University of Delaware

SC17 Denver, CO | hpc connects.

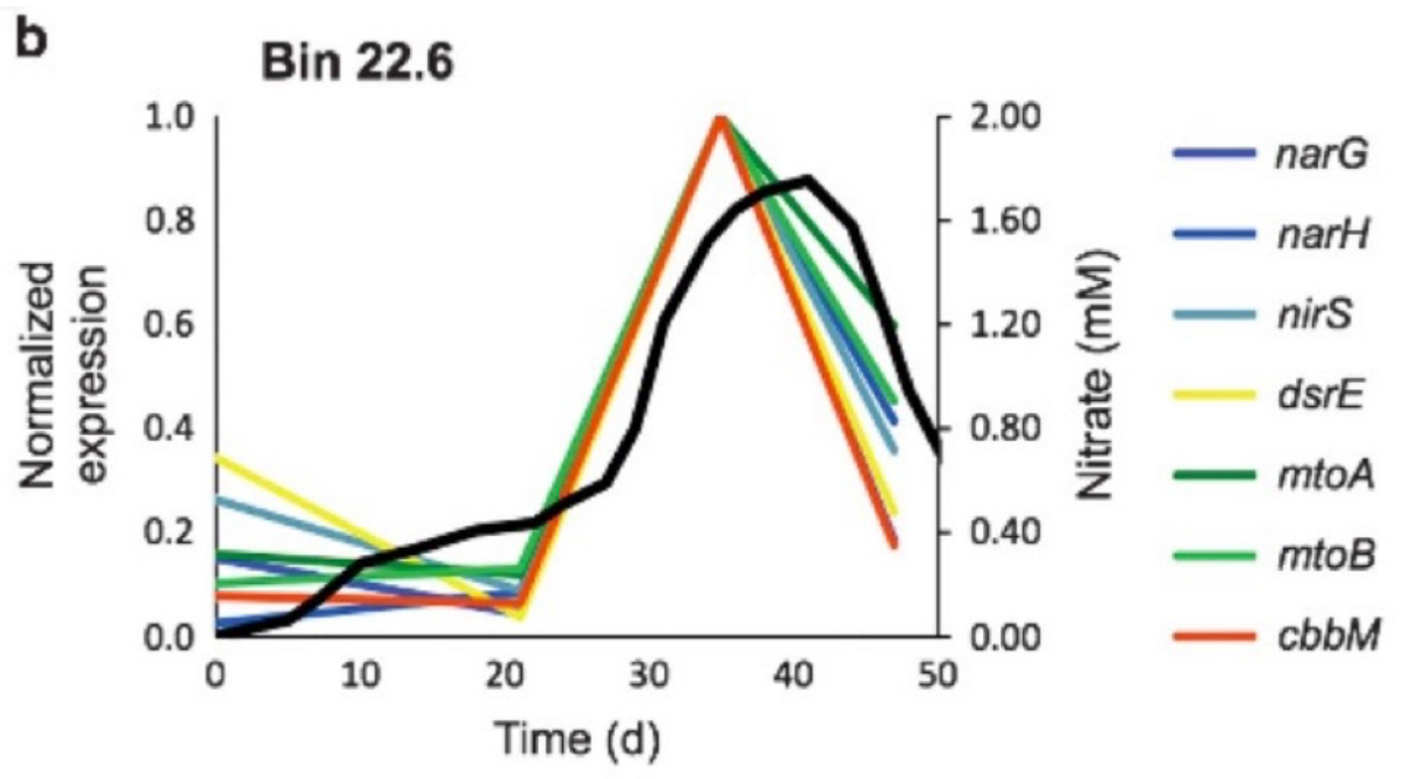
Validation of the Short Time-series Expression Miner (STEM) on Iron Cycling in a Shallow Alluvial Aquifer

Paul Soper, Michela Taufer, Clara Chan, and Stephen Herbein

Background

- We have time expression data for Zetaproteobacteria, iron-oxidizing bacteria found at deep-sea hydrothermal vents
- Using the STEM software to analyze the data produced interesting graphs, but it was unclear what they meant
- Our data are similar to those analyzed in the paper *Metatranscriptomic evidence of pervasive and diverse chemolithoautotrophy relevant to C, S, N and Fe cycling in a shallow alluvial aquifer* by Jewell *et al.*
- We want to use this better-understood system to see whether STEM correctly clusters genes involved in iron oxidation

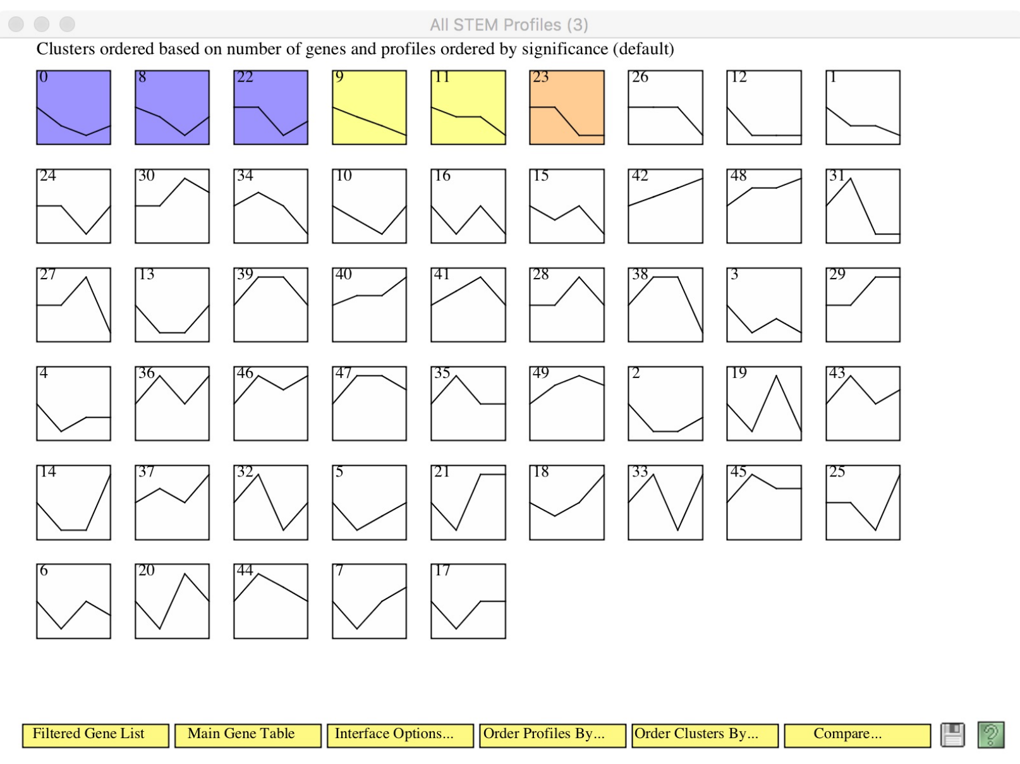
Can STEM replicate bin 22.6 of Jewell *et al.*?



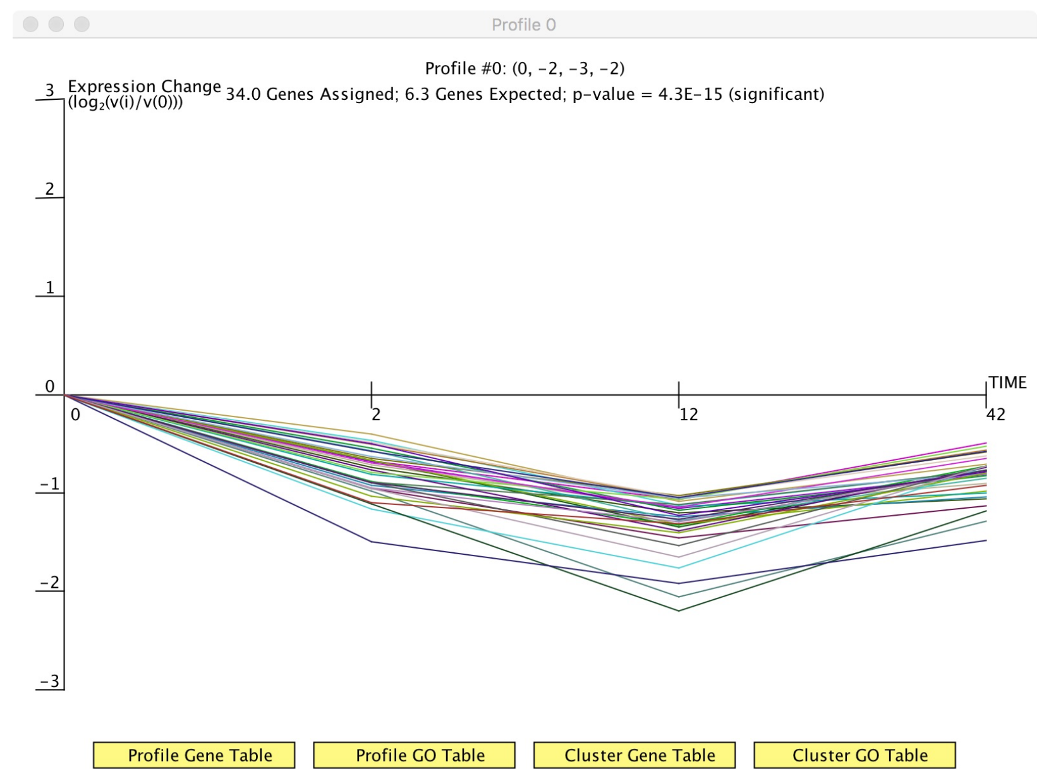
Methodology – Tools

- STEM was developed at Carnegie-Mellon University by Jason Ernst and Ziv Bar-Joseph
- Written in java, and available free for non-commercial use
- Although it can exploit gene annotations and get location data, these were not available
- STEM generates an exhaustive set of profiles, then selects the ones that match the data
- Values are the log of the ratio of a gene's expression at time t > 0 to its expression at t = 0

What are these results trying to tell us?


All STEM Profiles (3)

Clusters ordered based on number of genes and profiles ordered by significance (default)


Profile 0

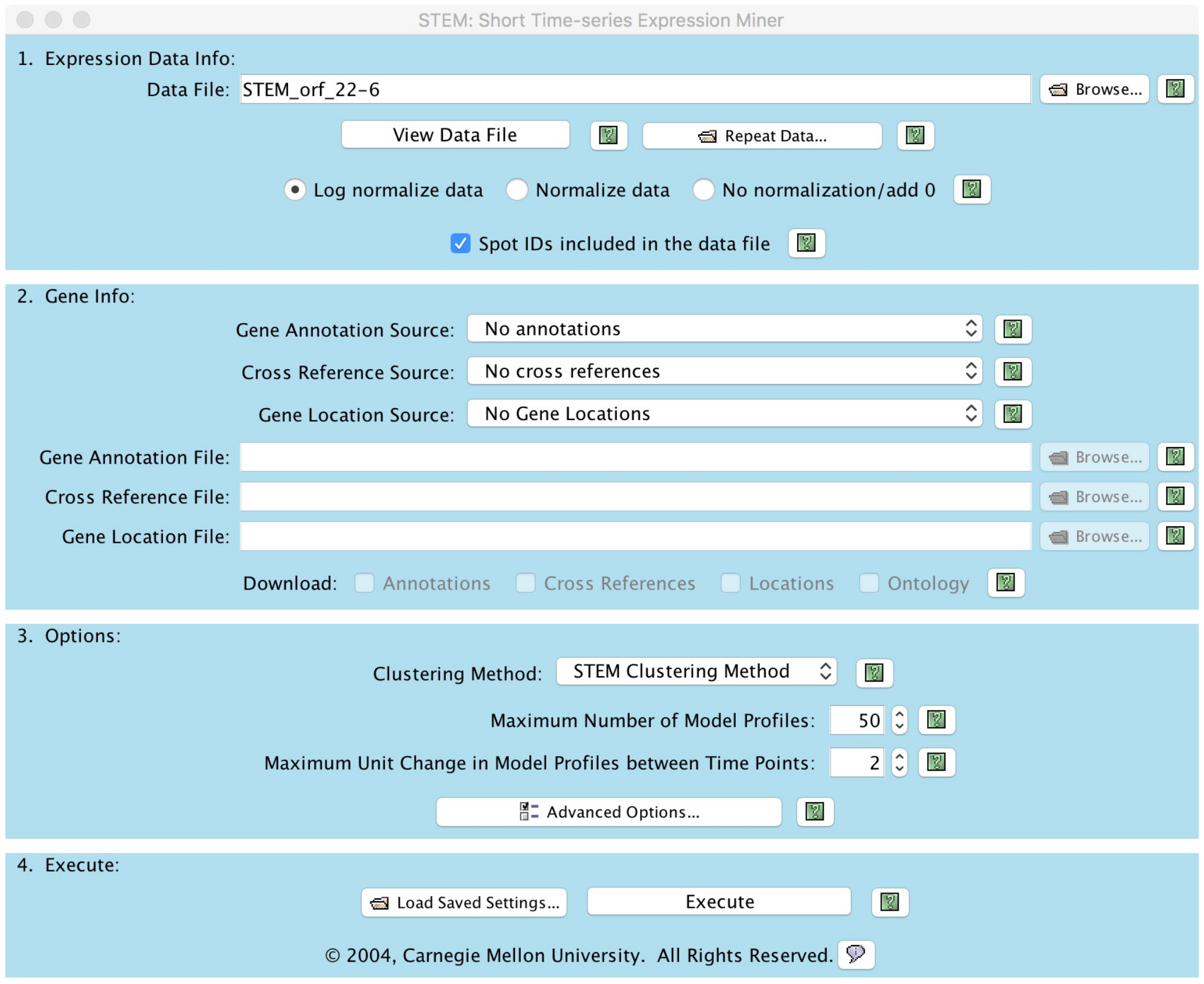
Profile #0: (0, −2, −3, −2)

Expression Change (log₂(v(i)/v(0))) 34.0 Genes Assigned; 6.3 Genes Expected; p-value = 4.3E−15 (significant)

To find out, we ran STEM on a known system.

Methodology – Data

- Gene expression data were downloaded from the Supplemental Information provided for the Jewell paper
- Some open reading frames (contiguous stretches of RNA) were labelled with the corresponding gene
- We filtered the data to include only bins 22.6 and 22.9, which were processed separately


STEM: Short Time-series Expression Miner

1. Expression Data Info:

Data File: STEM_orf_22-6 Browse...

View Data File | Repeat Data...

● Log normalize data ○ Normalize data ○ No normalization/add 0

☑ Spot IDs included in the data file

2. Gene Info:

Gene Annotation Source: No annotations

Cross Reference Source: No cross references

Gene Location Source: No Gene Locations

Gene Annotation File: Browse...

Cross Reference File: Browse...

Gene Location File: Browse...

Download: □ Annotations □ Cross References □ Locations □ Ontology

3. Options:

Clustering Method: STEM Clustering Method

Maximum Number of Model Profiles: 50

Maximum Unit Change in Model Profiles between Time Points: 2

Advanced Options...

4. Execute:

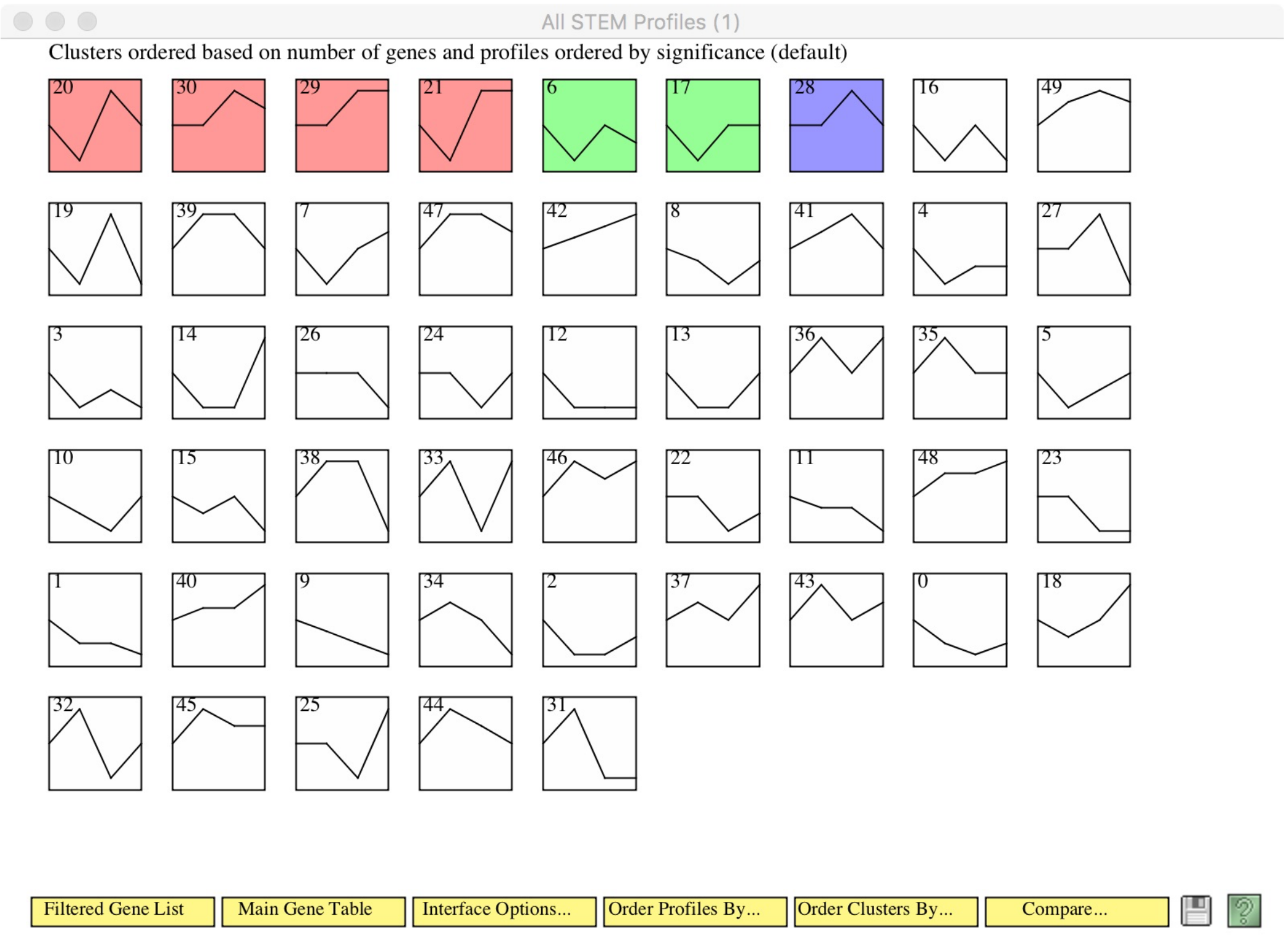
Load Saved Settings... | Execute

© 2004, Carnegie Mellon University. All Rights Reserved.

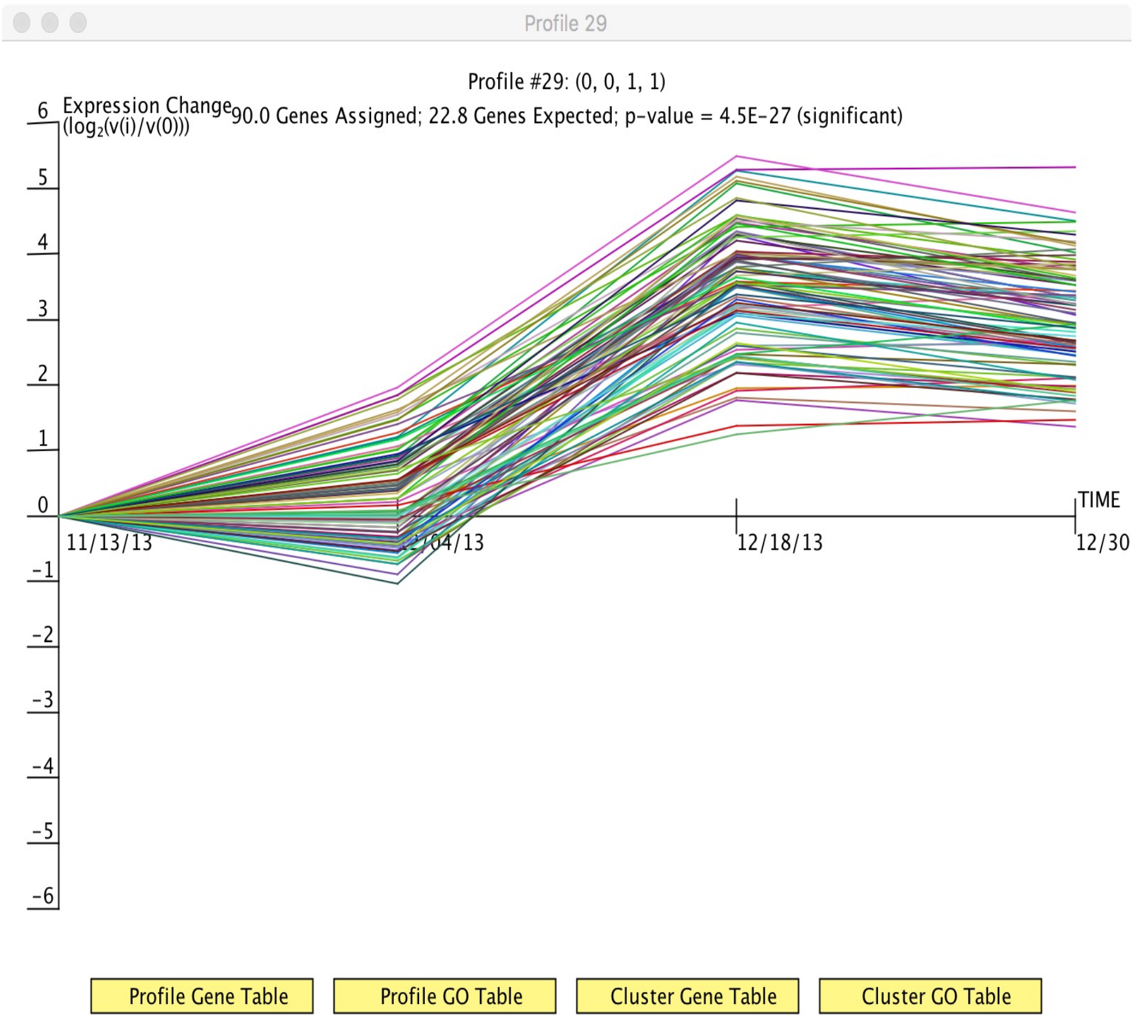
STEM's profiles and clusters are reasonable

Bin 22.6

gene	profile	cluster
dsrL	20	1
cyc2	29	1
mtoA	29	1
dsrO	29	1
mtoB	30	1


All STEM Profiles (1)

Clusters ordered based on number of genes and profiles ordered by significance (default)


Profile 29

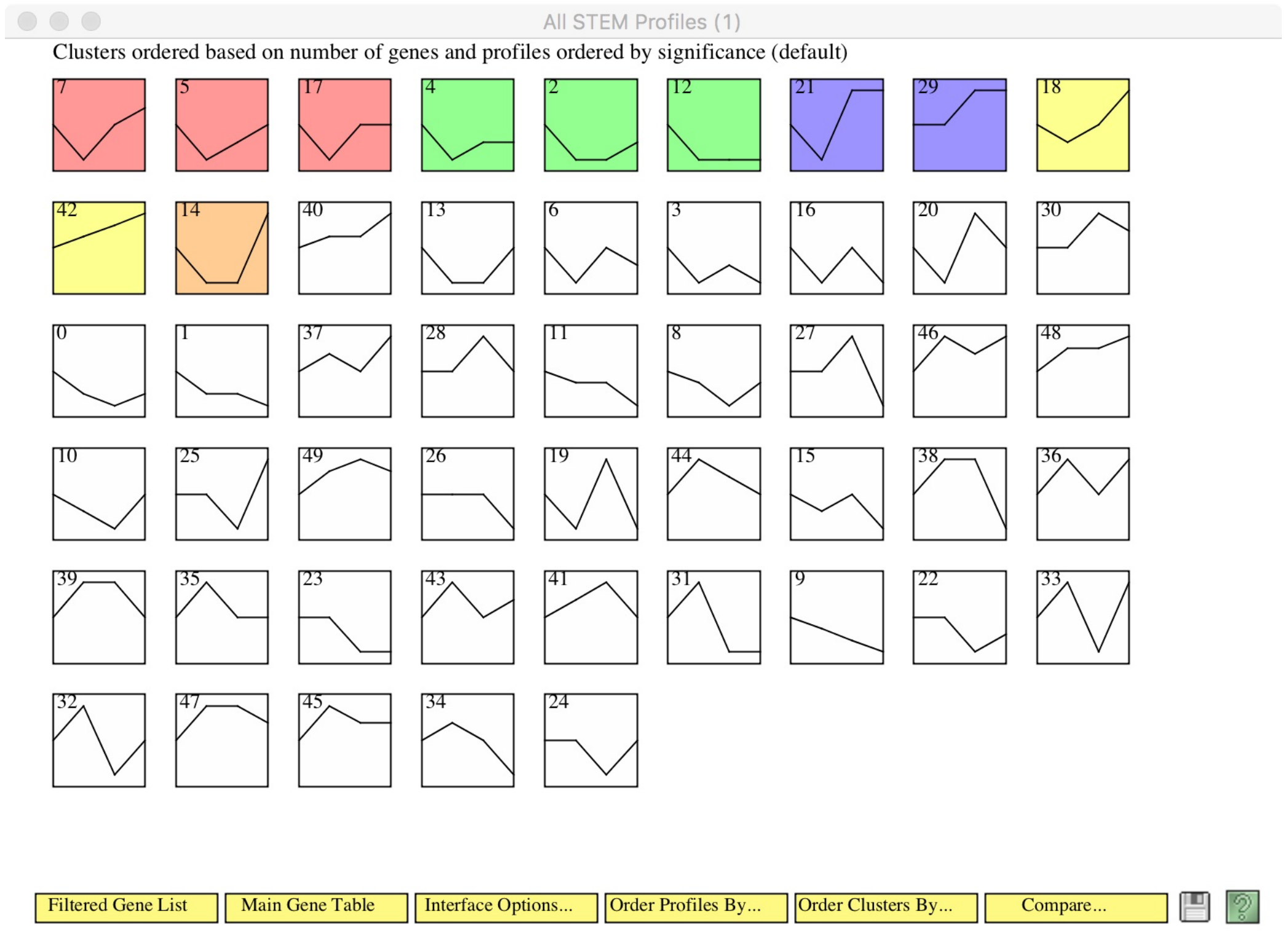
Profile #29: (0, 0, 1, 1)

Expression Change (log₂(v(i)/v(0))) 90.0 Genes Assigned; 22.8 Genes Expected; p-value = 4.5E−27 (significant)

- Statistically significant profiles have colored backgrounds
- Profiles with the same color are in the same cluster

Bin 22.9

gene	profile	cluster
dsrO	5	1
dsrK	5	1
dsrP	5	1
dsrL	7	1
mtoB	21	3
dsrF	21	3
cyc2	42	4


All STEM Profiles (1)

Clusters ordered based on number of genes and profiles ordered by significance (default)

- For bin 22.9, all of the identified genes appear in statistically significant clusters

Conclusions

- STEM allocates all known genes to statistically significant clusters
- For bin 22.6, agreement with Jewell is good (keeping in mind the differences in normalization and y scale)