GO HERE!

- http://hpc.ilri.cgiar.org/beca/training/AdvancedBFX2013/index2.html

J. Craig Venter™ INSTITUTE

Data Mining: Clustering and Statistical Analysis with MeV

Marcus Jones
Infectious Disease/Genomic Medicine
J. Craig Venter Institute
28 August 2013

J. Craig Venter™
INSTITUTE

Data Analysis Concepts

J. Craig Venter™ INSTITUTE

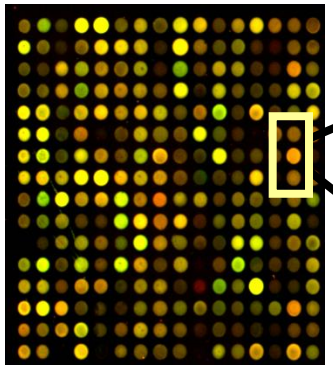
Multiple Sample Expression Analysis

Clustering, Statistical Analysis, and Beyond

- The last stage of analysis involves the analysis of multiple hybridizations

- Time Courses
- Distinct Experimental Groups
- Replicates of one direct measurement of interest

Expression Ratios



Dyes

Measured Intensities

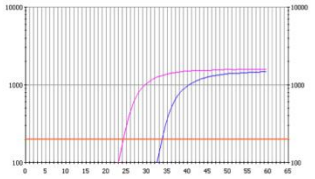
Cy5 → 216108

Cy3 → 77103

$$\frac{216108}{77103} = 2.798$$

Ratio

Genome Sequencing



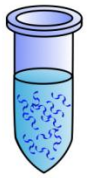
RNA

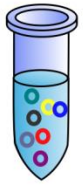
cDNA synthesis

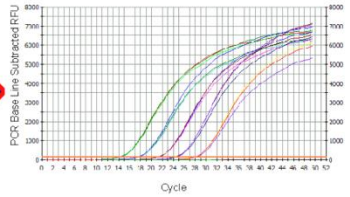
384-well qRT-PCR

Data analysis

***in vivo* Gene Expression**







J. Craig Venter™ INSTITUTE

Log_2 Expression Ratios: Benefits

Log ratios are easier to work with than regular ratios.

A five-fold change in expression level could be represented by one of two regular expression ratios:

$$\frac{500}{100} = \text{Ratio } 5.0 \quad \text{OR} \quad \frac{100}{500} = \text{Ratio } 0.2$$

Note the asymmetrical nature of the ratio values

However, if the regular expression ratios are converted to log expression ratios:

$$\frac{500}{100} = \text{Ratio } 5.0 \rightarrow \log_2(5.0) = \text{Log}_2 \text{ Ratio } 2.32$$

OR

$$\frac{100}{500} = \text{Ratio } 0.2 \rightarrow \log_2(0.2) = \text{Log}_2 \text{ Ratio } -2.32$$

The log ratios equal in magnitude and the sign indicates which channel had the higher intensity

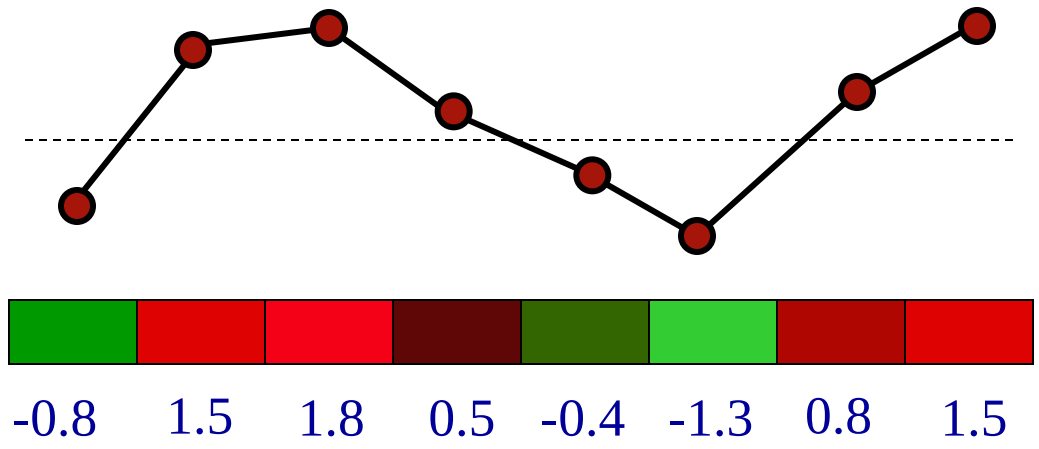
J. Craig Venter™ INSTITUTE

Expression Vectors: Shape

	Expt 1	Expt 2	Expt 3	Expt 4	Expt 5	Expt 6	Expt 7	Expt 8
Gene A	-0.8	1.5	1.8	0.5	-0.4	-1.3	0.8	1.5

Log_2 Ratios

Log2(cy5/cy3)



J. Craig Venter™ INSTITUTE

Expression Matrix

We can study the expression levels of multiple genes across a series of experimental conditions by stacking the corresponding expression vectors.

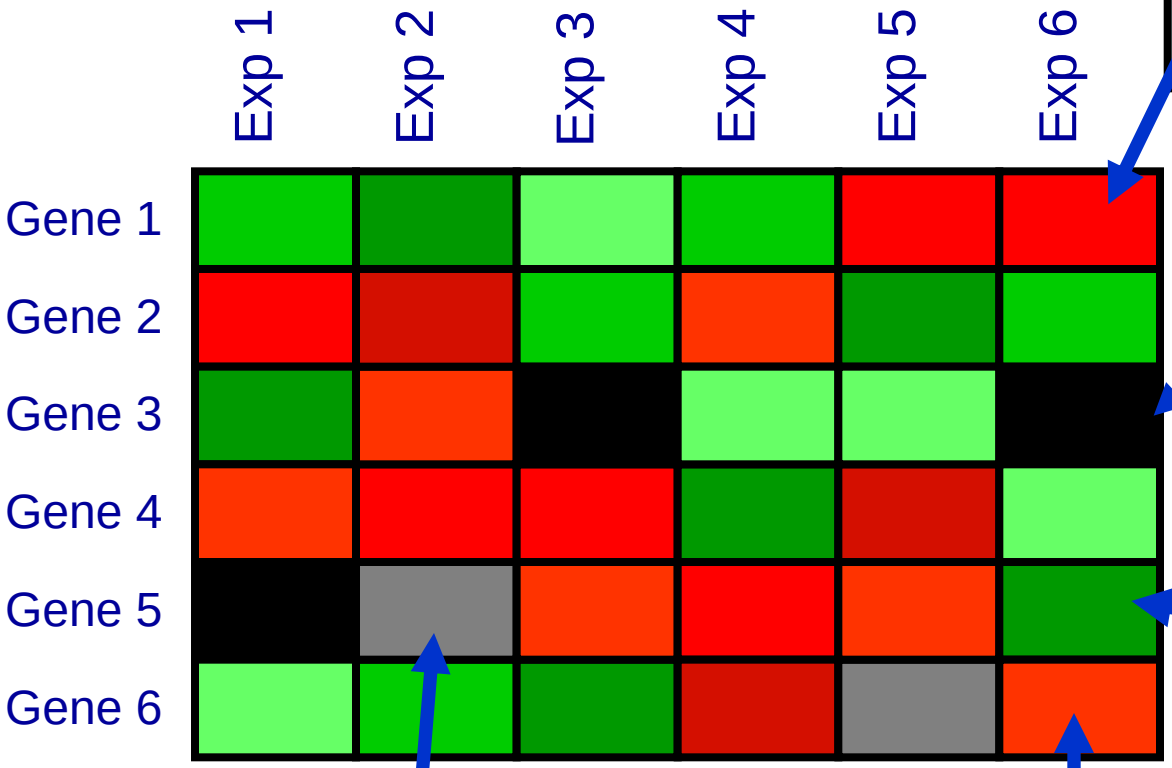
	Expt 1	Expt 2	Expt 3	Expt 4	Expt 5	Expt 6
Gene A	-2.32	-1.69	-0.87	-0.12	0.73	1.42
Gene B	2.71	2.09	1.24	0.70	0.25	0.08
Gene C	-1.55	-0.49	0.97	1.32	0.59	-0.38

This is called an *Expression Matrix.*

J. Craig Venter™ INSTITUTE

Expression Matrix: Just Add Color

Expression Matrices are commonly represented as a grid of red and green cells:



Each element is a log ratio: $\log_2$ (**Cy5 / Cy3**)

Black indicates a log ratio of ~ 0 (i.e. **Cy5 = Cy3** or **Cy5 ~= Cy3**)

Green indicates a negative log ratio (i.e. **Cy5 < Cy3**)

Gray indicates missing data

Red indicates a positive log ratio (i.e. **Cy5 > Cy3**)

J. Craig Venter INSTITUTE

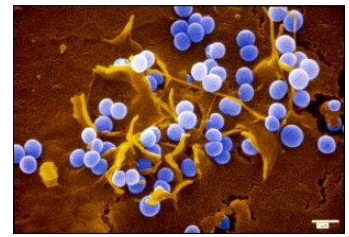
Basic Analysis Approaches

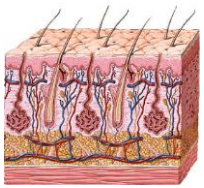
- **General Clustering**
  - For finding gene sets with coherent patterns of expression, e.g. hierarchical clustering, k-means, SOTA, SOM
- **Hypothesis Driven Analysis**
  - Statistical tests based on defined experimental design, e.g. t-test, Significance Analysis of Microarrays (SAM), ANOVA, 2-Factor ANOVA
- **Biological Role Identification**
  - Finding biological meaning from gene lists, EASE

What is RNASeq?

- Sample type
- Sample/Library preparation
- Instrument for processing
- Mapping vs *de novo* assembly
- RNASeq measurement values
  - RPKM
  - FPKM

Characterization of Host and Pathogen Expression





Total RNA

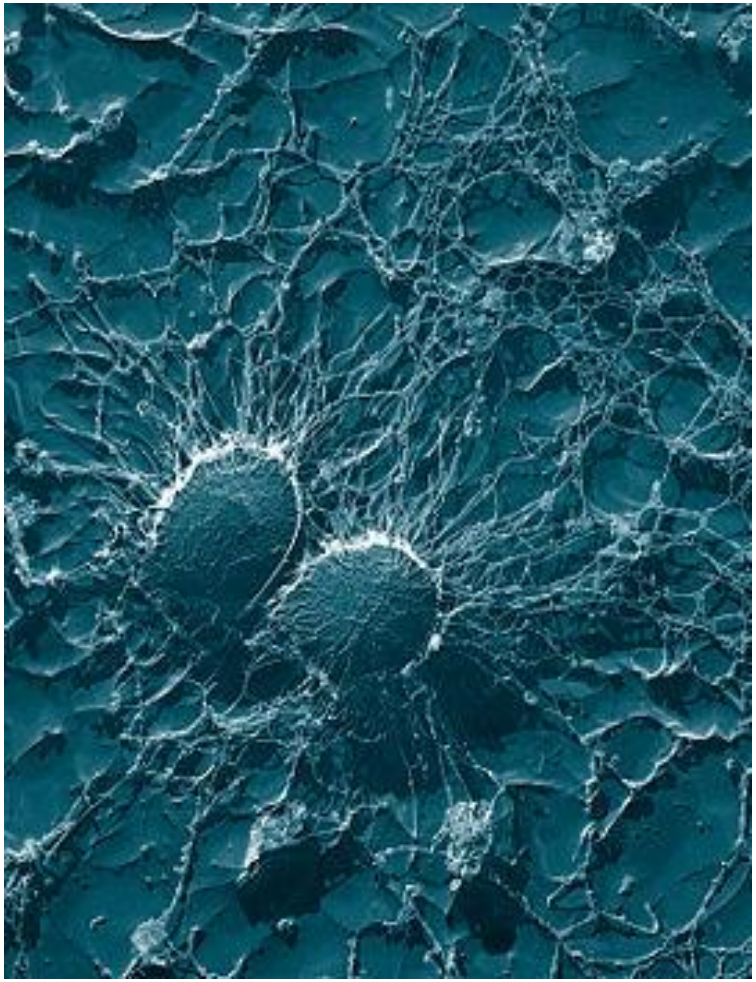
Enriched Bacterial mRNA

HiSeq 2000

J. Craig Venter™ INSTITUTE

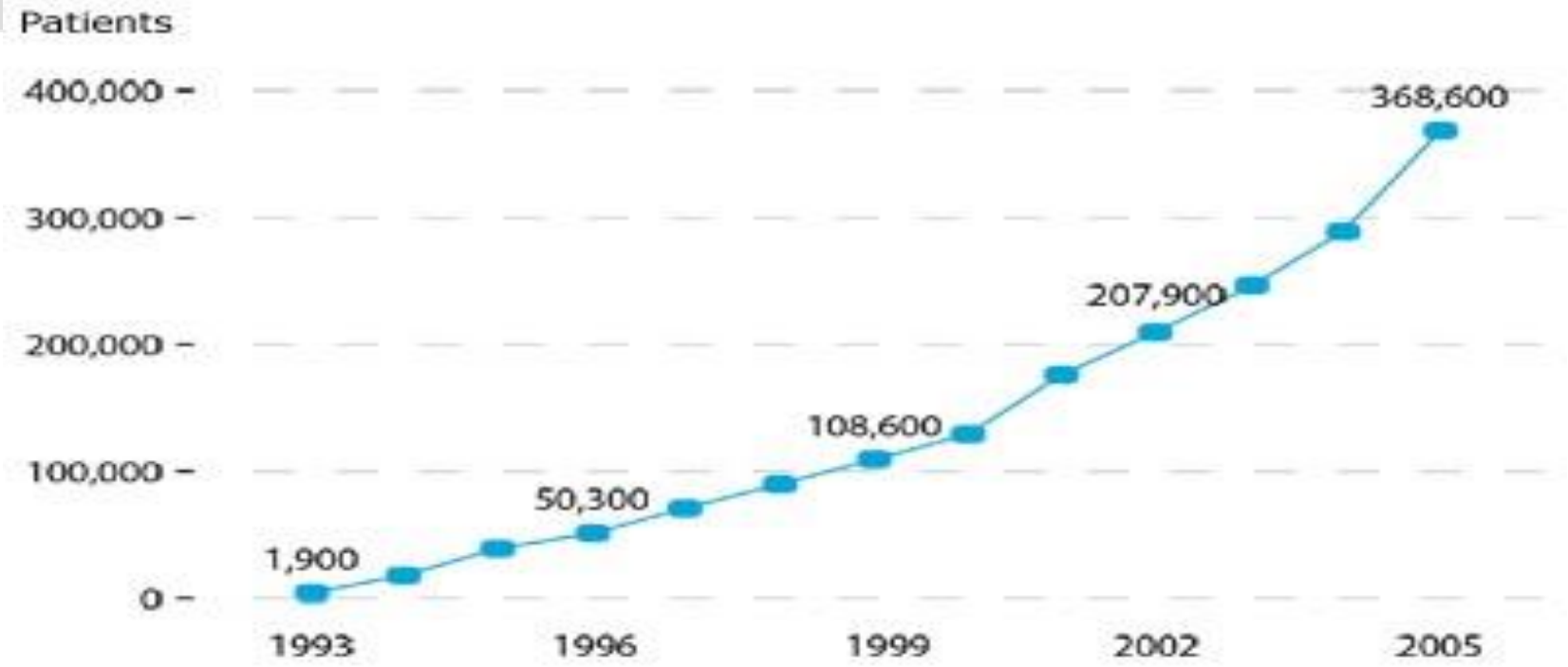
Staphylococcus aureus

- **Facultative anaerobe, Gram-positive**
- **Clinical presentation:**
  - **skin infections, abcess**
  - **pneumonia**
  - **meningitis**
  - **persistent infections**
  - **artificial joints, bone**



J. Craig Venter™ INSTITUTE

The Concern Over MRSA

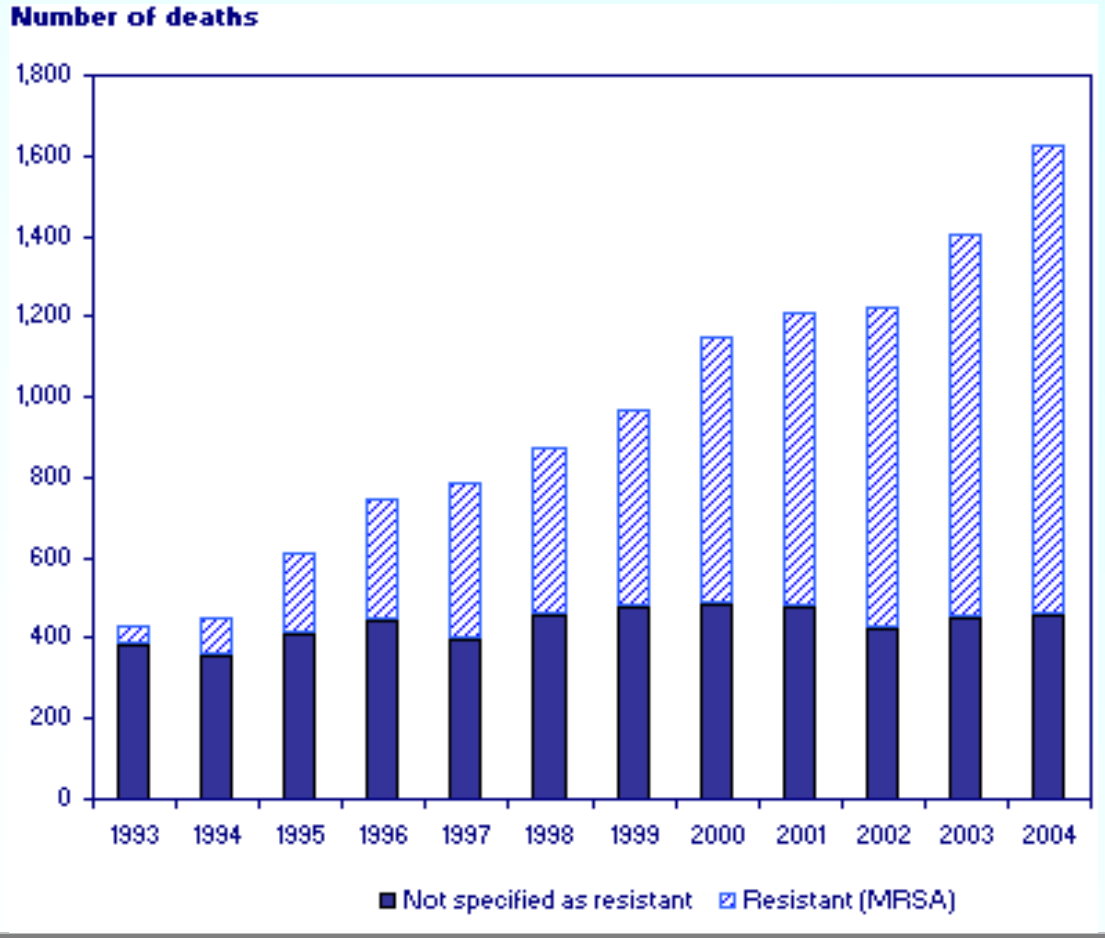


Hospital MRSA infections in the USA

Adapted from "Hospital stays with MRSA infections 1993-2005
Source: AHRQ, Center for Delivery, Organization and Markets,
Healthcare Cost and Utilization Project,
Nationwide Inpatient Sample, 1993-2005"

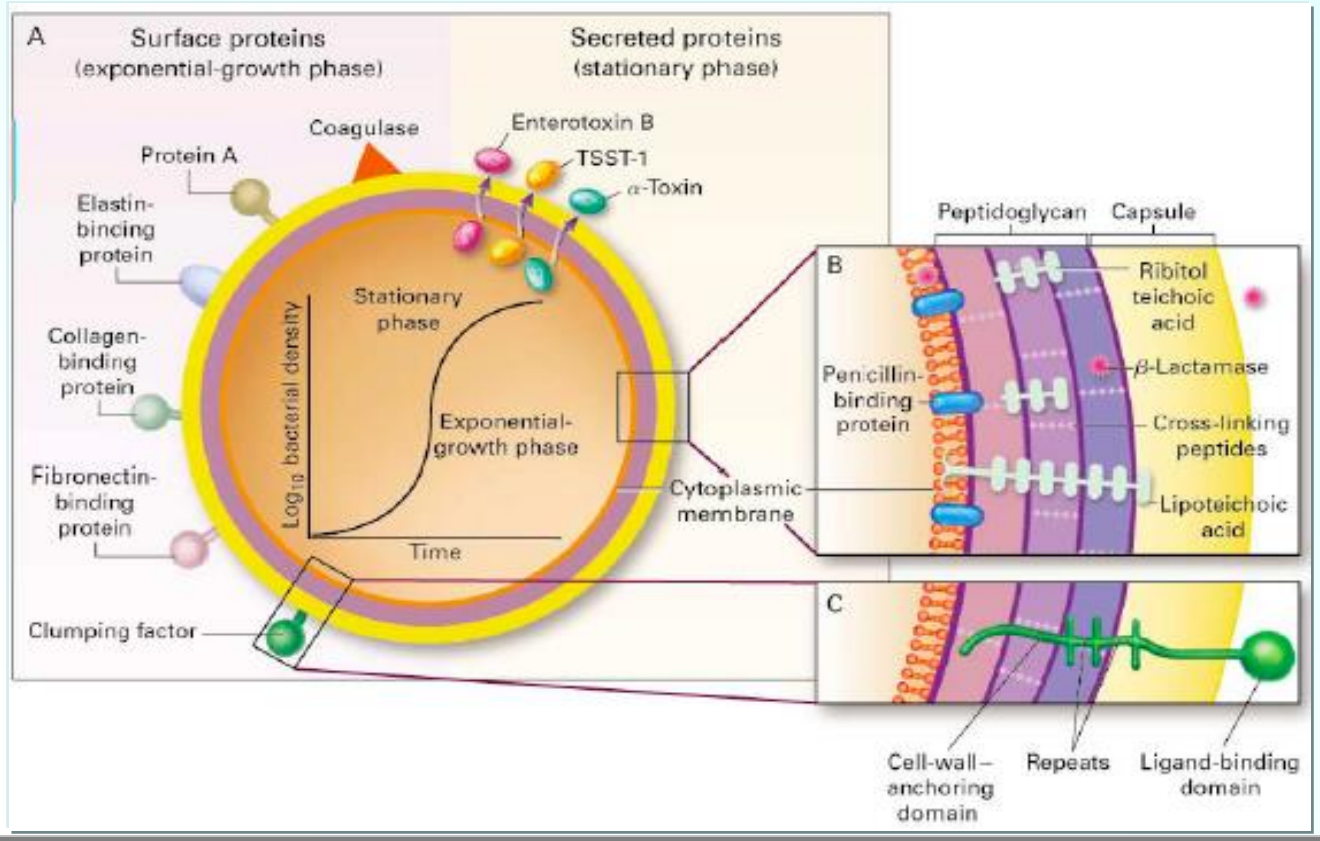
J. Craig Venter™
INSTITUTE

S. aureus Impact on Health Care



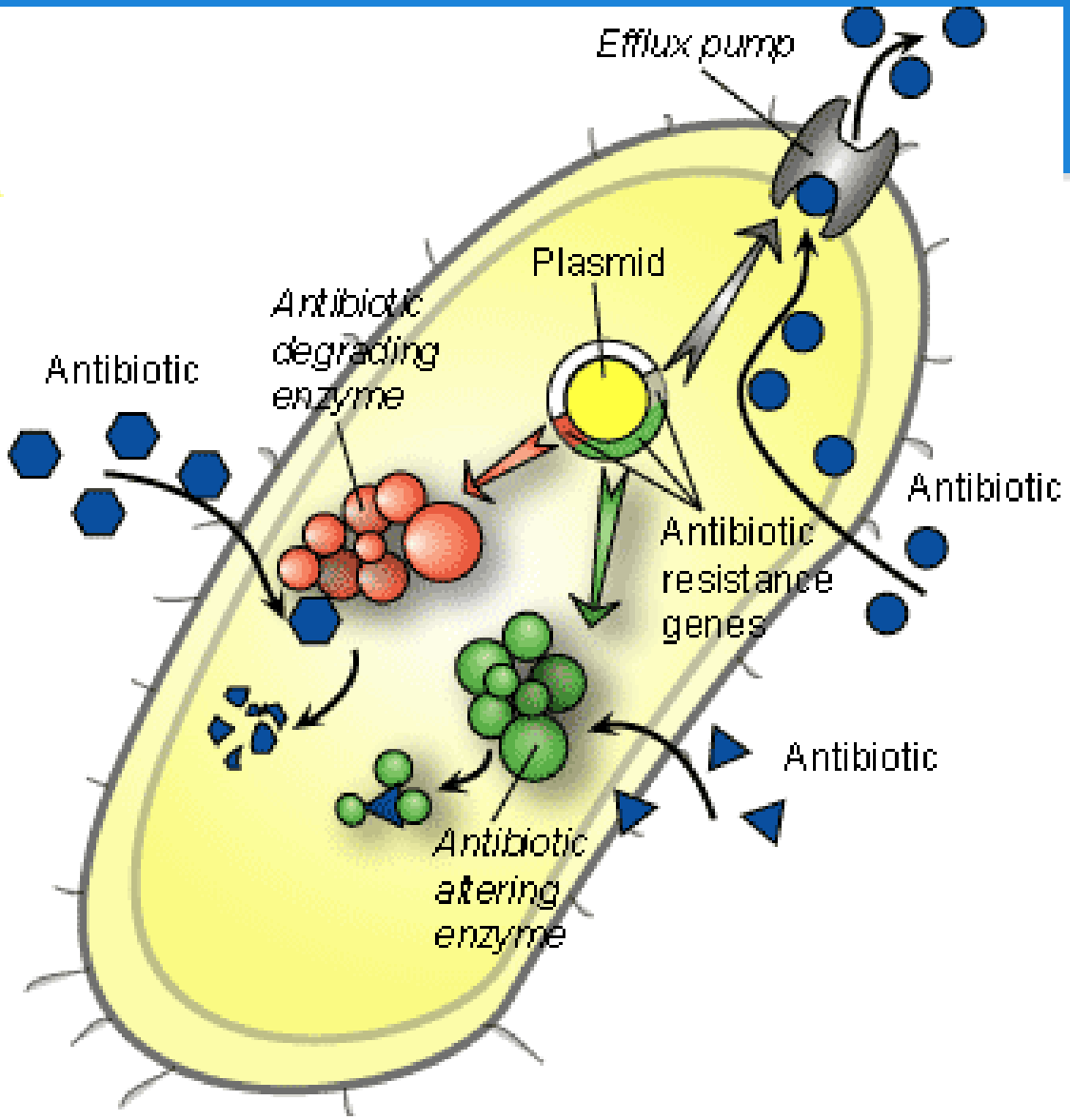
http://www.destiny-pharma.demon.co.uk/images/MRSA%20Deaths%202004.gif

Virulence Factors



http://cmgm.stanford.edu/biochem/biochem230/discussions2005/AntibioticsDiscussion.pdf

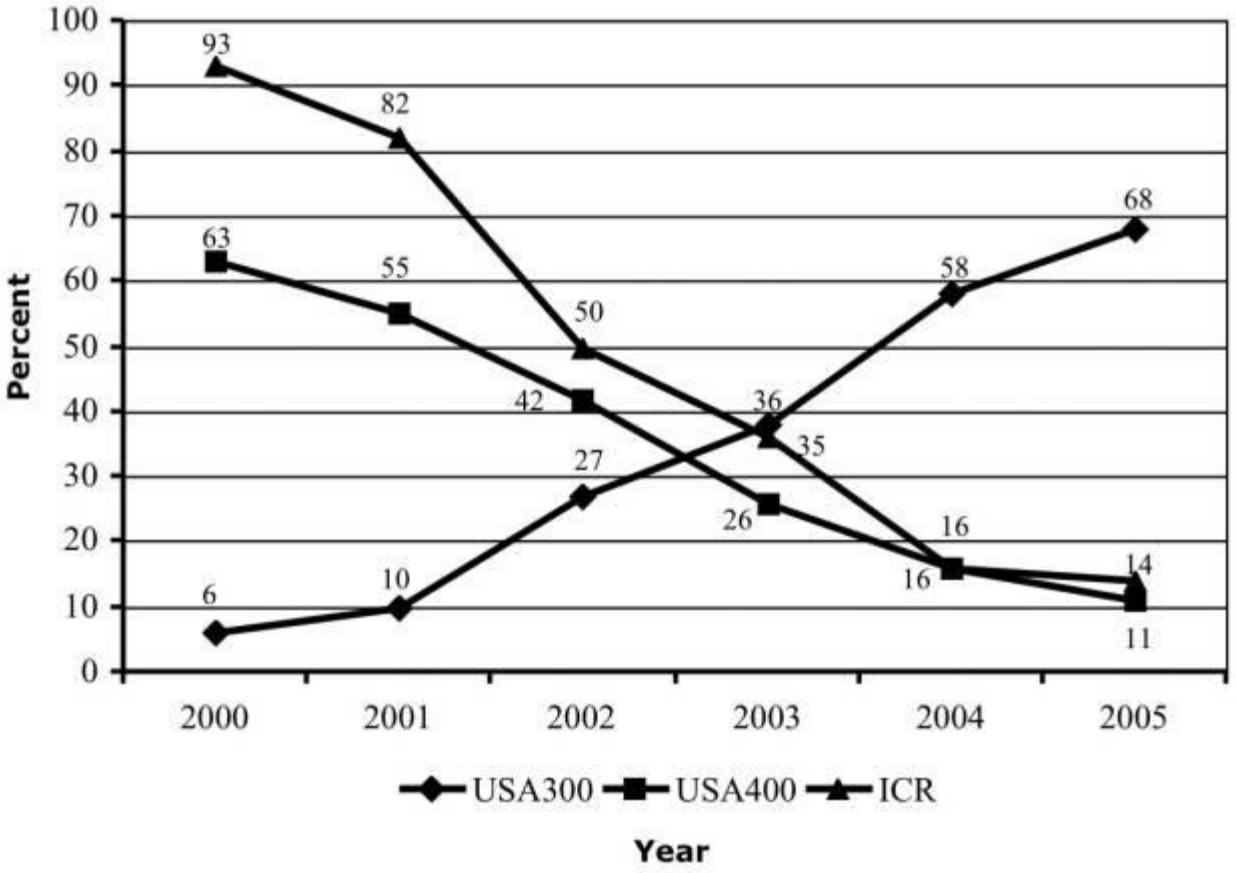
J. Craig Venter™ INSTITUTE



http://**www.bioteach.ubc.ca/Biodiversity/AttackOfTheSuperbugs/ResistanceMechanisms.gif**

J. Craig Venter™ INSTITUTE

S. aureus USA300 replacing USA400



Public Health Rep. 2009 May-Jun; 124(3): 427–435

J. Craig Venter INSTITUTE

Experimental Goal

- Understand the host-pathogen interaction during *S. aureus* infection.
- Identify host pathways differentially expressed during infection.
- Determine host tissue specific expression.
- Determine pathogen differential expression.

J. Craig Venter™ INSTITUTE

Steps in analysis

- Clustering to find expression trends
- Statistical analysis of significant genes
- Functional analysis of differentially expressed genes
- Visual representation of differentially expressed genes

Clustering

One goal may be to identify genes which have “similar” patterns of expression (i.e. similar expression vectors).

“Clustering Algorithms” are a popular method for doing this.

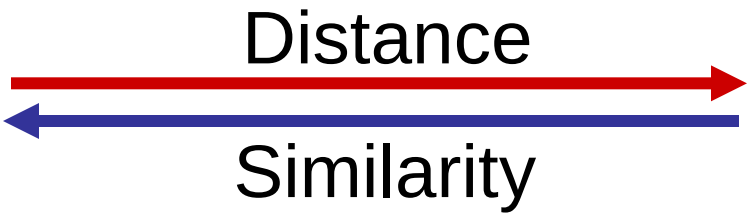
Some Clustering Algorithm Types:

* Agglomerative: Hierarchical Trees
* Divisive: *k*-means, Self-Organizing Maps
* Nonclustering: Principal Component Analysis

Now we just need to decide what it means to be “similar”…

Distance and Similarity

The ability to calculate a *distance* (or *similarity* - its inverse) between two expression vectors is fundamental to clustering algorithms.



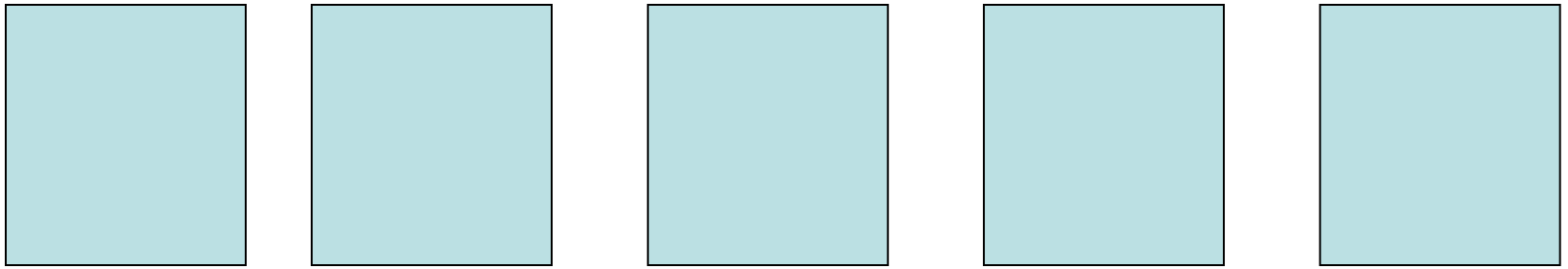
Distance between vectors is the basis upon which decisions are made when grouping similar patterns of expression.

Selection of a *distance metric* defines the concept of distance for a particular experiment.

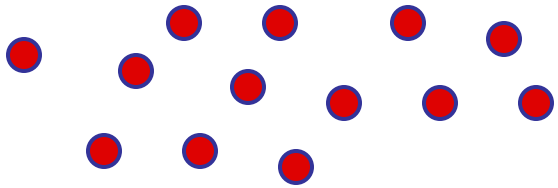
k-Means Clustering (KMC)

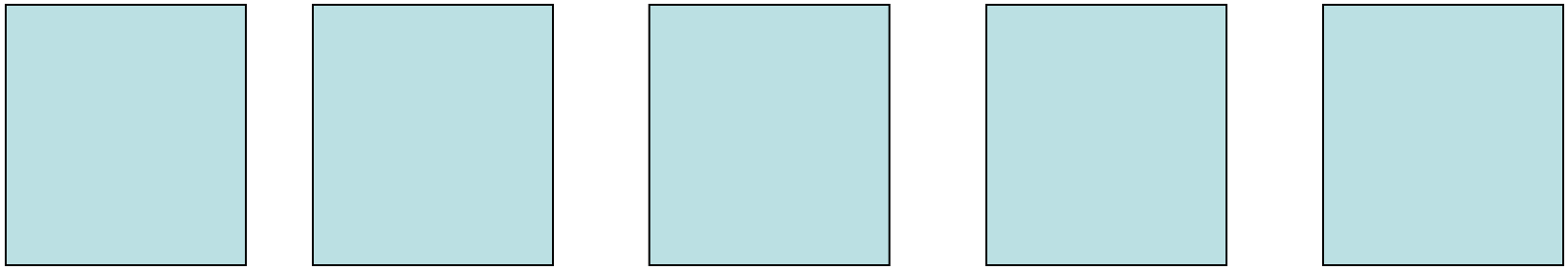
K-Means Clustering (KMC)

1. Specify number of clusters, e.g., 5.



2. Randomly assign genes to clusters.

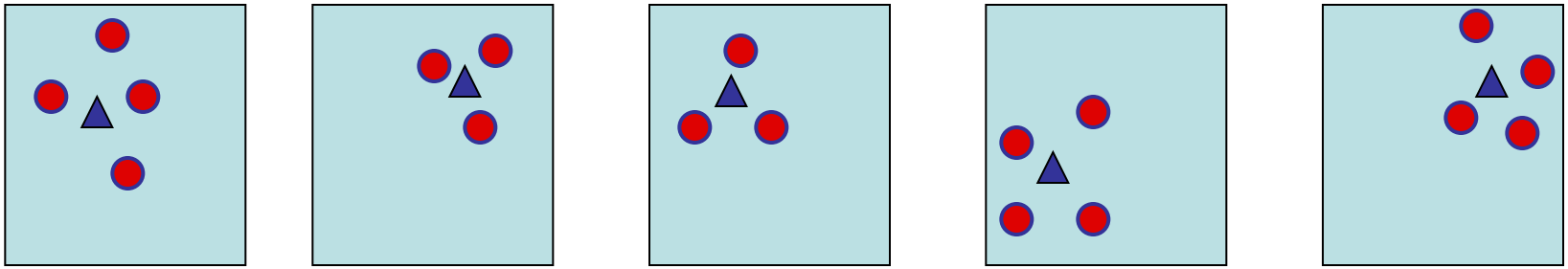




KMC, continued

3. Calculate mean / median expression profile of each cluster.

4. Select a gene and move it to the cluster having the closest mean profile.



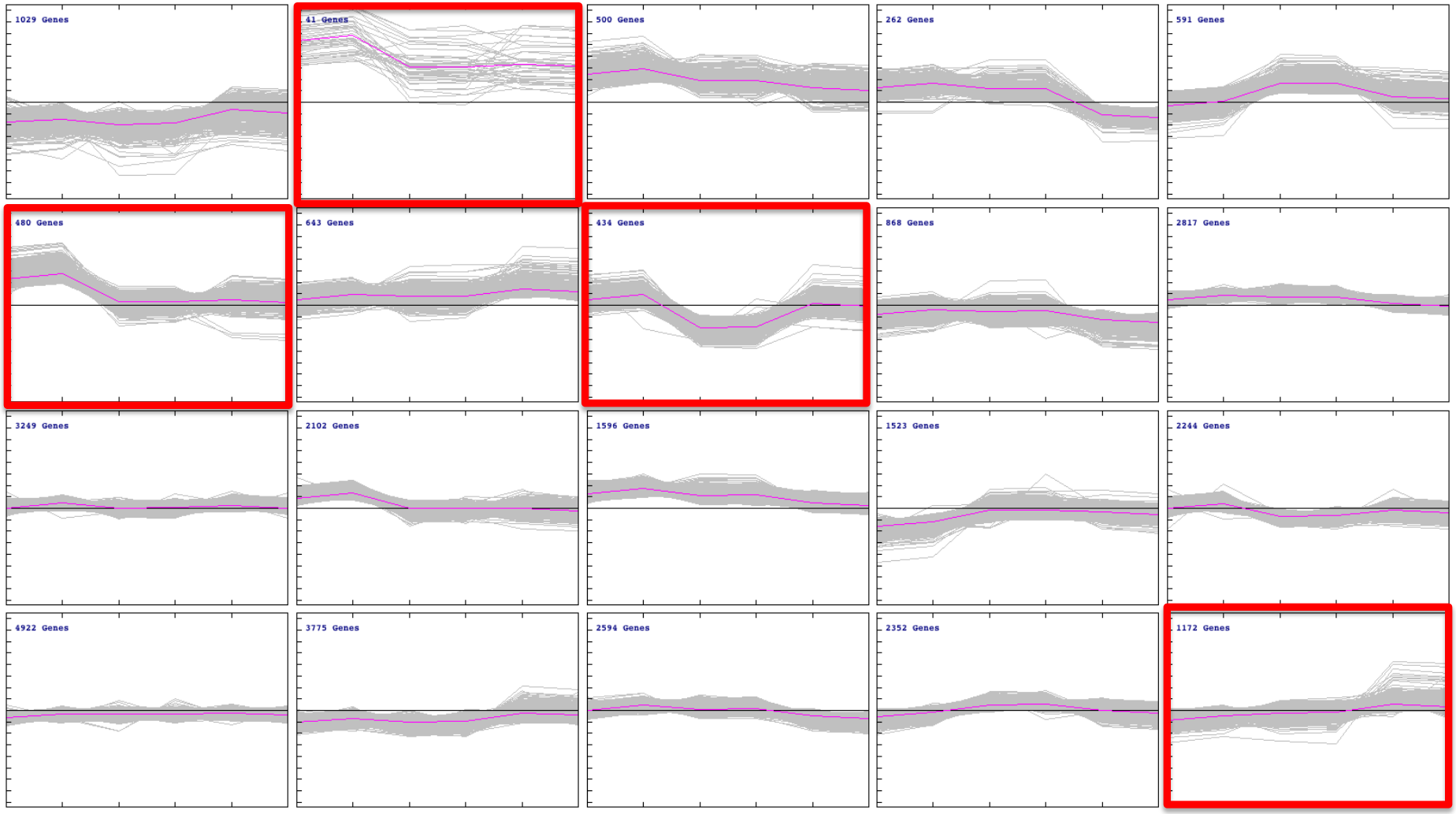
5. If the gene is shifted to a new cluster, recalculate means for the winning and losing clusters.

6. Repeat steps 4 and 5 until genes cannot be shuffled around any more, OR a user-specified number of iterations has been reached.

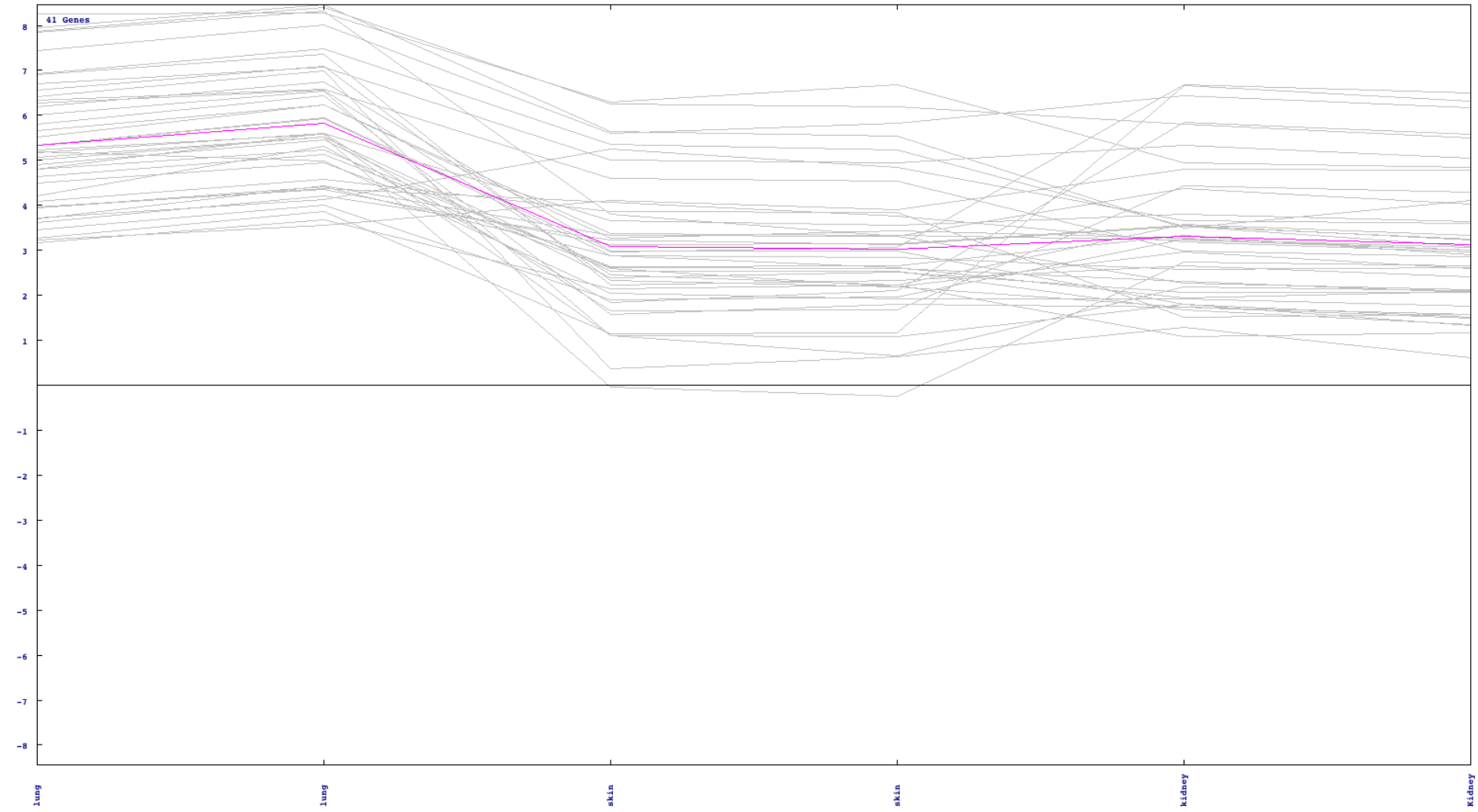
Hands On

- Cluster sample data by KMC analysis

KMC cluster analysis



J. Craig Venter™ INSTITUTE

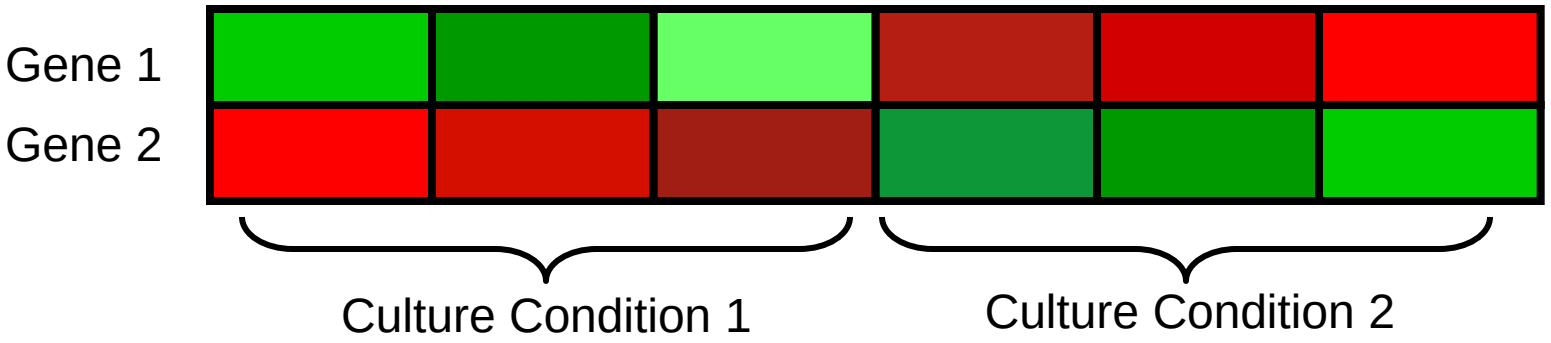


J. Craig Venter™
INSTITUTE

Statistical Methods

Statistical Methods - Overview

- Statistical tests can often be applied to data sets where replicate samples fall into distinct experimental groups.



- Stat tests can be used to find genes that are differentially expressed in accordance with the various conditions under study.
- Unlike general clustering, these tests can provide measures of confidence when reporting genes that are differentially expressed across experimental conditions.

Finding Significant Genes

- **Average Fold Change Difference for each gene suffers from being arbitrary and not taking into account systematic variation in the data**

Finding Significant Genes

- Assume we will compare two conditions with multiple replicate hybs for each condition
- Our goal is to find genes that have significantly different mean expression between these conditions
- These are the genes that we will use for later data mining such as biological role analysis

MeV's Analysis Modules

Hands On
and
Demonstration

J. Craig Venter™
INSTITUTE

Matching Methods to Designs

Experimental designs help to dictate which methods are appropriate to apply.

- One Sample T-test – Finds genes within a single experimental group that are over or under expressed relative to the reference sample.
- Two Sample T-test – Finds genes that have mean expression values that are different between two experimental groups.

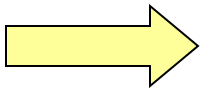
Matching Methods to Designs

- **ANOVA** – similar to t-test but used for the analysis of multiple experimental groups

  condition 1, condition 2, ...condition n

- **Two Factor ANOVA** – useful for the analysis of multiple experimental groups where two experimental factors are under study.

2x3 Design



	condition 1	*condition 2*	*condition 3*
strain 1			
strain 2			

Matching Methods to Designs

- SAM – Significance Analysis of Microarrays provides options to cover many experimental designs.

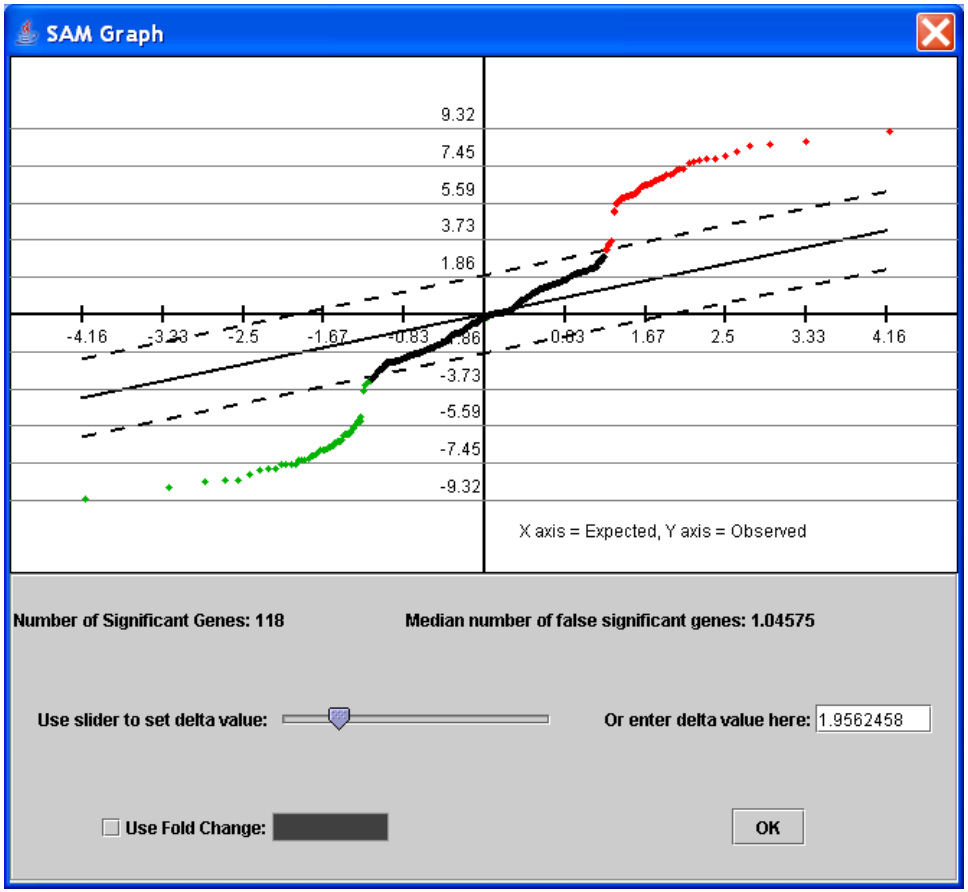
- One Class – Single experimental group of replicate hybs
- Two Class – Two experimental groups
- Two Class Paired – Two experimental groups where related samples across groups can be paired
- Multi-class – Multiple experimental groups

SAM

- Developed by Tusher and Tibshirani at Stanford to address the problems of multiple tests on type I error, false positives.
- Uses an estimated FDR as a the criterion for significance.
- Interactive means of selecting gene lists while monitoring FDR
- Provides testing modes that cover many common experimental designs.

J. Craig Venter™ INSTITUTE

SAM Graph



-plots expected vs. observed d-scores.

-slider alters the delta value (obs – exp, dashed lines), from the diagonal line (obs = exp).

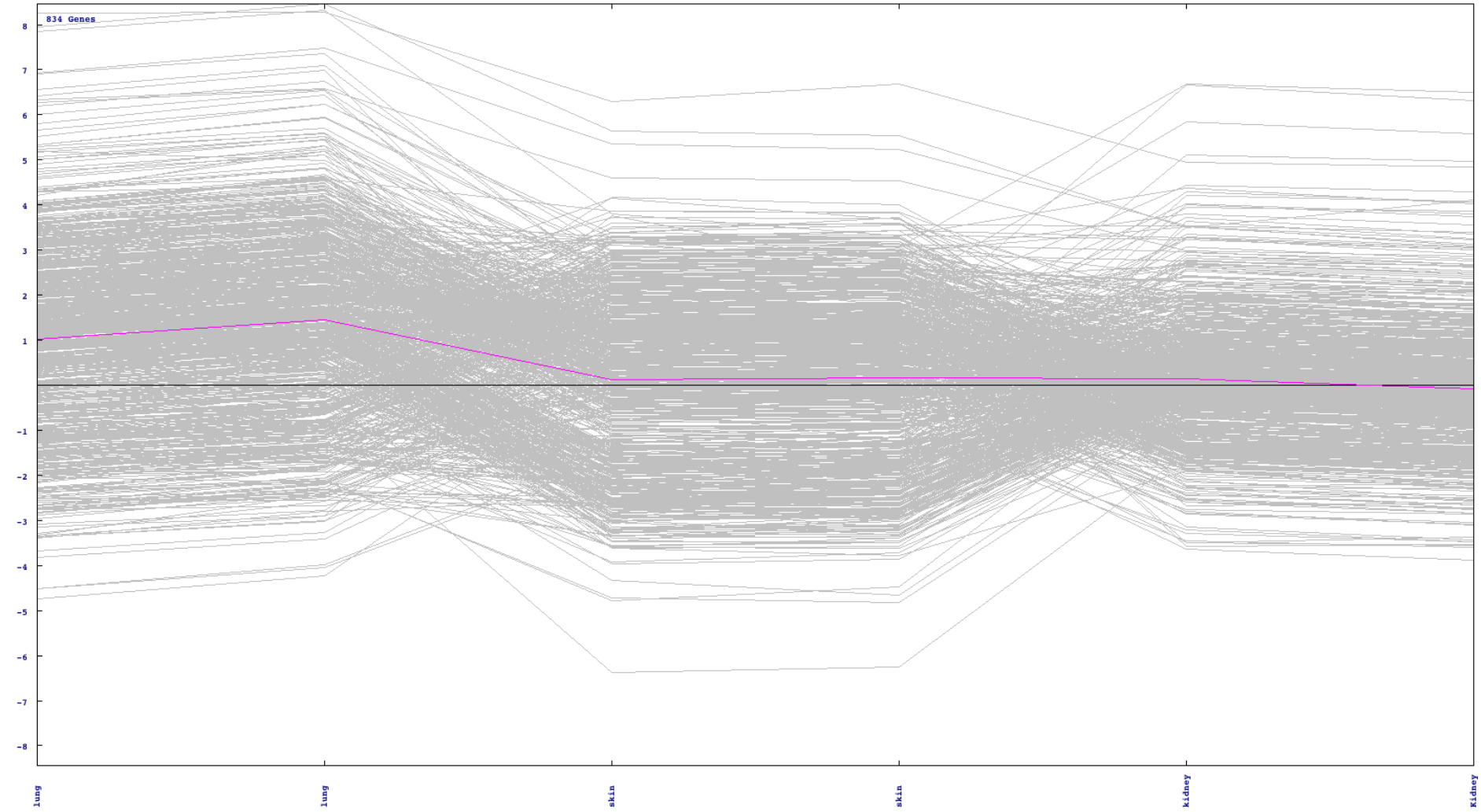
-the number of genes selected (outside the bounds) and the estimated number of false calls is reported as you adjust delta

J. Craig Venter™ INSTITUTE

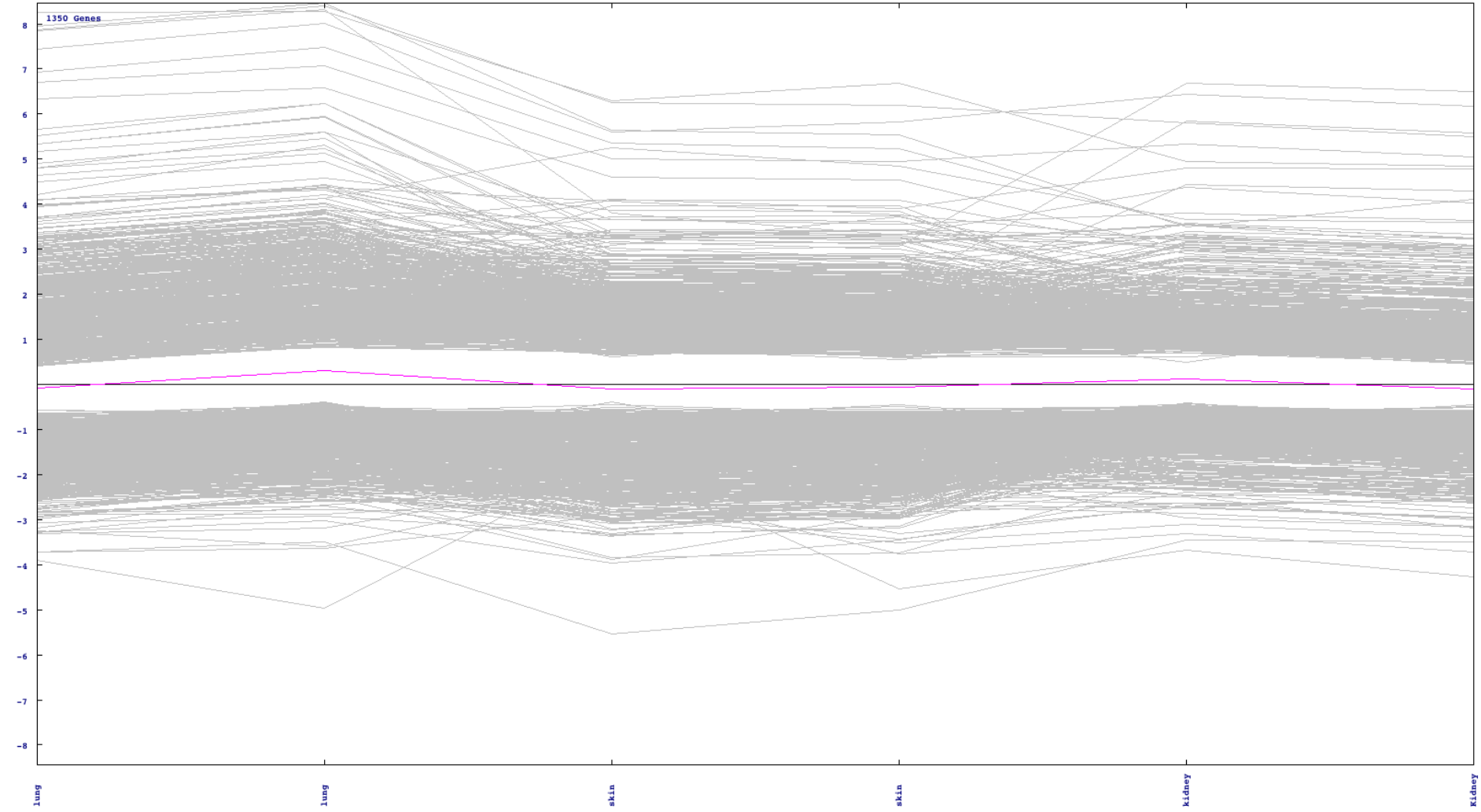
Hands On SAM

J. Craig Venter™
INSTITUTE

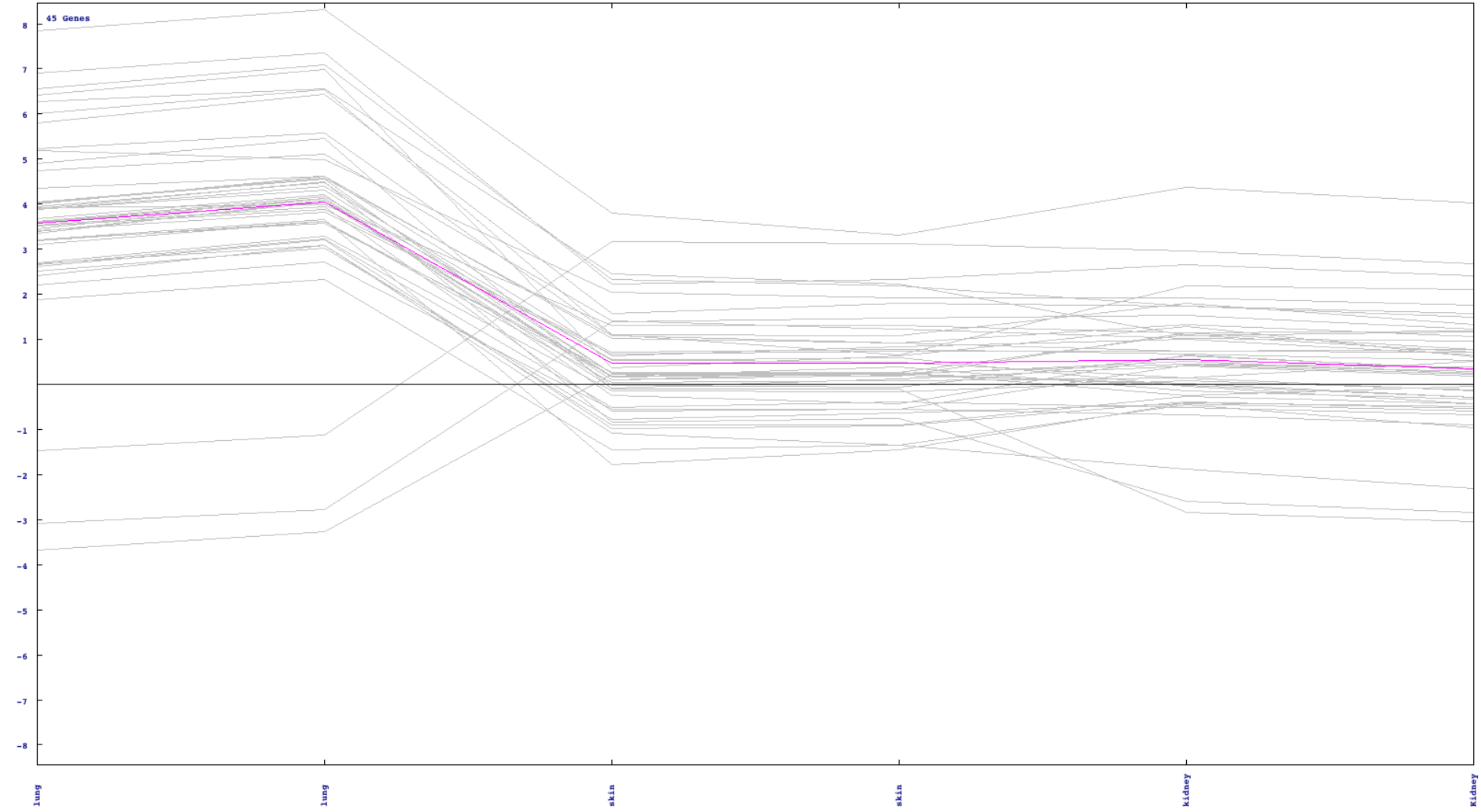
Genes Unique to Infection Models



Genes Shared in Infection Models



Lung Specific Genes



Statistical Significance and Biological Significance

Statistical tests provide lists of genes that show significant changes in expression, however at least two important considerations remain:

What is the magnitude of the expression change? Relatively small changes may be significant due to low variability of the measurements.

Is there a common biological system implicated by multiple genes that show changes in expression?

Expression Analysis Systematic Explorer (EASE)

Exploring Prevalent Biological Roles within Gene Lists

J. Craig Venter™ INSTITUTE

EASE

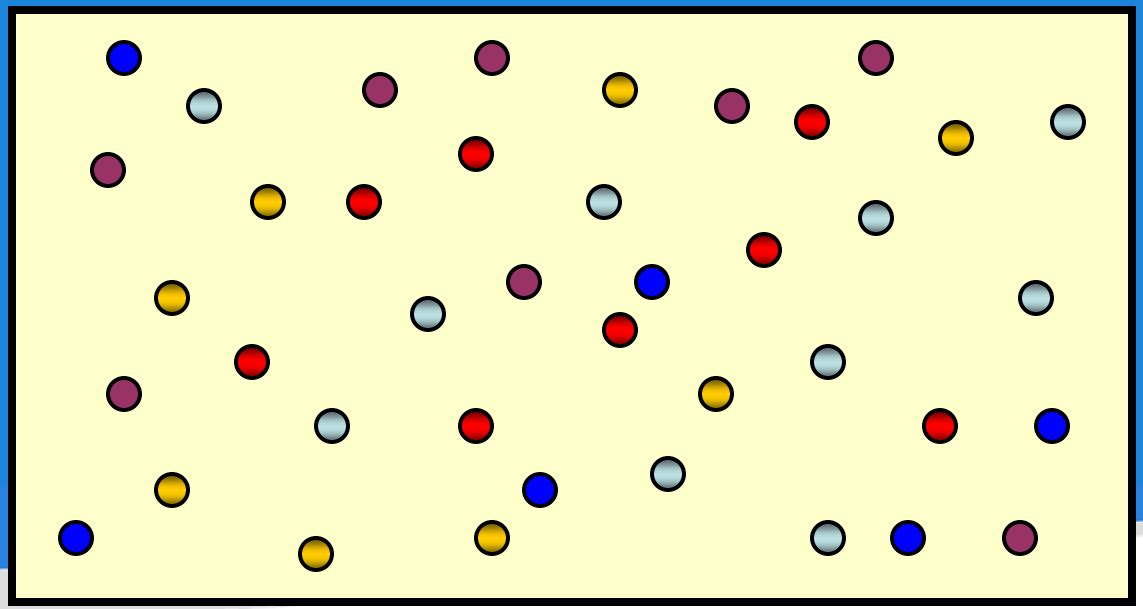
Expression Analysis Systematic Explorer

EASE analysis identifies prevalent biological themes within gene clusters.

The significance of each identified theme is determined by its prevalence in the cluster and in the gene population of genes from which the cluster was created.

Diverse Biological Roles

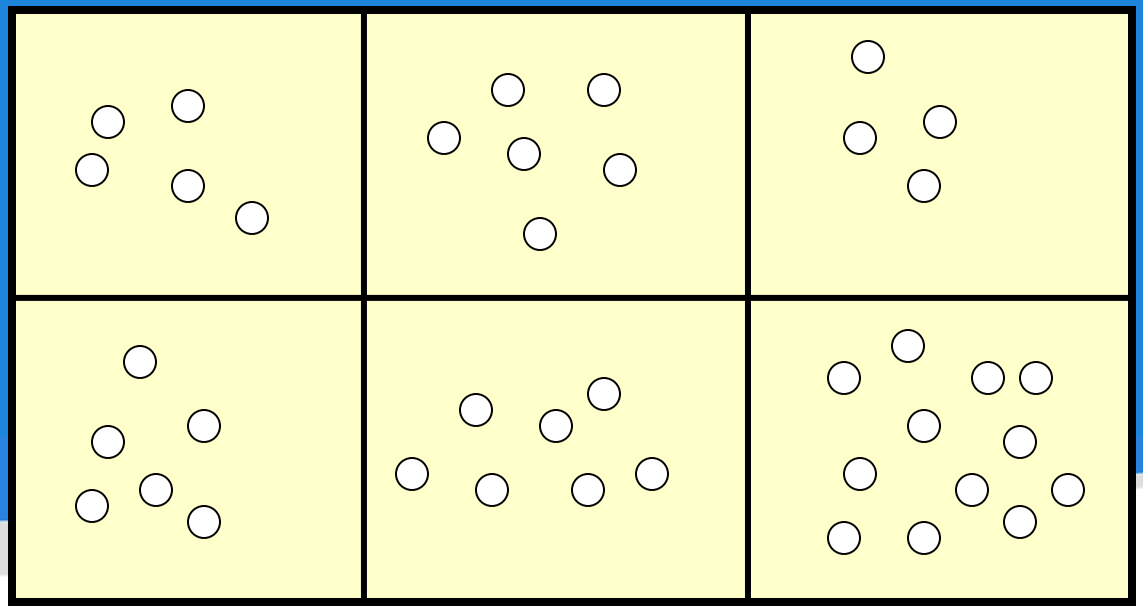
Consider a population of genes representing a diverse set of biological roles or themes shown below as different colors.



J. Craig Venter™ INSTITUTE

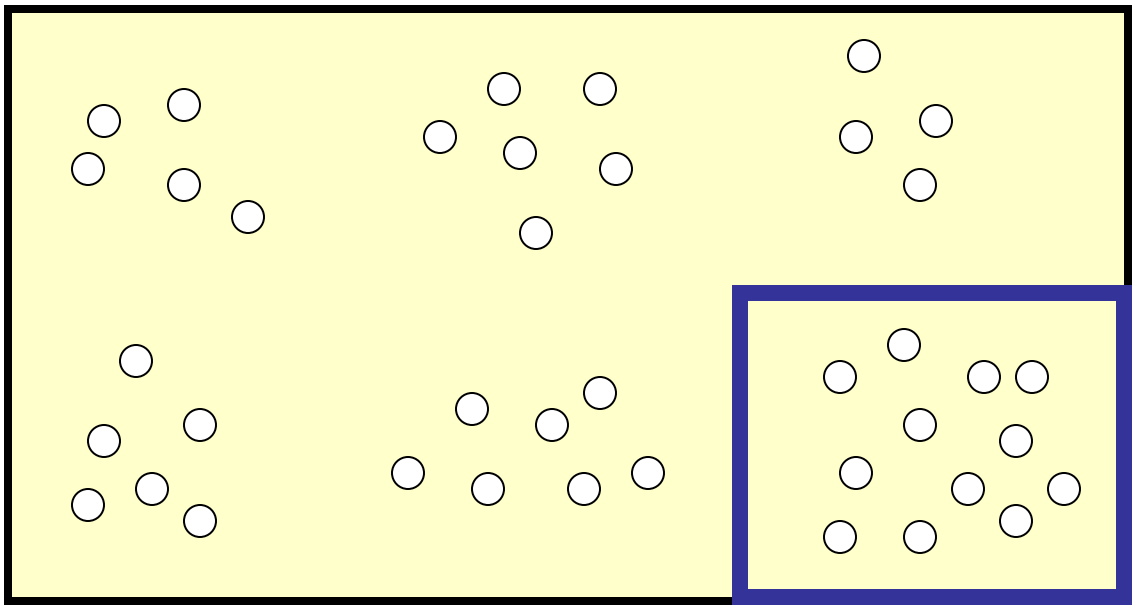
Many algorithms can be applied to expression data to partition genes based on expression profiles over multiple conditions.

Many of these techniques work solely on expression data and disregard biological information.



Consider a particular cluster…

-What are the some of the predominant biological themes represented in the cluster and how should significance be assigned to a discovered biological theme?

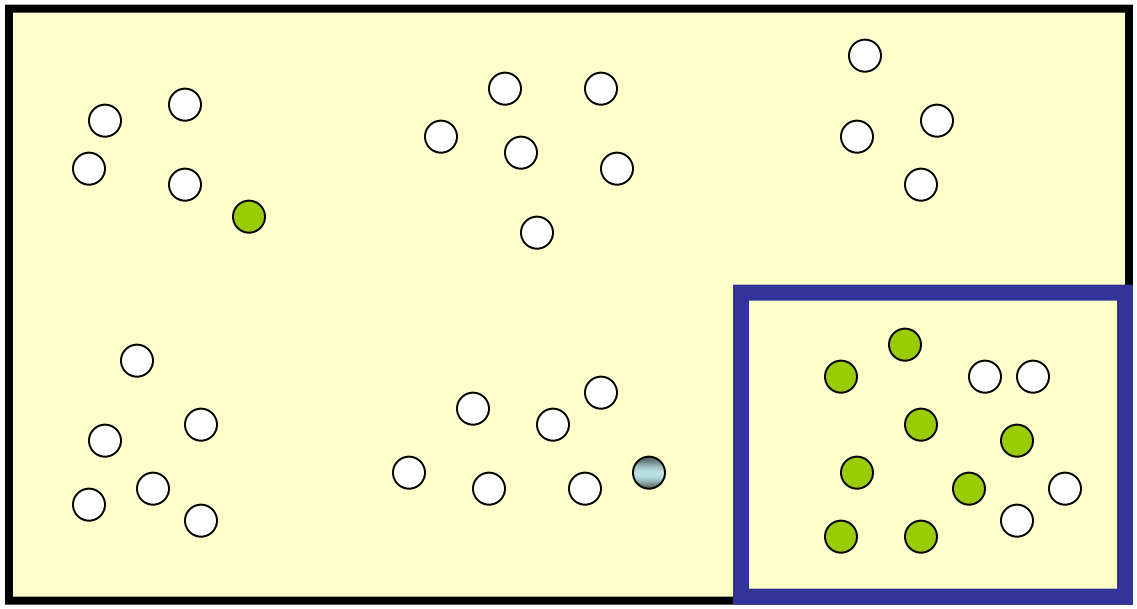


J. Craig Venter™ INSTITUTE

Example:

Population Size: 40 genes
Cluster size: 12 genes

10 genes, shown in green, have a common biological theme and 8 occur within the cluster.



J. Craig Venter™ INSTITUTE

EASE Results

- Consider all of the Results

  EASE reports all themes represented in a cluster and although some themes may not meet statistical significance it may still be important to note that particular biological roles or pathways are represented in the cluster.

- Independently Verify Roles

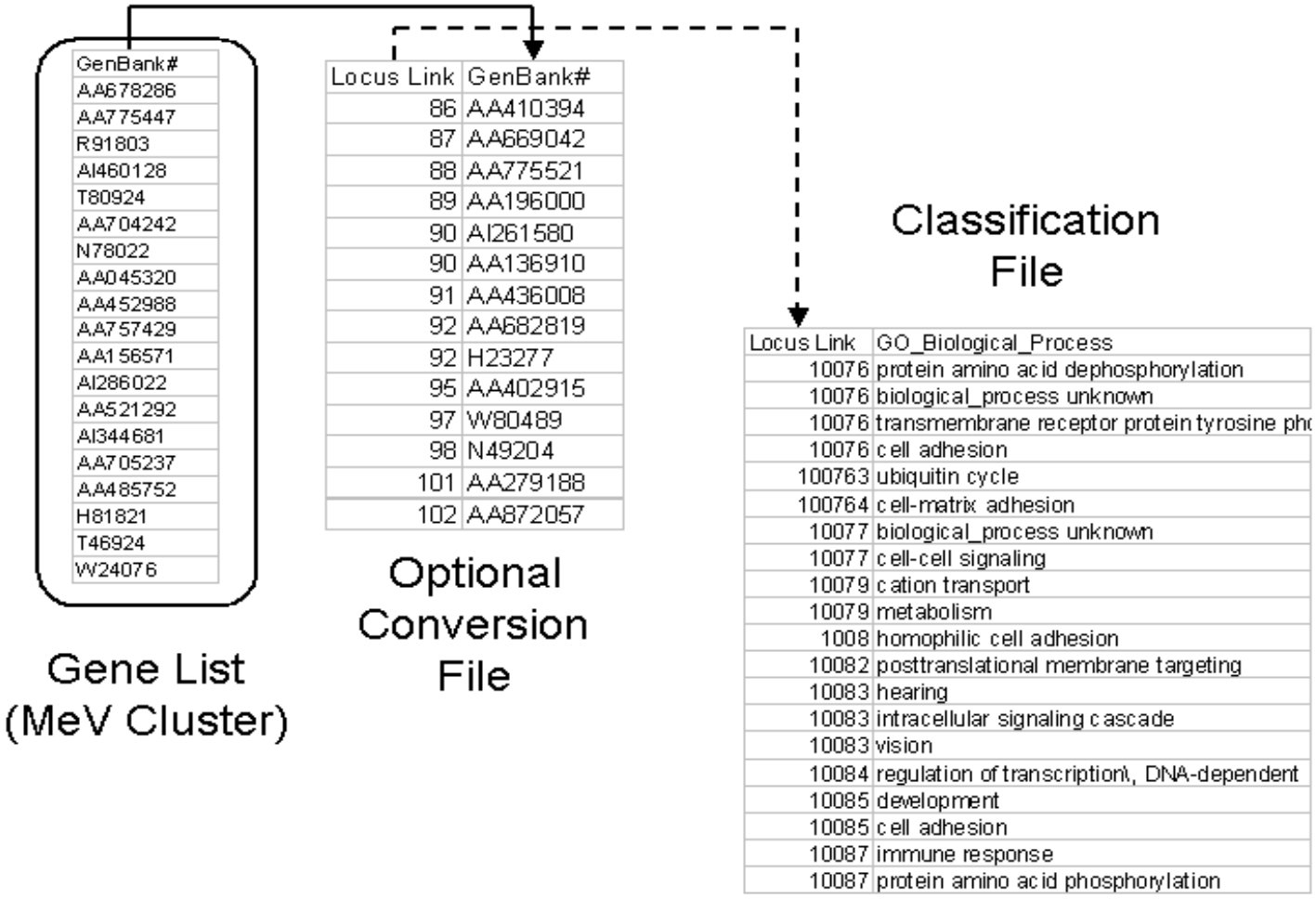
  Once found, biological themes should be independently verified using annotation resources.

Basic EASE Requirements

Annotation keys; identifiers for each gene must be loaded with the data into MeV.

EASE file system; EASE uses a file system to link annotation keys to biological themes.

EASE File System



GenBank#
AA678286
AA775447
R91803
AI460128
T80924
AA704242
N78022
AA045320
AA452988
AA757429
AA156571
AI286022
AA521292
AI344681
AA705237
AA485752
H81821
T46924
W24076

Gene List (MeV Cluster)

Locus Link	GenBank#
86	AA410394
87	AA669042
88	AA775521
89	AA196000
90	AI261580
90	AA136910
91	AA436008
92	AA682819
92	H23277
95	AA402915
97	W80489
98	N49204
101	AA279188
102	AA872057

Optional Conversion File

Classification File

Locus Link	GO_Biological_Process
10076	protein amino acid dephosphorylation
10076	biological_process unknown
10076	transmembrane receptor protein tyrosine pho
10076	cell adhesion
100763	ubiquitin cycle
100764	cell-matrix adhesion
10077	biological_process unknown
10077	cell-cell signaling
10079	cation transport
10079	metabolism
1008	homophilic cell adhesion
10082	posttranslational membrane targeting
10083	hearing
10083	intracellular signaling cascade
10083	vision
10084	regulation of transcription\, DNA-dependent
10085	development
10085	cell adhesion
10087	immune response
10087	protein amino acid phosphorylation

J. Craig Venter™ INSTITUTE

EASE

Expression Analysis Systematic Explorer

Hosack et al. Identifying biological themes within lists of genes with EASE. Genome Biol., 4:R70-R70.8, 2003.

NIAID provided the foundation Java classes upon which the MeV version was built.

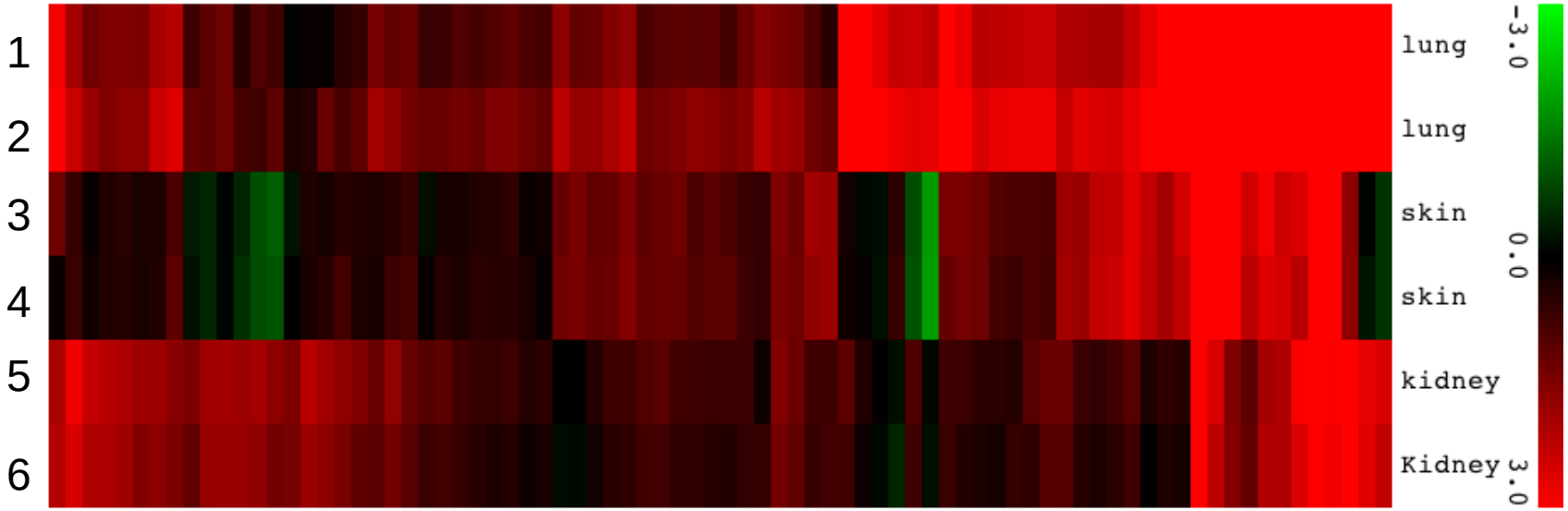
J. Craig Venter™ INSTITUTE

Role Category Breakdown

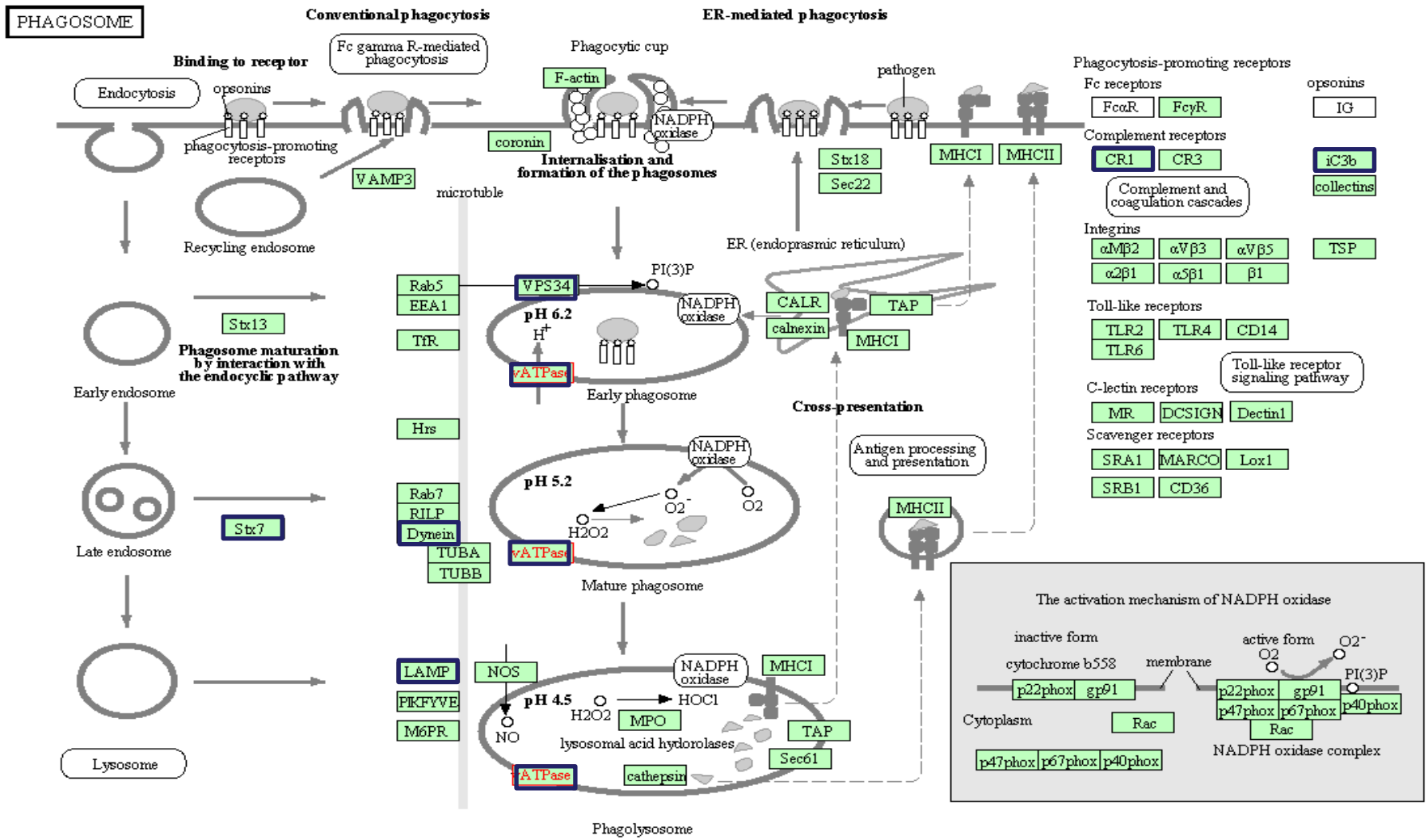
Term	Pop. Hits	Pop. Size
signal transduction	1940	14845
transcription, DNA-dependent	1830	14845
regulation of transcription	1728	14845
biological_process	1695	14845
transport	1683	14845
G-protein coupled receptor protein signaling pathway	1596	14845
sensory perception of smell	1079	14845
multicellular organismal development	895	14845
regulation of transcription, DNA-dependent	785	14845
metabolic process	645	14845
oxidation-reduction process	604	14845
protein phosphorylation	575	14845
ion transport	573	14845
cell differentiation	559	14845
protein transport	508	14845
modification-dependent protein catabolic process	490	14845
cell cycle	479	14845
transmembrane transport	469	14845
cell adhesion	468	14845
apoptosis	426	14845
proteolysis	408	14845
positive regulation of transcription from RNA polymerase II promoter	341	14845
immune response	305	14845
translation	277	14845
mRNA processing	266	14845
cell division	265	14845
intracellular signal transduction	244	14845
lipid metabolic process	234	14845
response to DNA damage stimulus	230	14845
spermatogenesis	216	14845
negative regulation of transcription from RNA polymerase II promoter	215	14845
response to stimulus	213	14845
nervous system development	210	14845
DNA repair	208	14845
RNA splicing	203	14845
small GTPase mediated signal transduction	202	14845
chromatin modification	195	14845
in utero embryonic development	192	14845
mitosis	190	14845
intracellular protein transport	171	14845
carbohydrate metabolic process	167	14845
positive regulation of cell proliferation	163	14845
potassium ion transport	161	14845
vesicle-mediated transport	158	14845
heart development	148	14845
inflammatory response	142	14845

J. Craig Venter™ INSTITUTE

Heat Map



J. Craig Venter™
INSTITUTE



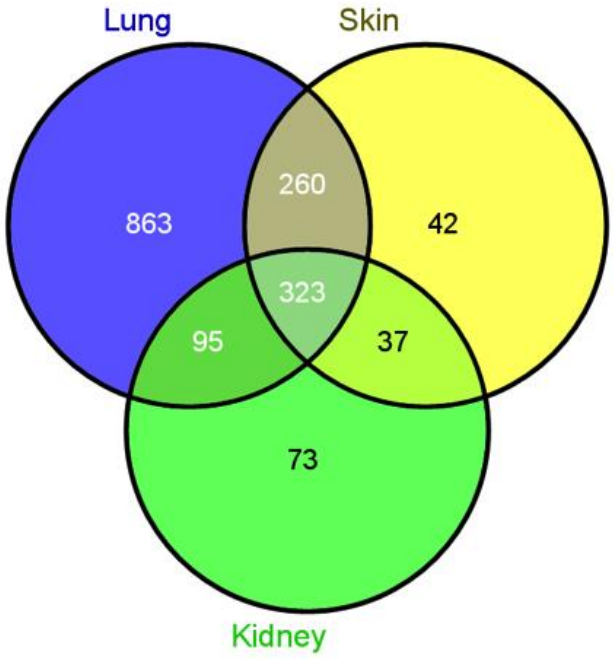
Neutrophil Evasion

Locus	Symbol	USA300_LUNG	USA300_SKIN	USA300_KIDNEY
SAUSA300_0108	sok	-6.30	N/A	-0.23
SAUSA300_0113	SpA	N/A	N/A	-11.43
SAUSA300_0379	ahpF	9.964	N/A	-6.430
SAUSA300_0380	ahpC	2.150	N/A	-2.384
SAUSA300_0398	ssl-3/4	0.67	0.98	-0.96
SAUSA300_0404	ssl-10	-1.79	-6.31	-5.06
SAUSA300_0407	ssl-11	-0.05	-5.08	-7.65
SAUSA300_0883	eap	6.64	-1.51	-2.77
SAUSA300_1052	ecb	-1.65	-4.85	-4.83
SAUSA300_1055	efb	4.85	-6.54	-13.44
SAUSA300_1058	hla	8.49	N/A	-4.10
SAUSA300_1067	PSM beta	N/A	N/A	-6.95
SAUSA300_1255	mprF	6.68	1.56	-7.97
SAUSA300_1381	lukF	7.87	4.10	3.84
SAUSA300_1382	lukS	-4.32	-3.65	-3.40
SAUSA300_1445	scpA	-6.37	-7.04	-10.33
SAUSA300_1768	lukD	6.65	4.26	-5.69
SAUSA300_1769	lukE	1.61	-5.29	-7.09
SAUSA300_1890	Staphopain	-5.63	-5.60	-6.99
SAUSA300_1919	SCIN	14.44	5.25	-5.23
SAUSA300_1920	chips	12.17	9.12	-1.99
SAUSA300_1922	sak	5.28	4.24	-2.05
SAUSA300_1974	lukG	6.97	-2.63	-5.62
SAUSA300_1975	lukH	8.52	1.50	-2.86
SAUSA300_2364	SBI	1.70	4.46	4.16
SAUSA300_2365	hlgA	19.85	5.49	5.46
SAUSA300_2366	hlgB	17.73	0.59	0.89
SAUSA300_2367	hlgC	6.54	-3.38	-3.62
SAUSA300_2504	OatA	4.15	-5.30	-1.63
SAUSA300_2572	aureolysin	-8.64	5.51	-5.39

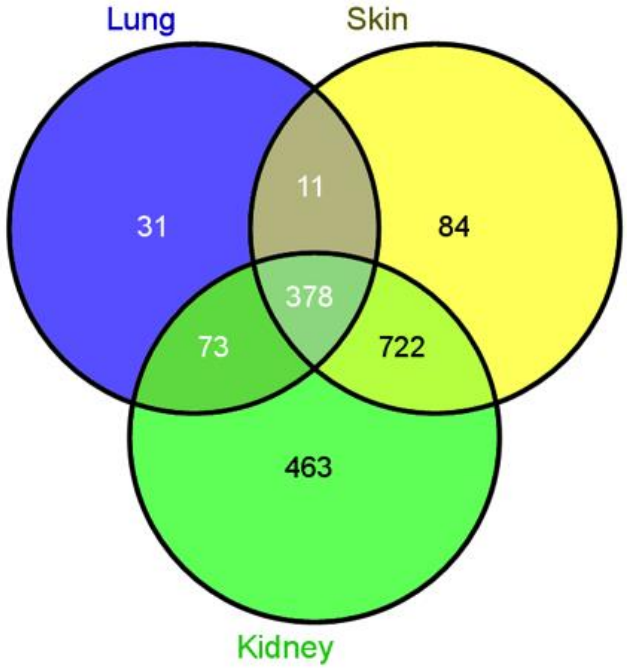
J. Craig Venter™ INSTITUTE

S. aureus shared expression

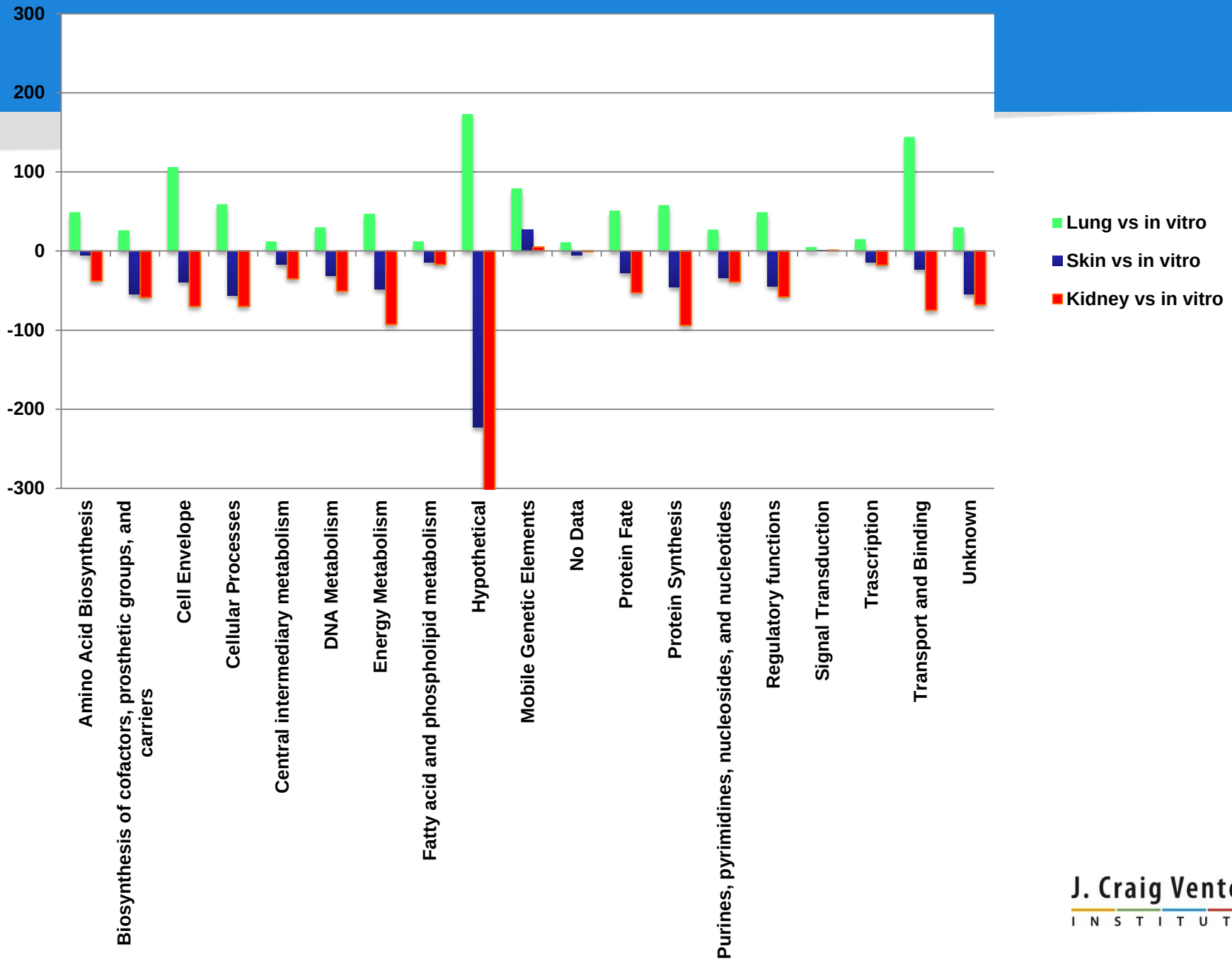
Up regulated



Down regulated

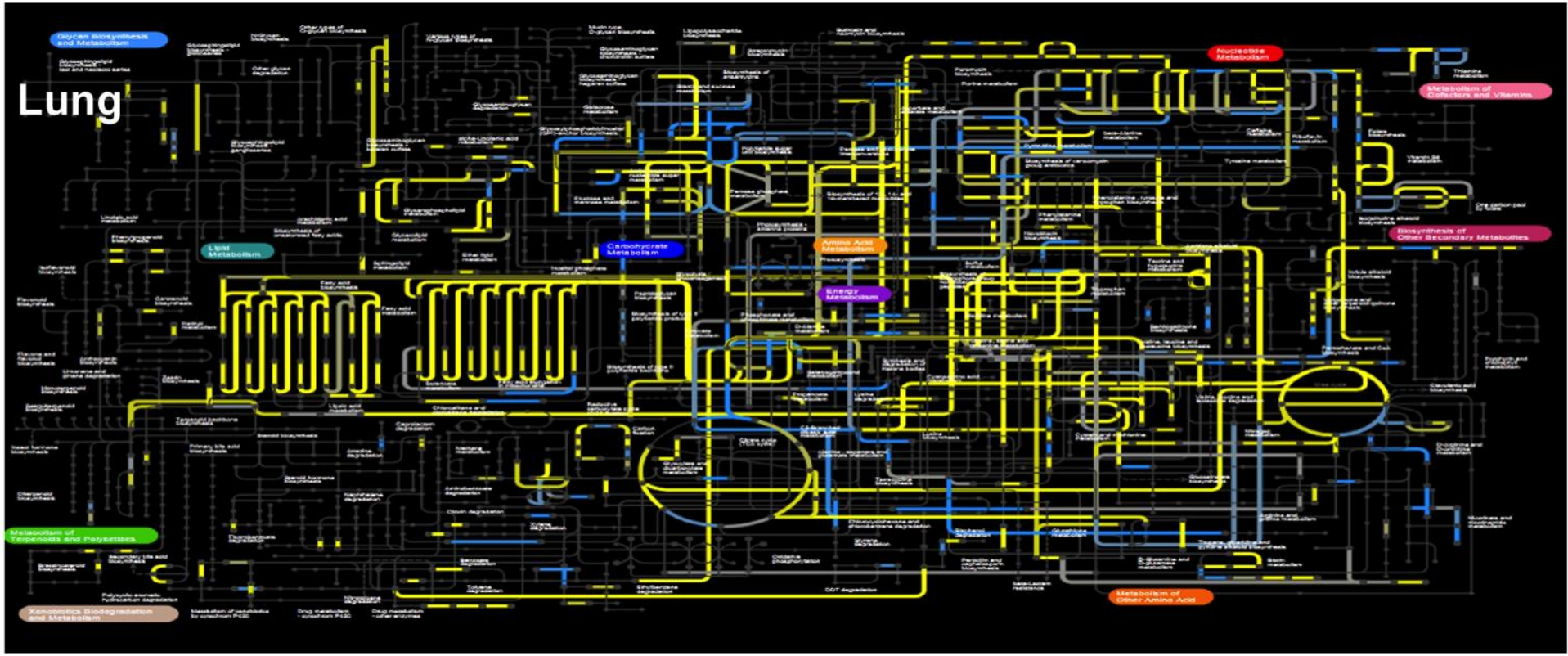


J. Craig Venter™ INSTITUTE



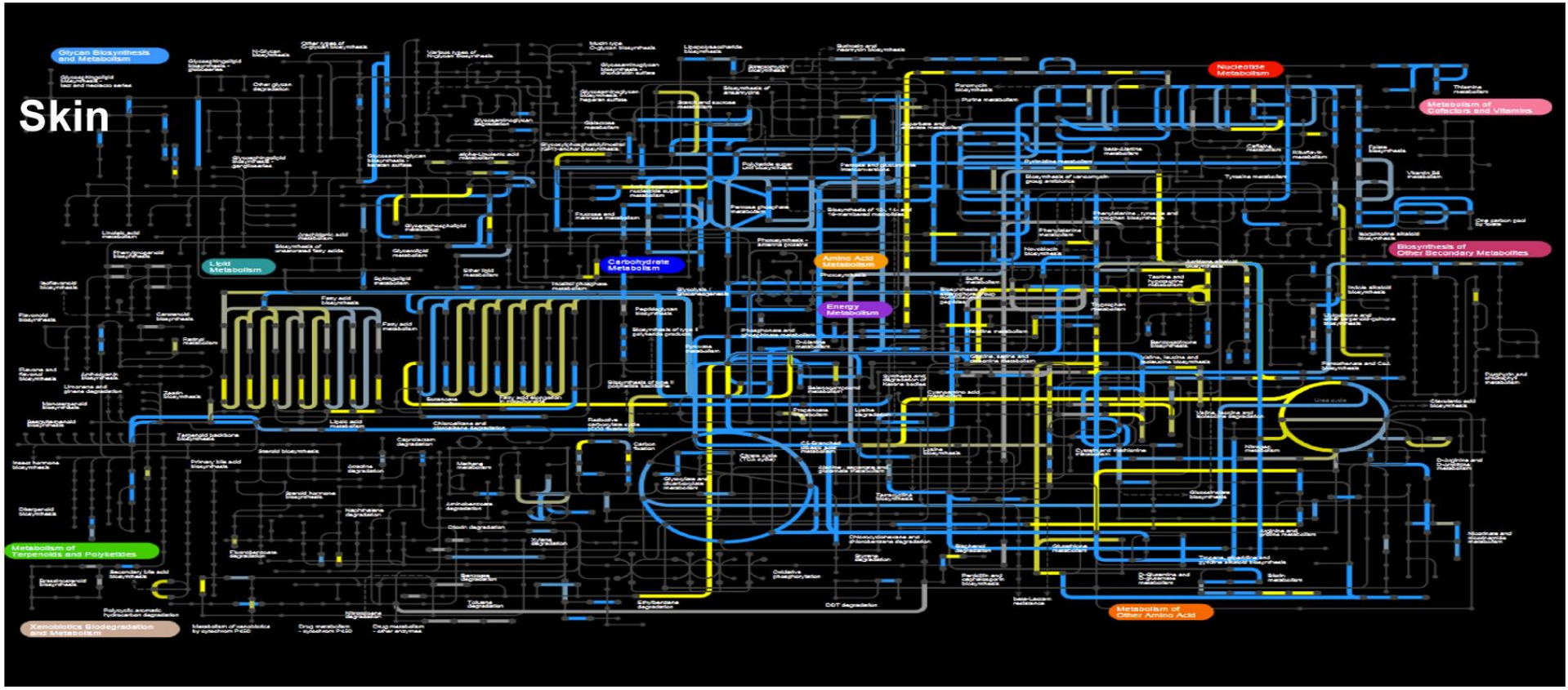
J. Craig Venter INSTITUTE

Metabolic Analysis Lung Model



J. Craig Venter™ INSTITUTE

Metabolic Analysis Skin Model



J. Craig Venter™ INSTITUTE

Linear Expression Maps (LEM)

Expression Organized by Chromosomal Location

J. Craig Venter™ INSTITUTE

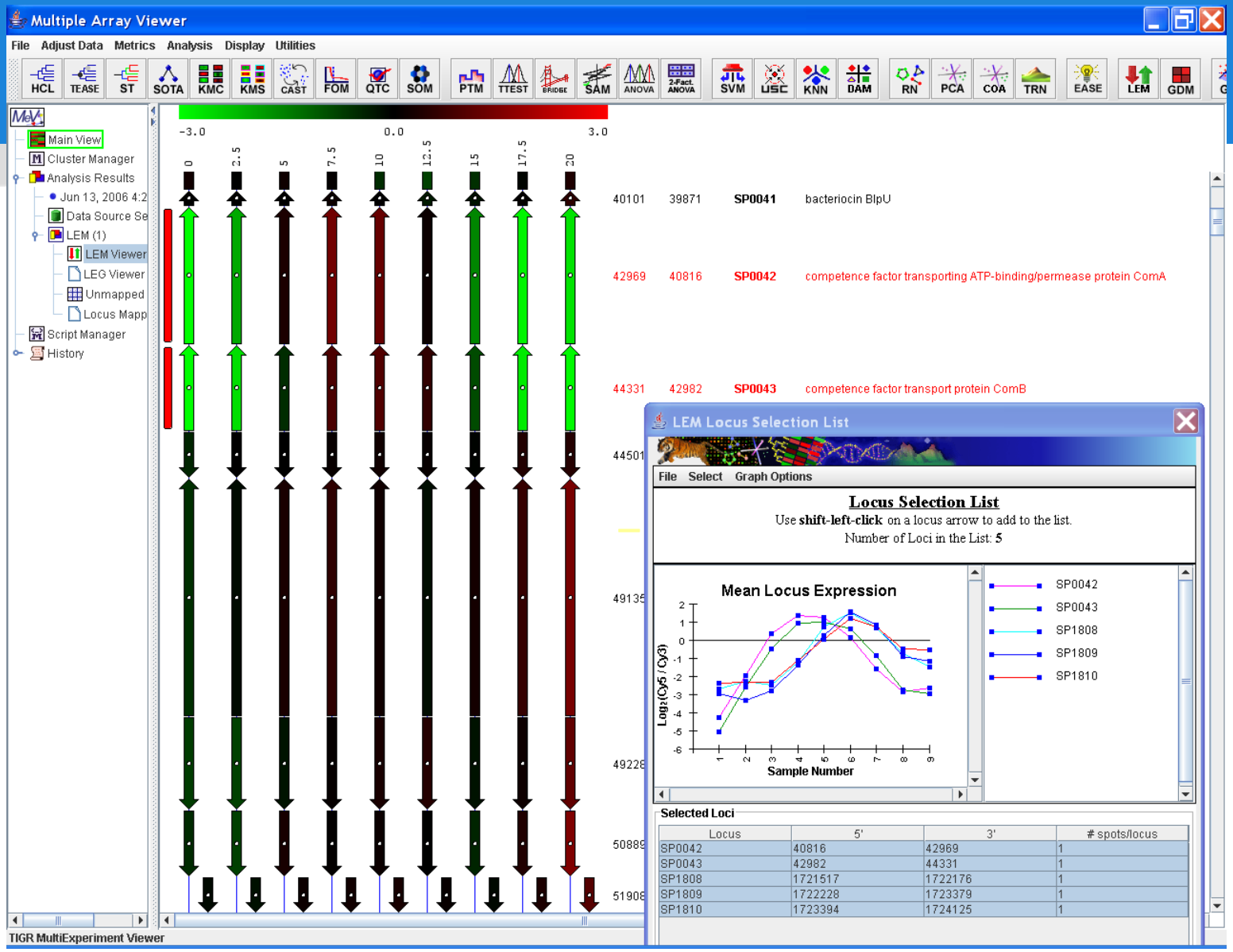
LEM

- Linear Expression Maps organize expression by locus location on the chromosome or plasmid

- LEMs provide a means to navigate over the genome to find contiguous loci displaying similar patterns of expression.

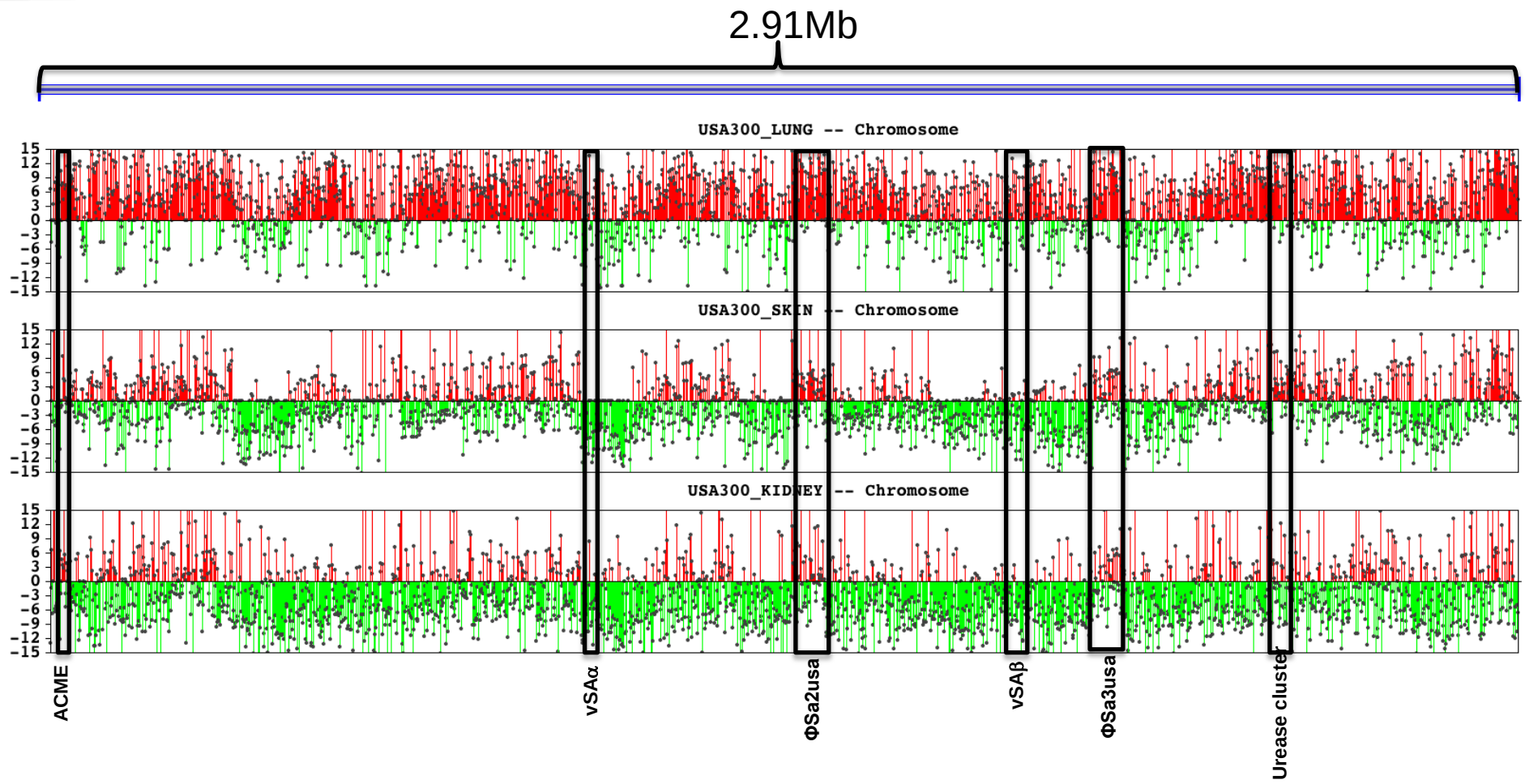
LEM Requirements

- Locus IDs - each gene of interest should have a gene identifier that can be mapped to the genome
- Chromosomal Location – a separate coordinate file or information in the loaded annotation file should provide a chromosome id and location for each locus.



J. Craig Venter™ INSTITUTE

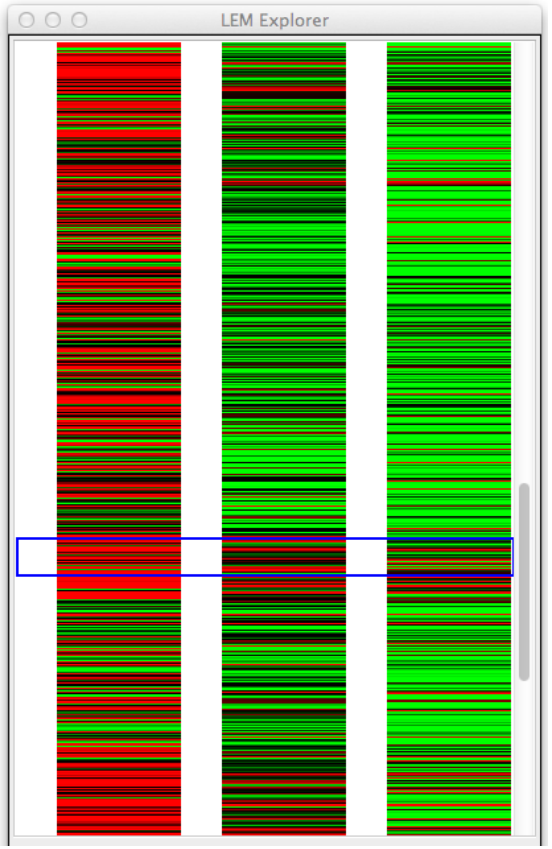
LEG Analysis



J. Craig Venter INSTITUTE

LEM Analysis

2086896	2087345	SAUSA300_1920	chemotaxis-inhibiting protein CHIPS
2087440	2087736	SAUSA300_1921	truncated amidase
2088425	2088916	SAUSA300_1922	staphylokinase precursor
2089107	2089862	SAUSA300_1923	autolysin
2090666	2090962	SAUSA300_1925	phiPVL ORF17-like protein
2091020	2091307	SAUSA300_1926	phi77 ORF044-like protein
2091354	2091506	SAUSA300_1927	phi77 ORF109-like protein
2091496	2095281	SAUSA300_1928	"phi77 ORF002-like protein, phage minor structural protein"
2095297	2096781	SAUSA300_1929	"phi77 ORF004-like protein, putative phage tail component"
2096778	2101307	SAUSA300_1930	"phi77 ORF001-like protein, phage tail tape measure protein"
2101364	2101525	SAUSA300_1931	phi77 ORF100-like protein
2101552	2101902	SAUSA300_1932	hypothetical protein
2101952	2102092	SAUSA300_1933	hypothetical protein
2102218	2102862	SAUSA300_1934	"phi77 ORF020-like protein, phage major tail protein"
2103267	2103671	SAUSA300_1935	phi77 ORF029-like protein
2103668	2104030	SAUSA300_1936	hypothetical protein
2104288	2104572	SAUSA300_1937	phi77 ORF045-like protein
2104592	2105737	SAUSA300_1938	"phi77 ORF006-like protein, putative capsid protein"
2105761	2106498	SAUSA300_1939	"phi77 ORF015-like protein, putative protease"
2106482	2107669	SAUSA300_1940	phage portal protein



J. Craig Venter INSTITUTE

Useful Open Source Tools

- Comprehensive Microbial Resource
  - (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi)
- MeV
  - (http://sourceforge.net/projects/mev-tm4/)
- GLAMM
  - (http://www.microbesonline.org/cgi-bin/glamm)
- MicrobesOnline
  - (http://www.microbesonline.org/)
- KEGG Database
  - (http://www.kegg.jp/)

Acknowledgements

- These projects have been funded with federal funds from the National Institute of Allergy and Infectious Diseases at the National Institutes of Health.
- JCVI Faculty and Staff

J. Craig Venter™ INSTITUTE

SAM d-score

The d-score is analogous to a t-value (in t test) and includes a term representing group difference (r_i) and variance (s_i).

$$d_i = \frac{r_i}{s_i + s_0}$$

r_i is termed the *quantitative response.* It's definition varies depending on the type of test. e.g. for two class unpaired:

$$r_i = \bar{x}_i - \bar{y}_i$$

s_i ; Variance term, def. varies with test type

s_0 ; "exchangeability factor" previously called the "fudge factor"

J. Craig Venter INSTITUTE

d-score for gene i

The key point is that a large d-score represents a large response and tight variance. Perhaps significant…

$$d_i = \frac{r_i}{s_i + s_0}$$

How large is "large"? It's relative….

An "expected" mean d-score is generated for a set of random data permutations. The "observed" d-score is compared to the expected d-score. The larger the difference (observed – expected), the greater the significance of the observed d-score.

J. Craig Venter™ INSTITUTE