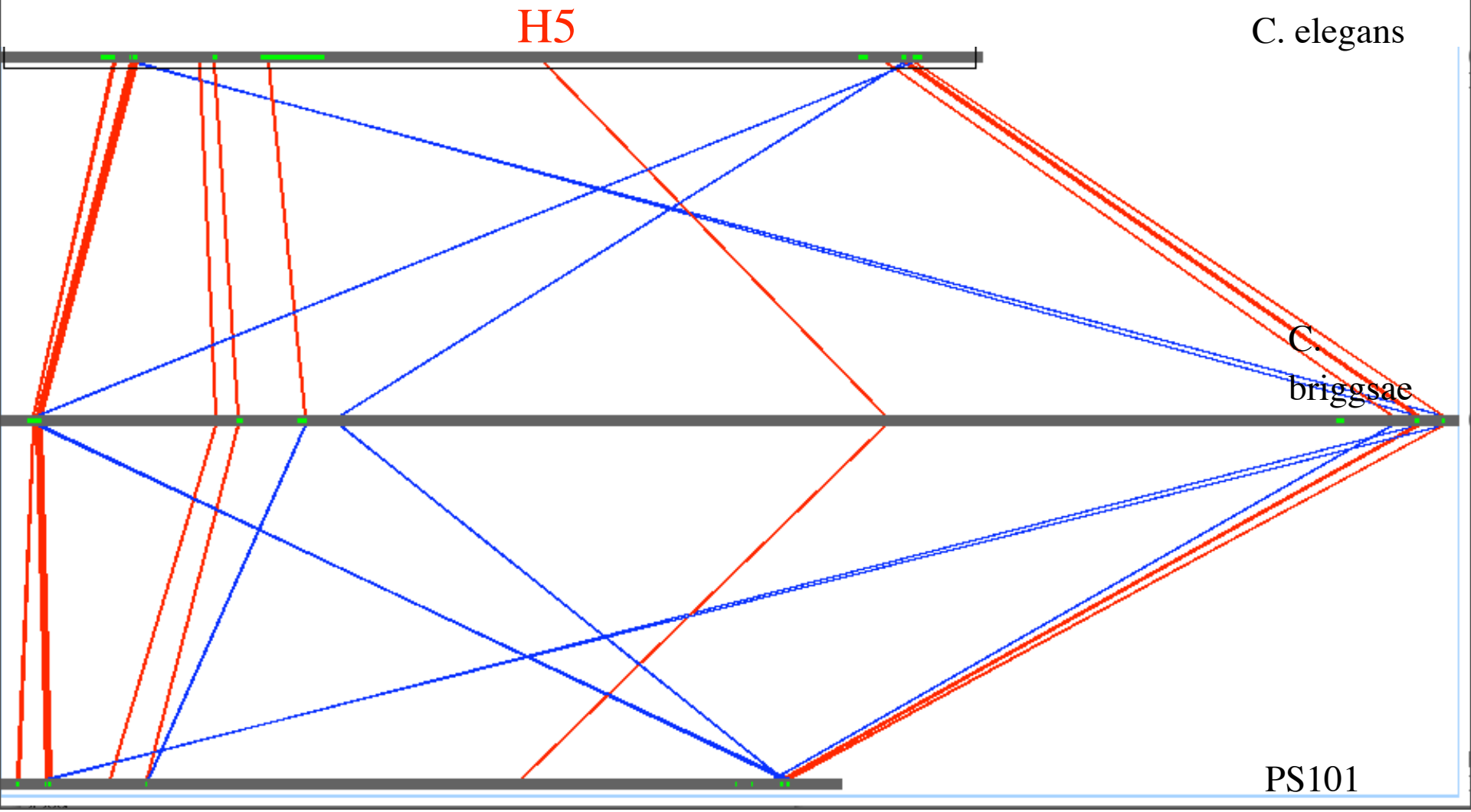
- Window: 30 Threshold: 23



Lin-39 (- strand)



Ceh-13 (+ strand)

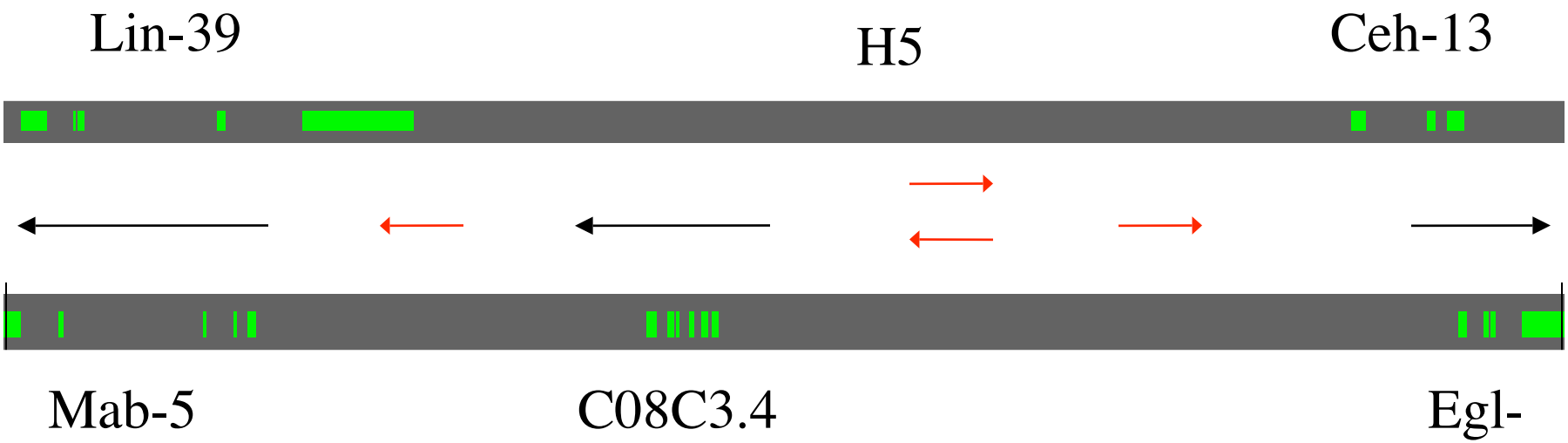


in vivo vs. in silico

- Mussa element H7 reflects the element found to drive ceh-13 expression by Streit et al.
- Element H8 reflects the conserved enhancer found by the same group

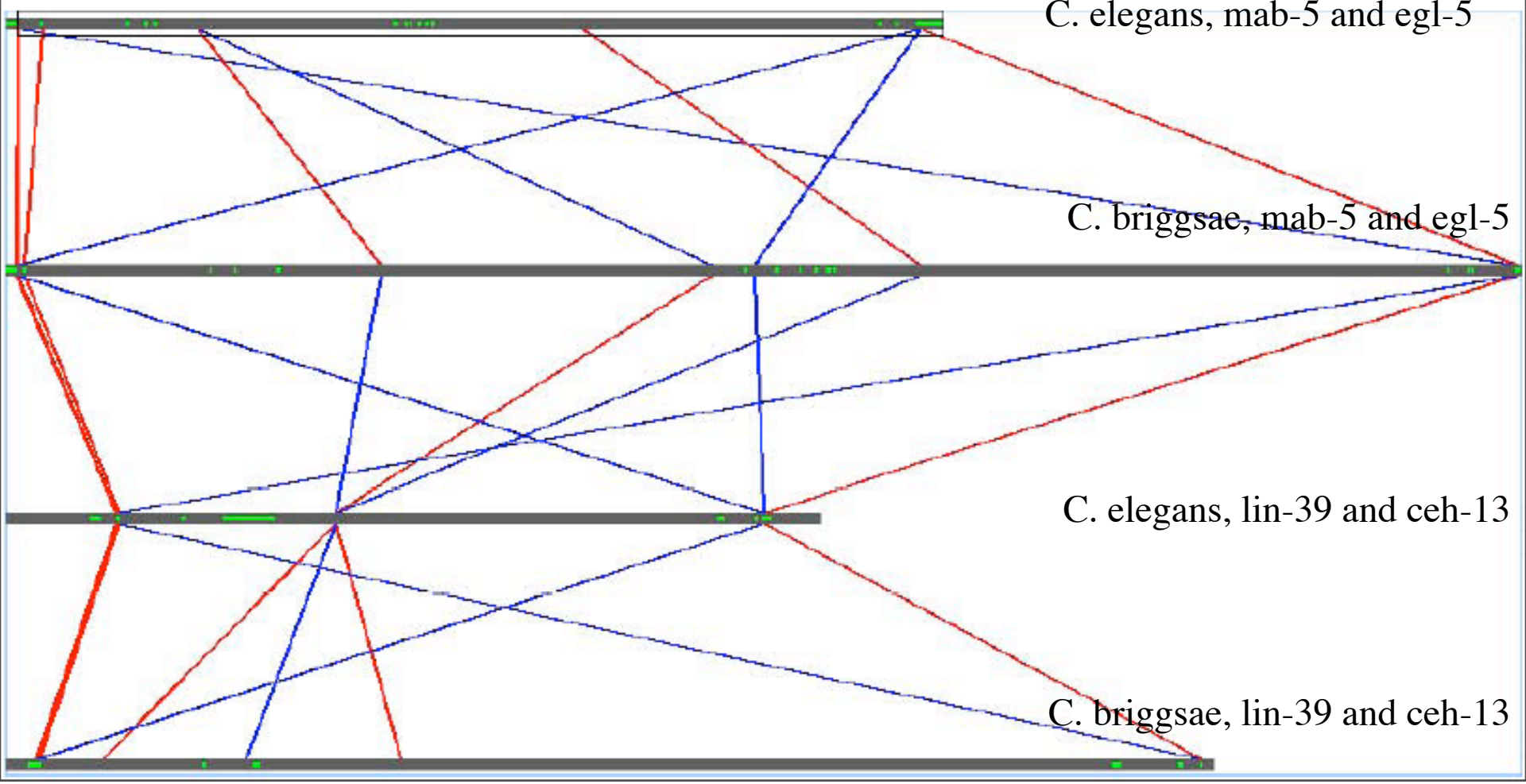
Question of direction

- Cis-regulatory enhancers may be split where the egl-5/mab-5 region is split by the additional gene, C08C3.4
- Enhancers may travel in either direction, or both directions, regardless of location
- H5 is at the center, being the most ambiguous location



4-way: egl-5 and mab-5 vs. lin-39 and ceh-13

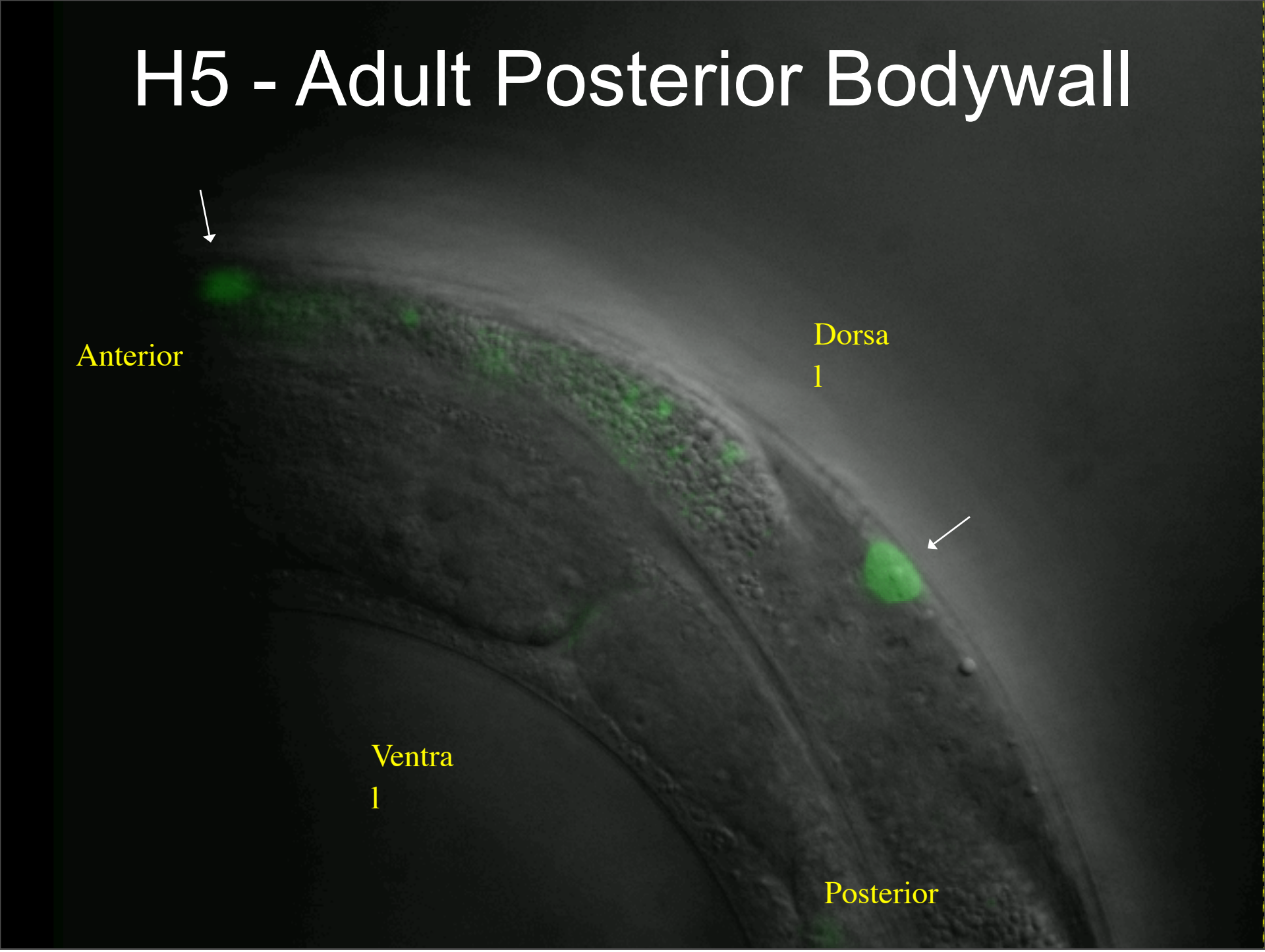
Window: 30 Threshold: 23



Status of Transgenics

- 15 of 16 cloned into GFP-LacZ vector
  - Nuclear localized
  - ∆pes-10 basal promoter
  - All elements but L2
- 12 attempted injections
- 2 stable transgenic strains (mutliple lines each)
  - H5 has anterior body wall muscle expression
  - H9 has no expression at low magnification

H5 - Adult Posterior Bodywall



Conclusions

- Mussa appears useful in finding real enhancers
- Agreement between *in vivo* experimental results on elements and computational predictions
- Testing for repressors or combinatorial effects will be necessary for examining negative results

Part II

‘Dicer’ Sequence Comparison Program

Dicer Program:

- Python program to analyze non-coding regions surrounding genes
  - Looks at small conserved segments as a companion to other seqcomp programs (FRII and Mussa)
  - Is designed to address insertions and deletions
  - Adjustable size of motif may be close to binding region of transcription factors
  - Noise issues: may reduce noise that plagues Seqcomp programs

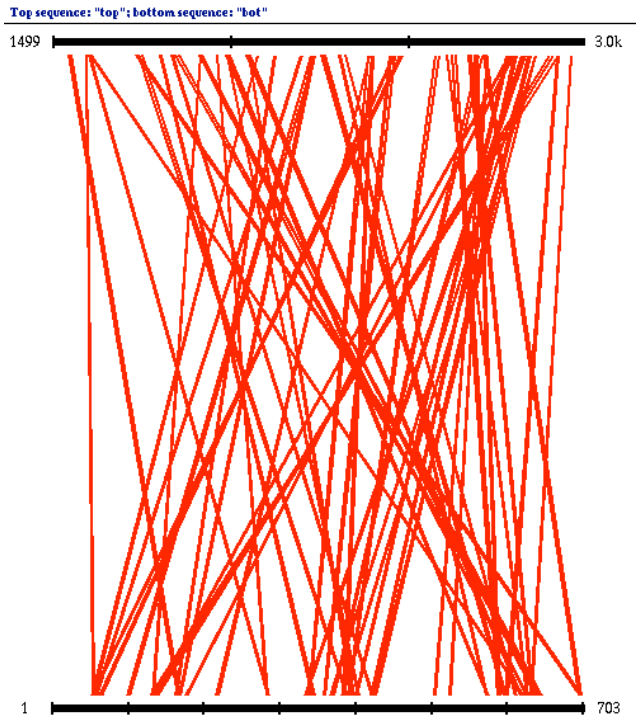
Dicer Program: Methodology

- Establishes all possible *n* bp elements (windows) from the regions in an array (single or multiple frames)
- Compares these elements with those from a second sequence
- Discards any overlapping matches (leaves only unique hits)
- Displays all matches
- A series of parallel matches may denote an interesting region

Counting array

[1, 4, 7, 7, 7, 6, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 3, 3, 3, 3, 1, 0, 0, 0, 0, 1, 2,
4, 6, 7, 6, 5, 3, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 3, 3, 3, 3, 3, 1, 1, 1, 0, 0, 0, 1, 2, 2, 2, 2, 1, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 2, 4, 7, 8, 8, 7, 4, 2, 1, 1, 1, 1, 0, 0, 1,
1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 2, 5, 8, 9, 8, 7,
4, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 4, 5, 5, 4, 4, 2, 1, 1, 1, 1, 2, 2, 2, 1,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 3, 4, 4, 3, 2, 1, 0, 0, 0, 0, 0, 1, 1, 2, 2, 2, 4, 6, 8, 9, 10, 7,
5, 3, 2, 1, 1, 1, 0, 0, 0, 0, 0, 0, 2, 5, 8, 9, 9, 7, 4, 1, 0, 0, 0, 0, 1, 3, 4, 4, 4, 4, 2, 1, 1, 1, 0,
0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0]

Dicer results: Egl-5/Mab-5

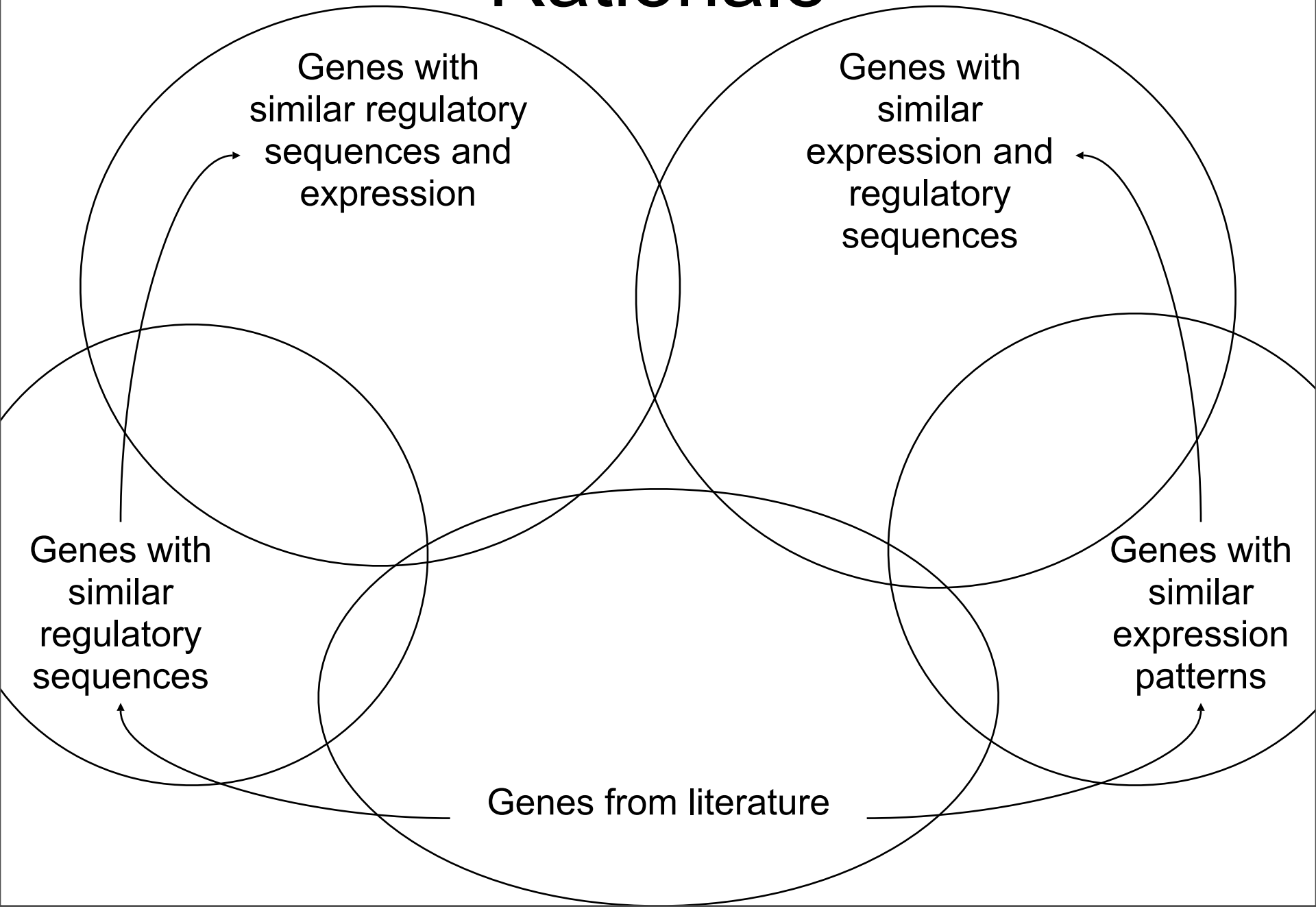


- 1.5 kb X 700 bp analysis shown
- Noise level comparison
  - For window size of 5, standard technique returns 207,084 hits and precise technique returns 3,385 hits
  - The smaller the window size, the greater the magnitude reduction in number of hits for egl-5/mab-5 C. elegans/C. briggsae comparison

Part III

Gene Circuit Evolution Update

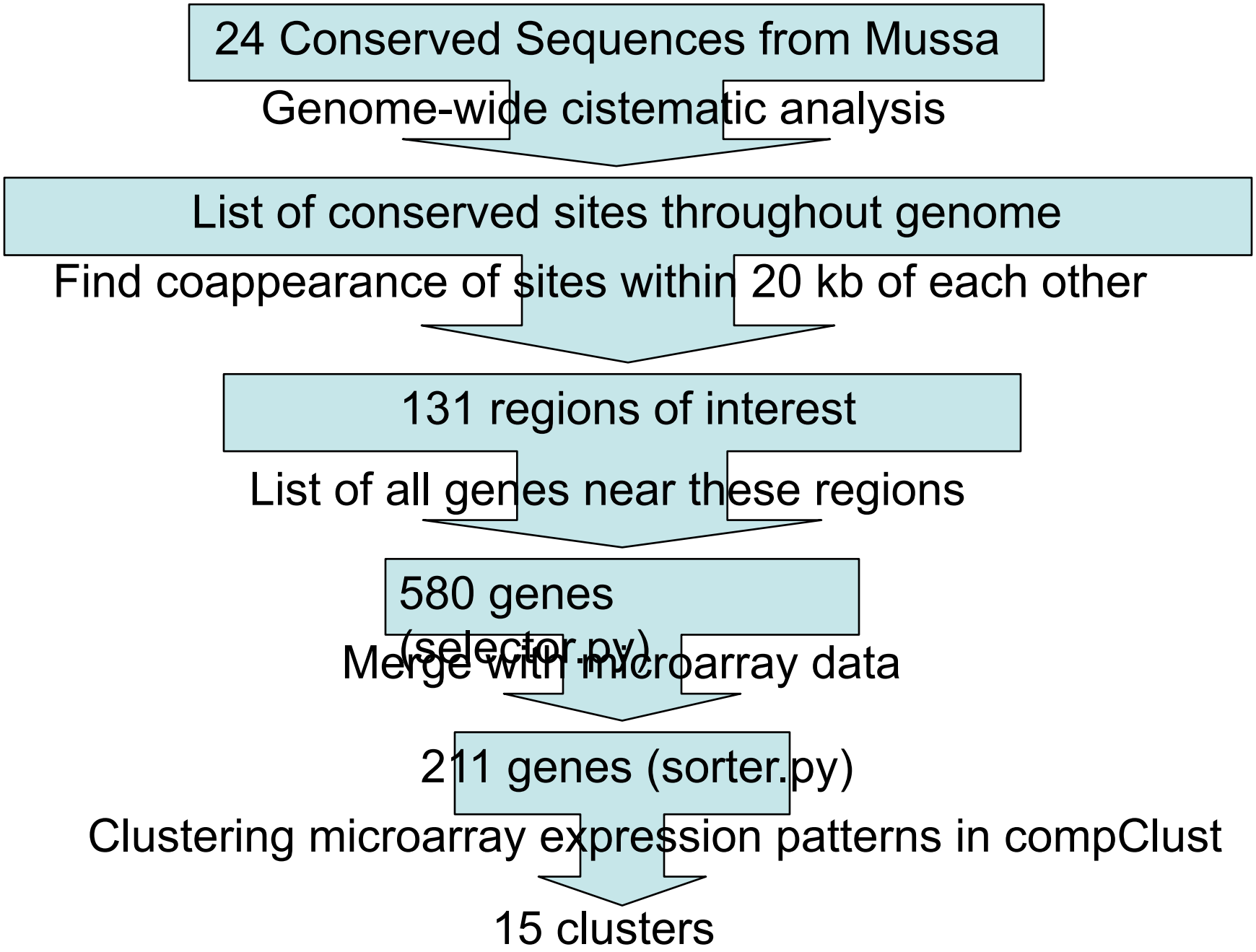
Rationale


Genes with similar regulatory sequences and expression
Genes with similar expression and regulatory sequences
Genes with similar regulatory sequences
Genes with similar expression patterns
Genes from literature

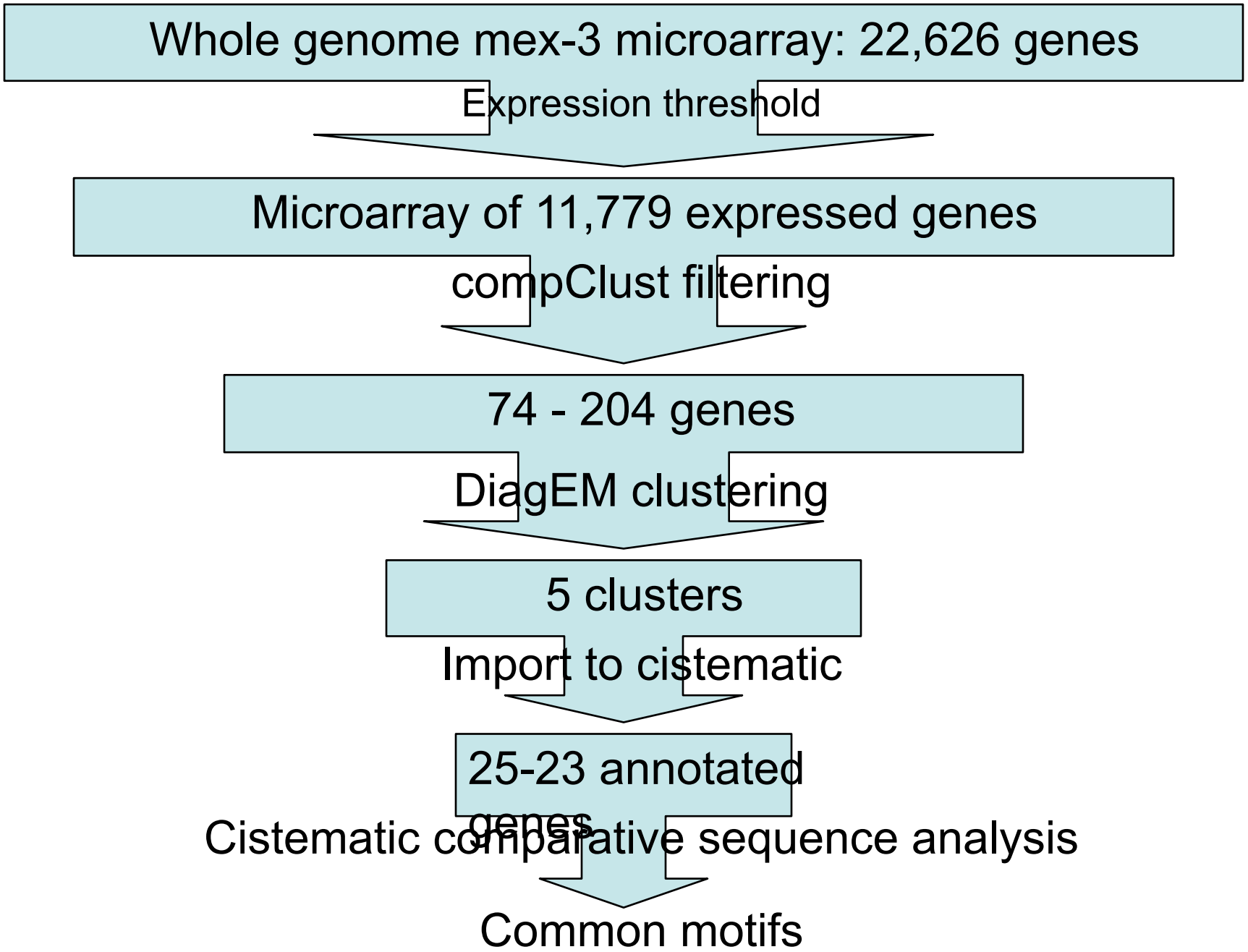
Gene Lists

- List of genes from literature:
  - Short list of genes (hlh-1, hnd-1, mex-3, etc.)
- List of genes from expression data:
  - List of genes with same expression patterns as literature genes within tissue subset
- List of genes from genomic data:
  - List of genes with same regulatory regions as literature genes
  - Refined list of genes with same regulatory regions and expression patterns as literature genes
- List of overlapping genes
  - List of genes with same regulatory regions, expression patterns, and tissue expression

Gene List Refinement: Genomic Data ==> Expression Data



Gene List Refinement: Expression ==> Genome



Advantages of muscle tissue vs. anchor cells

Aspect	Muscle tissue	Anchor cell
Interphyla comparisons	Conservation of tissue and factors, perhaps of elements	No known comparable circuits
Interspecies comparisons	Few differences between close species	Several interesting differences between close species
Target proteins	Good targets (MyoD, Mef2, mafbx, myostatin, calcineurin, NFAT, GATA, brachyury)	Very good targets (egl-17, lin-3, bac-1)
Epistatic foundation	Very poor in worm, strong in mouse	Strong in worm
Circuit noise	Cell fate/differentiation primary circuit	Numerous quasi-independent circuits
Cell picking	Known protocol	Original technique
Preciseness of signal	Good, cells non-homogenous	Very precise, only individual variations (can do paired t-test type analysis)
Statistical power	Very good	Adequate
Signaling comparisons	General pathway, few cell-cell direct effects	Effector and affected cells: Primary vs. secondary or AC vs. VPC
Internal tissue comparisons	Striated vs. cardiac muscle	None
Potential revelations: comparisons	Nature of nematode muscle differentiation	Keys in species specific differences
Potential revelations: mutations		
Potential revelations: regulation	Considerable differences with similar results	Considerable differences with similar results
Circuit foundations	Strong vertebrate foundation, poor nematode foundation	Good epistatic basis
mRNA source	Whole tissue availability	Single cell only
Fate determination/ Differentiation	Somewhat linear	Several parallel, possibly interlinked, processes

Part IIIa

Anchor Cell Specification in *Caenorhabditis elegans* vs. *C. briggsae*

List of Literature Genes

Wormbase ID	Gene	Protein	Wormbase ID	Gene	Protein
R107.8	Lin-12	Notch			
ZK662.4	Lin-15		F38G1.2	Lin-3	EGF
Y71F9B.5	Lin-17	Wnt			
ZK1067.1	Let-23	EGFR	F38G1.2	Egl-17	
F11C1.6	Nhr-25	FTZ-F1 (NHR)	EGAP1.3	Zmp-1	
K08F8.6	Let-19		C27A12.5	Ceh-2	
E01A2.3	Lin-44		ZK112.7	Cdh-3	
C07H6.7	Lin-39		T02B4.6		
M05B5.5	Hlh-2	Daughterless	B0034.1		
ZK792.6	Let-60	Ras	F47B8.6		

Plays role in anchor cell specification

Plays role in anchor cell function

Anchor Cell Picking

- Anchor Cell Background:
  - Critical for induction of vulva formation
  - Present for 10 hours during larval development
  - Approximately 1 hour between closest time points
  - AC/VU decision not well understood
  - Signals VPC's to induce differentiation into vulval cells
- Differences between species:
  - Wild-type worms identical
  - Bac-1 mutants in *C. briggsae* have reversed cell

Part IIIb

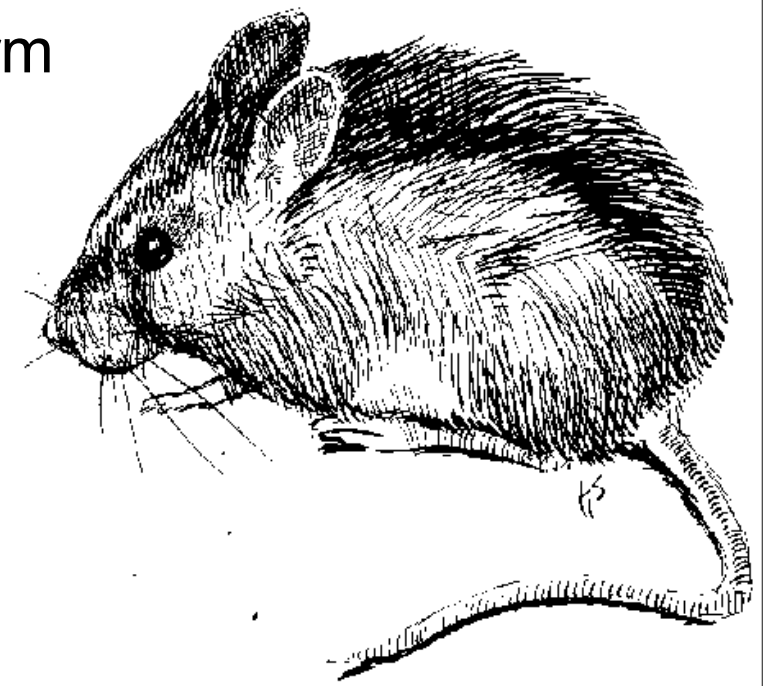
Muscle Cell Specification and Differentiation

Model Organisms: Why mouse and worm

Ideal model organisms for this study must meet 5 criteria:

1. Reliably sequenced genome

2. Closely related species have sequenced genomes



http://home.teleport.com/~hsimante/pictures/mouse.gif

http://www.desc.med.vu.nl/NL-taxi/ICE/C_elegans1.jpg

3. Some tissue extraction or culturing must be possible

4. Powerful epistatic models

5. Microarrays or qRT-PCR need to be available

Partial List of Literature Genes

- Nematode
  - CeHDA-7 (class IIHDA)
  - MEF-2 (MADS box)
  - Calcineurin
  - NFAT
  - DY3.6(Mafbx)
  - Brachyury (T-box)
  - W05F2.7
  - Snail (Zn finger)
  - ZNF510
  - Scratch (Zn finger)
  - Calcineurin B isoform, protein phosphatase 2B)
  - Calcineurin
  - Pax3
  - unc-120 (mads box, srf)
  - myo-3 (myosin activator)
  - hnd-1 (bHLH)
- Mammal
  - MyoD
  - myogenin
  - Mrf4
  - Myf5
  - Pax3
  - MEF
  - Brachyury (T-box)
  - Calicneurin
  - NFAT
  - Snail
  - Slug
  - Scratch
  - Capsulin
  - MurF
  - Myostatin
  - Mafbx

Use of mex-3; elt-1 mutant

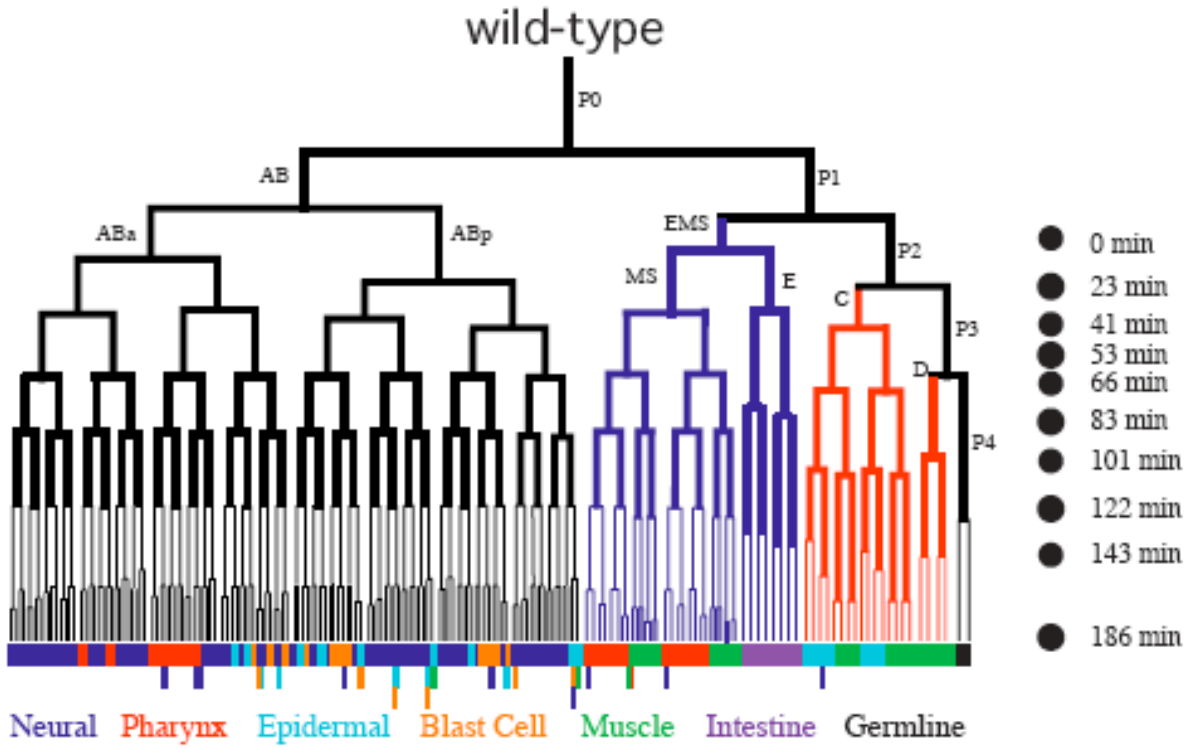
Take mex-3; dpy-5 homozygotes

RNAi with elt1

Eggs should all be mex-3; dpy-5; elt-1

Little impact from dpy-5 expected

Eggs should be all muscle lineage



Part IIIc

Data Analysis

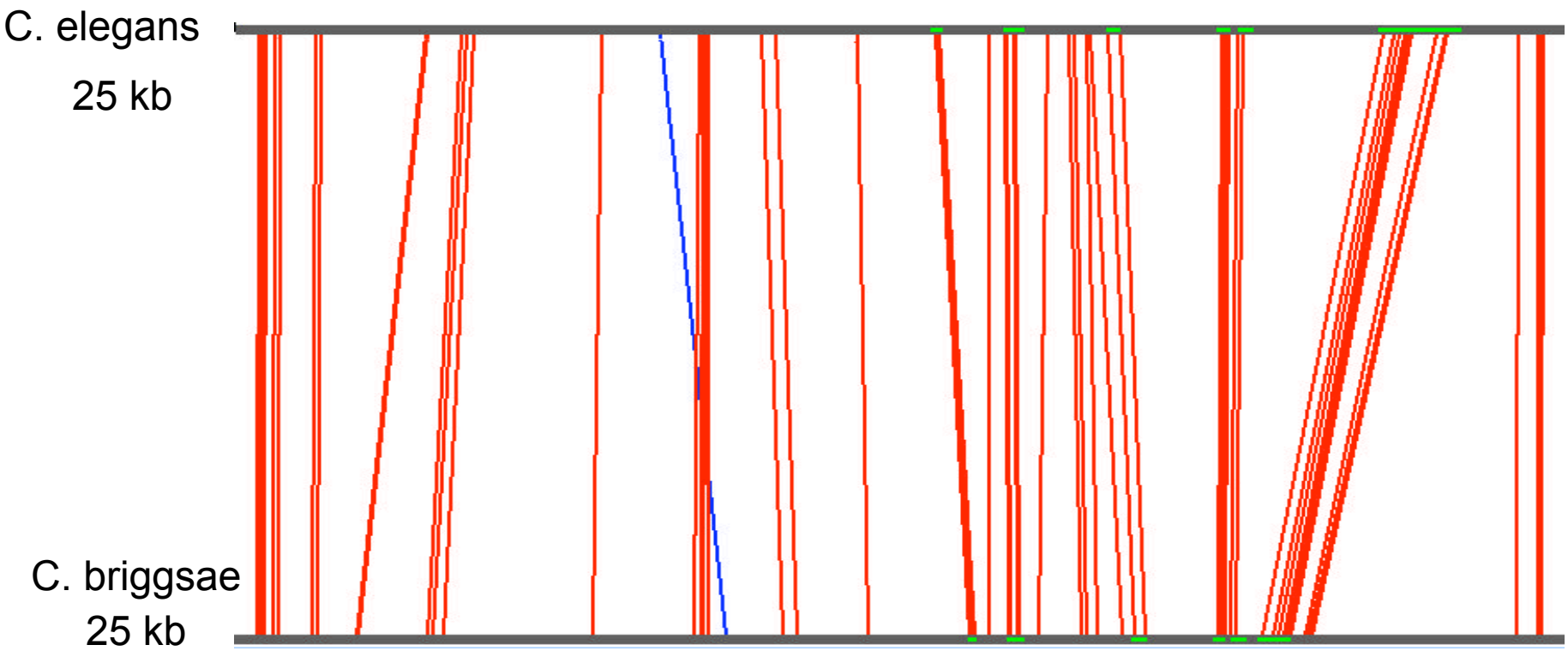
i. From Genomic Data

ii. From Expression Data

Gathering Conserved Sequences in Mussa

- Vertebrates and nematodes are too distant for direct comparisons of conserved sequences that may be cis-regulatory elements.
- Instead of inter-phyla comparisons, intra-phyla comparisons yield conserved cis-regulatory elements that will also be more useful in finding the phyla-specific elements in muscle development.

Mafbx (Muscle Atrophy F-box factor)



A comparison using mafbx, a muscle inhibitor, is shown as an example of Mussa analysis used to obtain conserved elements. Each of the conserved sequences, denoted by red lines, was recorded.

Using Cistematic to find genome-wide patterns of conserved sequences

- The conserved elements from Mussa are fed into Cistematic, a multi-functional python program, to find where else in the genome these elements exist
- The cistematic output provides information on the location of the match, the nearby genes, and all gene ontology information for these genes.
- Once the matches are found, Cistematic can find coappearances of matches or if the matches correspond to matches in other species
- From this data, a list of interesting genes, those with an enrichment of conserved sequence matches nearby, may be generated
- From the conserved sequences found in the Mafbx Mussa analysis, a list of 580 genes was generated

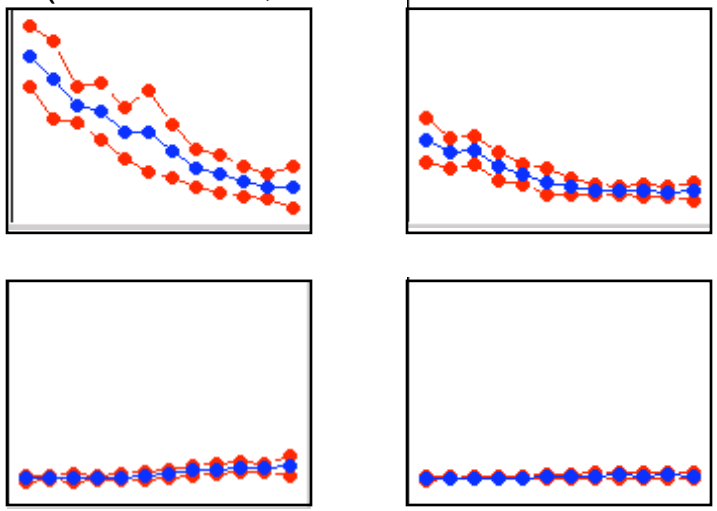
Sample cistematic output

element	genome	chrom.	site/orientation	nearest gene information					gene name	sequence	gene info		
1-LOC-a1	1 celegans	I	8798924 R	8806909	8809743 F	celegans	F36A2.3	aaattacaagtgc	WBGene00009453	8152	P		
10-LOC-a10	1 celegans	I	8787226 R	8771112	8773733 F	celegans	DY3.2	tgacacattcaaa	WBGene00003052	3	P		
13-LOC-a13	1 celegans	I	8797056 R	8809872	8812820 R	celegans	F36A2.2	taagtgcagctga	WBGene00009452	5554	F		

CompClustTk: Grouping genes by expression patterns

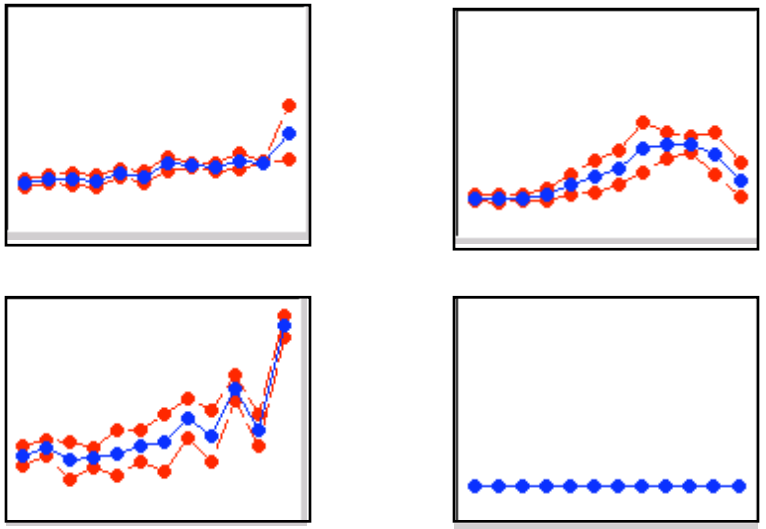
Whole microarray clustering

(15 clusters, 100 iter. in Kmeans)

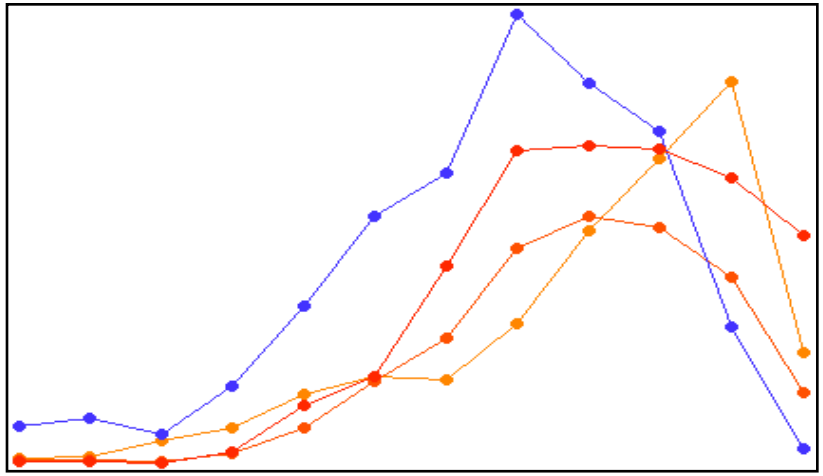


Clustering of selected genes

(15 clusters, 100 iter. in Kmeans)



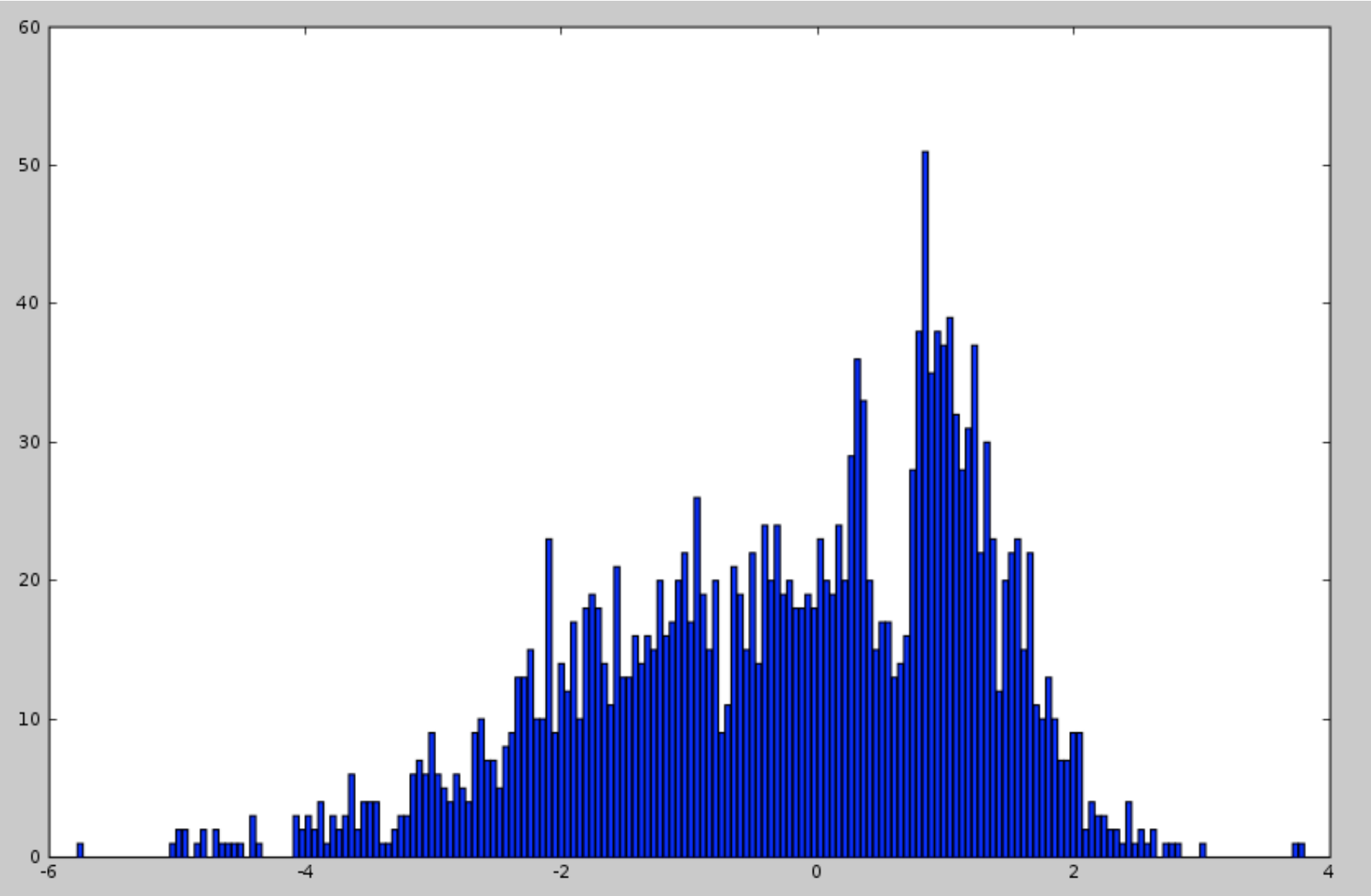
-Clustering of whole genome
-Clustering of selected 211 genes from mafbx mussa-cistematic pathway
-By selecting genes that share expression clusters with known muscle regulators, a refined list of interesting genes may be generated



From Expression Data: compClust/Cistematic

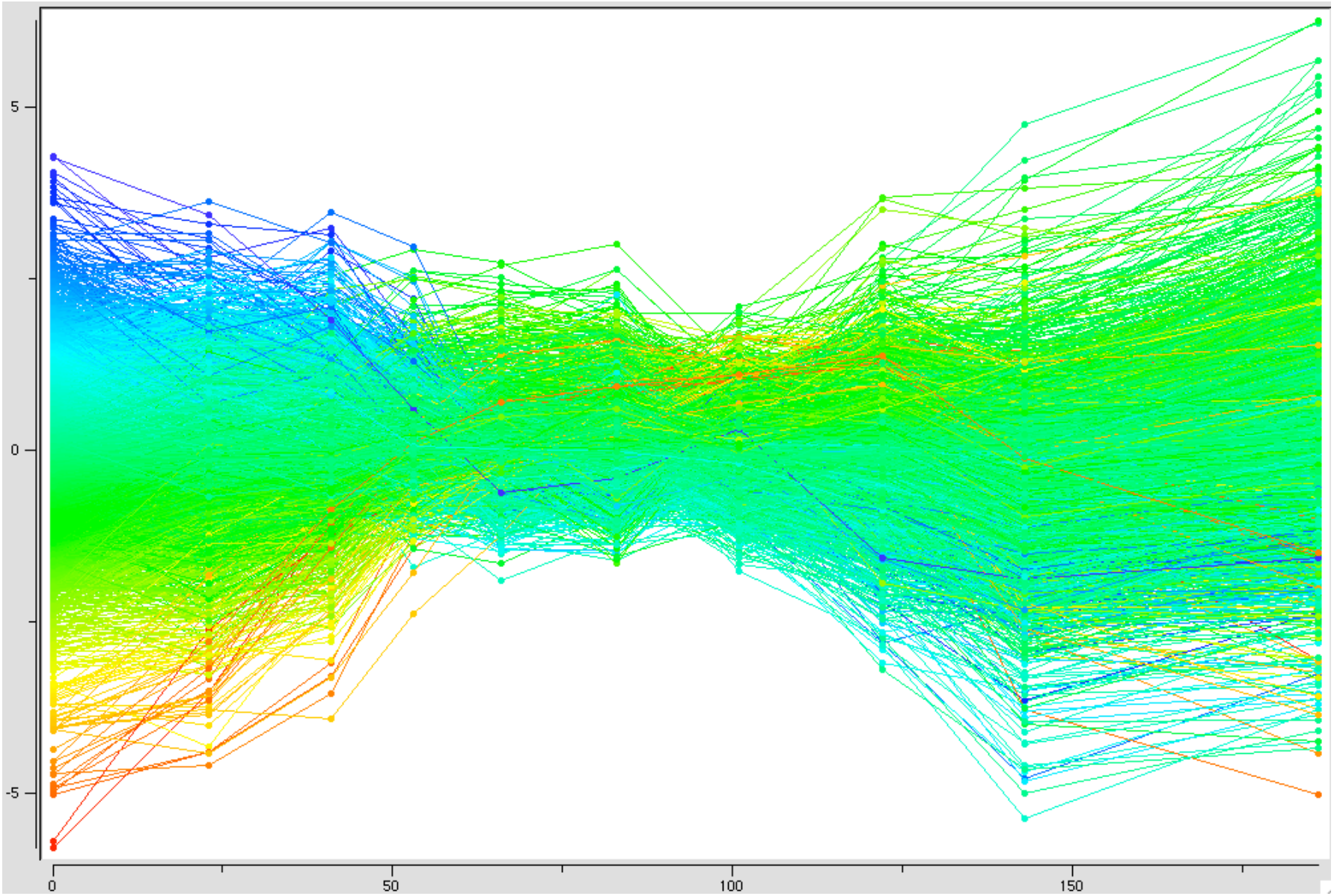
- Successes:
  - Interesting data
  - Can extract compClust data to Cistematic
  - Results match those soon to be published
- Problems:
  - BLT vs. matplotlib
  - Affy annotations
  - Limited homologies
  - Independent cistematic programs
- To do:
  - Analysis of mex-3 data normalized to N2
  - Analysis of C2C12 data
  - Update worm annotations
  - Work out kinks in cistematic integration

Gene distribution



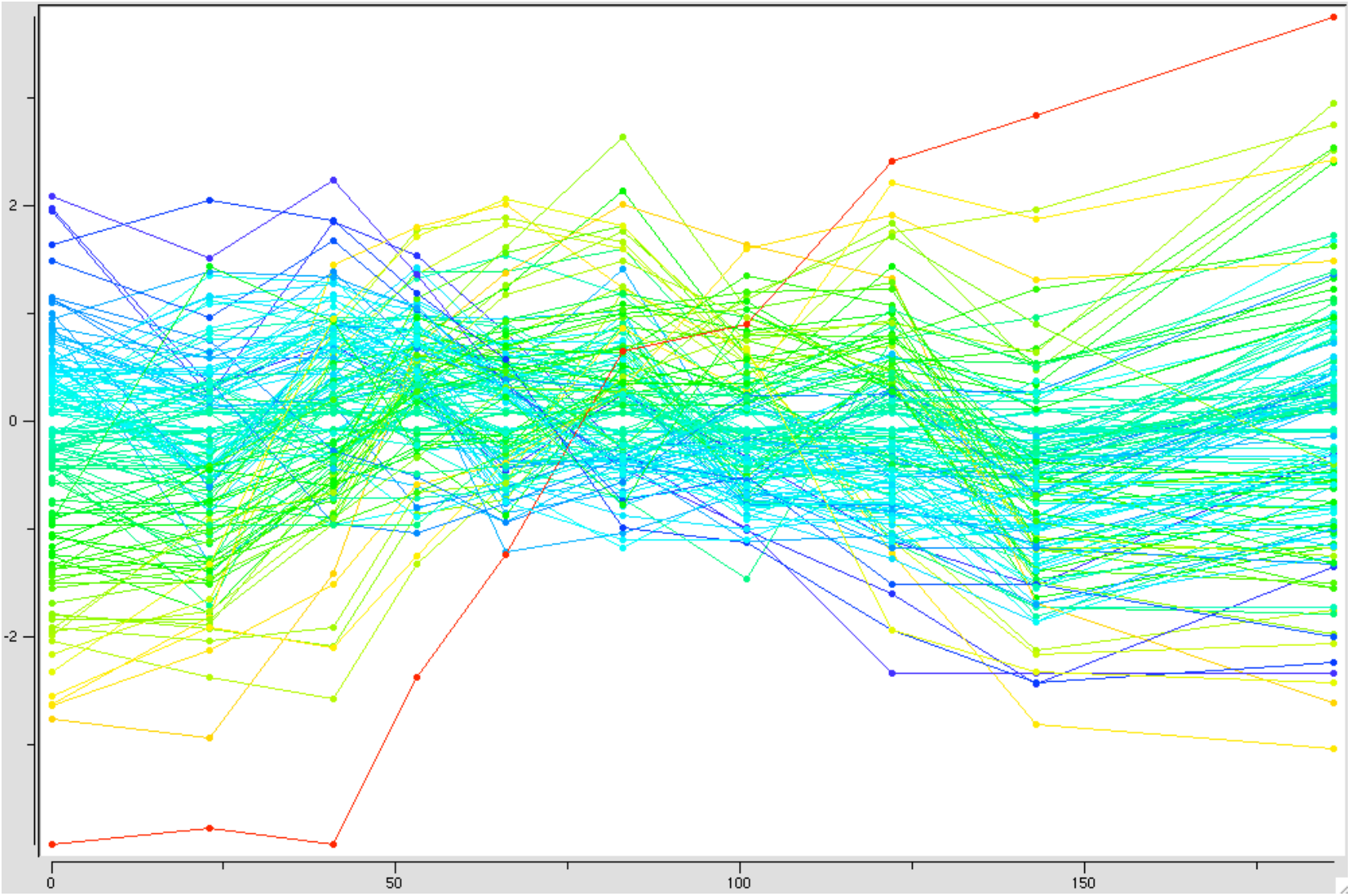
Distribution of log normal gene expression

Normalized Data Set. Unfiltered



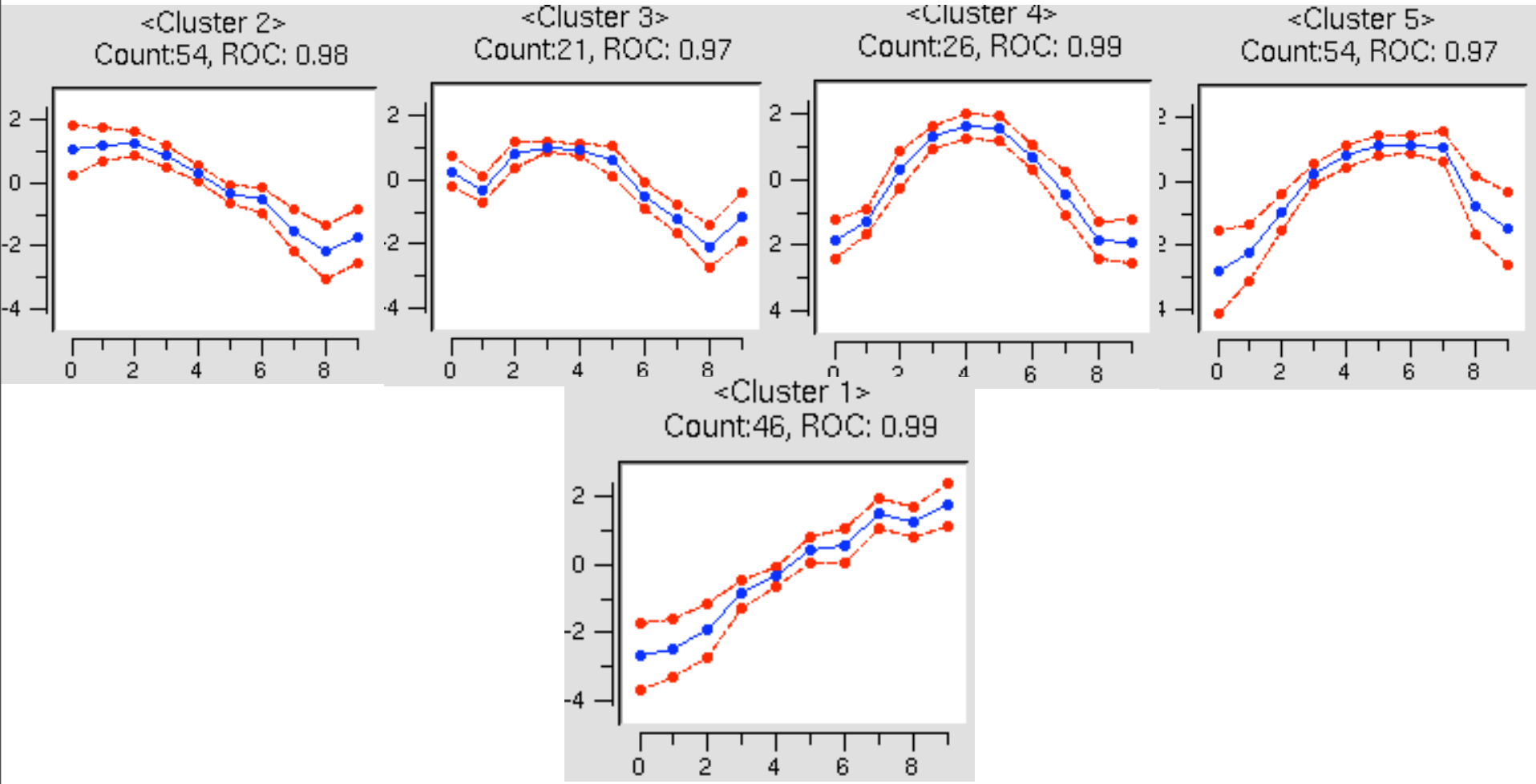
11778 genes

Filtered Data Set



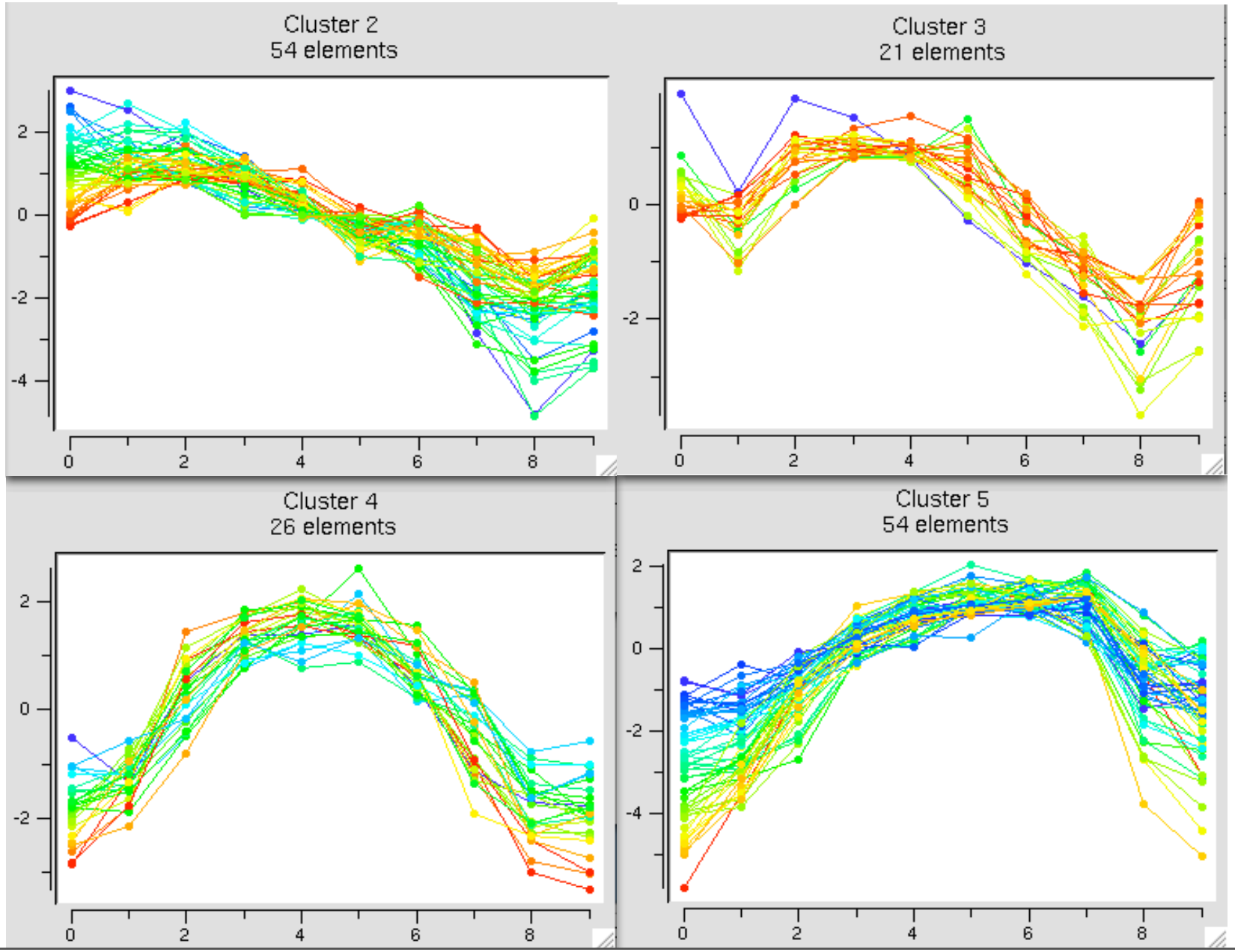
204 genes at 0, 23, 41, 53, 66, 83, 101, 122, 143, 186 min

DiagEM clustering of mex-3 embryonic expression data



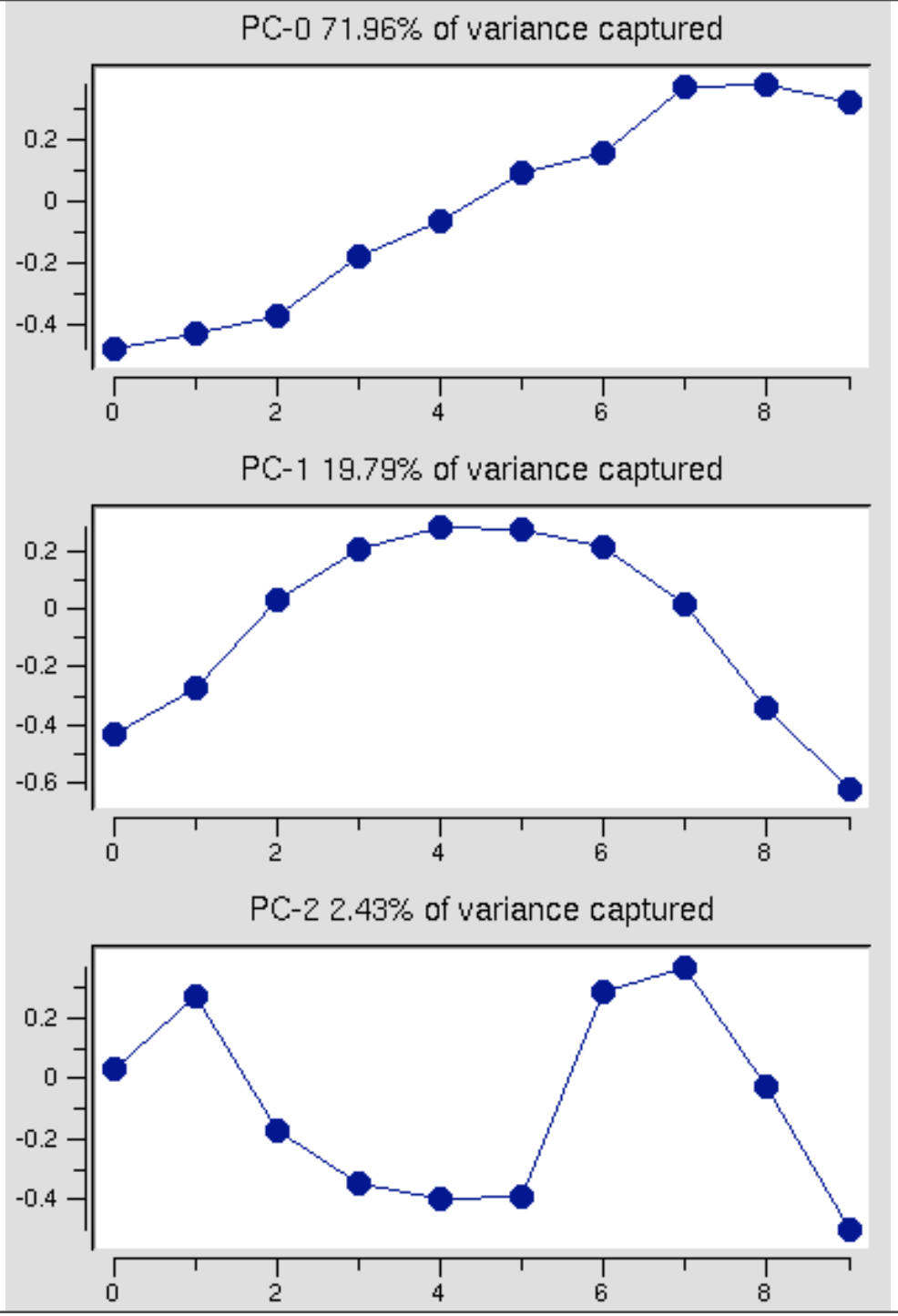
- Apparent temporal progression of expression peaks (2, 3, 4, 6 ==> 41 min, 53 min, 66 min, 101 min)

DiagEM Clustering

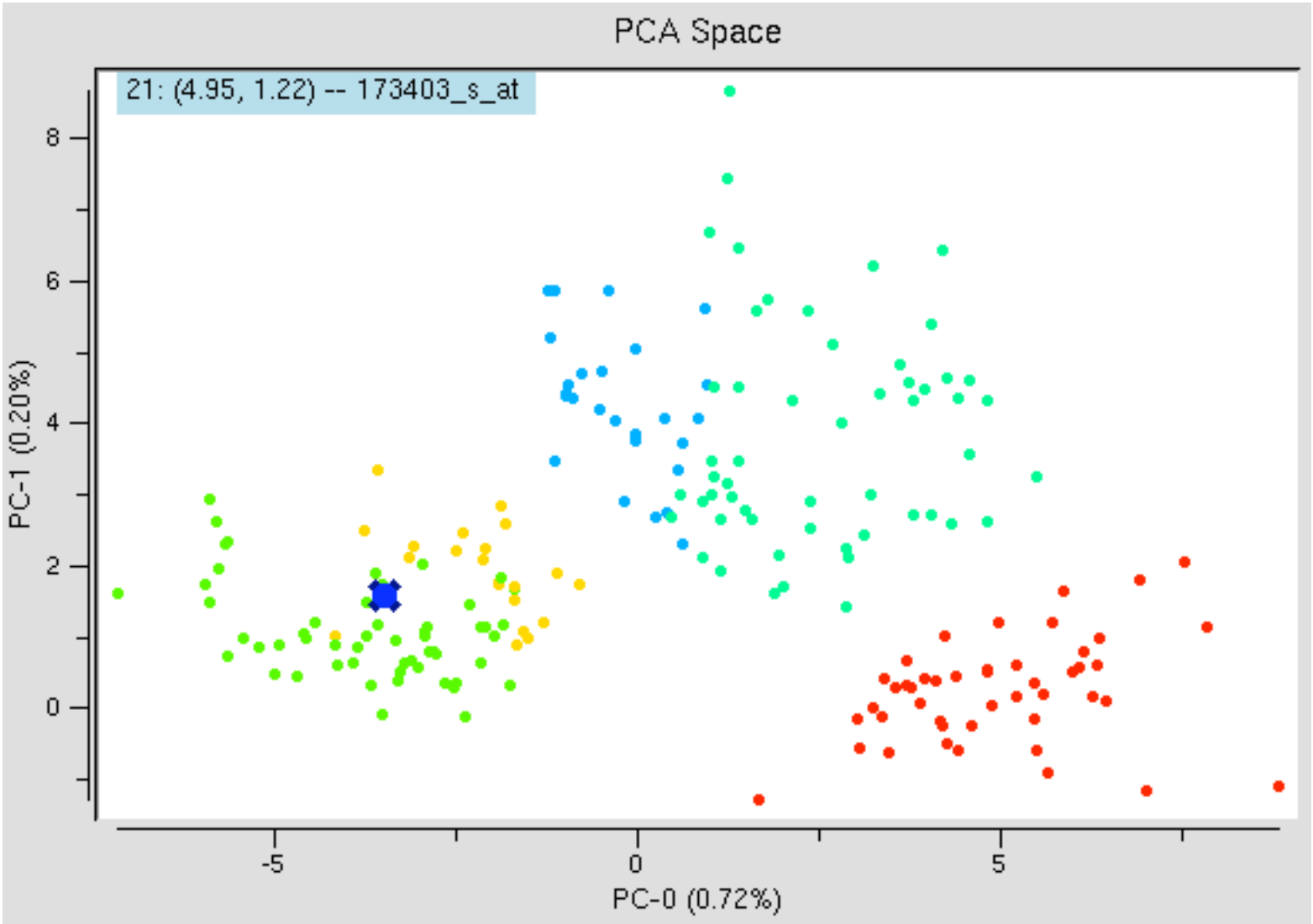


Principal Components

- Considerable drop-off in component contribution
  - 72% contribution from first component
  - 20% contribution from second component
  - 2.5% contribution from third component
- First two cover maternal to zygote shift and peaking of expression

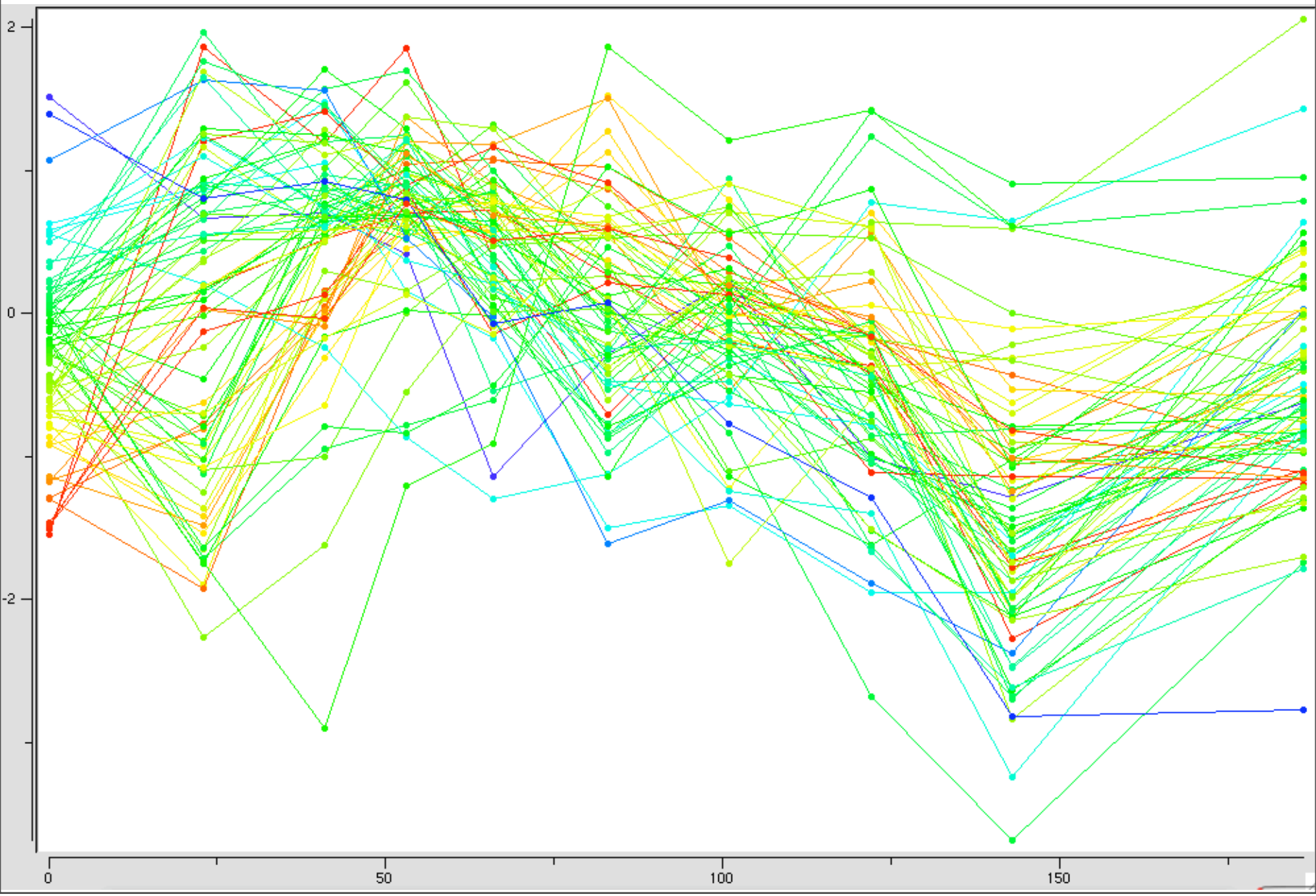


PCA space

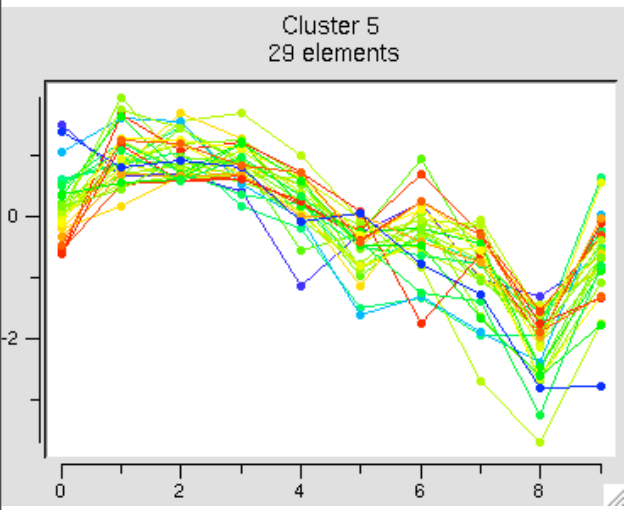


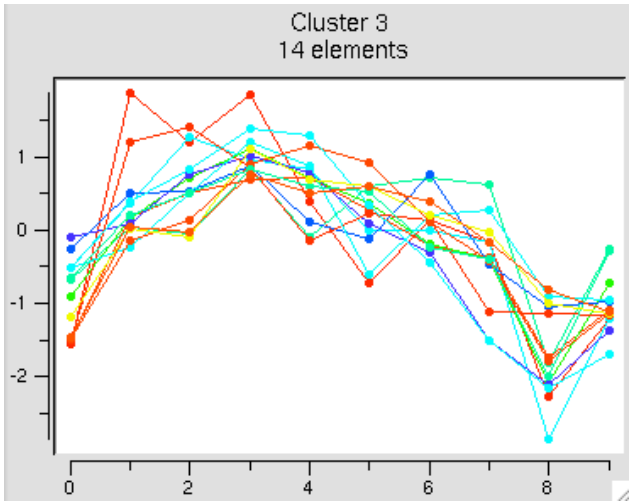
- PCA (principal component analysis)

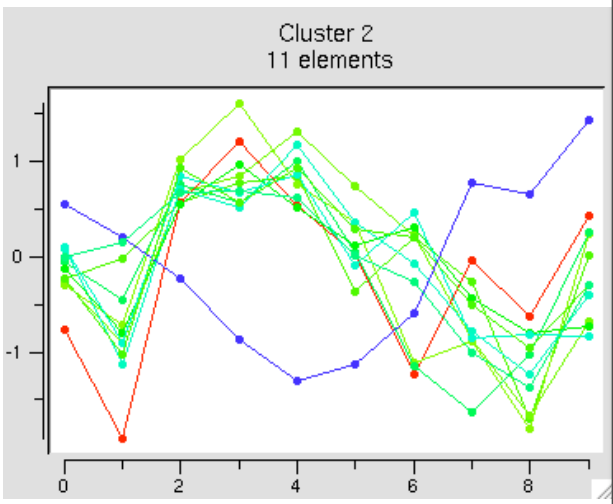
Muscle-Enhanced Filtered Data Set

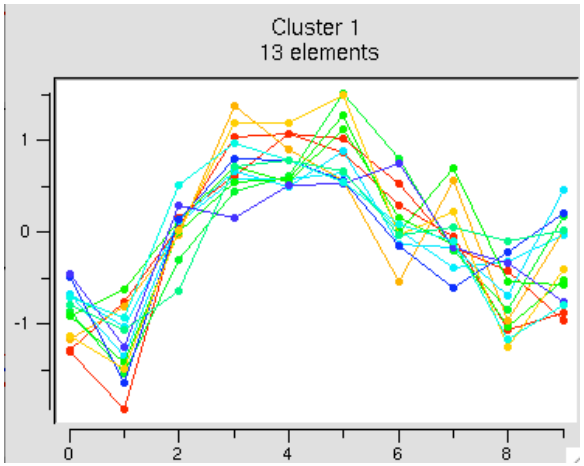


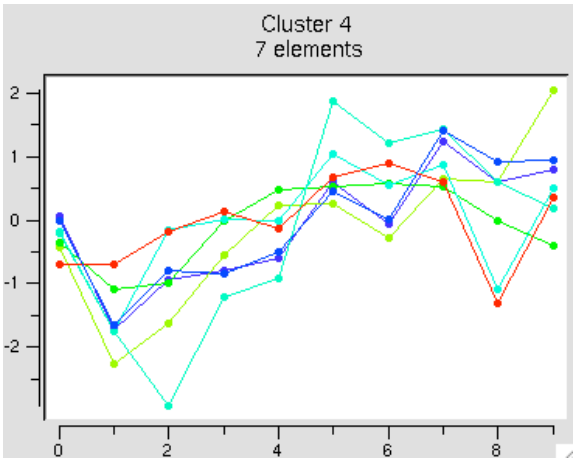
DiagEM Clustering of Muscle-Enhanced Genes







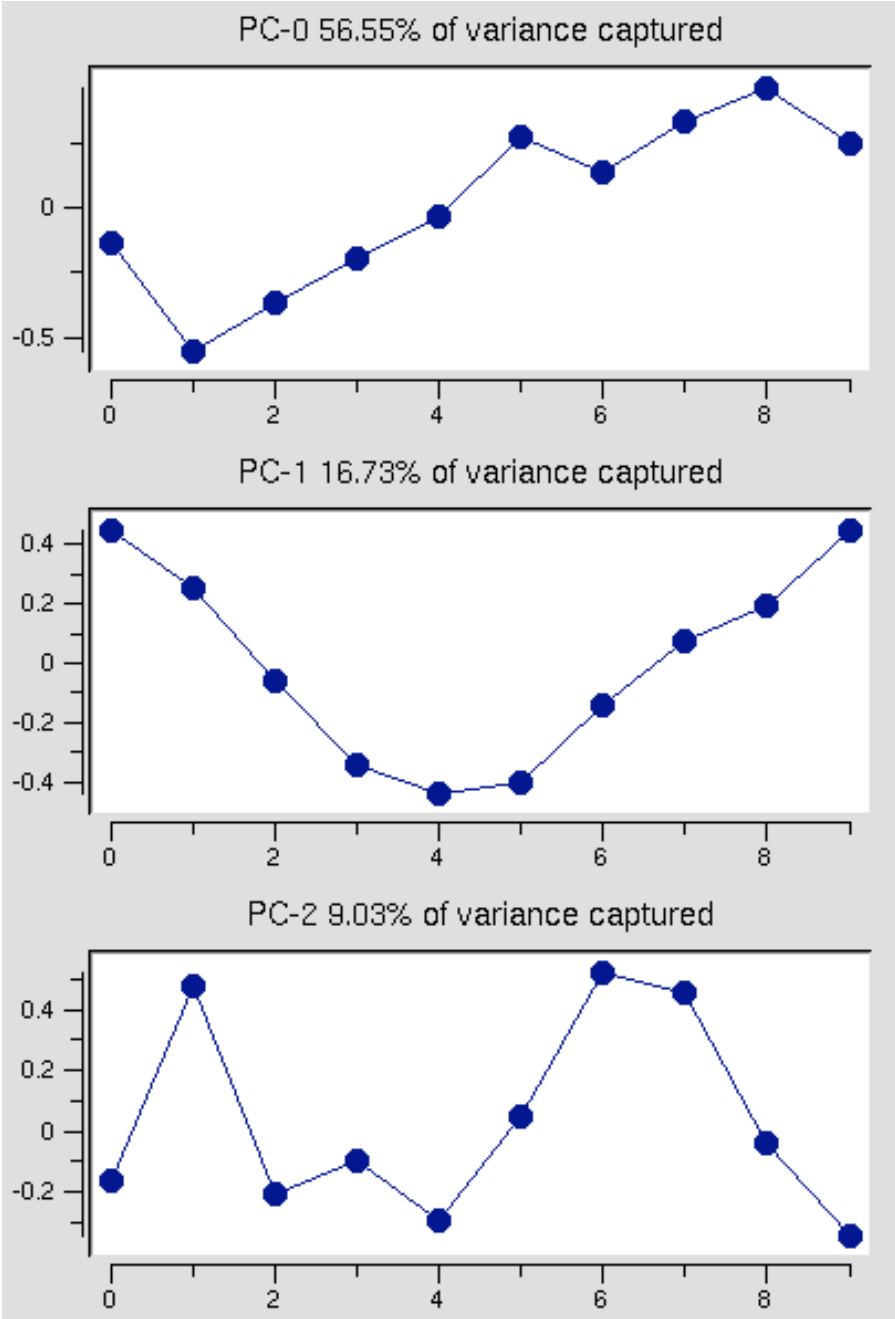




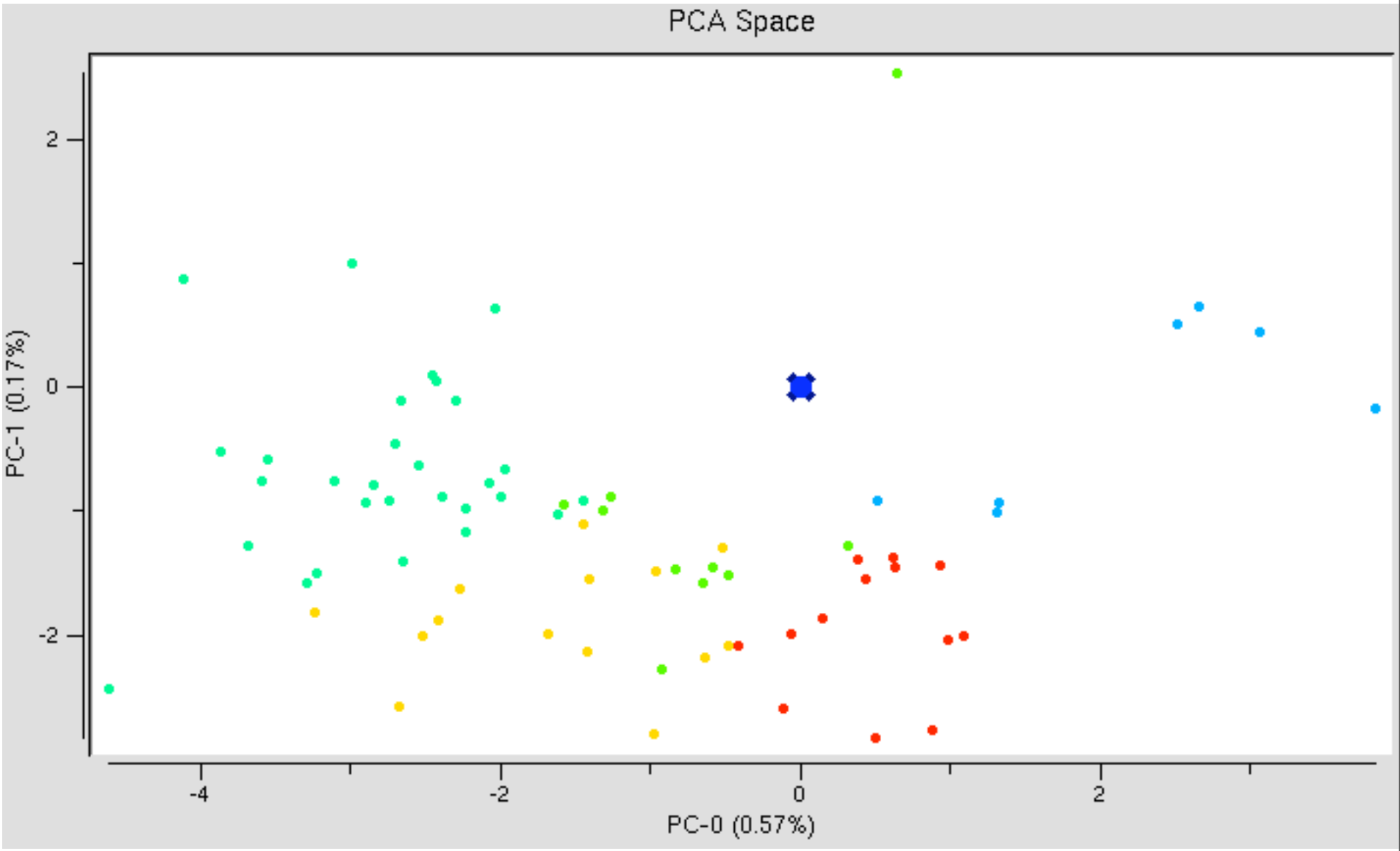
Clusters somewhat noisier with a significant weight toward maternal genes

Principal Components in Muscle-Enhanced Genes

- Less severe drop-off in component contribution
  - 57% contribution from first component
  - 17% contribution from second component
  - 9% contribution form third component
- Inverted secondary component
- Multiple possible causes for differences
  - Smaller data set causing statistical strain
  - Ratio of mex-3 to N2 behaves nonlinearly
  - Muscle genes inherently



PCA Space of Muscle-Enhanced Genes



Cistematic Example Output

Cistematic: Experiment Summary

Experiment: ZK512.3-mussa **Type:** orthology **Analysis:** default

Matching Motifs on Genes

Gene ID	Start	Visual	Stop
('cbriggsae', 'CBG06939')	0		3500
('celegans', 'ZK512.3')	0		3500

Motifs

Hover on top of motif to see consensus. Click on Run checkbox to toggle entire runs. Click on motif checkbox to toggle display of an individual motif.

RunID 1

Display ☑

1-meme-1 1-meme-2 1-meme-3 1-meme-4 1-meme-5

Gene hits that were fully annotated

- Muscle data
  - Isw-1
  - Dif-1
  - Eif-3.C
  - Hmp-2
  - Mes-4
  - Cdk-7
  - Smc-3
- Normalized muscle data
  - Fkh-3
  - Fkh-4
  - Ntp-1
  - Aps-2
  - Prx-12
  - Dnc-1
  - Plp-1
  - Cdd-2
  - Emb-30
  - Dad-1
  - Par-5
  - Rme-2
  - Pas-3
  - Oxi-1
  - Ceh-11

Conclusions and Future Work

- Both techniques can acquire an interesting set of genes
  - Implement Kmeans
  - Implement Signature Algorithm
  - Which method will be more interesting?
  - Will the gene lists largely converge?
- Annotations need considerable work
  - Acquire affy list from Wormbase
  - Address homolog libraries and multiple matches
- Extend available muscle expression data
  - Utilize C2C12 expression data
  - Measure extended time course of expression for

Acknowledgements

- C. Titus Brown
- Hunter Group
- Sternberg Lab
- Wold Lab