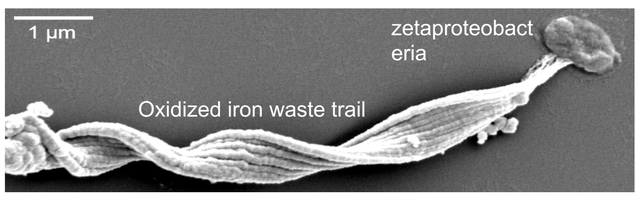
Clustering Temporal Gene Expressions of Iron-Oxidizing Zetaproteobacteria

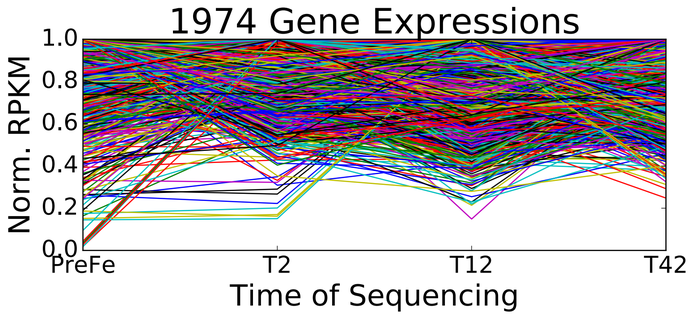
Students: Stephen Herbein and Sean McAllister
Advisors: Michela Taufer and Clara Chan

Zetaproteobacteria in a Nutshell

- Zetaproteobacteria are an iron-oxidizing bacteria commonly found at deep-sea hydrothermal vents
- The bacteria play a big role in the rusting of ship hulls, metal pilings, and pipelines
- Scientists want to better understand the genes involved in the oxidation of iron



Gene Expression Sampling

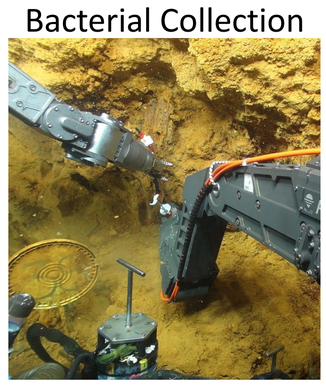


Can we find patterns in the gene expressions to determine which genes are involved in iron oxidation?

Data Acquisition & Cleaning

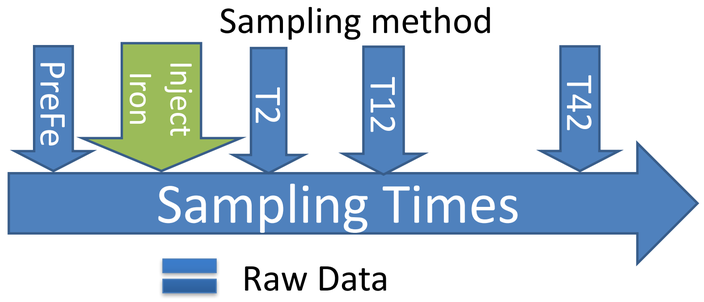
- Bacterial samples were collected from the Loihi Seamount in Hawaii
- The samples were preserved on site by Sean McAllister and Anna Leavitt
- The samples were preserved at 4 different times
  - 1 time before the injection of iron
  - 3 times after the injection of iron (2, 12, and 42 minutes)
- Raw data consists of 4 read counts over time for each gene, along with metadata associated with each gene (e.g. ID, length)
- Normalize raw reads for read length and number of sequence reads using Reads Per Kilobase per Million (RPKM)
- Normalize based on each gene's maximal expression over time
- Remove genes with any raw reads = 0, all raw reads < 50, or any RPKM > 10000

Bacterial Collection



Bacterial Preservation

Sampling method



Raw Data

Gene Information		Raw Reads			
ID	Length	PreFe	T2	T12	T42
2264875794	2217	227	236	40	83
2264875795	1092	124	100	21	60
2264875796	657	106	91	15	42
2264875797	582	216	158	27	65
2264875798	582	182	138	27	59
2264875793	105	3	6	1	0

Normalization and Cleaning

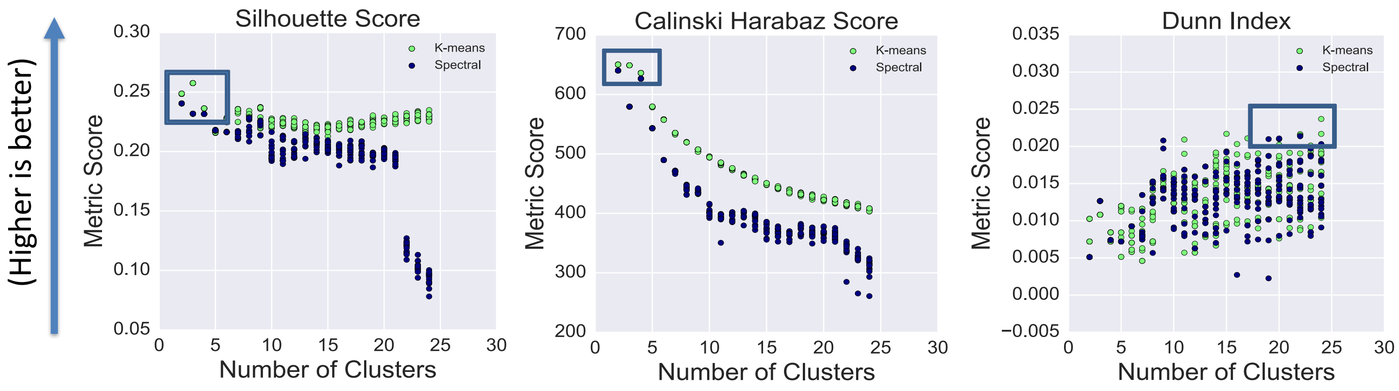
	Normalized RPKMS			
Gene ID	PreFe	T2	T12	T42
2264875794	78.9%	100.0%	69.3%	72.7%
2264875795	82.0%	80.6%	69.2%	100.0%
2264875796	95.5%	100.0%	67.4%	95.4%
2264875797	100.0%	89.2%	62.3%	75.9%
2264875798	100.0%	92.4%	74.0%	81.7%
~~2264875793~~	~~41.0%~~	~~100.0%~~	~~68.2%~~	~~0.0%~~

Clustering Methods

	K-means	Spectral
Use-cases	General-purpose, even cluster size, flat geometry	Few clusters, even cluster size, non-flat geometry
Scalability	Very large # of samples, medium # clusters	Medium # samples, small # clusters
Distance Metric	Euclidian Distance	Graph Distance
Parameters	Number of clusters	Number of clusters

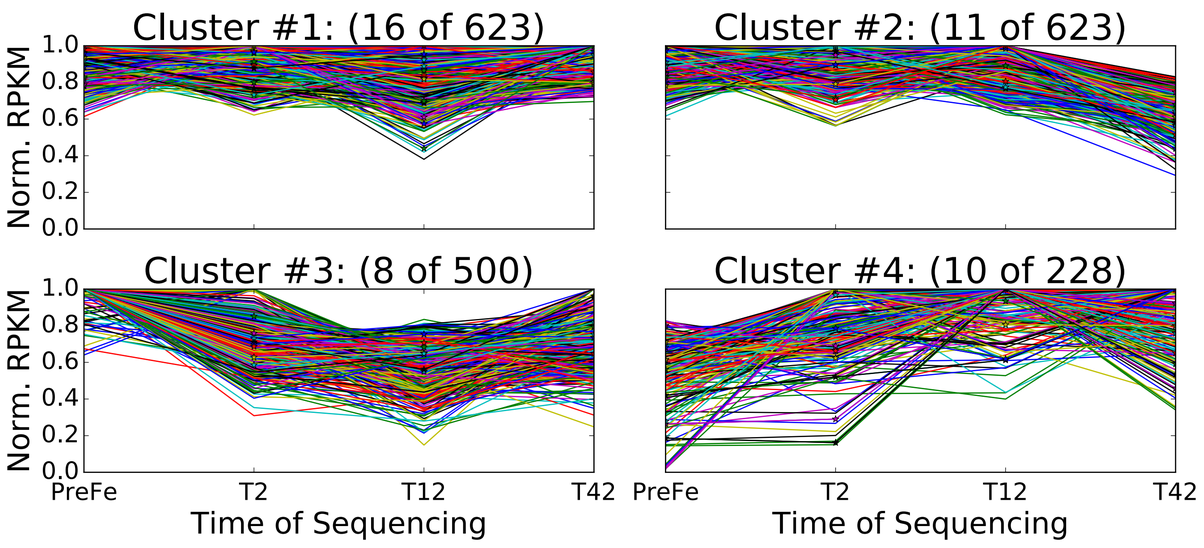
How do we determine the optimal number of clusters?

Determining the Optimal Number of Clusters



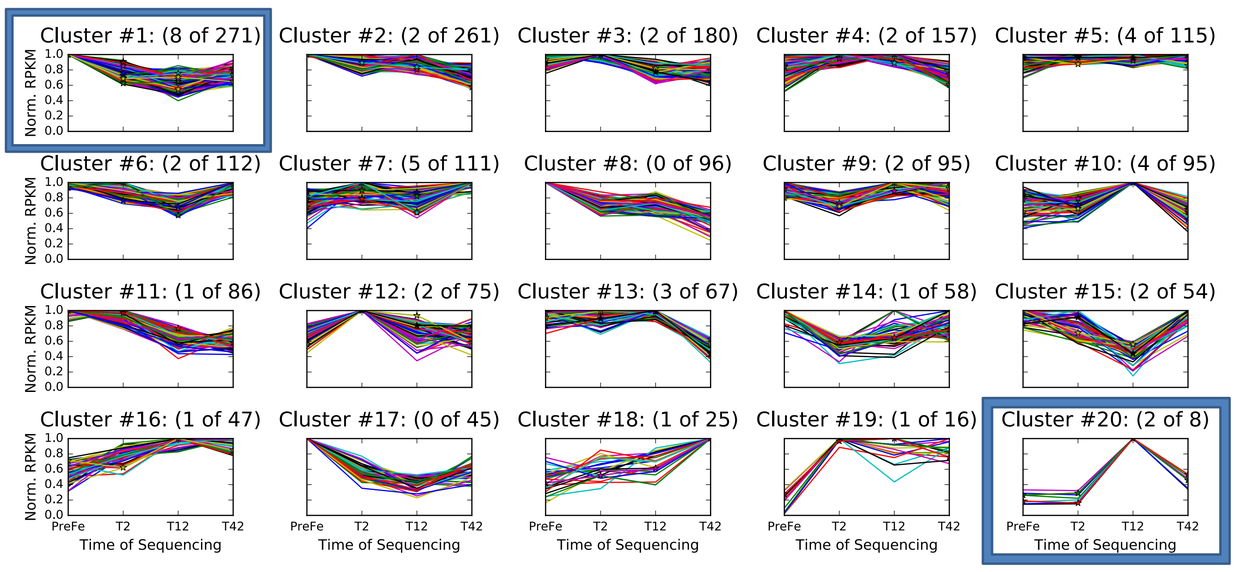
The metrics suggest that using either 4 or 20 clusters is optimal

4 Clusters Under Spectral Clustering



- Results of 4 clusters with k-means are very similar to Spectral

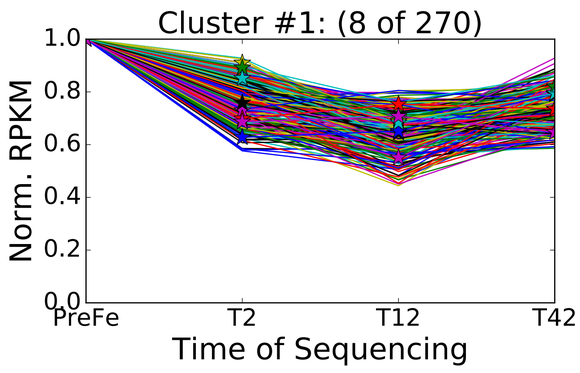
20 Clusters Under Spectral Clustering

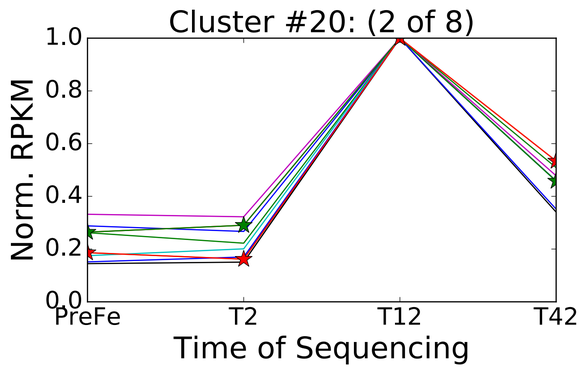


- Results of 20 clusters with k-means are visually noisier than Spectral

Conclusions

- 4 clusters result in high metric scores, but the clusters are too large to be useful for the next steps in scientific inquiry
- 20 clusters provide a finer level of detail and help narrow the search down to a smaller number of genes
- Cluster #1 has a large number of genes that were pre-selected by the scientists as interesting
- Cluster #20 has a high percentage of interesting genes





The optimal number of clusters depends on the scientific objective

NSF OCE-1155290 & NASA NNX10AN63H