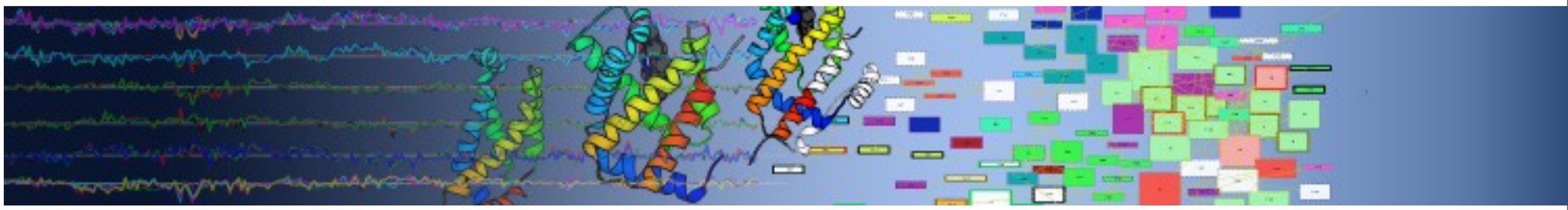
Reconstructing Biological Networks from Data: Part 1 - cMonkey



Richard Bonneau

bonneau@nyu.edu

http://www.cs.nyu.edu/~bonneau/

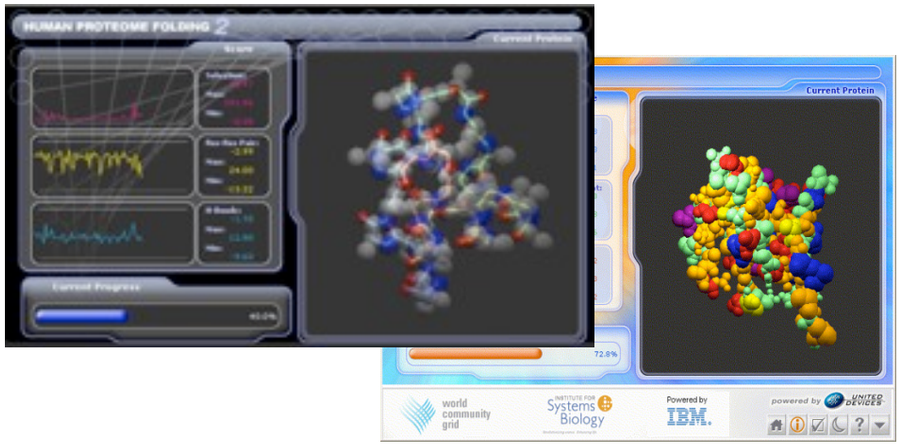
New York University,

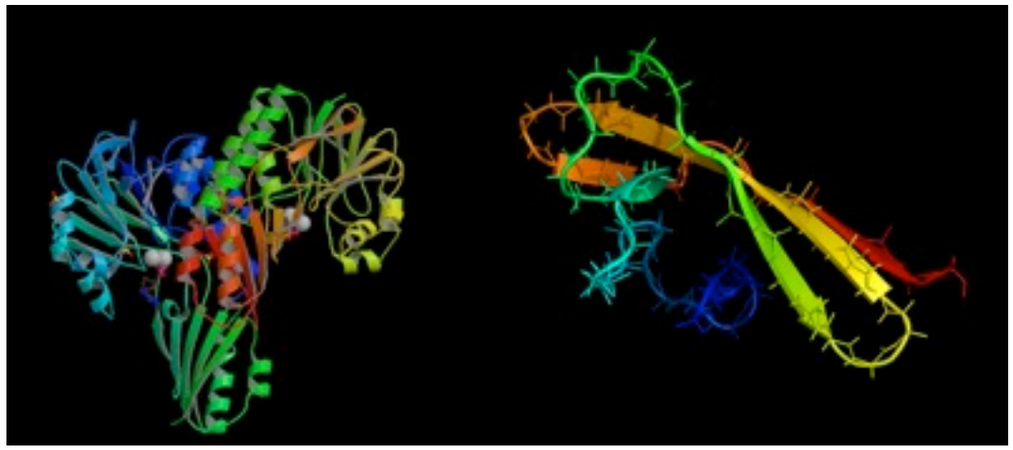
Dept. of Biology &

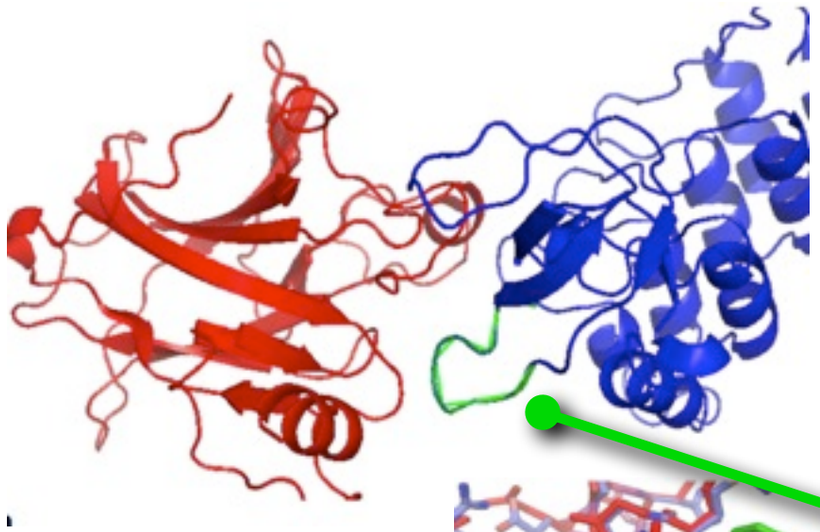
Computer Science Dept.

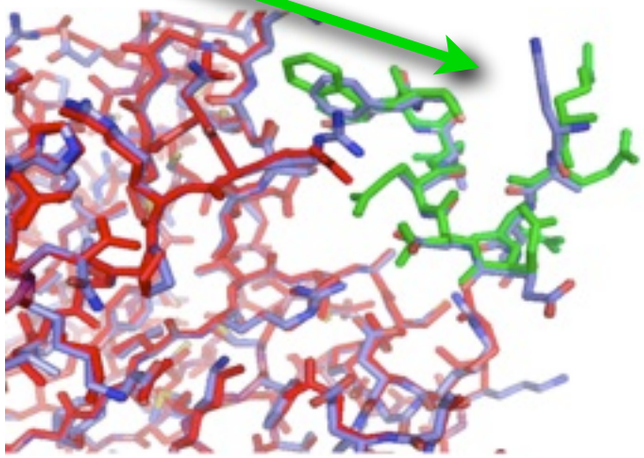
CENTER FOR GENOMICS
AND SYSTEMS BIOLOGY
NEW YORK UNIVERSITY

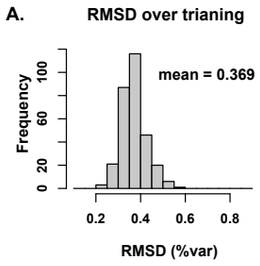
COURANT INSTITUTE

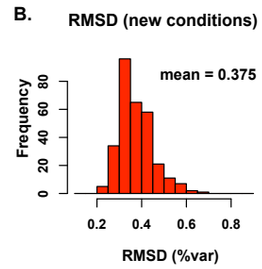

HUMAN PROTEOME FOLDING 2
Current Protein
Current Progress
Current Protein
72.8%
world community grid
Systems Biology
Powered by
IBM.
powered by UNITED DEVICES



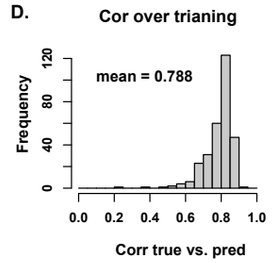


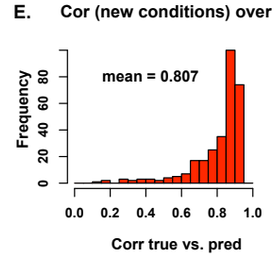


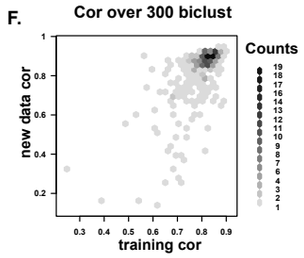

A.
RMSD over trianing
mean = 0.369
Frequency
RMSD (%var)

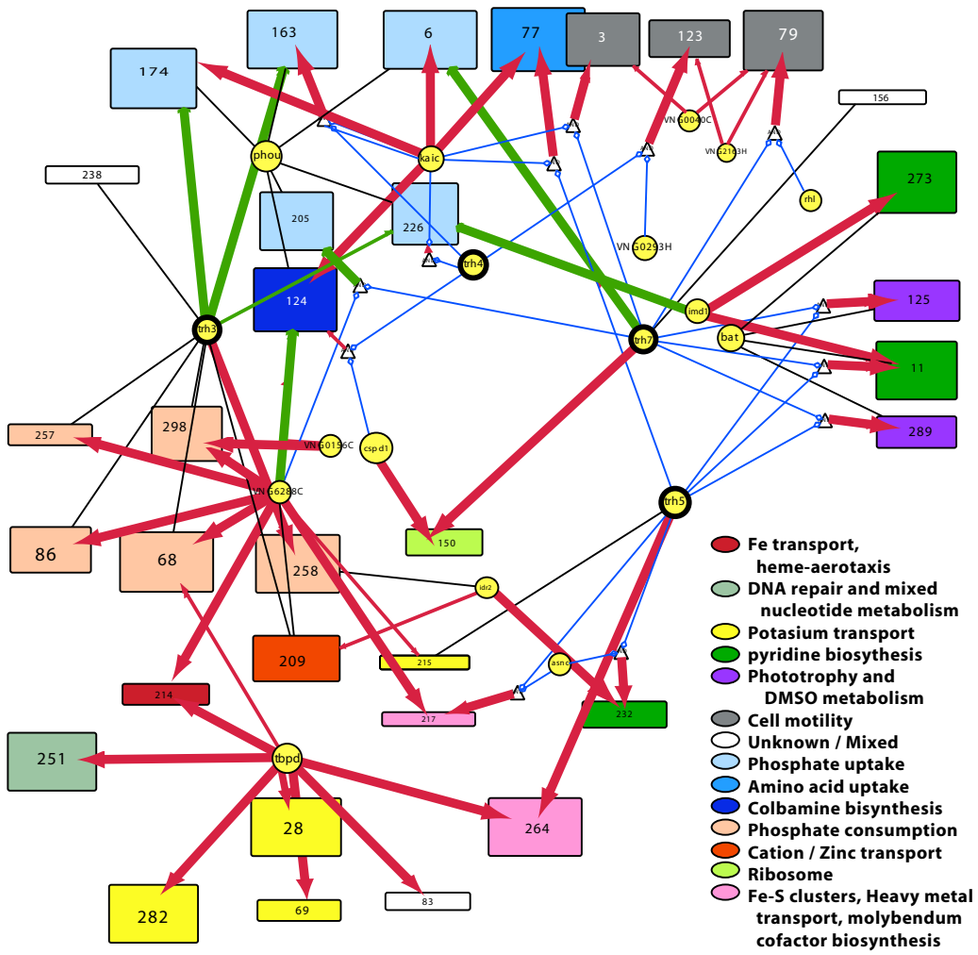

B.
RMSD (new conditions)
mean = 0.375
Frequency
RMSD (%var)


C.
RMS error over 300 biclusters
Counts
new data error
training error


D.
Cor over trianing
mean = 0.788
Frequency
Corr true vs. pred


E.
Cor (new conditions) over
mean = 0.807
Frequency
Corr true vs. pred


F.
Cor over 300 biclust
Counts
new data cor
training cor


163
6
77
3
123
79
174
156
phou
kaiC
VNG0004C
VNG2213H
238
205
226
rhl
273
VNG0293H
trh4
124
trh3
imd1
trh7
bat
125
11
289
257
298
VNG0136C
cspd1
VNG6238C
trh5
86
68
150
258
glo2
209
215
asnC
214
217
232
251
tbpd
28
264
282
69
83
Fe transport, heme-aerotaxis
DNA repair and mixed nucleotide metabolism
Potasium transport
pyridine biosythesis
Phototrophy and DMSO metabolism
Cell motility
Unknown / Mixed
Phosphate uptake
Amino acid uptake
Colbamine bisynthesis
Phosphate consumption
Cation / Zinc transport
Ribosome
Fe-S clusters, Heavy metal transport, molybendum cofactor biosynthesis

References

Bonneau, R*, Facciotti, MT, Reiss, DJ, Madar A., et al. , Baliga, NS*. A predictive model for transcriptional control of physiology in a free living cell. (2007) Cell. Dec 131:1354-1365.

cMonkey biclustering and co-regulated modules:
David J Reiss, Nitin S Baliga, Bonneau R. (2006) Integrated biclustering of heterogeneous genome-wide datasets. BMC Bioinformatics. 7(1):280.

Jochen Supper, Claas aufm Kampe, Dierk Wanke, Kenneth W. Berendzen, Klaus Harter, Richard Bonneau, and Andreas Zell. Modeling gene regulation and spatial organization of sequence based motifs. 8th IEEE international conference on BioInformatics and BioEngineering (BIBE 2008) [In Press].

network inference:
Bonneau R, Reiss DJ, Shannon P, Hood L, Baliga NS, Thorsson V (2006) The Inferelator: a procedure for learning parsimonious regulatory networks from systems-biology data-sets de novo. Genome Biol. 7(5):R36.

Bonneau, R. Learning biological networks: from modules to dynamics (2009). Nature Chemical Biology

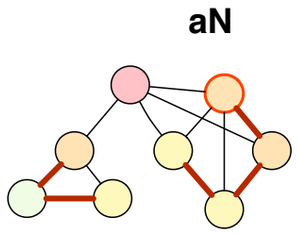
Aviv Madar, Alex Greenfield, Harry Ostrer, Eric Vanden-Eijnden and Richard Bonneau, The Inferelator 2.0: a scalable framework for reconstruction of dynamic regulatory network models. IEEE-ECMB09, In Press

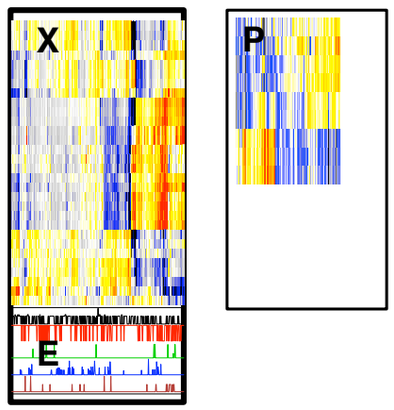
visualization:
Iliana Avila-Campillo*, Kevin Drew*, John Lin, David J. Reiss, Richard Bonneau. BioNetBuilder, an automatic network interface. Bioinformatics. (2007) Bioinformatics. Feb 1;23(3):392-3.

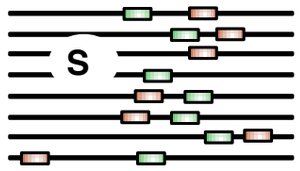
Shannon P, Reiss DJ, Bonneau R, Baliga NS (2006) The Gaggle: A system for integrating bioinformatics and computational biology software and data sources. BMC Bioinformatics. 7:176.


My mentors
the PDB,
genomics,
NCBI,
genomes,
etc!
ME

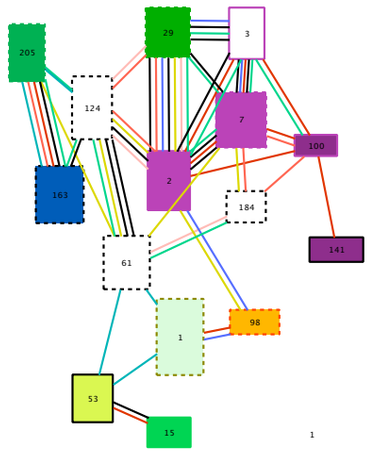
A. Data



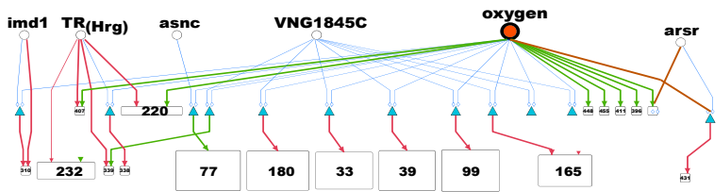




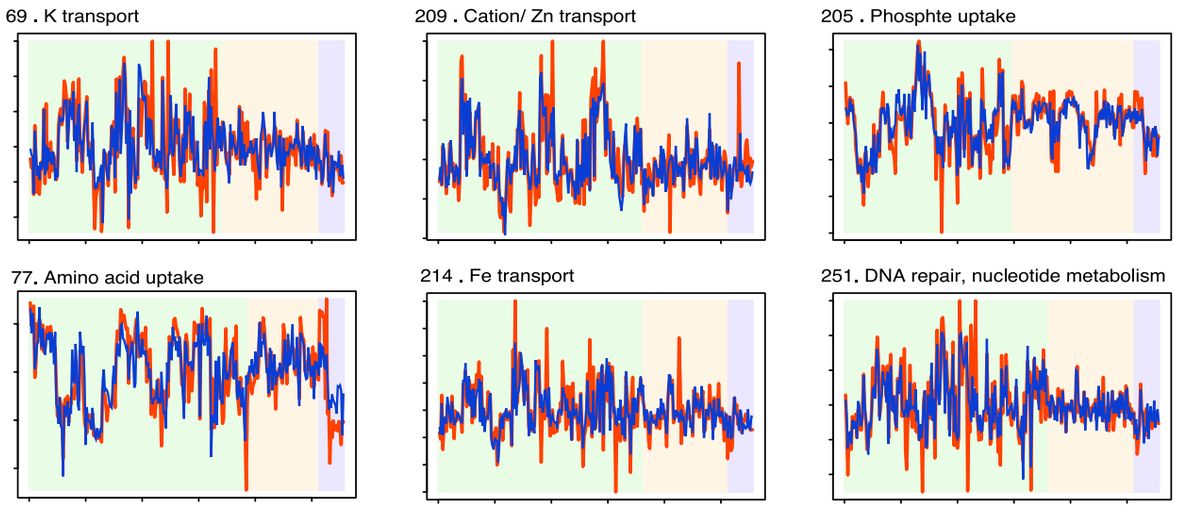
B. (Bi)clustering



C. Dynamical network model



D. Prediction



OVERVIEW

1. CO-REGULATED MODULES (INTEGRATE DATA TYPES).

2. LEARN TOPOLOGY AND DYNAMICS WITH GREEDY / LOCAL APROX. (INFERELATOR 1.0, 1.1)

3. IMPROVING PERFORMANCE OVER MULTIPLE TIME-SCALES (INFERELATOR 2.X)

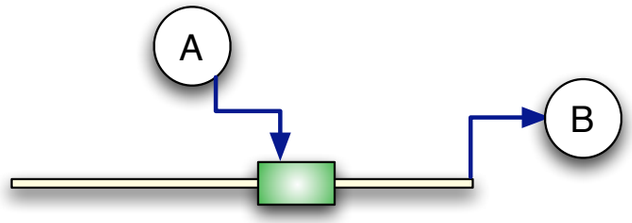
MAIN RESULTS:

- SURPRISING PREDICTIVE PERFORMANCE FOR PROKARYOTIC NETWORKS, T-CELL AND MACROPHAGE DIFFERENTIATION EE NETWORKS

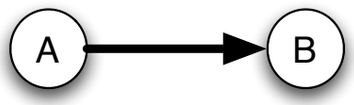
- LONGER TIME SCALE STABILITY

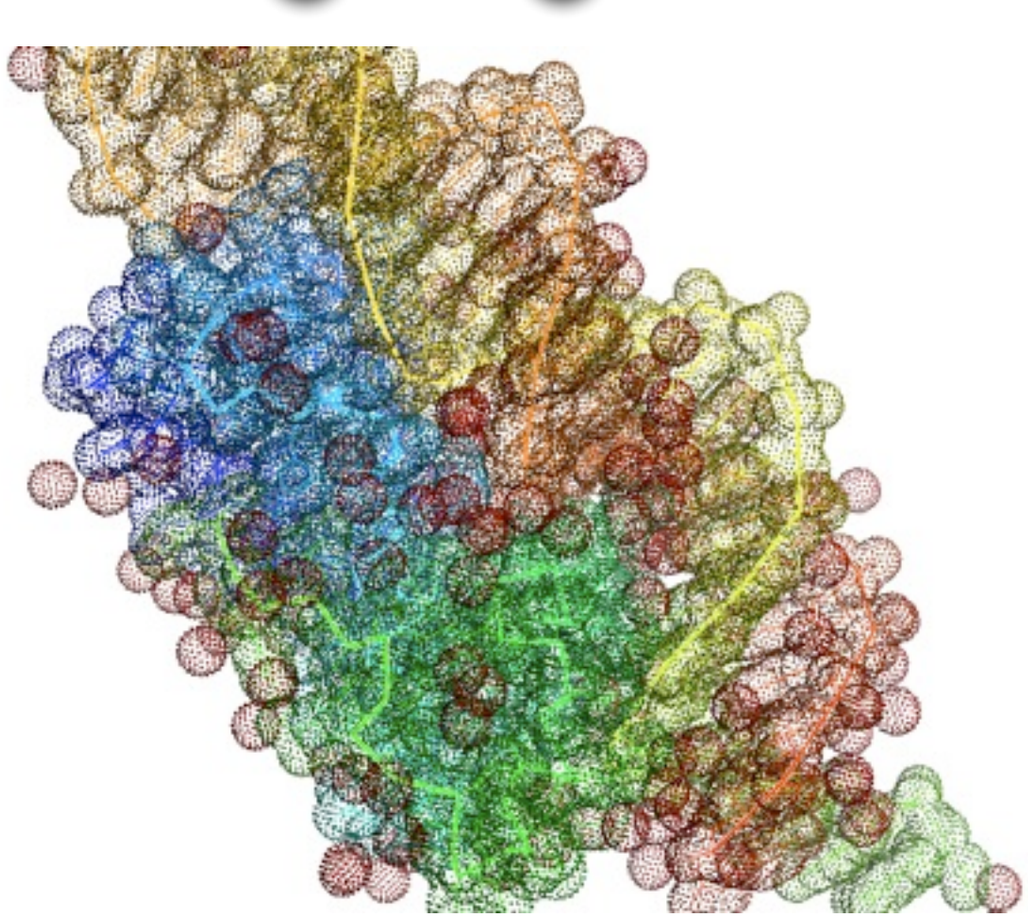
- MODEL FLEXIBILITY

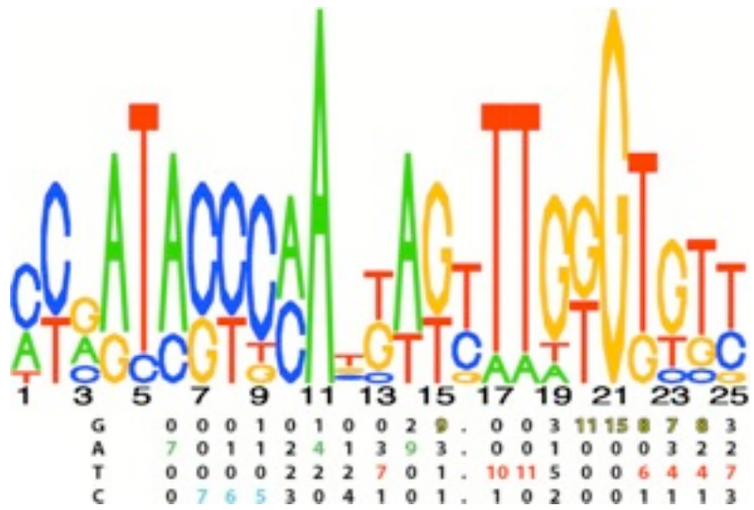
TRANSCRIPTIONAL REGULATION



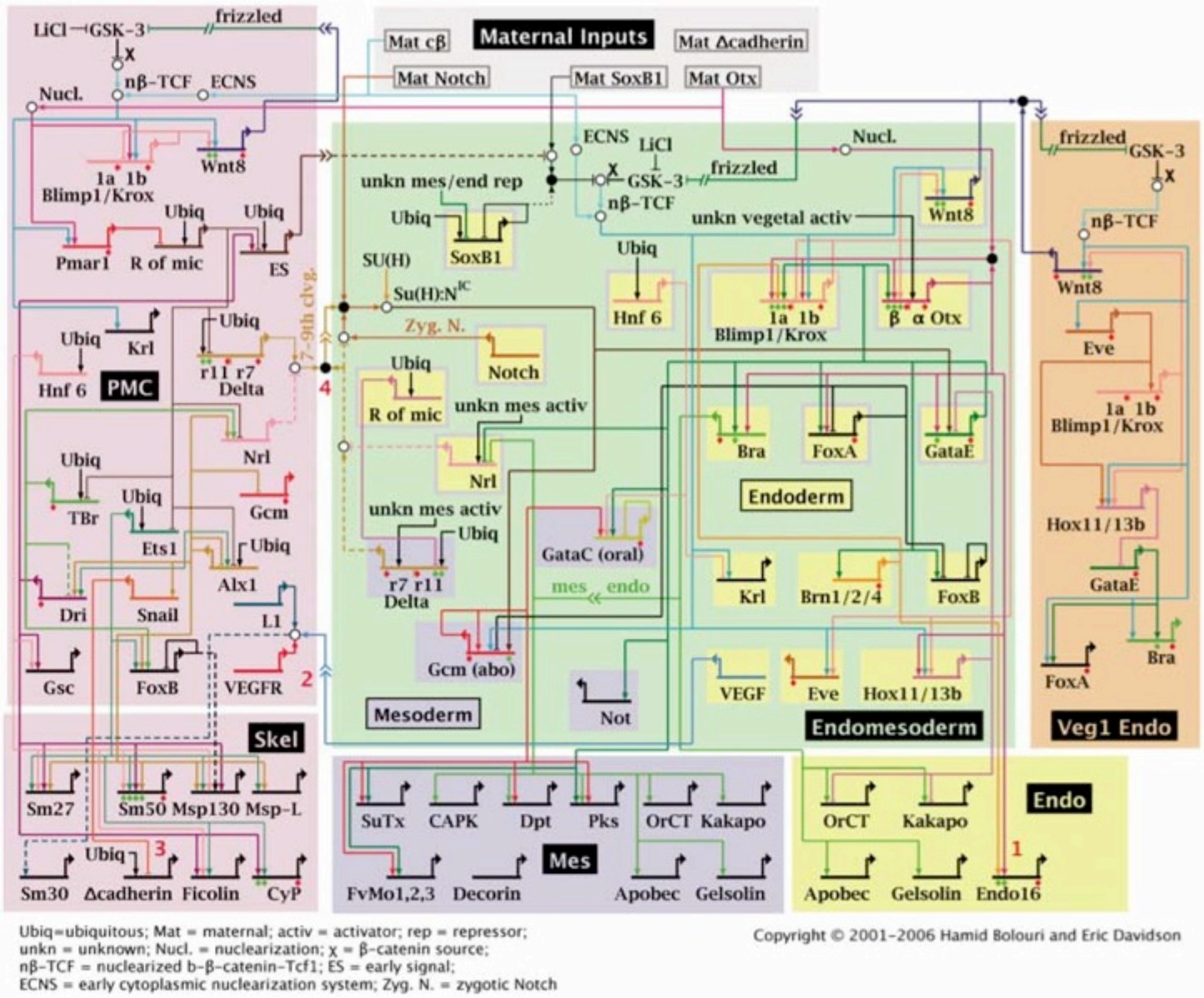
OR





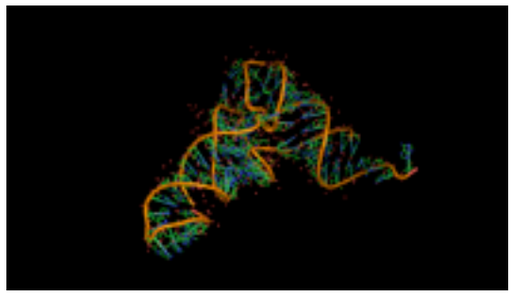


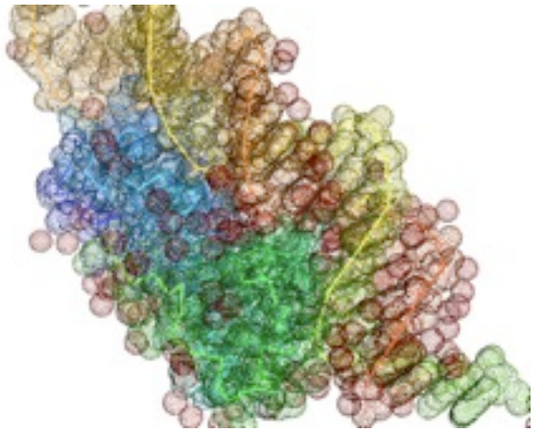
TRANSCRIPTIONAL NETWORKS CONTROLLING DEVELOPMENT

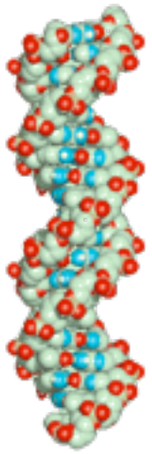


Ubiq=ubiquitous; Mat = maternal; activ = activator; rep = repressor;
unkn = unknown; Nucl. = nuclearization; χ = β-catenin source;
nβ-TCF = nuclearized b-β-catenin-Tcf1; ES = early signal;
ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

BOLOURI, DAVIDSON



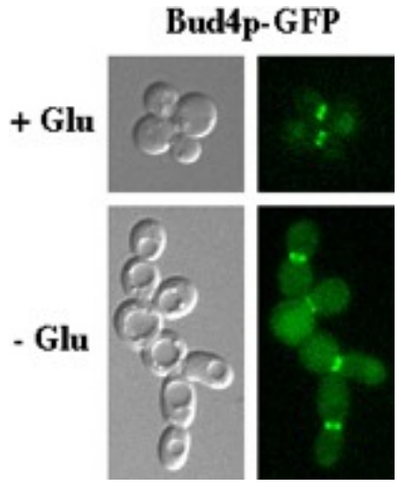




DNA → RNA → PROTEIN

PHENOTYPE

Automated microscopy, etc.



http://www.molbio.uoregon.edu/images/research/spragueg1.jpg

Gene sequencing
Whole Genome assembly

ChIP-chip
TF-DNA Binding
experiments

MICROARRAYS

cDNA
ESTs

Libraries of
Functional RNA

PROTEIN
SEQUENCE
DATABASES,

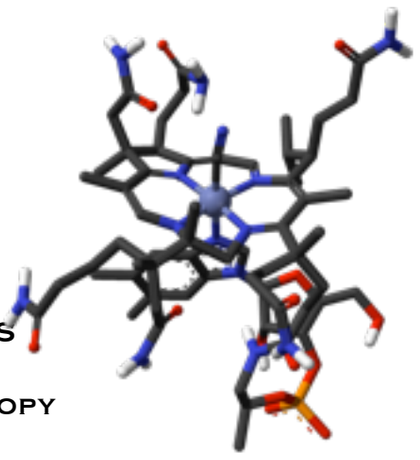
Protein structures,

Proteomics

Protein-
protein
interactions

Metabolomics

Mass-spectroscopy
NMR
Chromotography



Genotype & sequencing

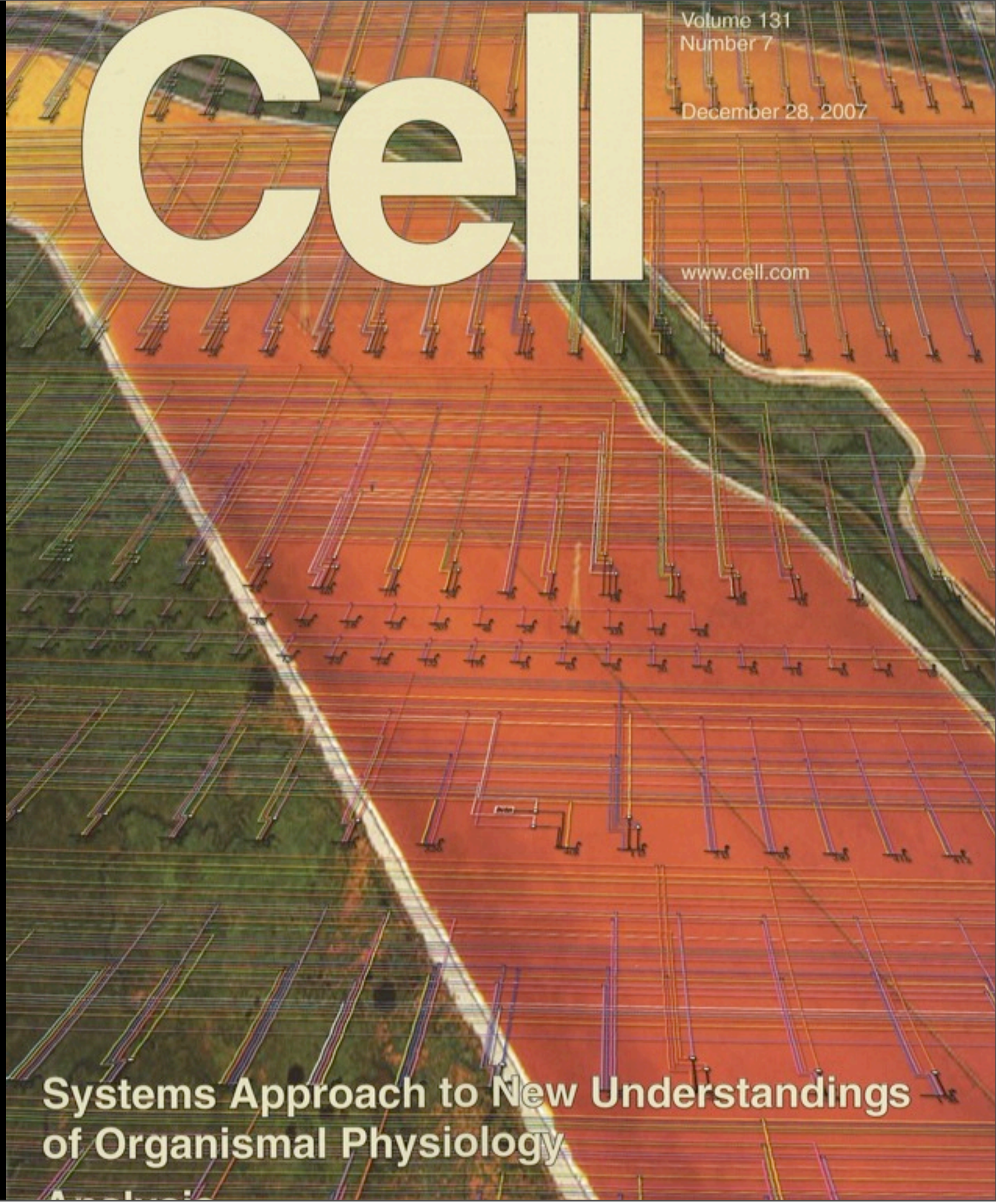
Measuring affinities / binding

Measuring Levels

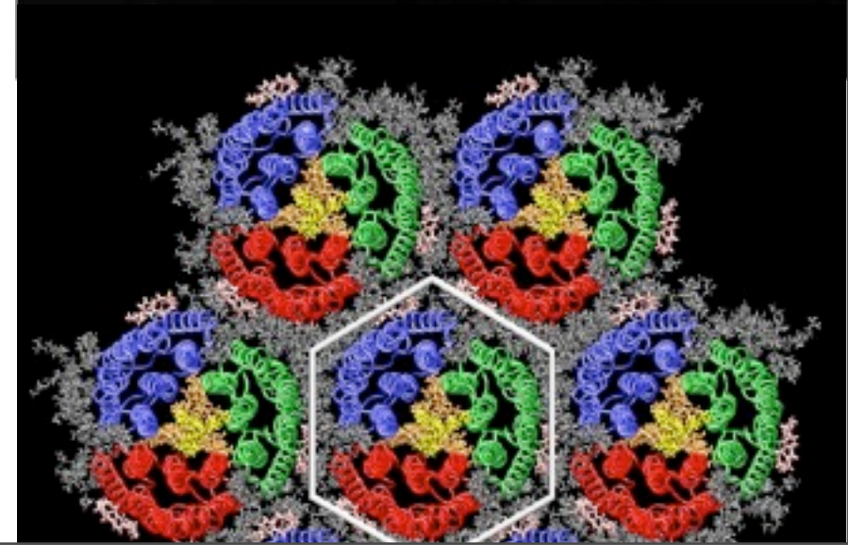
Assaying functional outcome

algorithms:
David J. Reiss (cMonkey)
Vesteinn Thorsson (Inferelator)
Richard Bonneau

functional genomics:
Marc T. Facciotti
Amy Schmid,
Kenia Whitehead
Min Pan, Amardeep Kaur,
Leroy Hood
Nitin S. Baliga



An example : Halobacterium



Why halobacterium:

- if your friends are working on halo ... (Hood, Baliga)
- not a "model" system (originally)
- high IQ
- diverse environment
- small genome
- good genetics, cultivable, etc.
- a very tough extremophile, bioengineering

Data collection and modeling effort

- genome and genome annotation
- microarrays
- genetic and environmental perturbations
- proteomics
- ChIP-chip
- some protein-protein

Halobacterium dataset including

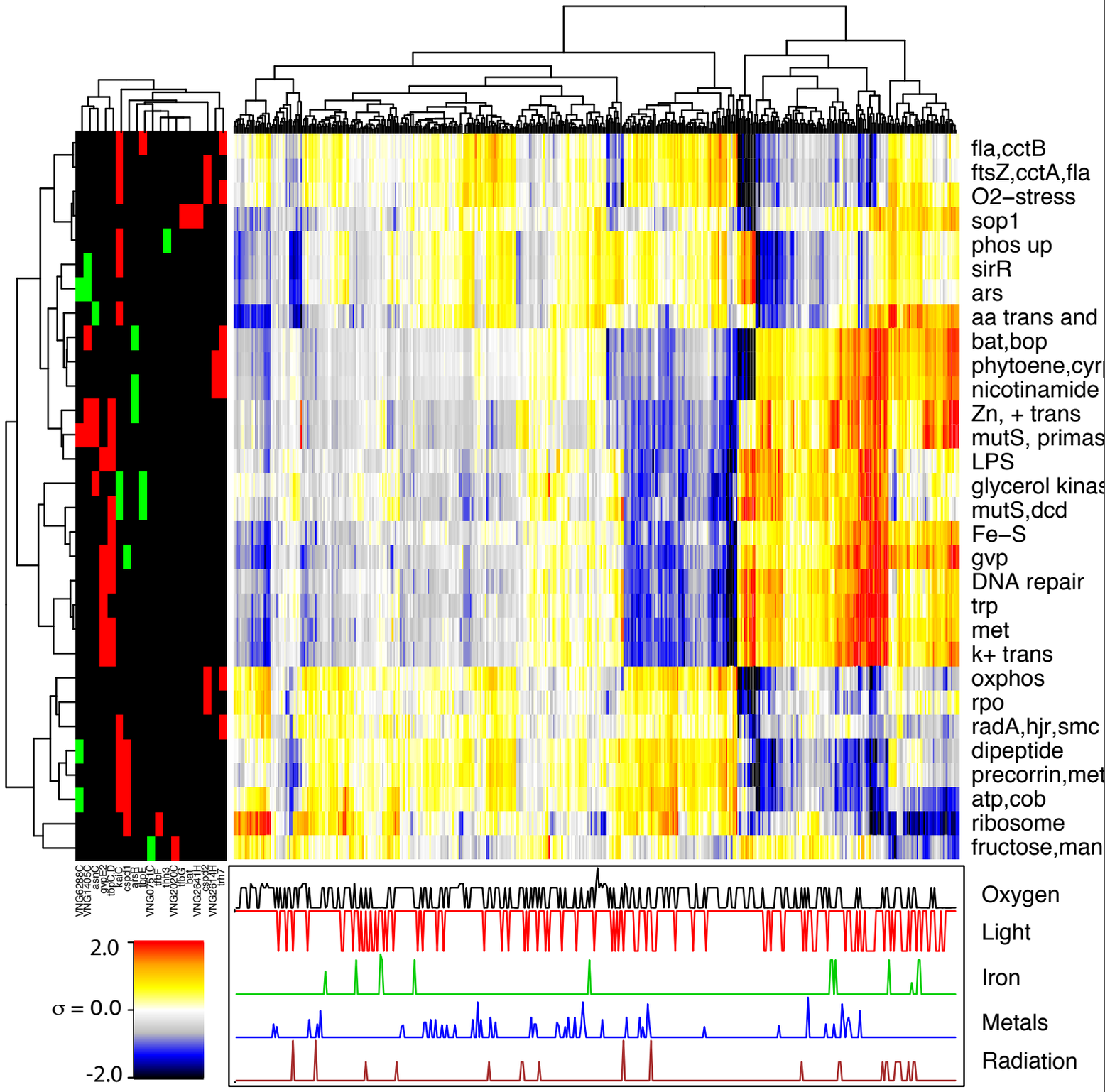
>800 microarrays
time series
knock outs

ChIP-chip experiments

proteomics

phenotype

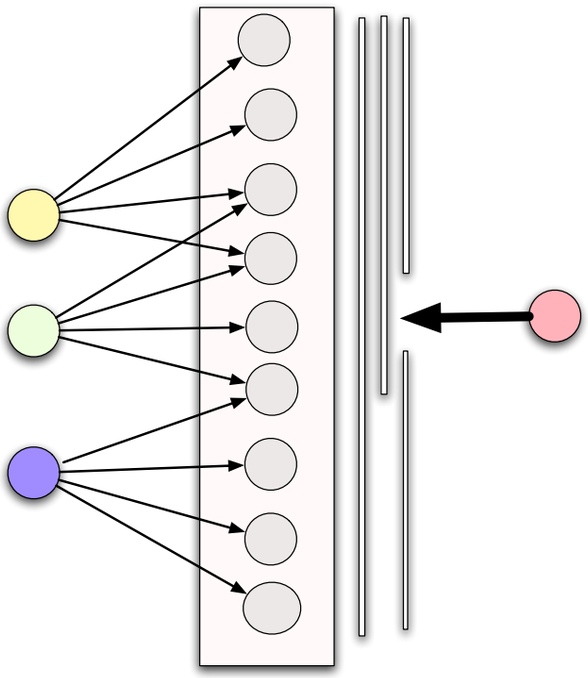
among the most complete prokaryotic datasets



M. Facciotti, N. Baliga

min pan, Kenia Whitehead, Amy Schmid

Our approach



Biological motivation:

Co-regulation dramatically reduces complexity of network inference, and unlike simple co-expression has direct mechanistic relevance to biological control.

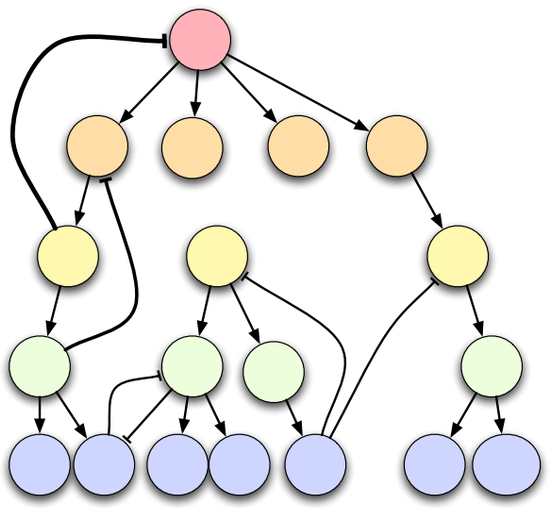
Time (explicit learning/modeling of kinetic parameters) helps even in our current state of affairs.

Model must be capable of modeling interactions with bio relevant functional forms.

Experimental design is key

cMonkey:

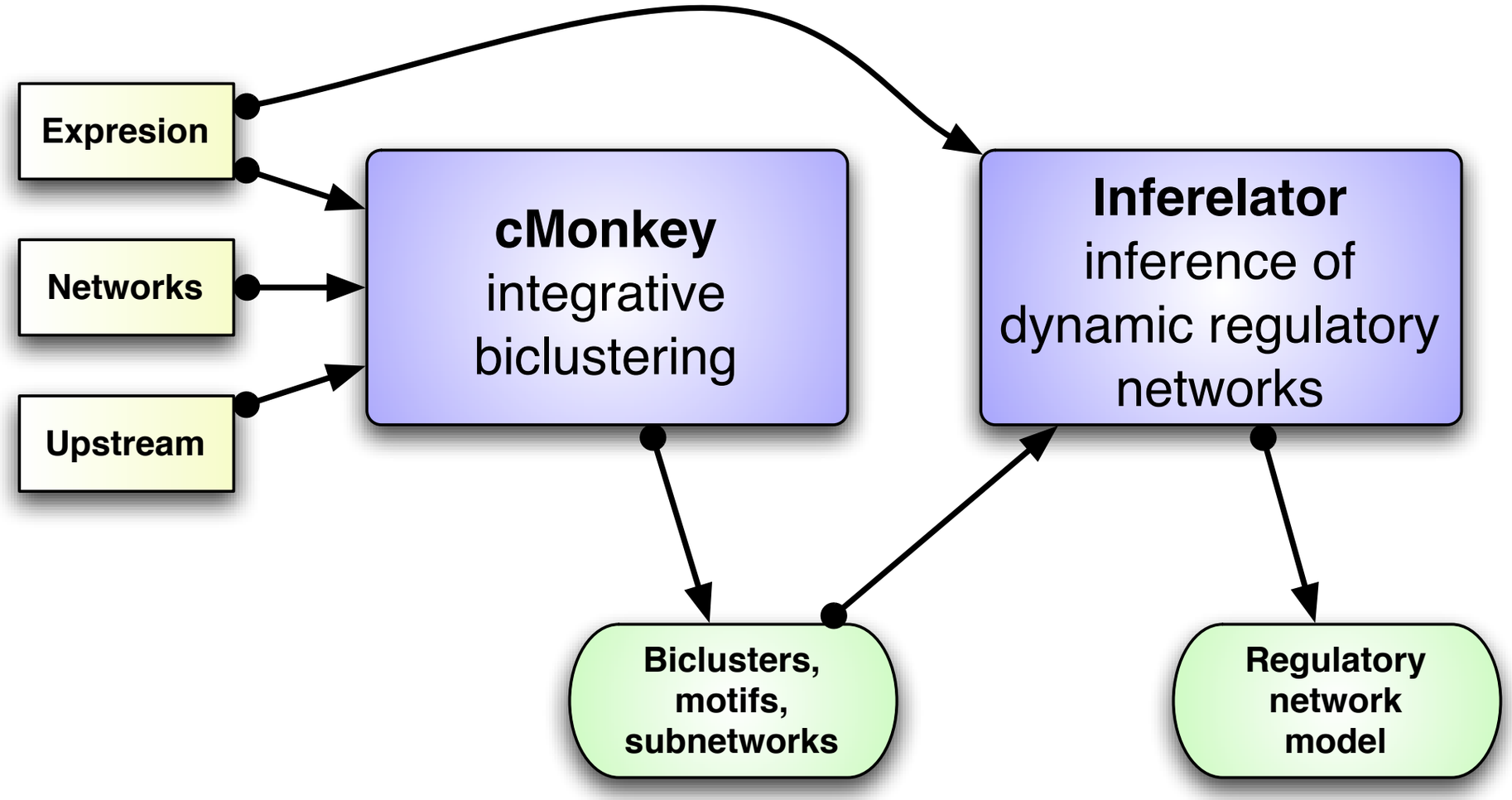
- ★ integrate data-types other than expression to constrain search for co-regulated modules
- ★ avoid lossey transformations of the data and derive joint P of gene given bicluster and all datatypes
- ★ derive framework with eye toward flexibility (new datatypes)



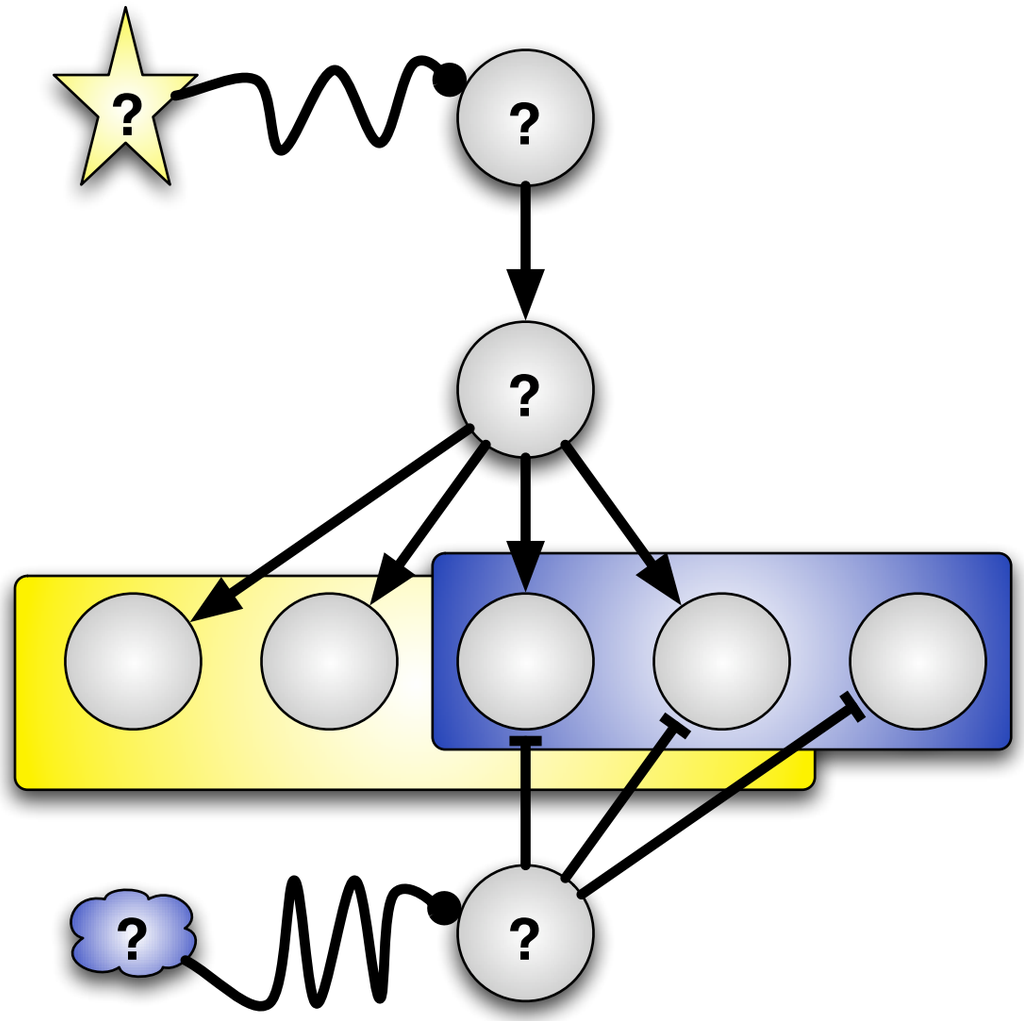
Inferelator:

- ★ frame parameterization of global set of ODEs as regression problem
- ★ interactions: map problem onto tropical semi-ring

OVERVIEW:


Expresion
Networks
Upstream
cMonkey
integrative
biclustering
Inferelator
inference of
dynamic regulatory
networks
Biclusters,
motifs,
subnetworks
Regulatory
network
model

Learning co-regulated groups:



What is Biclustering?

- Concurrent clustering of both rows & conditions
- Given an $n \; x \; m$ matrix, A, find a set of submatrices, B_k, such that the contents of each B_k follow a desired pattern, i.e. gene co-expression.

Based on lecture notes from Kai Li:
http://www.cs.princeton.edu/courses/archive/spr05/cos598E/Biclustering.pdf

Reasons to Bicluster in Biology & Bioinformatics

- Genes not regulated under all conditions

  $\Longrightarrow$ patterns of correlation may exist only under subsets of conditions

- Genes can participate in multiple modules or processes

  $\Longrightarrow$ exclusive clustering algorithms (HAC, K-means) will miss valid clusterings

I. cMonkey:

integrative biclustering

Biological motivation:

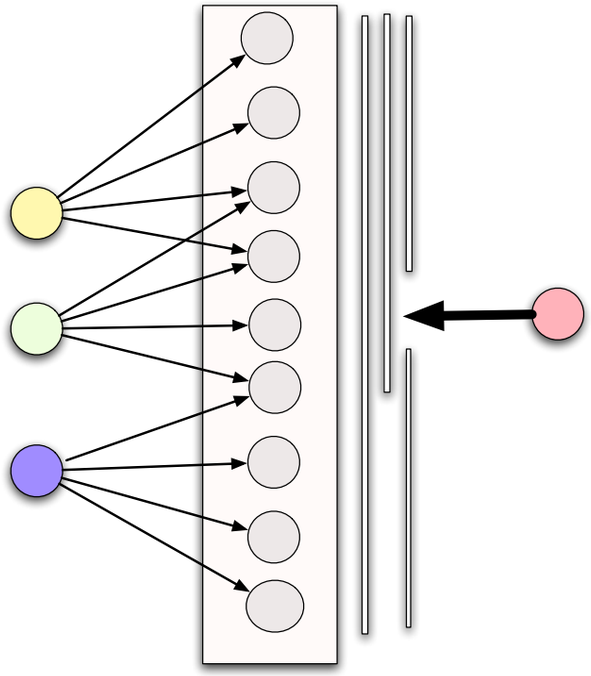
Co-regulation dramatically reduces complexity of network inference, and unlike simple co-expression has direct mechanistic relevance to biological control.

strategy:

- ★ integrate data-types other than expression to constrain search for co-regulated modules
- ★ avoid lossey transformations of the data and derive joint P of gene given bicluster and all datatypes
- ★ derive framework with eye toward flexibility (new datatypes)

Challenges:

- overlapping (genes participate in multiple functions)
- diverse data types
- mix of well studied and completely unknown genes
- **many think of this as a solved problem...why?**
- Resultant models are a complex low-level abstraction of the systems behavior (functional modules, complexes, annotations, etc. are linked to clusters).



Dave Reiss

Peter Waltman

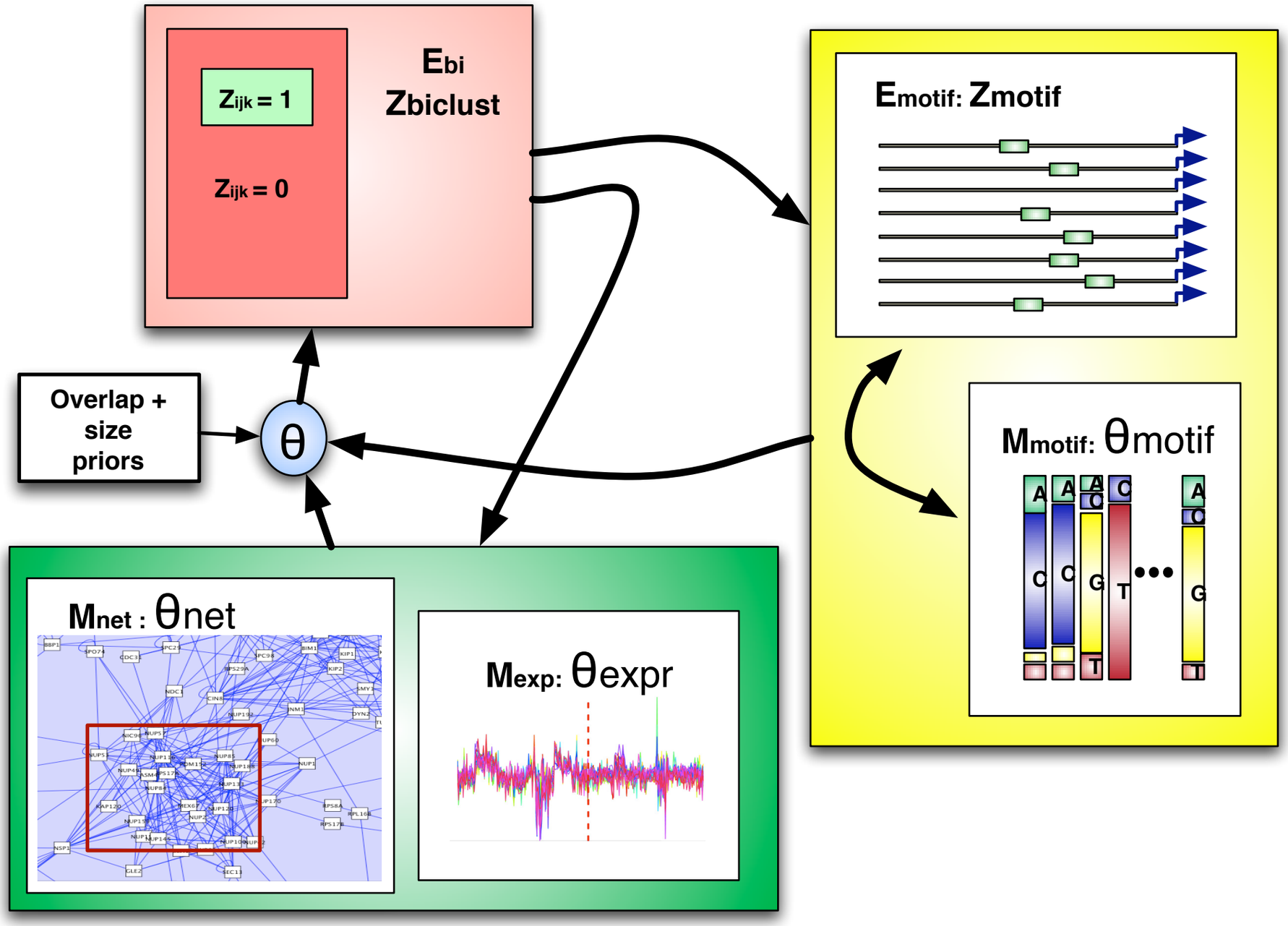
cMonkey:
MCMC optimization of a multi-data likelihood

other data:
exp-like:
[GWA,
Copy number,
phenotype]

nets:
[chip-seq,
etc.]

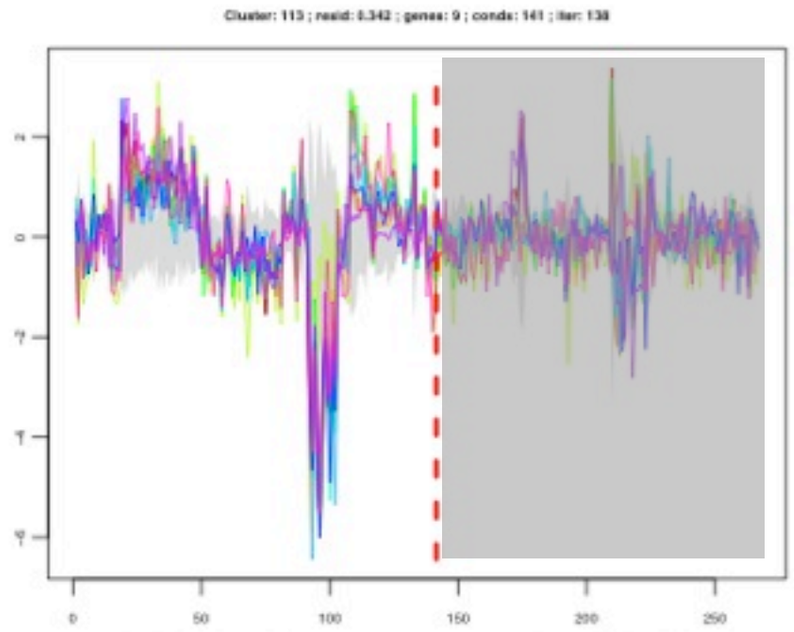
seq:
[UTR,known sites]

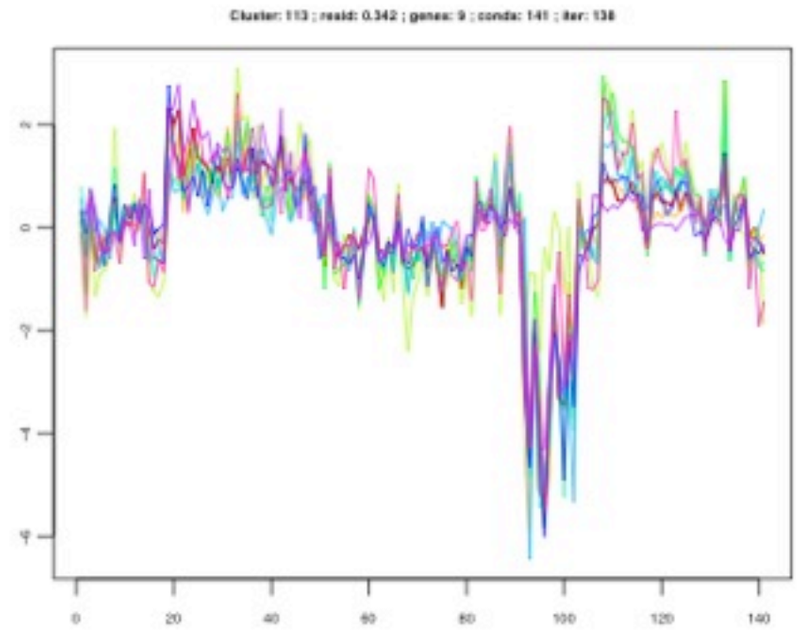
The real advantage of cMonkey is its lack of lossy transformations



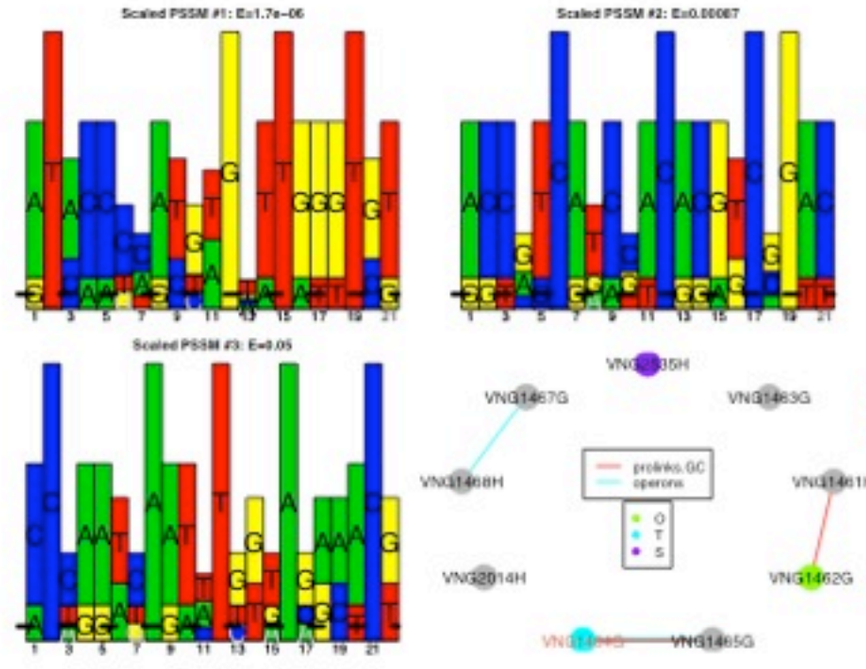
Archaea: bop/bat-associated regulon
[Halobacterium NRC-1]

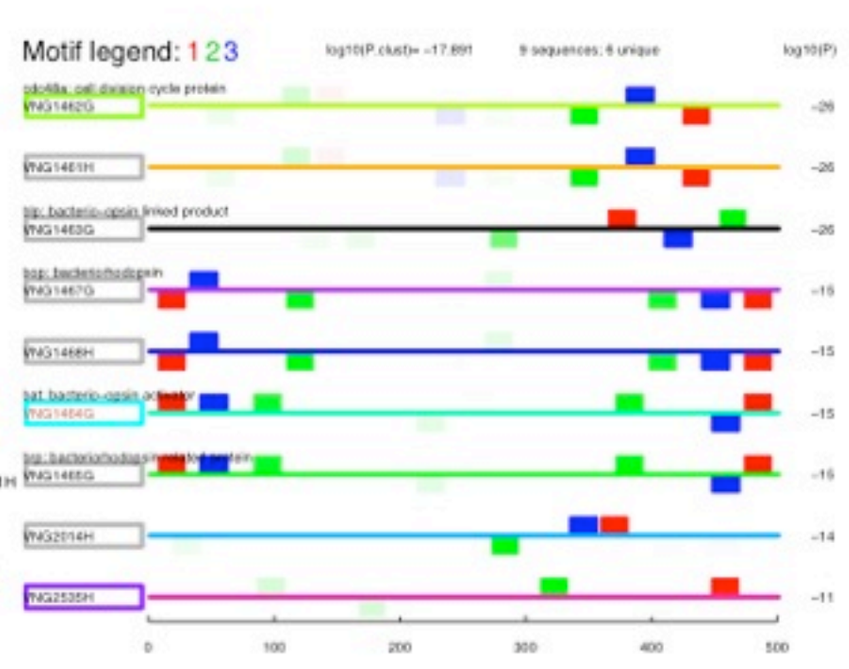
Expression





Motifs



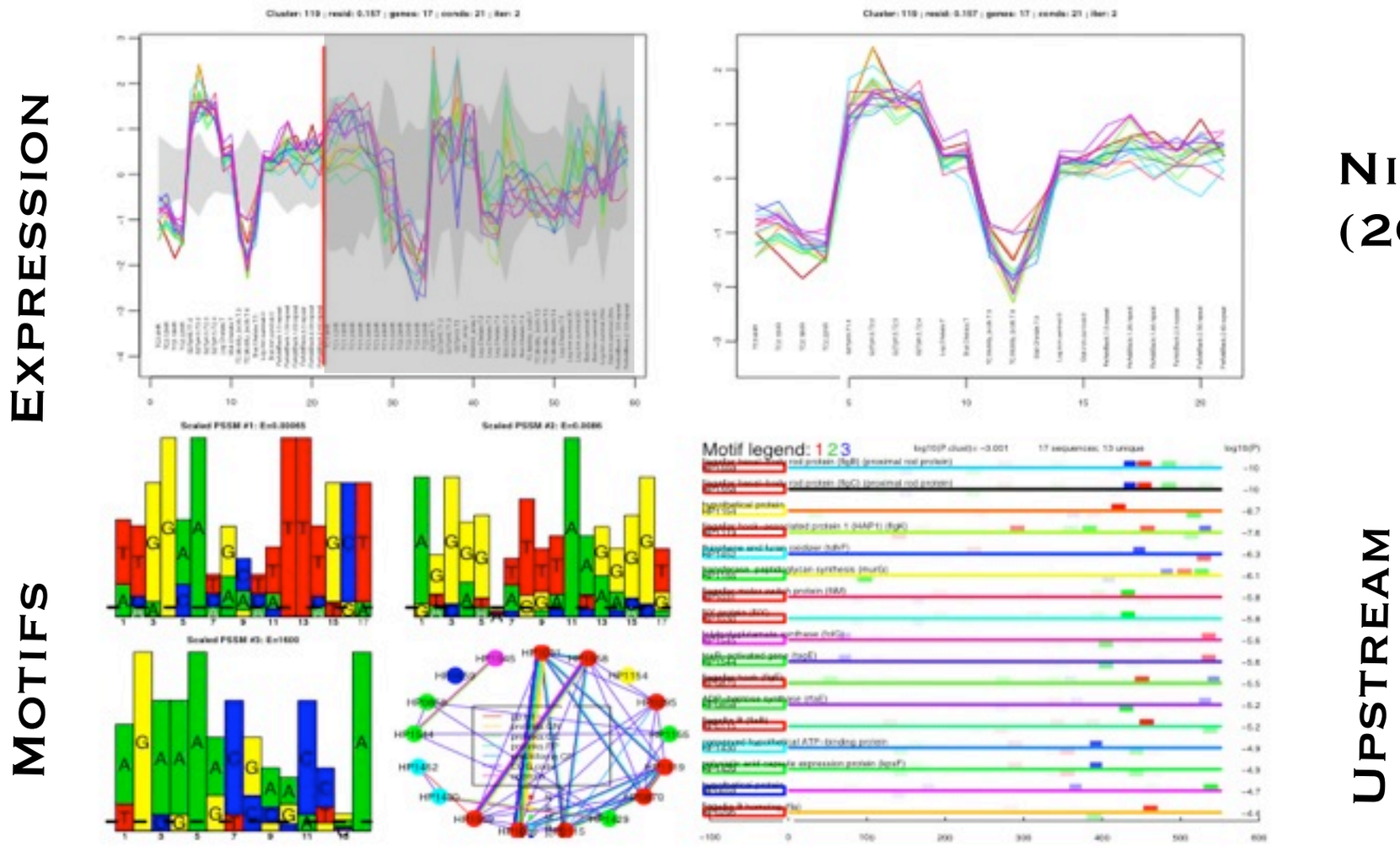


Upstream

Baliga, et al.
(1999,2000)

MULTI-BICLUSTERING: MULTISPECIES

BACTERIA: RPON-ASSOCIATED FLAGELLAR REGULON [H. PYLORI]
---> [ALSO IN E. COLI]

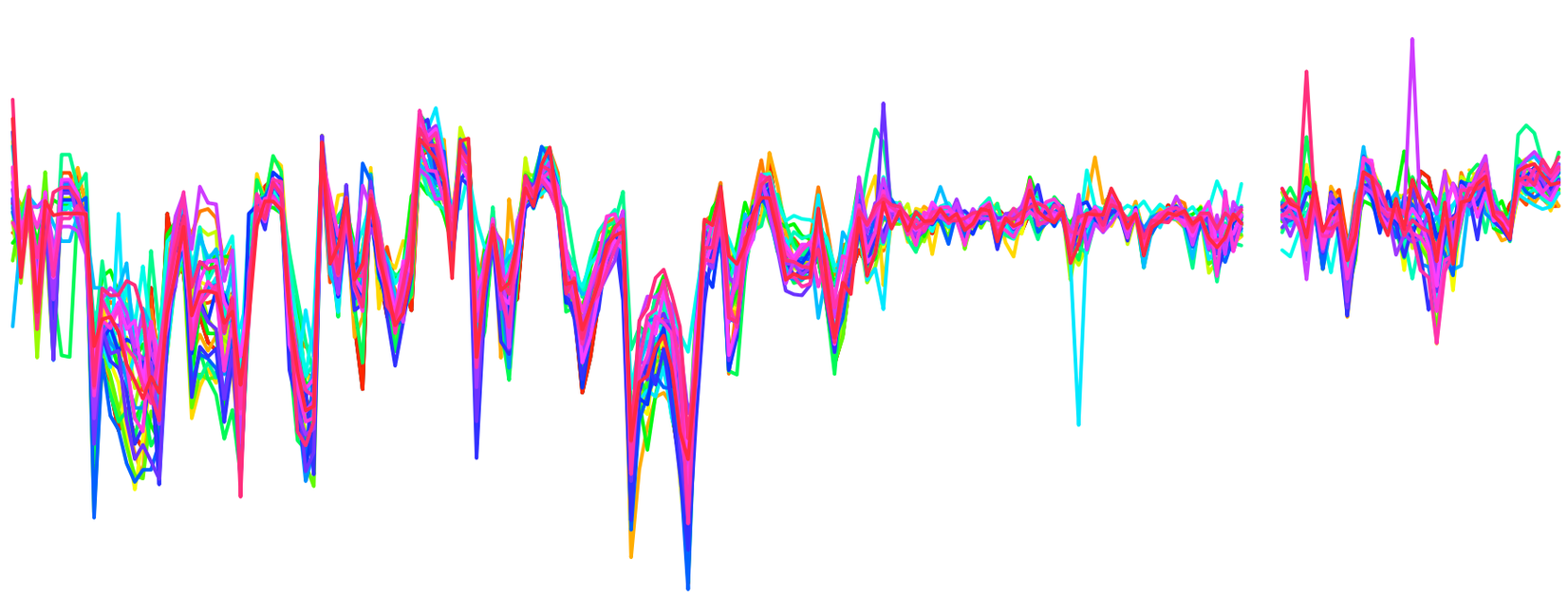


NIEHUS, ET AL. (2004)

W/ PATRICK EICHENBERGER, NYU; HARRY OSTRER, NYU-MED
ERIC ALM, MIT, BROAD

SCORE COMPONENT I:
R, EXPRESSION [LEVELS]

$$p(x_{ij}) = \frac{1}{\sqrt{2\pi(\sigma_j + \epsilon)^2}} \exp\left[-\frac{1}{2}\left(\frac{x_{ij} - \bar{x}_{jk} + \epsilon}{\sigma_j + \epsilon}\right)^2\right]$$



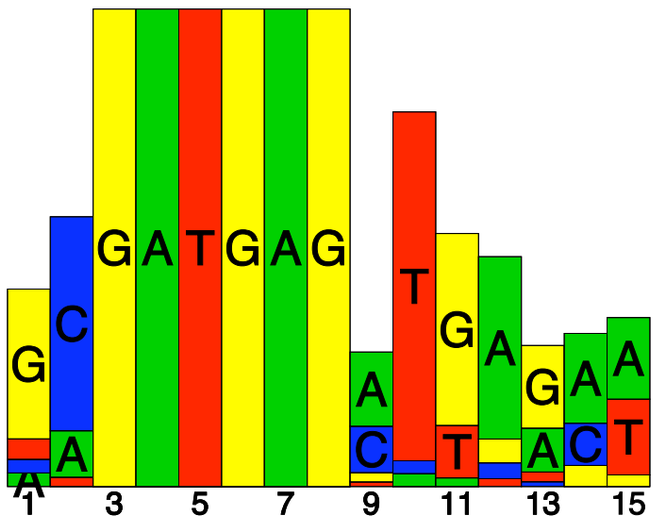
Reiss, Shannon, Baliga, Bonneau, 2006

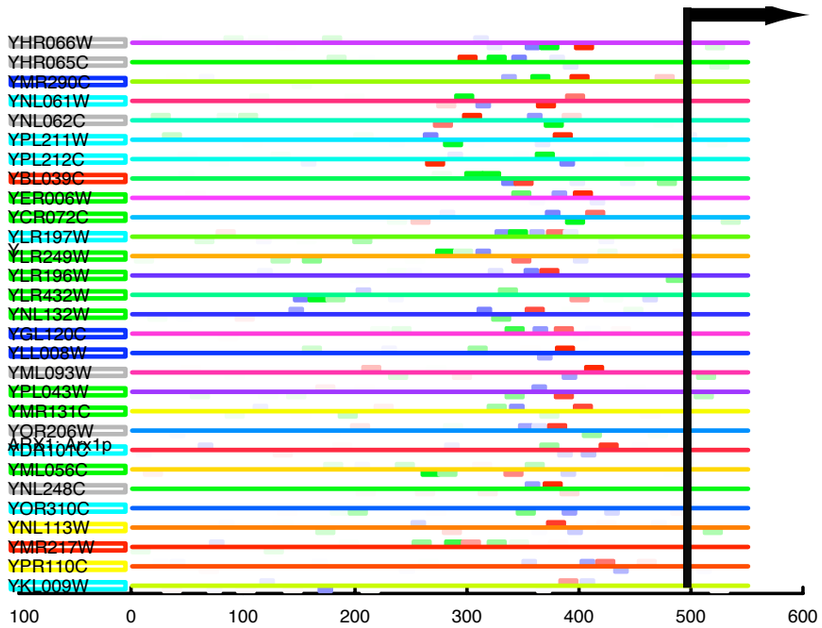
SCORE COMPONENT II:

P, MOTIF DETECTION AND CO-OCCURANCE

[SHORT SEQUENCES]

motif models:
MEME , Weeder, known
->
cis, trans, UTR



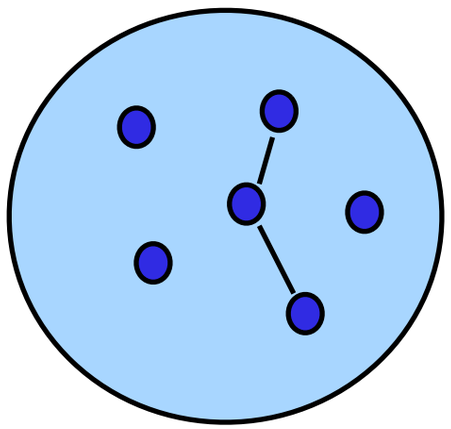


Reiss, Shannon, Baliga, Bonneau, 2006

SCORE COMPONENT III:

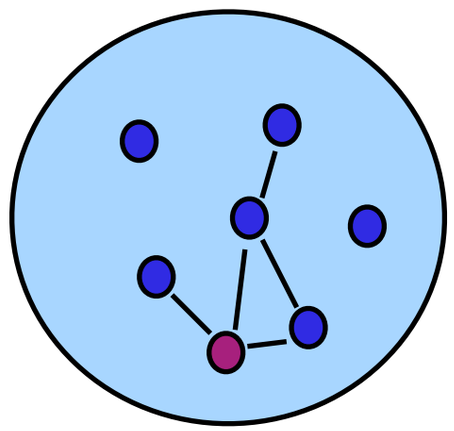
q, NETWORKS [ASSOCIATIONS]

Before adding gene



p = 0.16

After adding gene



p = 0.012

Reward addition of genes to bicluster that share edges with other genes in bicluster.

Hypergeometric distribution to derive *p*-values:

$$p(n_{i\to \mathbf{I}_k} | n_{i\to \mathbf{I}'_k}, n_{\mathbf{I}_k\to \mathbf{I}_k}, n_{\mathbf{I}_k\to \mathbf{I}'_k}) = \frac{\left(\begin{array}{c} n_{i\to \mathbf{I}_k} + n_{\mathbf{I}_k\to \mathbf{I}_k} \\ n_{i\to \mathbf{I}_k} \end{array} \right) \left(\begin{array}{c} n_{i\to \mathbf{I}'_k} + n_{\mathbf{I}_k\to \mathbf{I}'_k} \\ n_{i\to \mathbf{I}'_k} \end{array} \right)}{\left(\begin{array}{c} n_{i\to \mathbf{I}_k} + n_{\mathbf{I}_k\to \mathbf{I}_k} + n_{i\to \mathbf{I}'_k} + n_{\mathbf{I}_k\to \mathbf{I}'_k} \\ n_{i\to \mathbf{I}_k} + n_{i\to \mathbf{I}'_k} \end{array} \right)}$$

cMonkey continued

- Combine 3 likelihoods into a joint log-likelihood:

$$g_{ik} = r_0 \, log(\tilde{r}_{ik}) + s_0 \, log(\tilde{s}_{ik}) + \sum_{n \in N} q_0^n \, log(\tilde{q}_{ik}^n)$$

where r_0, s_0 and q_0 are *"mixing parameters"* –

Pre-selected and set according to an annealing schedule

- Logistic regression to discriminate between genes in/ out of bicluster:

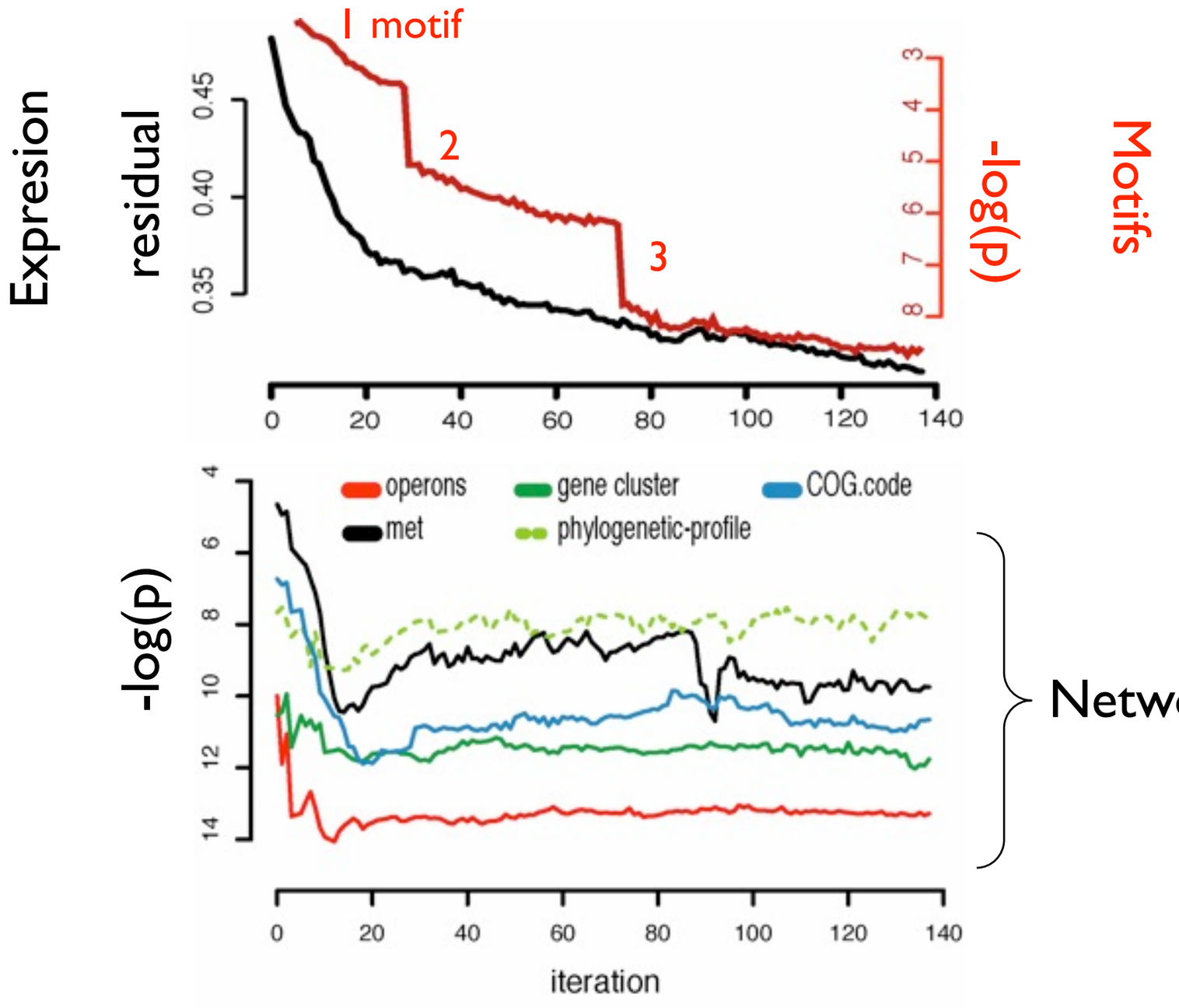
$$\pi_{ik} \equiv p(y_{ik} \mid \mathrm{X}_k, S_i, \mathrm{M}_k, \mathrm{N}) = \frac{e^{(\beta_0 + \beta_1 g_{ik})}}{1 + e^{(\beta_0 + \beta_1 g_{ik})}}$$

where $p(y_{ik}=1)$ indicates likelihood of membership of gene i to cluster k

- Monte Carlo, annealing of the biclusters:

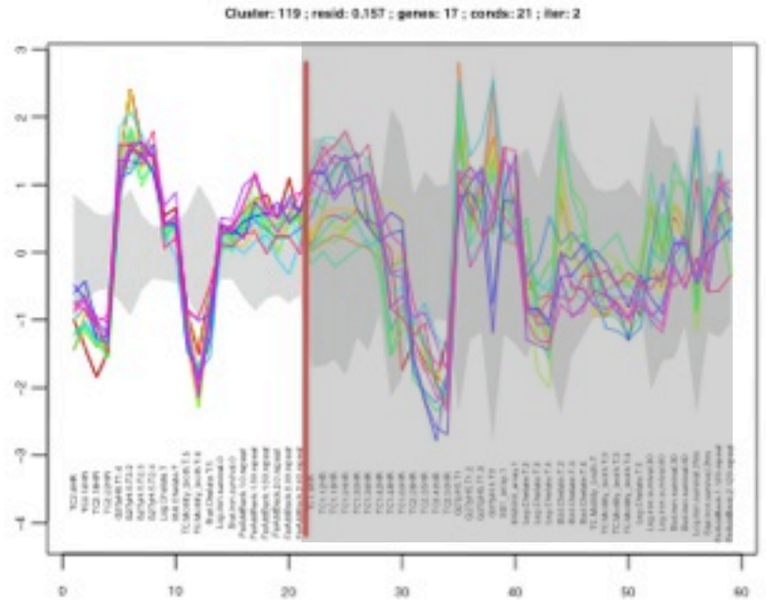
$$p(add_{ik} \mid \pi_{ik}) = e^{\frac{-(1-\pi_{ik})}{T}} \; ; \quad p(drop_{ik} \mid \pi_{ik}) = e^{\frac{-\pi_{ik}}{T}}$$

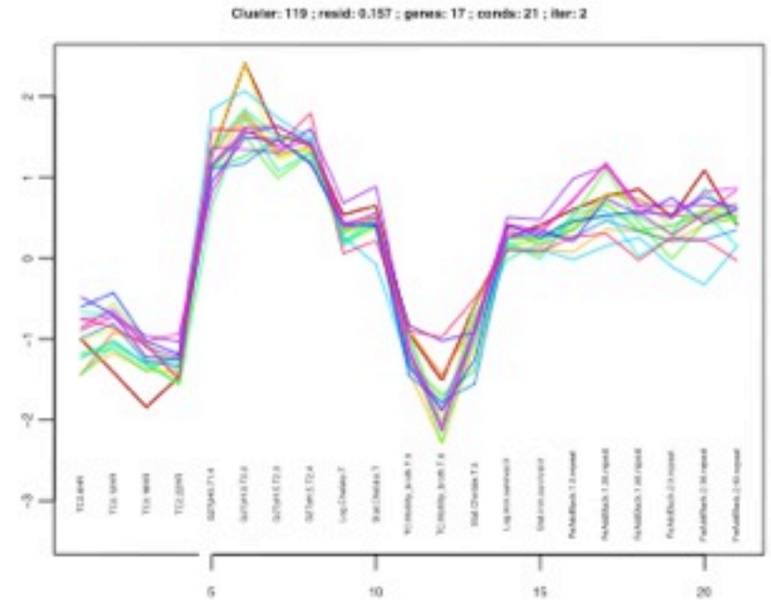
OPTIMIZATION OF SCORE ELEMENTS



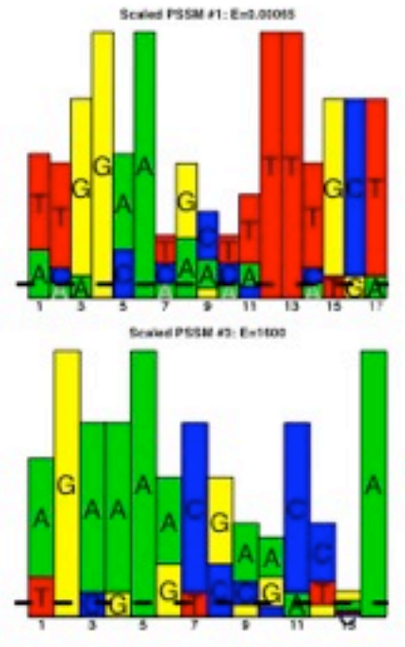
Bacteria: RpoN-associated flagellar regulon [H. pylori] ---> [also in E. coli]

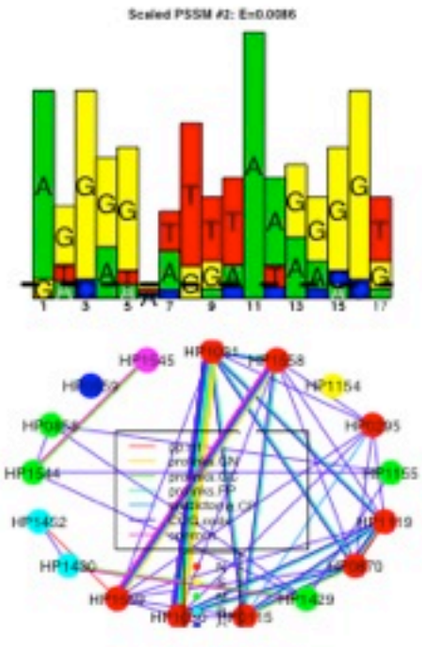
Expression





Motifs

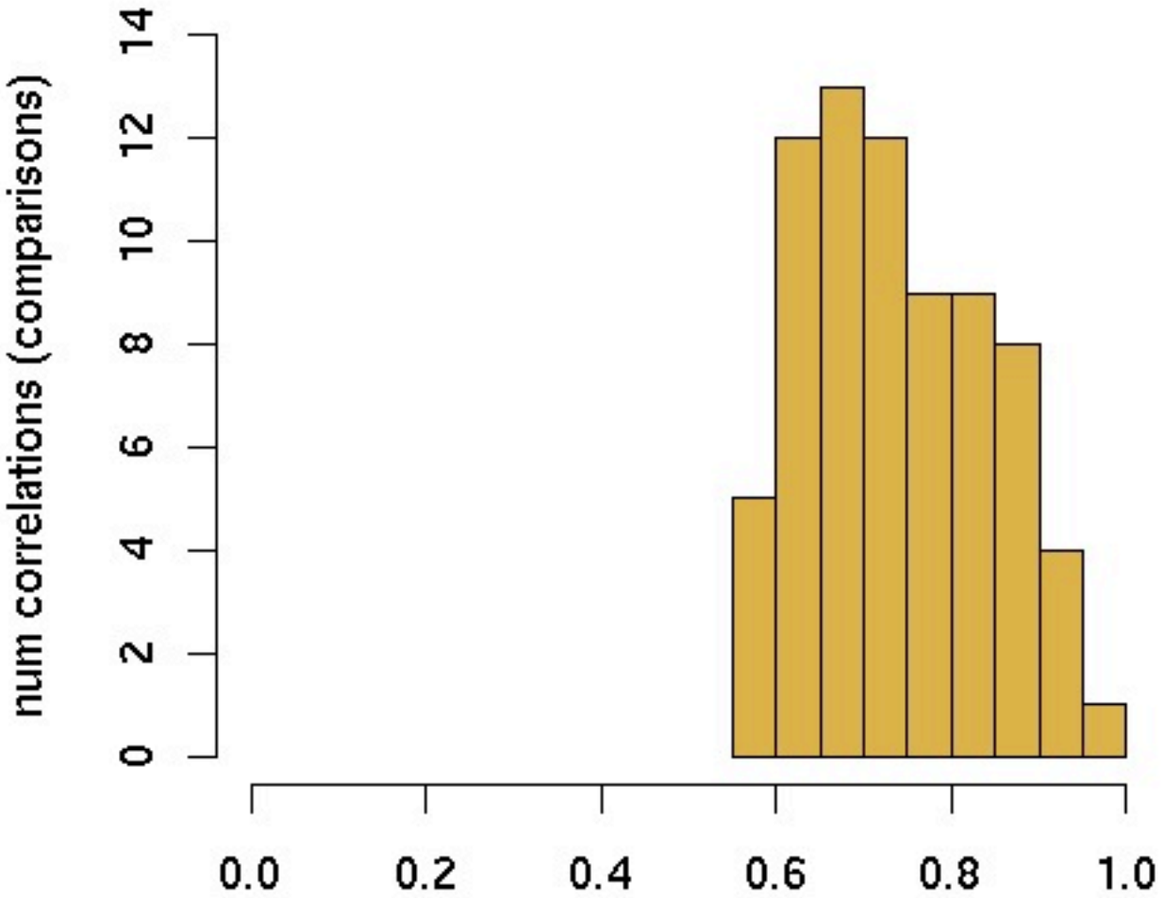




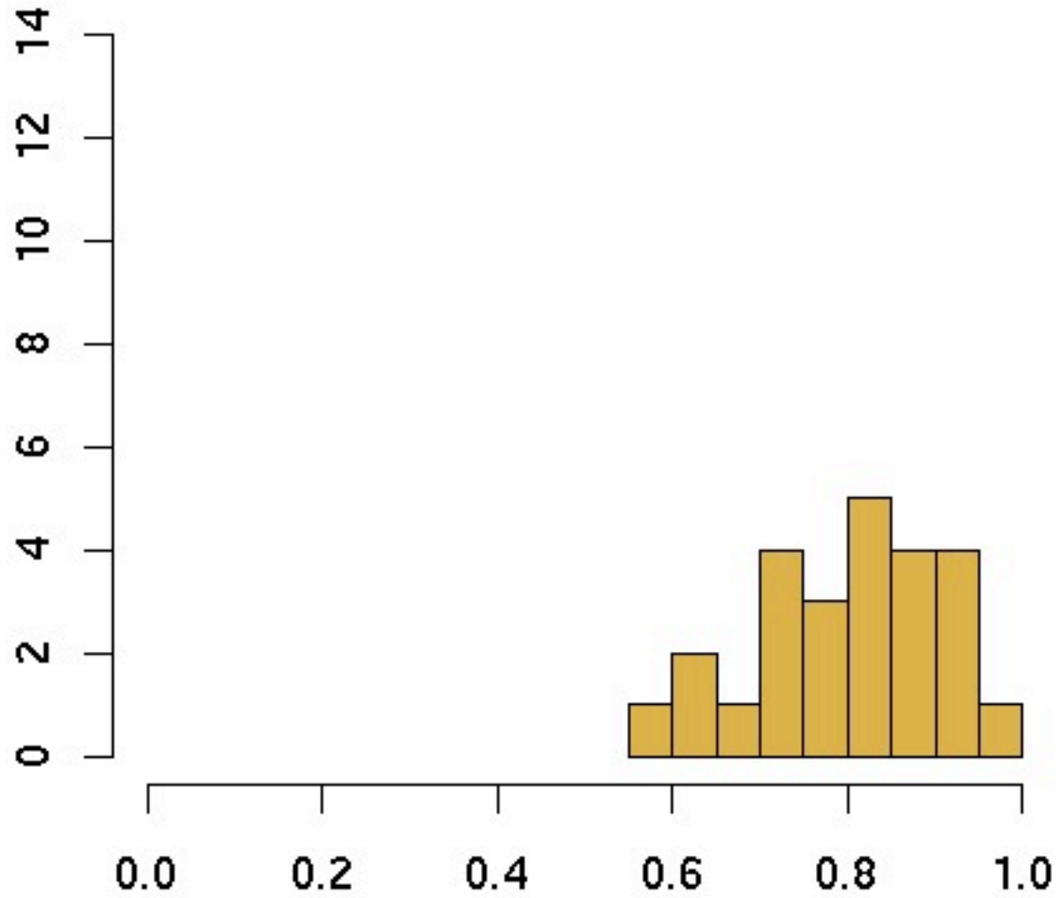
Upstream

Niehus, **et al.** (2004)

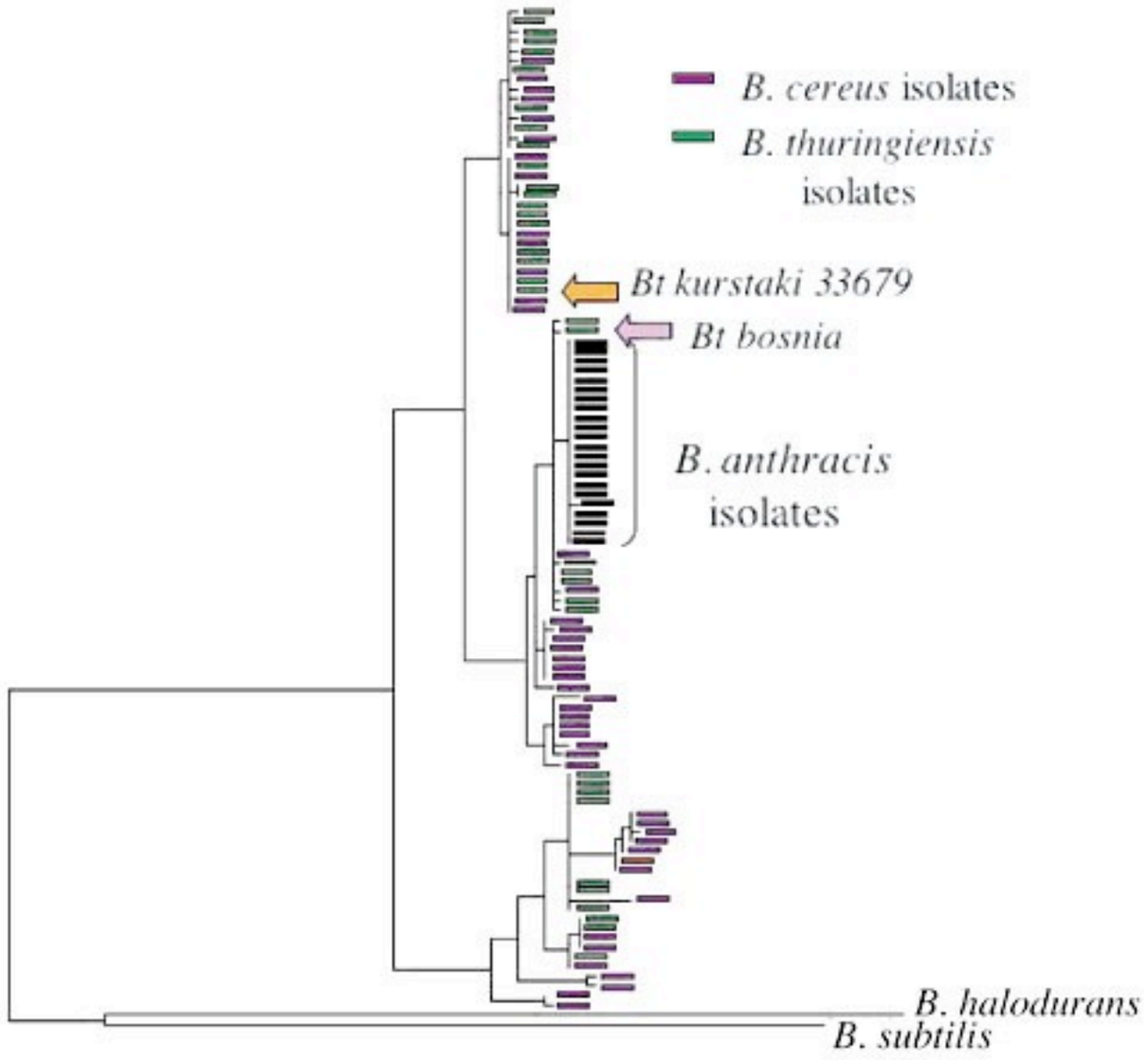
E. coli motif comparisons

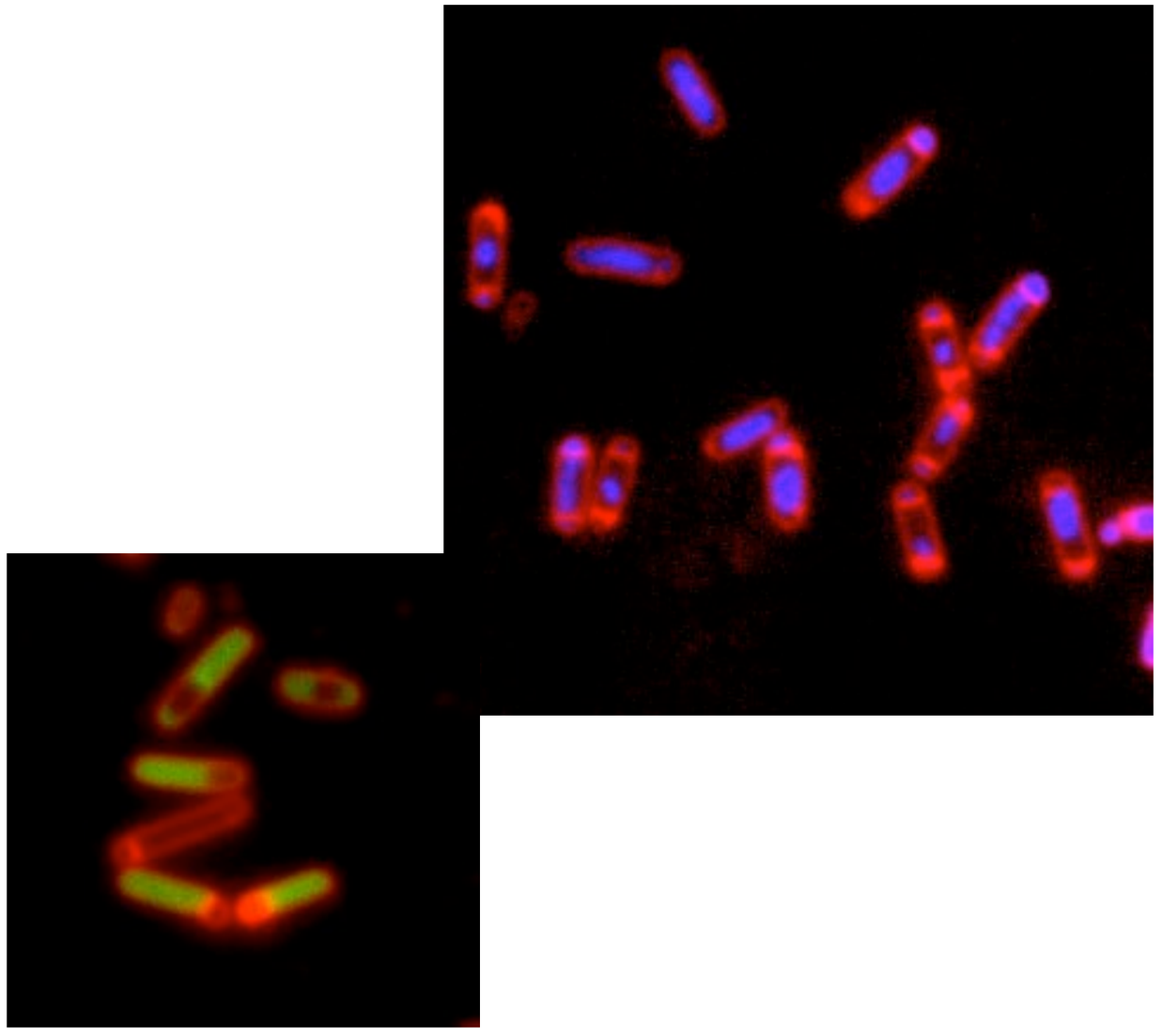


B. subtilis motif comparisons



MULTI-BICLUSTERING: MULTISPECIES



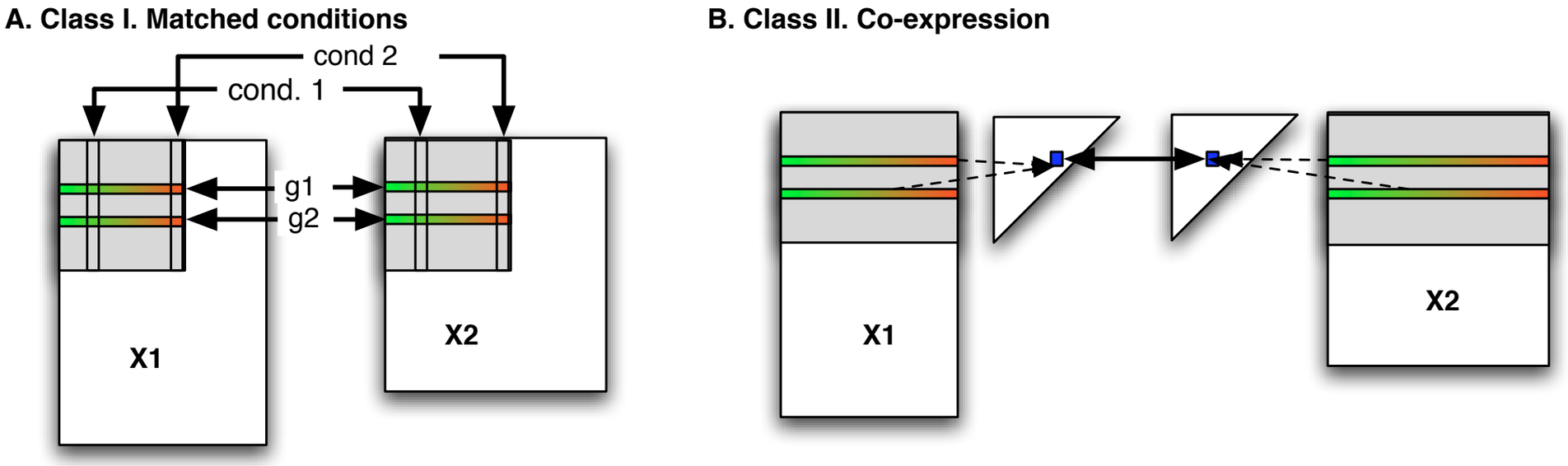


w/ Patrick Eichenberger, NYU
w/ Eric Alm, MIT, Broad
w/ Harry Ostrer

Previous Multi-Species Comparisons

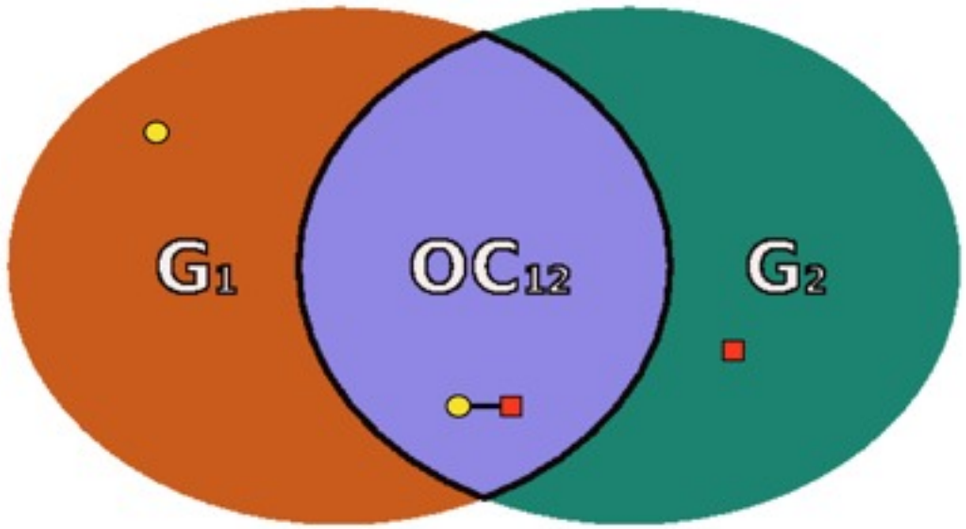
- McCarroll, Murphy, Zou, et al (2004, Nature Genetics)
- Ihmels, Bergmann, Berman, Barkai (2005, PLoS Genetics)
- **Tirosh, Barkai (2007, Genome Biology)**

3 classes of multi-species comparisons



Proposed Multi-species cMonkey model

- Given 2 genomes, G_1 & G_2:
  - OC_1 & OC_2 as the set of genes in G_1 & G_2 with orthologs in the other
  - Define OC_{12} as the set of all putative orthologous pairs, including:
    - 1-to-1
    - 1-to-many
    - many-to-many



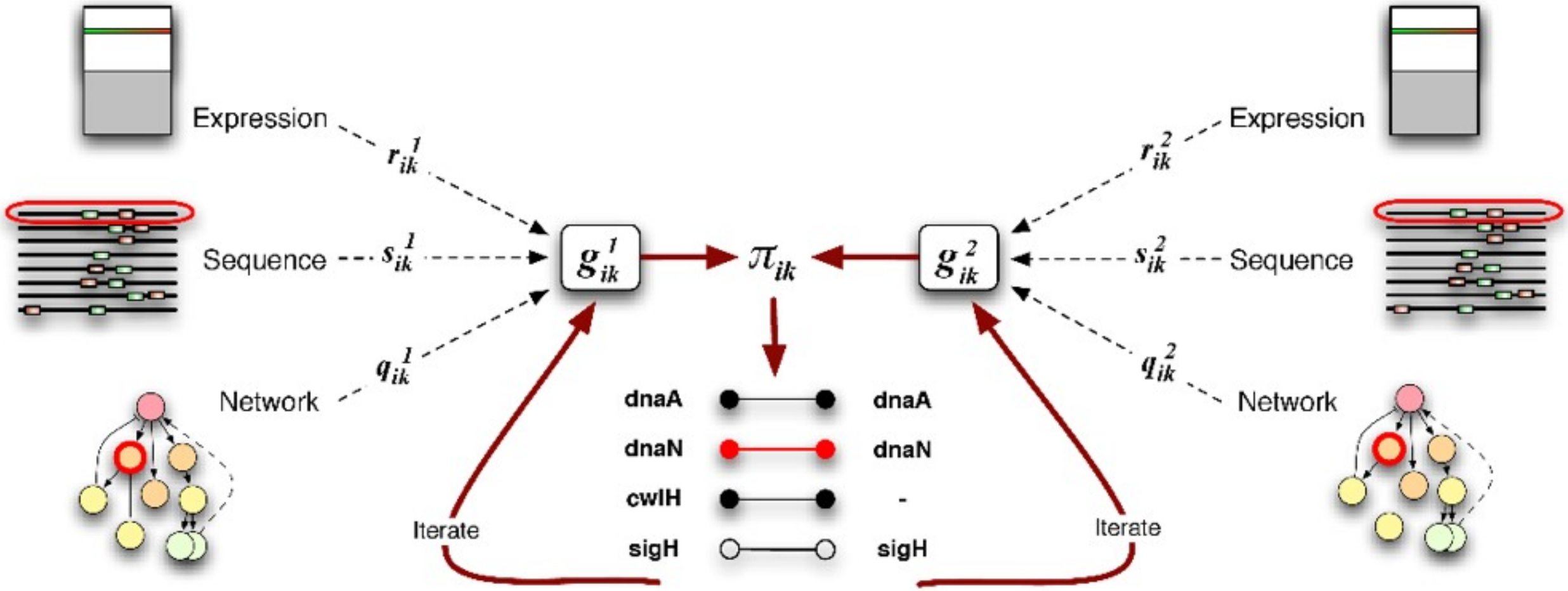
Proposed Multi-species cMonkey model

Algorithm outline:

- Shared-space search: optimize biclusters in OC_{12} space
  - Optimize each OC_{12} bicluster within "species data space"

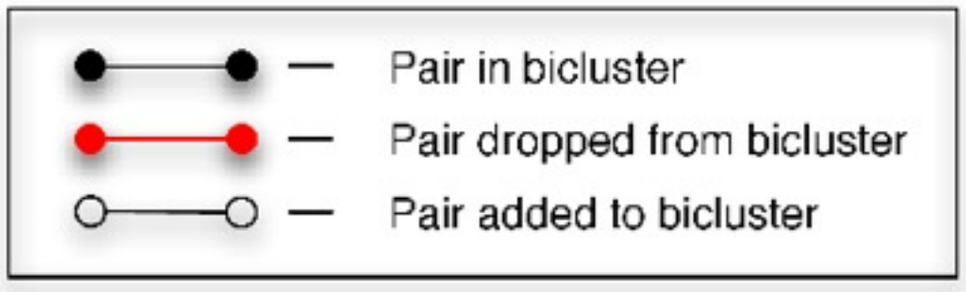
    don't merge data
  - Add/drop a gene-pair from OC_{12} based on evolving single species models
    - *What to do if a gene exhibits correlation to bicluster in one species, but its ortholog in other does not? (**answer coming**)*

Shared Optimization



$$\pi_{ik} \equiv p\left(y_{ik}^1, y_{ik}^2 \mid g_{ik}^1, g_{ik}^2\right) \propto e^{\left(\beta_0 + \beta_1 \left(g_{ik}^1 + g_{ik}^2\right)\right)}$$

$$g_{ik}^j \equiv r_0 \log\left(\tilde{r}_{ik}^j\right) + s_0 \log\left(\tilde{s}_{ik}^j\right) + \sum_{n \in N} q_0^n \log\left(\tilde{q}_{nik}^j\right)$$

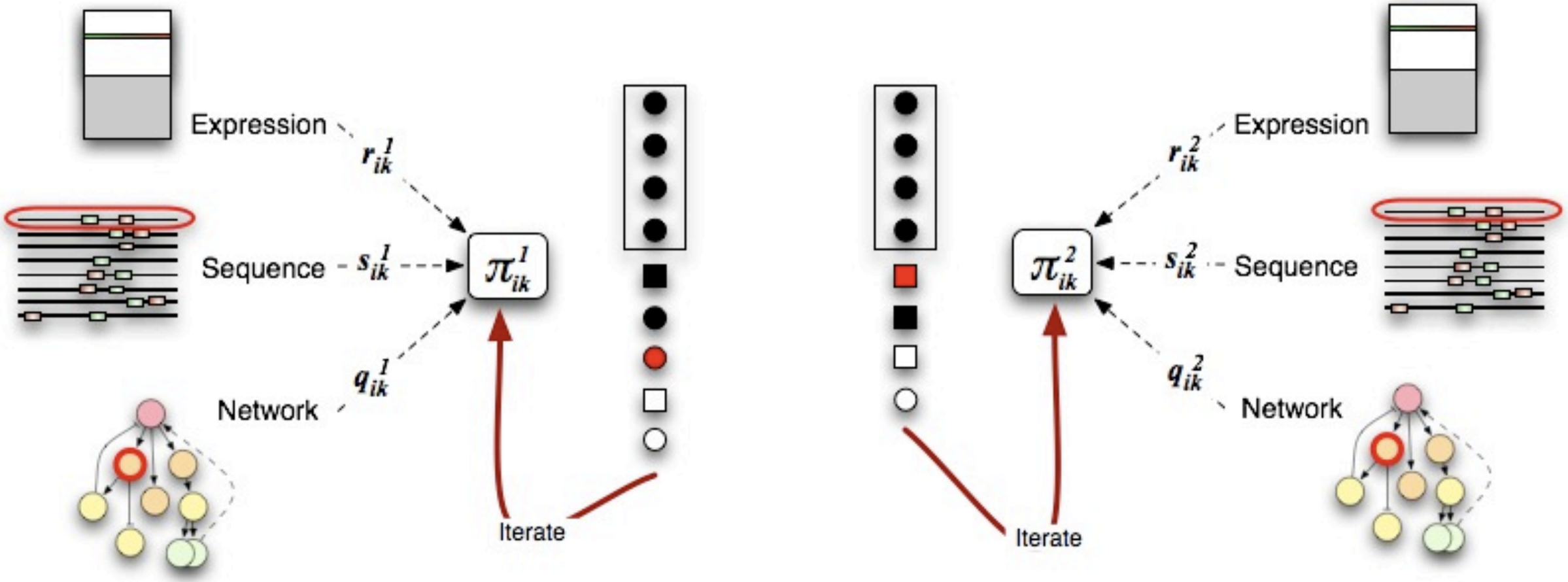


Proposed Multi-species cMonkey model

Algorithm outline:

- Elaborate: optimize OC_{12} biclusters in each organism's "species space"
  - Seed with genes from pairs in the OC_{12} biclusters
  - Use original single-species *cMonkey* to optimize the OC_{12} seeds:
    - Cannot drop genes from original OC_{12} gene-pairs
    - Allow genes from entire genome to be added, i.e. species-specific, "orthologous core" and paralogs.

Elaboration



$$\pi_{ik}^{j} \equiv p\left(y_{ik} \mid \mathbf{X}_{k}^{j}, S_{i}^{j}, \mathbf{N}_{ik}^{j}\right) \propto e^{\left(\beta_0 + \beta_o g_{ik}^{j}\right)}$$

- Orthologous core gene, from original shared bicluster
- Orthologous core gene, in elaborated bicluster
- Orthologous core gene, added to elaborated bicluster
- Orthologous core gene, added to elaborated bicluster
- Species-specific gene in bicluster
- Species-specific gene dropped from bicluster
- Species-specific gene added to bicluster

Proposed Multi-species cMonkey model

Algorithm outline:

- Extend: find new biclusters for G_j in its own "species space"
  - Seed & optimize new clusters following original *cMonkey* single-species model
  - Allow extend step to consider genes from orthologous core (OC)?:
    - Yes (*currently, we allow overlap potential to reduce over-sampling of explored modules*)
    - No (*possible future direction to force identification of species-specific modules*)

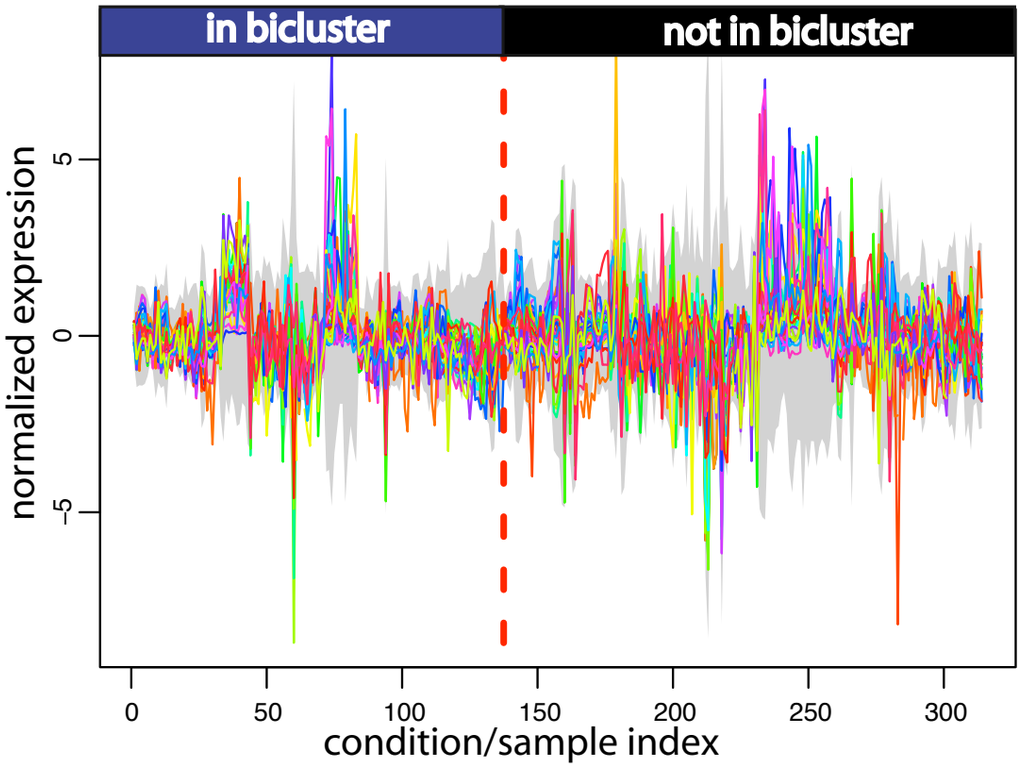
Species Analyzed

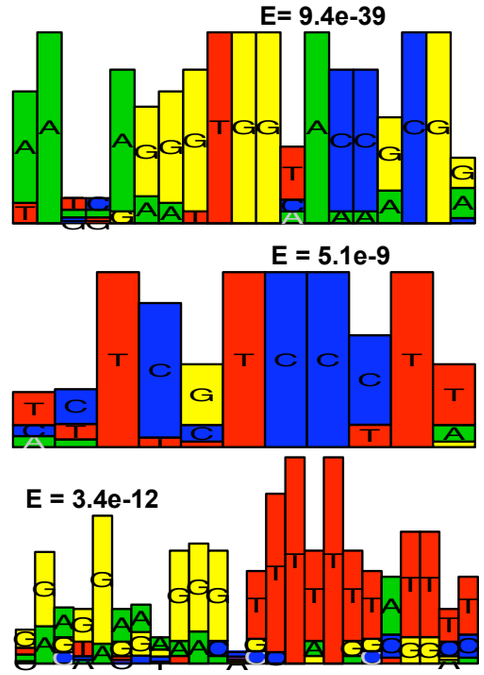
- Compared 3 bacterial species:
  - *Bacillus subtilis*
  - *Bacillus anthracis* (Anthrax)
  - *Listeria monocytogenes* (Listeriosis)
- 3 organisms → 3 pairings
  - Inparanoid to identify orthologs and orthologous families
  - 150 biclusters generated per pairing

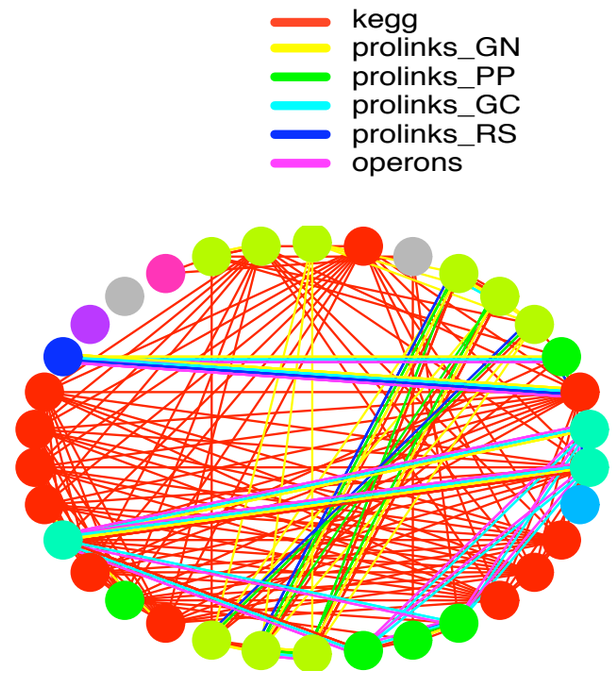
Number of:

	B. subtilis* – *B. anthracis	***B. subtilis* – *L. monocytogenes***	***B. anthracis* – *L. monocytogenes***
orthologous groups	2225	1439	1494
orthologous pairs	2443	1564	1690
unique genes (per organism)	***B. subtilis***: 2279/3928	***B. subtilis***: 1519/3928	***B. anthracis***: 1634/5865
	B. anthracis: 2339/5865	***L. mono***: 1478/2795	***L. mono***: 1537/2795

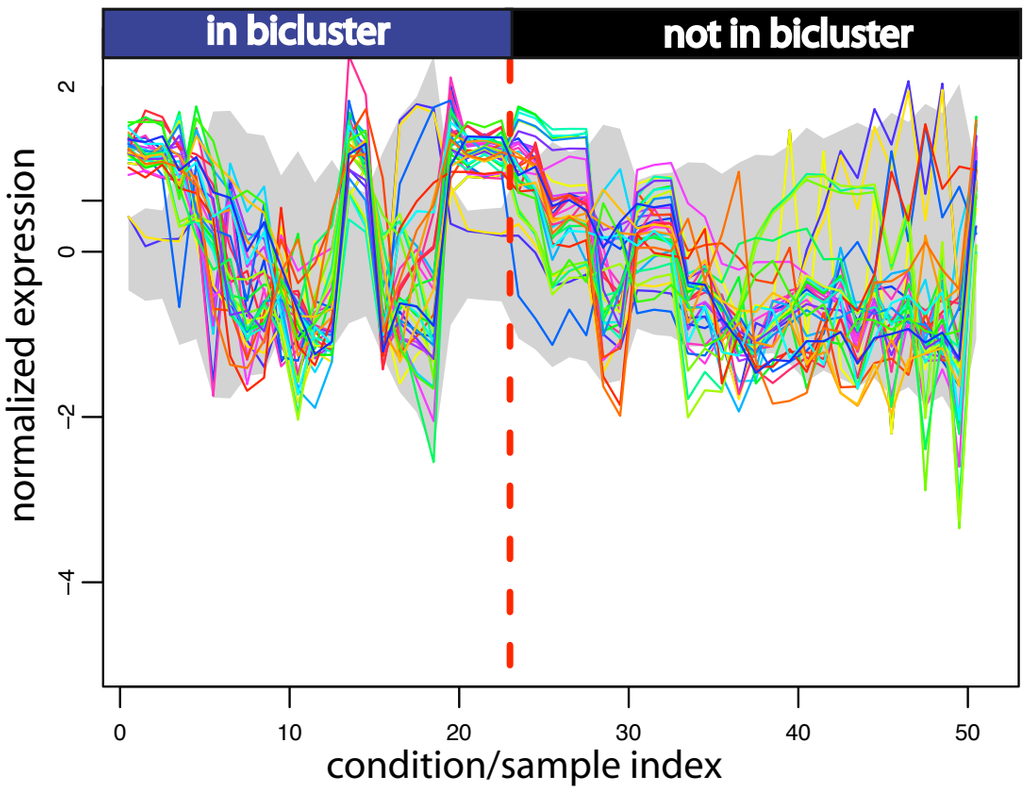
A. *B. subtilis:*

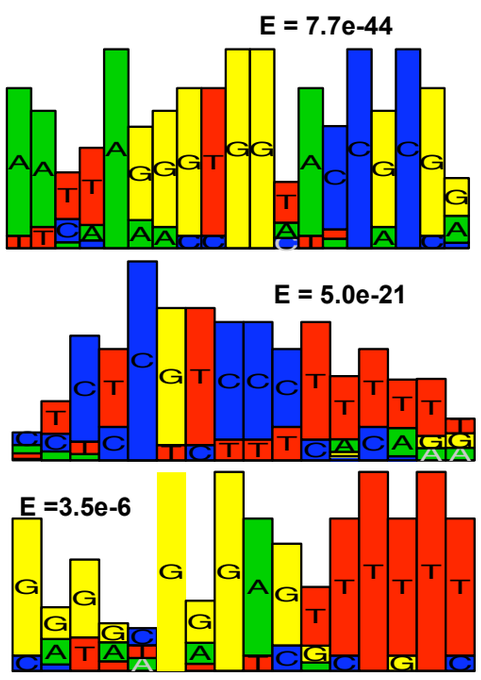


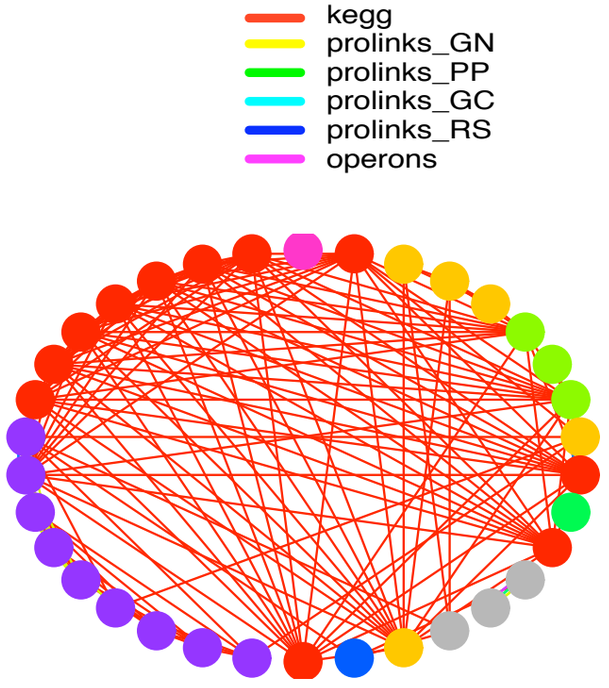




B. *B. anthracis:*





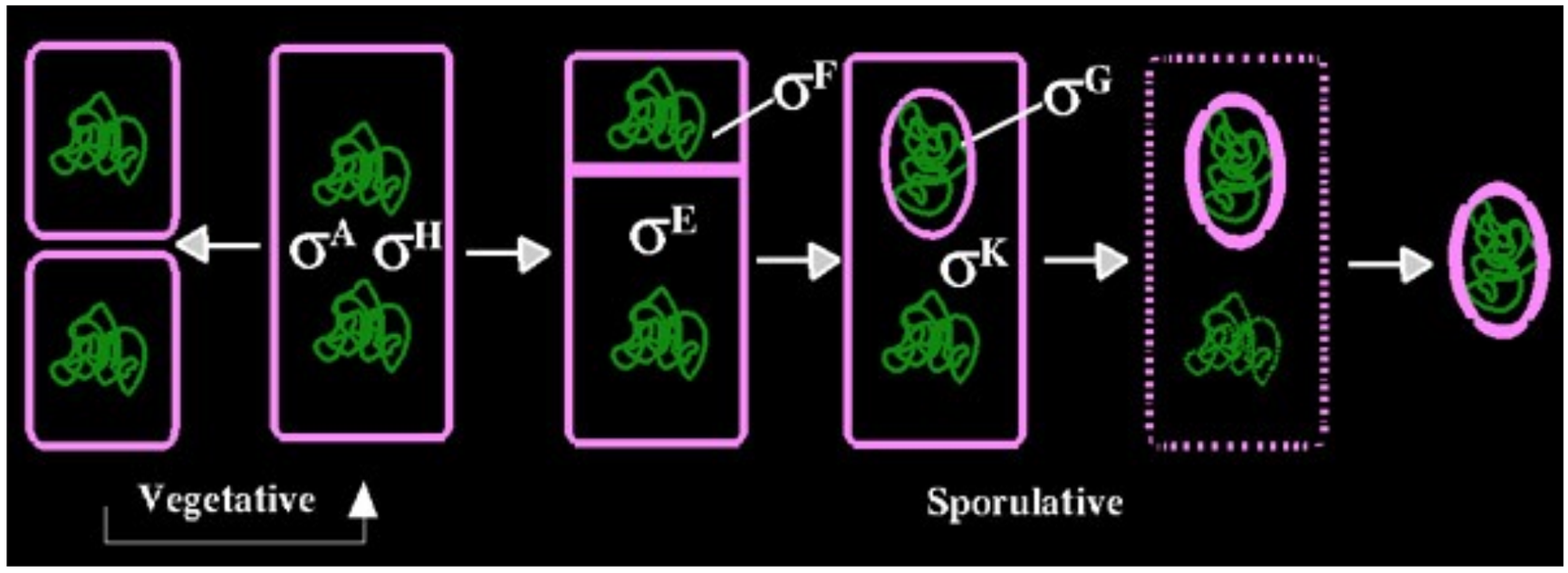


PETER WALTMAN

σ^E biclusters
(*B. subtilis – B. anthracis*)

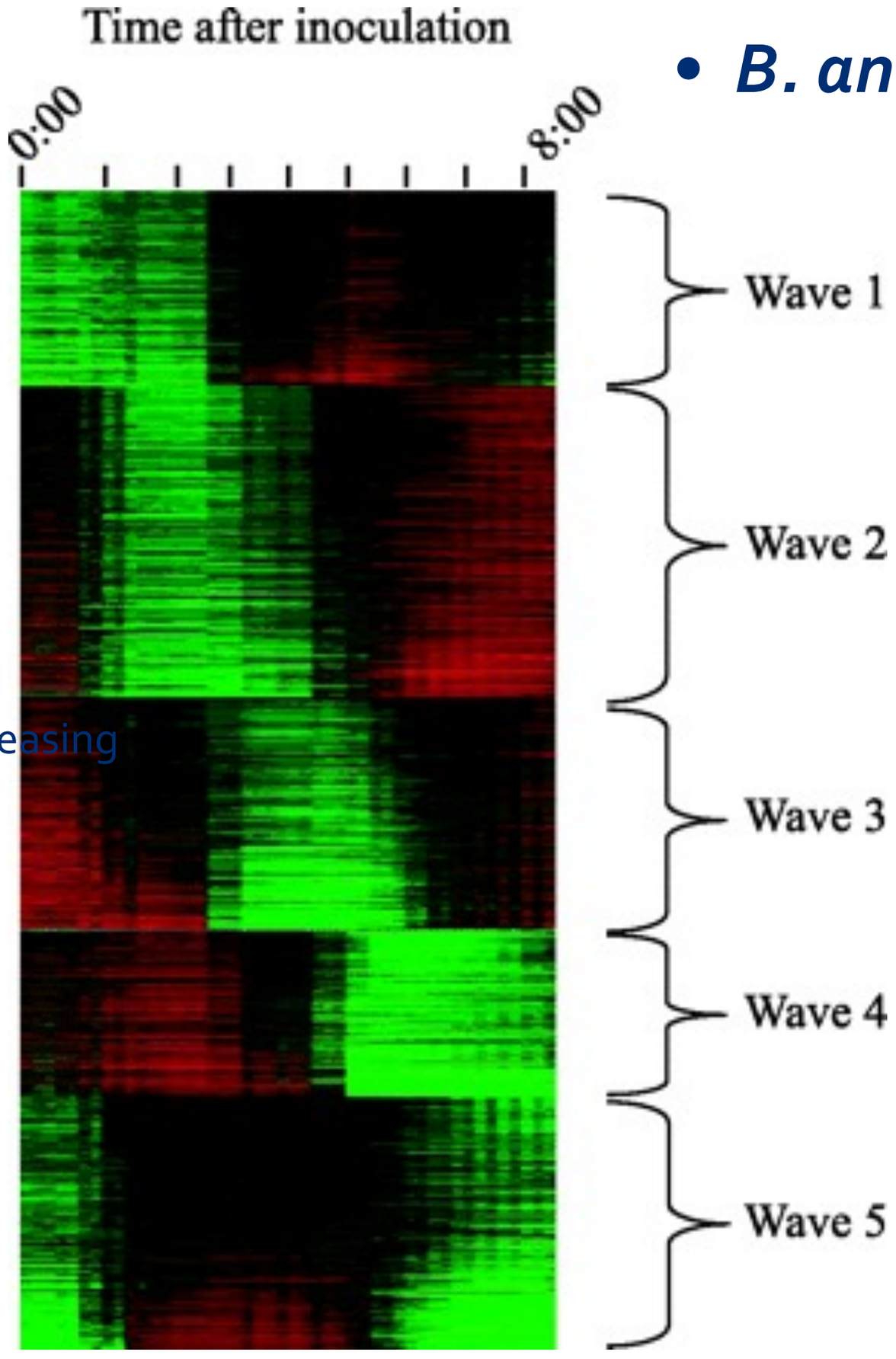
Significantly enriched for sporulation genes (σ^E regulated):

- Bicluster 17 (includes Metabolism, Glutamine Transport, Transporters genes)
- Bicluster 35 (includes Metabolism, Glutamine Transport, Detoxification, Transporters genes)
- Bicluster 84 (includes Metabolism, Glutamine Transport genes)



http://biology.kenyon.edu/courses/biol114/Chap11/spore_cycle.jpg

• *B. anthracis* Waves of Gene Expression



•Germination and early outgrowth

•Rapid Growth

•Rapid Growth and Responding to increasing toxic environment

•Sporulation and Oxidative Stress

•Sporulation and early germination and outgrowth

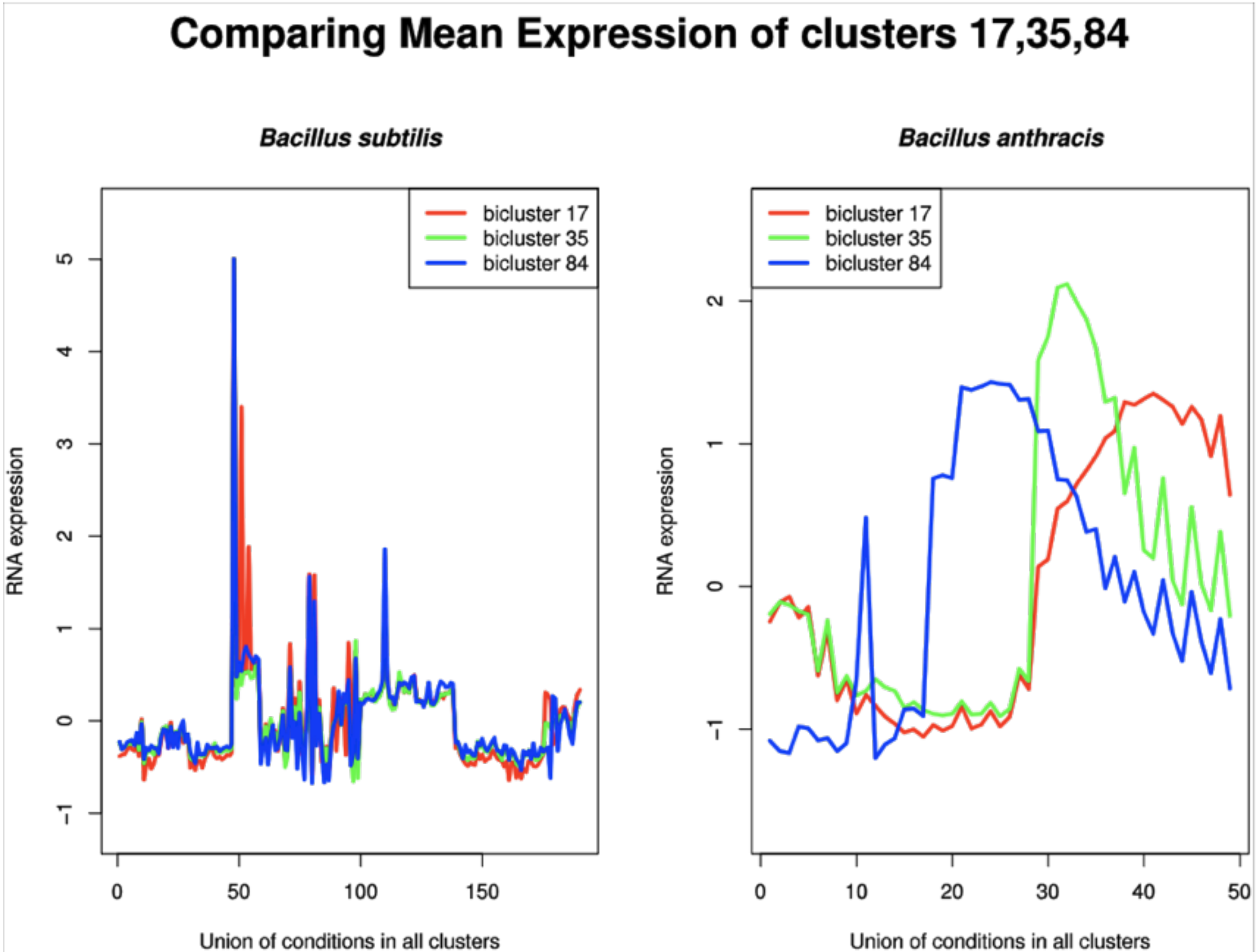
•Biclusters 84 : 3 into 4

•Biclusters 35: 4

•Biclusters 17 : 4 into 5

•(Bergmen *et al.*, 2006)

Sporulation Biclusters

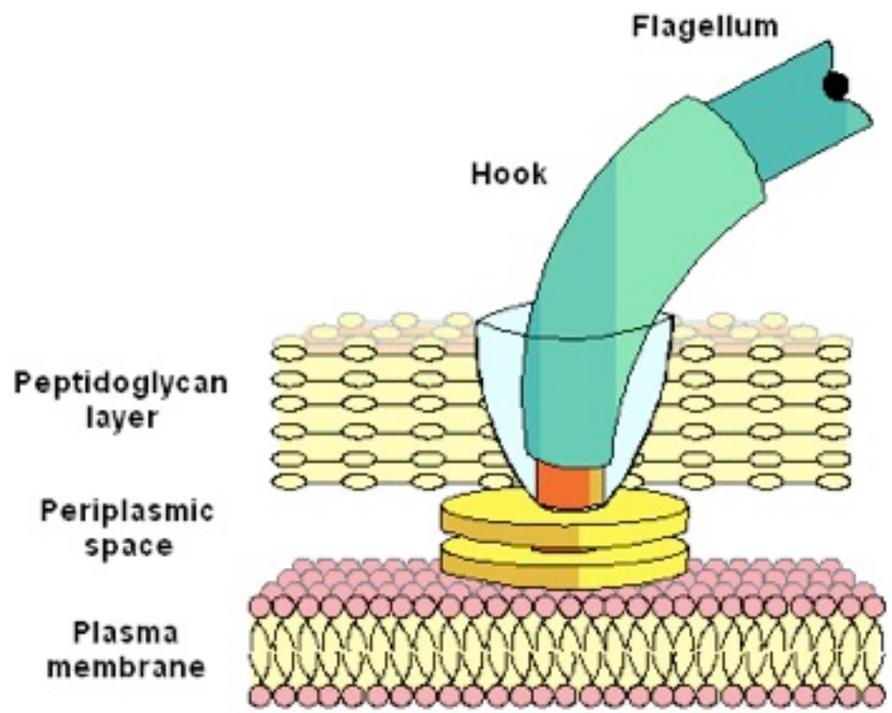


Flagellar Assembly Biclusters

- Flagellar Assembly biclusters for all 3 organisms
- *B. anthracis* thought to be non-motile:
  - Missing σ^D (flagellar TF in *B. subtilis*)
  - frameshifts to 4 critical flagellar genes:
    - *cheA*
    - *flgL*
    - *fliF* (MS ring)
    - *fliM* (C ring component)
- *B. anthracis* biclusters also enriched for:
  - Chemotaxis
  - Type III secretion system*

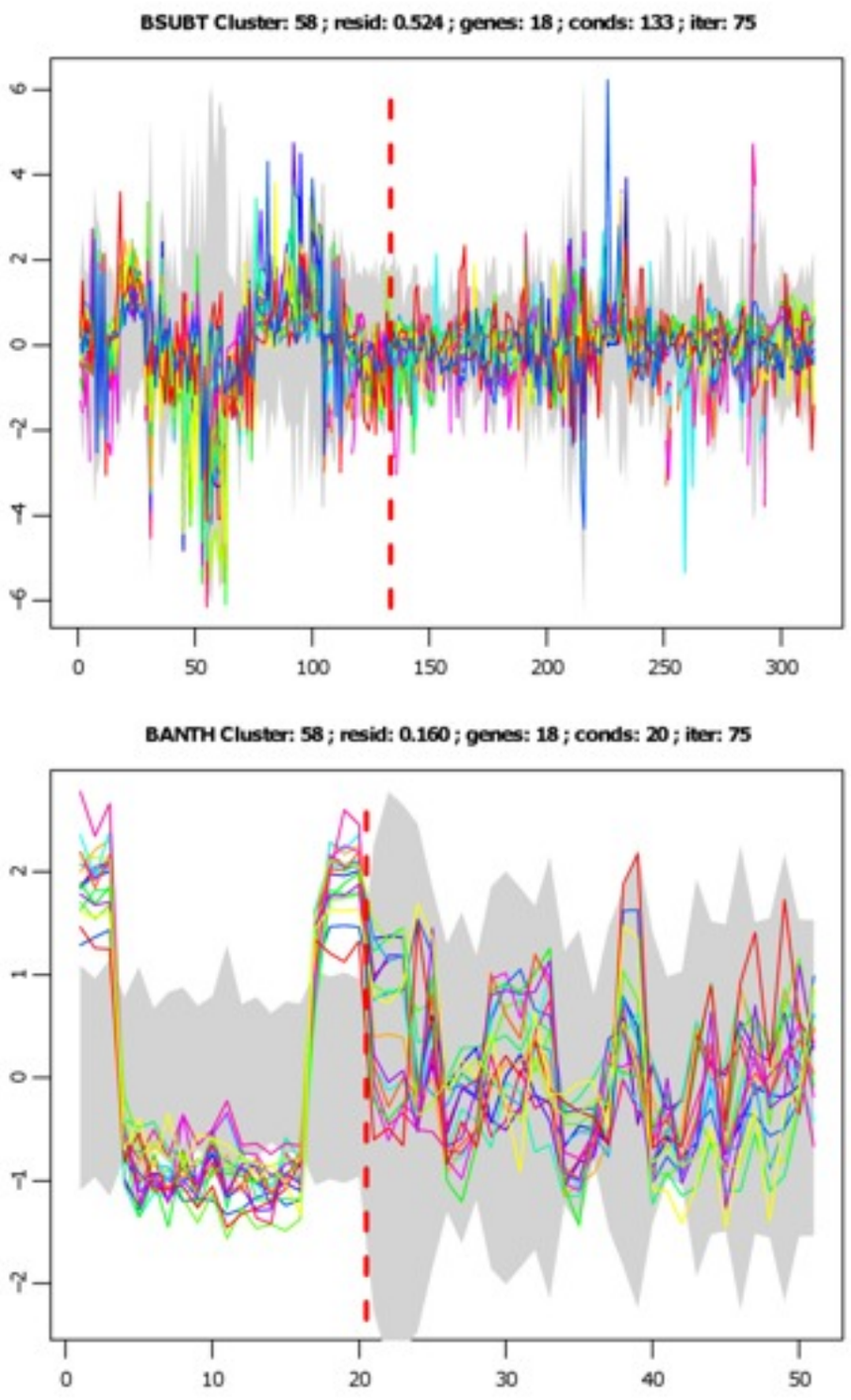
* *B. subtilis – B. anthracis* only

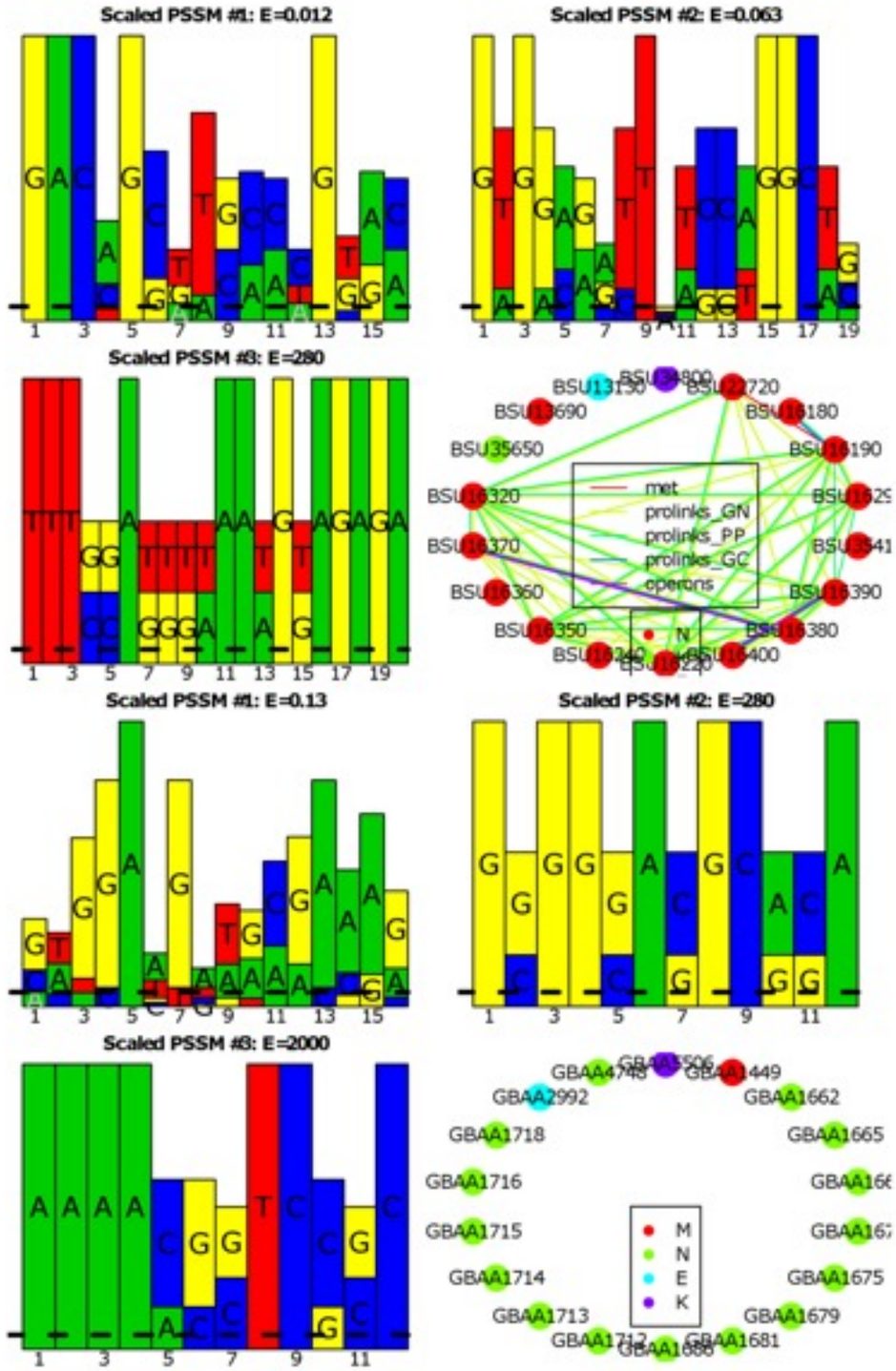
Ultrastructure of Gram-Positive Bacteirial Flagella



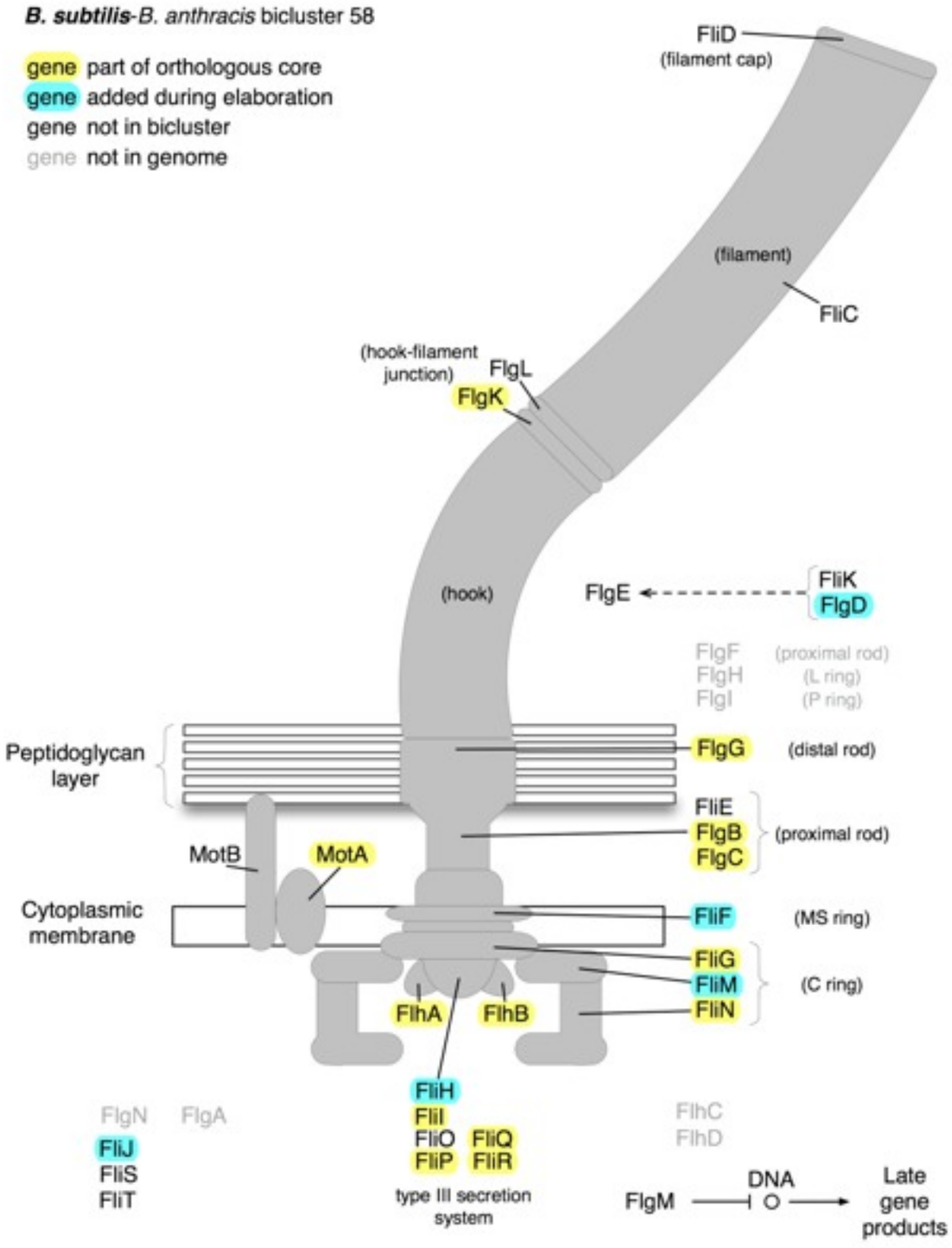
http://www.conceptdraw.org/sampletour/medical/GPositiveBFlagella.gif

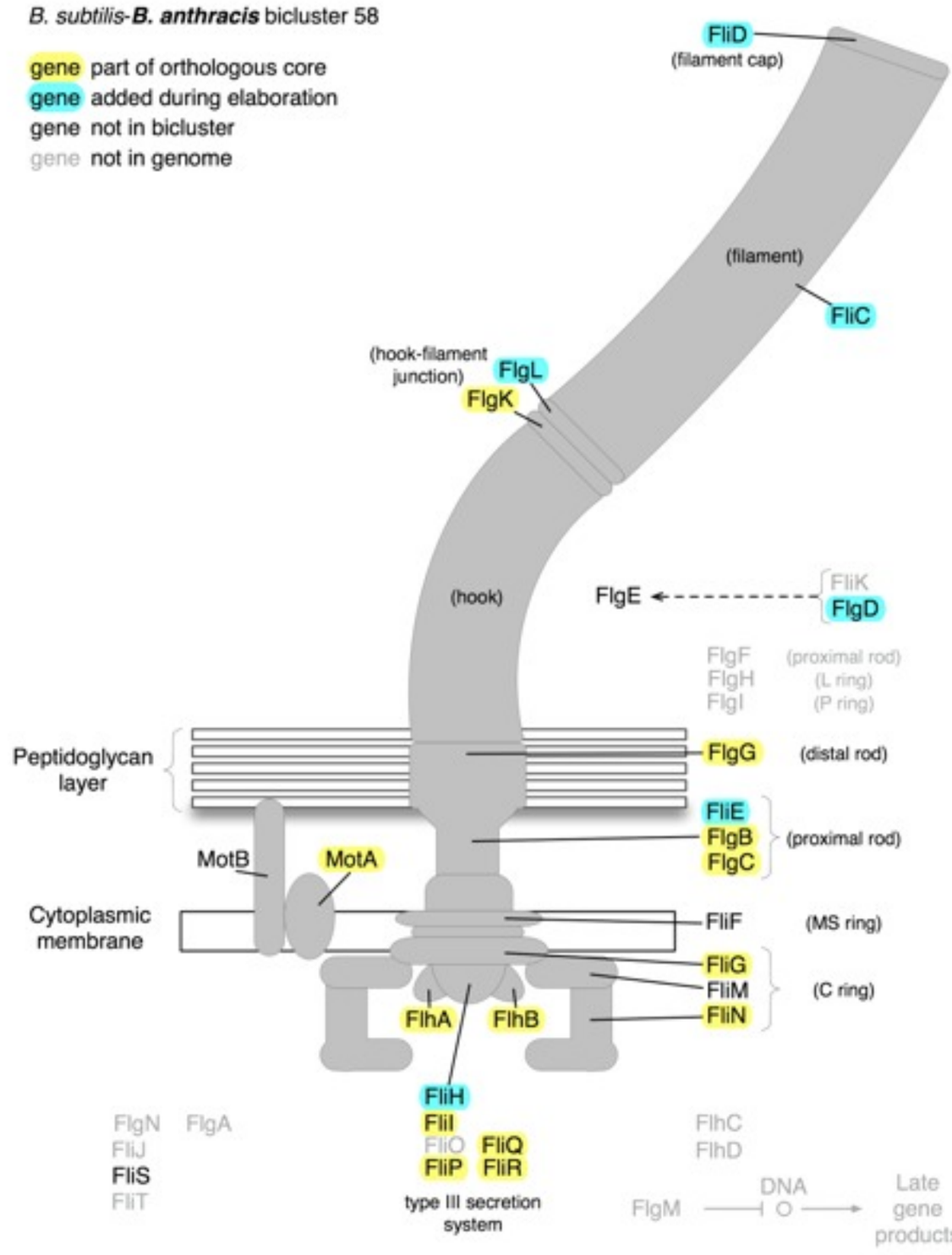
Shared *B. subtilis - B. anthracis* Flagellar Assembly Bicluster





Elaborated *B. subtilis - B. anthracis* Flagellar Assembly Bicluster





Globally Validating Multi-Species method

- Issues
  - No solved organism as validation - only partial solutions available
  - Large number of results (12) to validate:
    - 3 organism-pairs → 6 results (2 for each pair)
    - 2 steps (shared & elaboration opt's) → 12 total
  - No existing metric for measuring quality & conservation
    - Could we use either DCA or ECC for a metric?
      - DCA not a genuine clustering method & no metric provided
      - ECC gave inconsistent results in our own tests

How to measure or compare conservation & quality?

- Conservation metric
- Compare Shared & Elaboration optimizations with biclusters from *ideal* single-species *cMonkey*
  - Expression *(residuals)*
  - Networks *(association p-values)*
  - Sequence:
    - *Motif E-values*
    - *Sequence p-values*
  - Enrichments of:
    - GO terms
    - KEGG pathways

New Conservation Metric

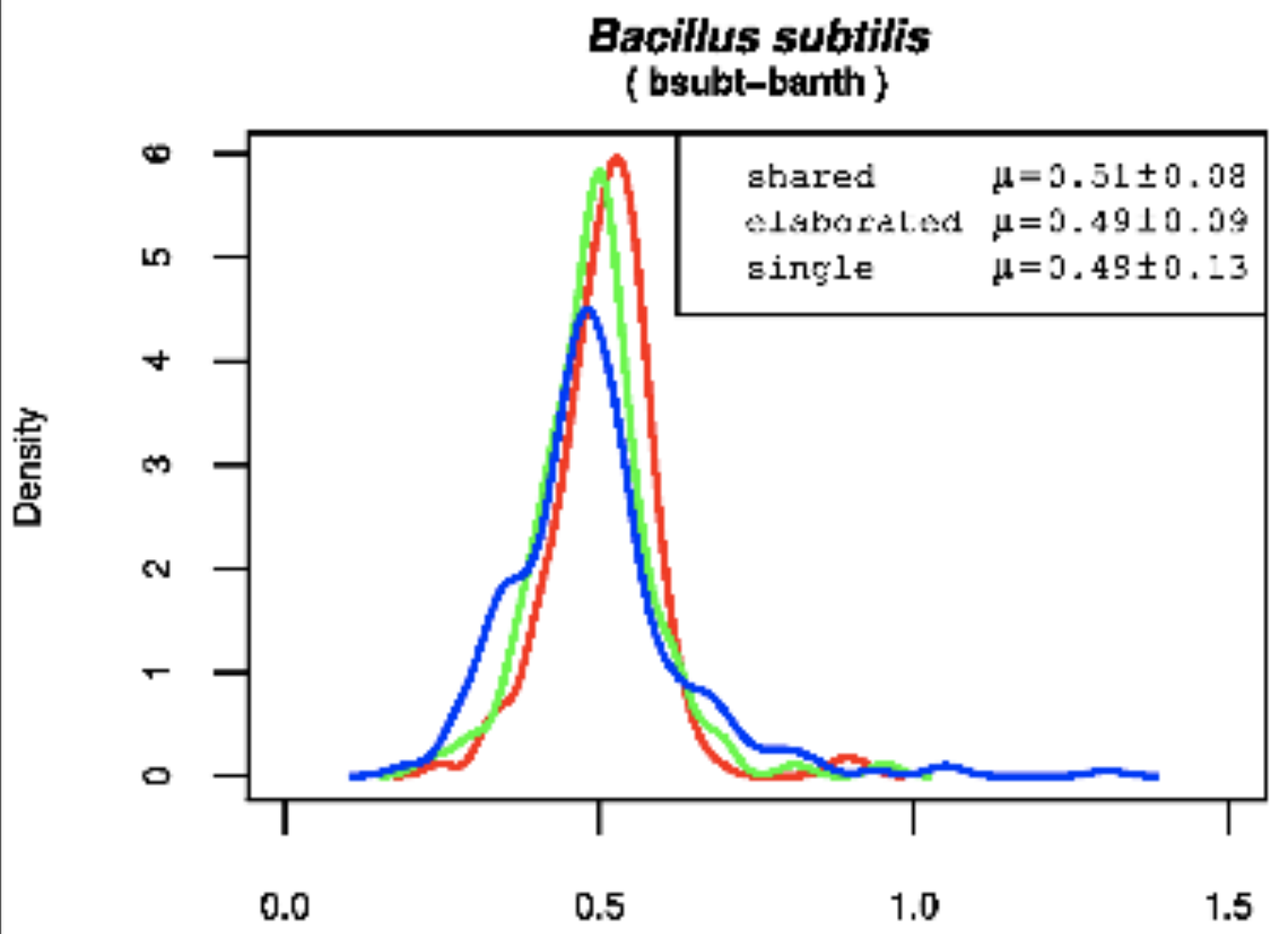
$$Cons\left(bicl_i^1, bicl_j^2\right) \equiv \frac{\left|matches\left(oc\left(bicl_i^1\right) oc\left(bicl_j^2\right)\right)\right|}{\left|oc\left(bicl_i^1\right)\right|}$$

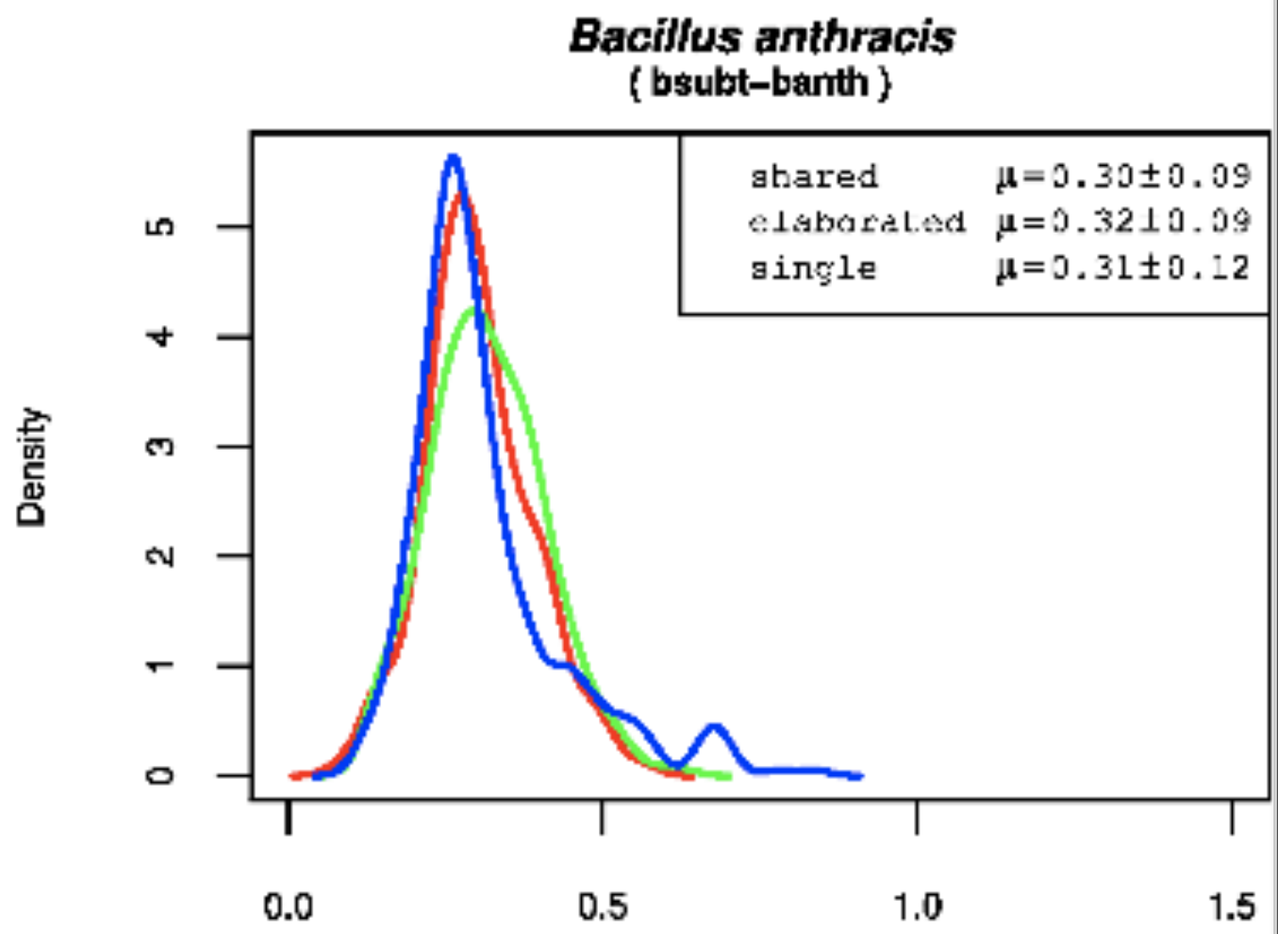
Find $max\left(Cons\left(bicl_i^1, bicl_j^2\right)\right)$ for each bicluster and average over all biclusters

Conservation Metric

	B. subtilis-*B. anthracis*	*B. subtilis*-*L. monocytogenes*	*B. anthracis*-*L. monocytogenes*
Single	**0.218**	**0.235**	**0.177**
Elaborated	**0.825**	**0.883**	**0.922**
Shared	**1**	**1**	**1**

Residuals





Shared

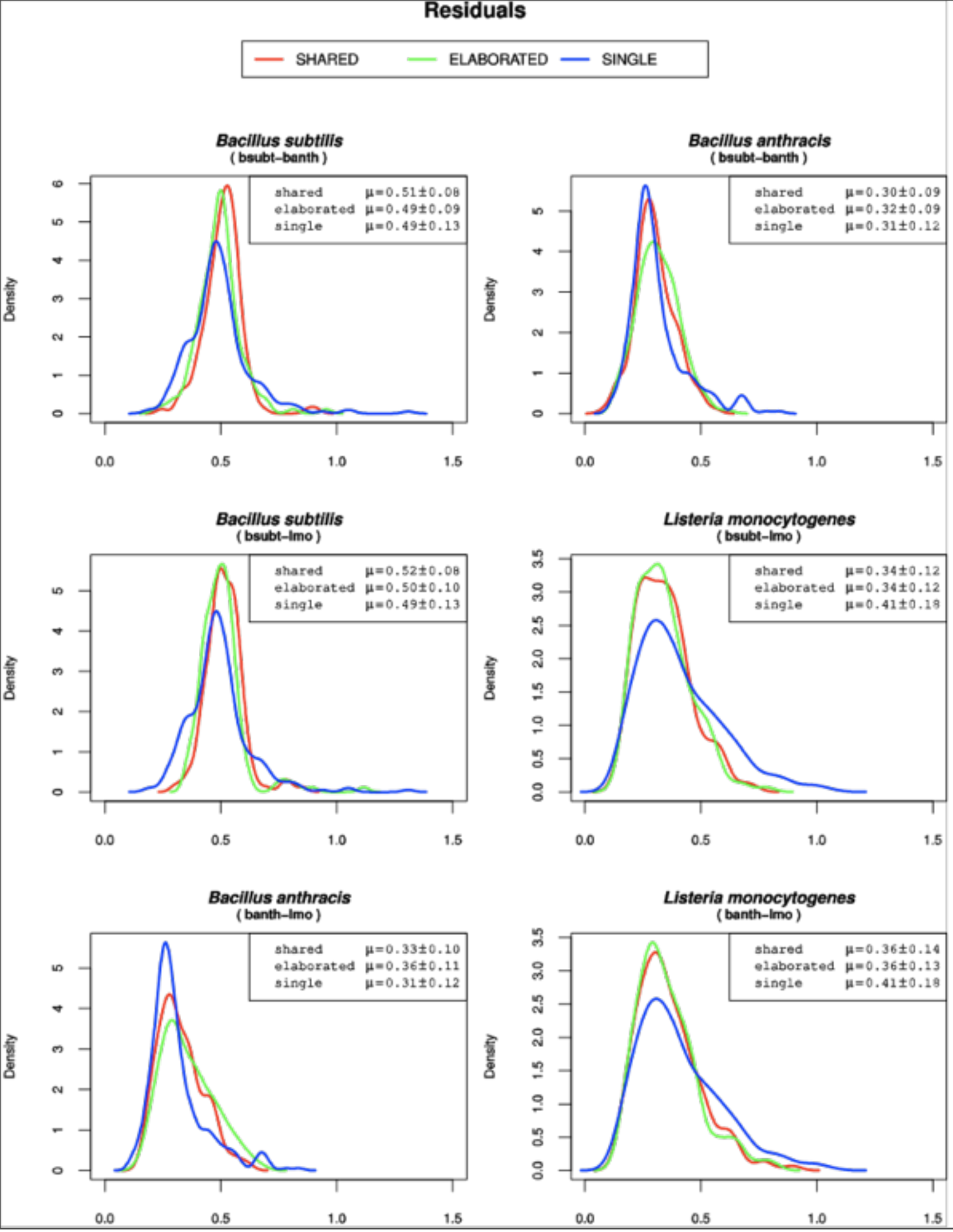
Elaborated

Single

P-values from two-sided Wilcoxen's rank test (α=0.01):

B. subtilis - B. anthracis

	B. subt.	*B. anth.*
Shared-Single	**1.03E-03**	**0.46**
Elaborated-Single	**0.27**	**0.04**



$$resid(I,J) = \frac{\frac{1}{|I||J|}\sum_{i\in I, j\in J,} abs\left(x_{ij} - x_{iJ} - x_{Ij} + x_{IJ}\right)}{\frac{1}{|I|}\sum_{i\in I} var_j(x_i)}$$

P-values from two-sided Wilcoxen's rank test (α=0.01):

B. subtilis - B. anthracis

	B. subt.	*B. anth.*
Shared-Single	1.03E-03	0.46
Elaborated-Single	0.27	0.04

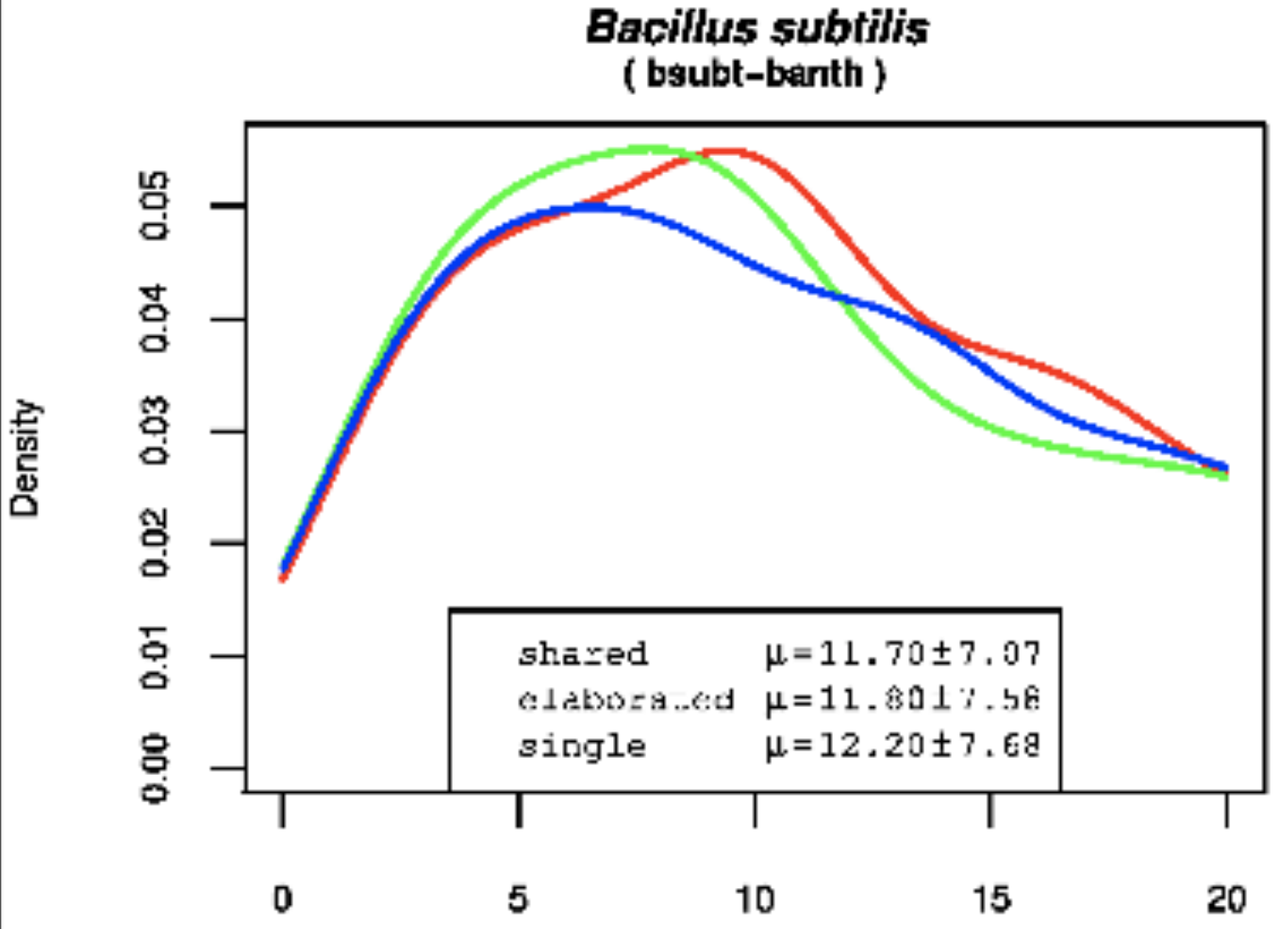
B. subtilis - L. monocytogenes

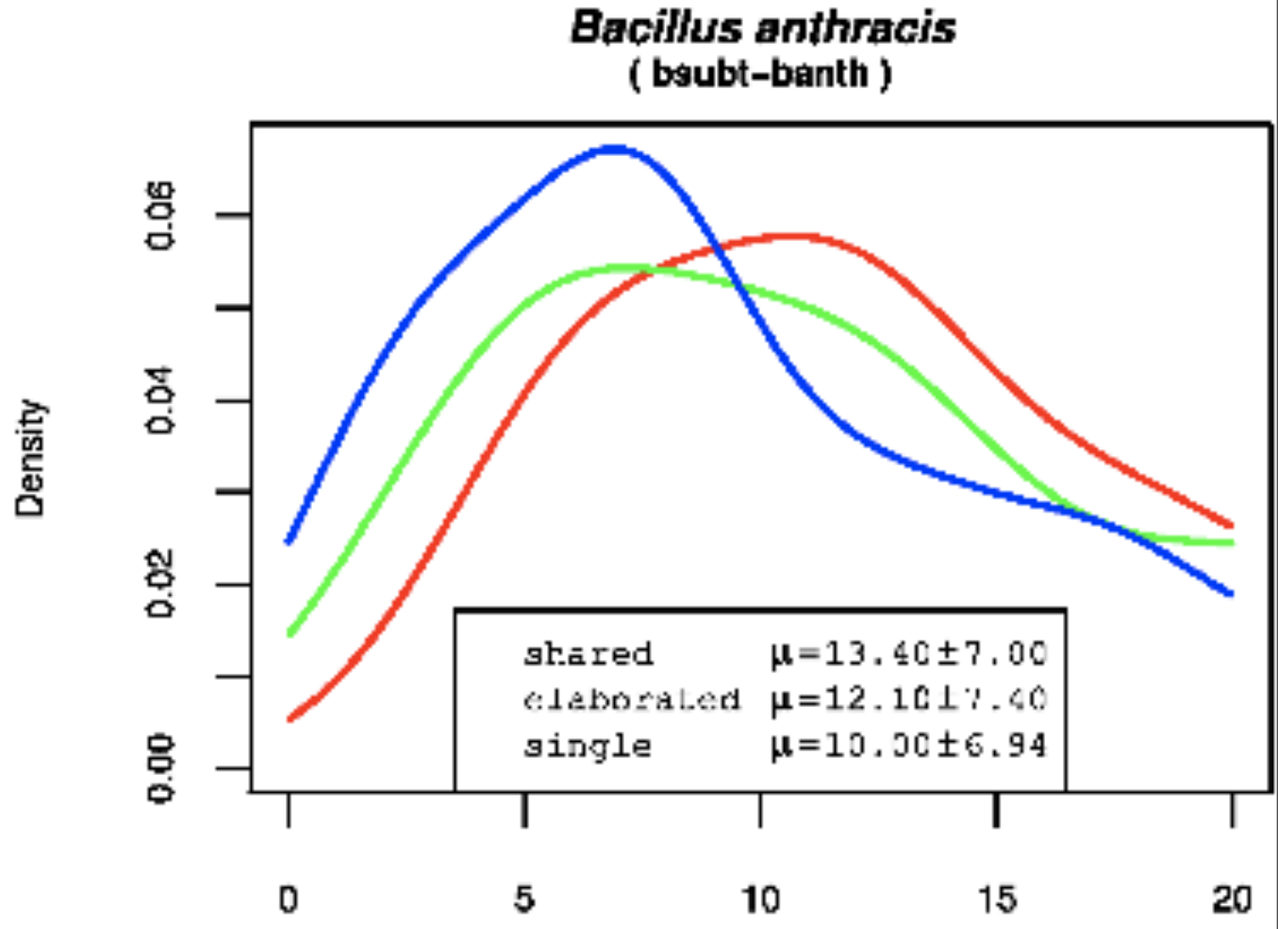
	B. subt.	*L. mono.*
Shared-Single	2.82E-04	7.81E-04
Elaborated-Single	0.09	3.98E-04

B. anthracis - L. monocytogenes

	B. anth	*L. mono*
Shared-Single	0.01	0.02
Elaborated-Single	1.68E-06	0.01

Association p-values (-log10)





Shared

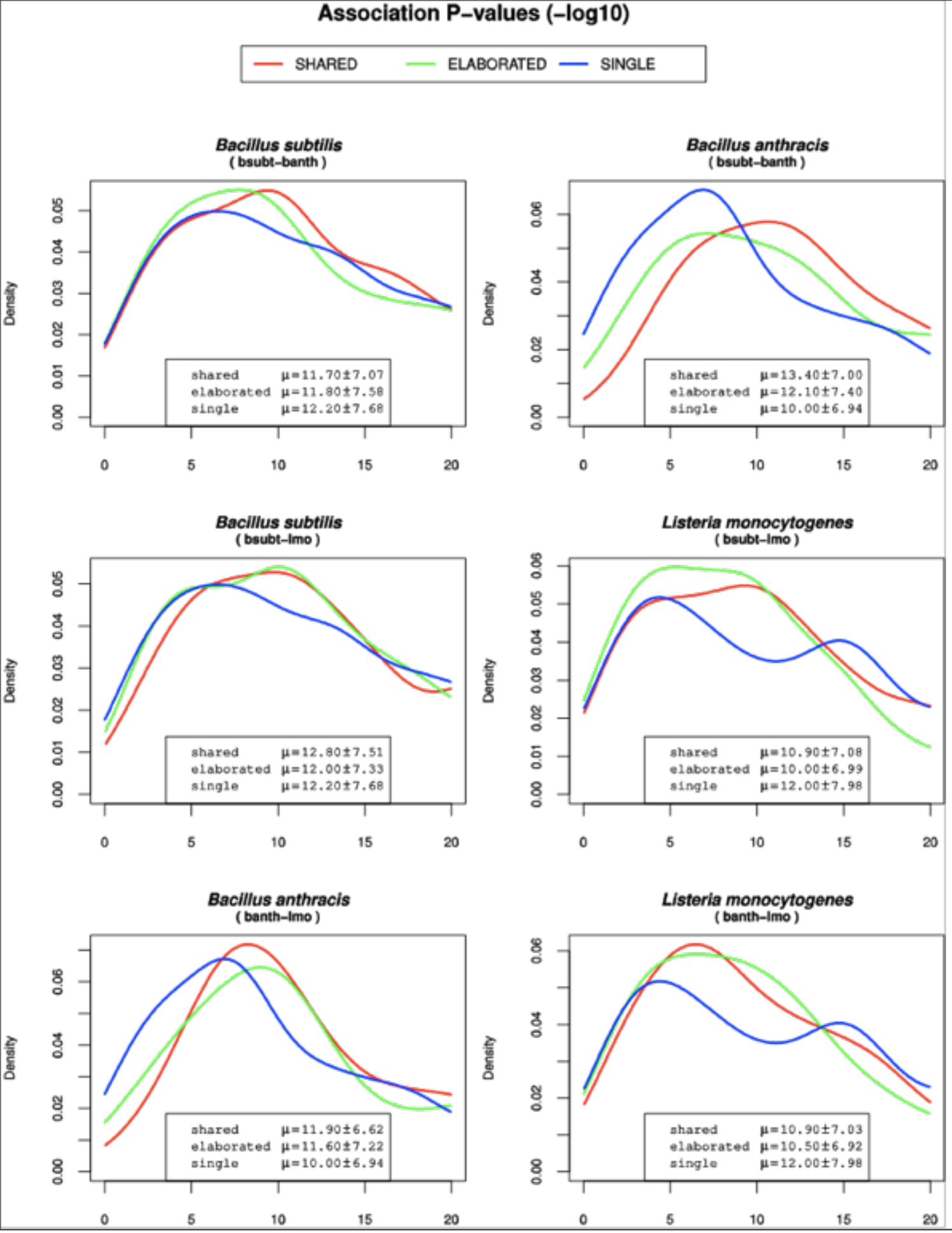
Elaborated

Single

P-values from two-sided Wilcoxen's rank test (α=0.01):

B. subtilis - B. anthracis

	B. subt.	*B. anth.*
Shared-Single	**0.40**	**1.52E-05**
Elaborated-Single	**0.34**	**0.01**



$$pvalue(b_k, N) = \frac{\binom{|N|}{n_{b_k \to b_k}} \binom{poss(G) - |N|}{poss(b_k) - n_{b_k \to b_k}}}{\binom{poss(G)}{poss(b_k)}}$$

P-values from two-sided Wilcoxen's rank test (α=0.01):

B. subtilis - B. anthracis

	B. subt.	*B. anth.*
Shared-Single	**0.40**	**1.52E-05**
Elaborated-Single	**0.34**	**0.01**

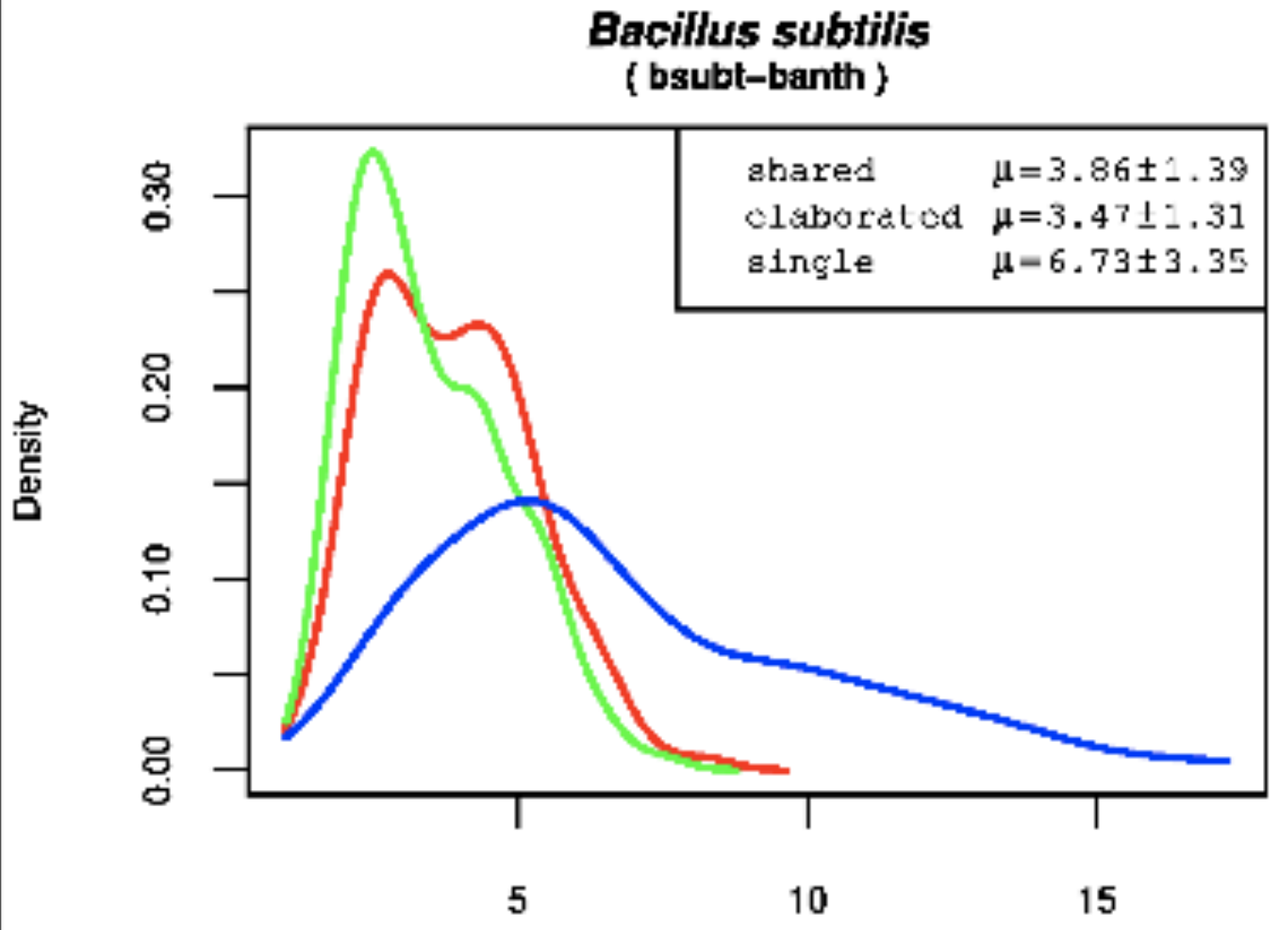
B. subtilis - L. monocytogenes

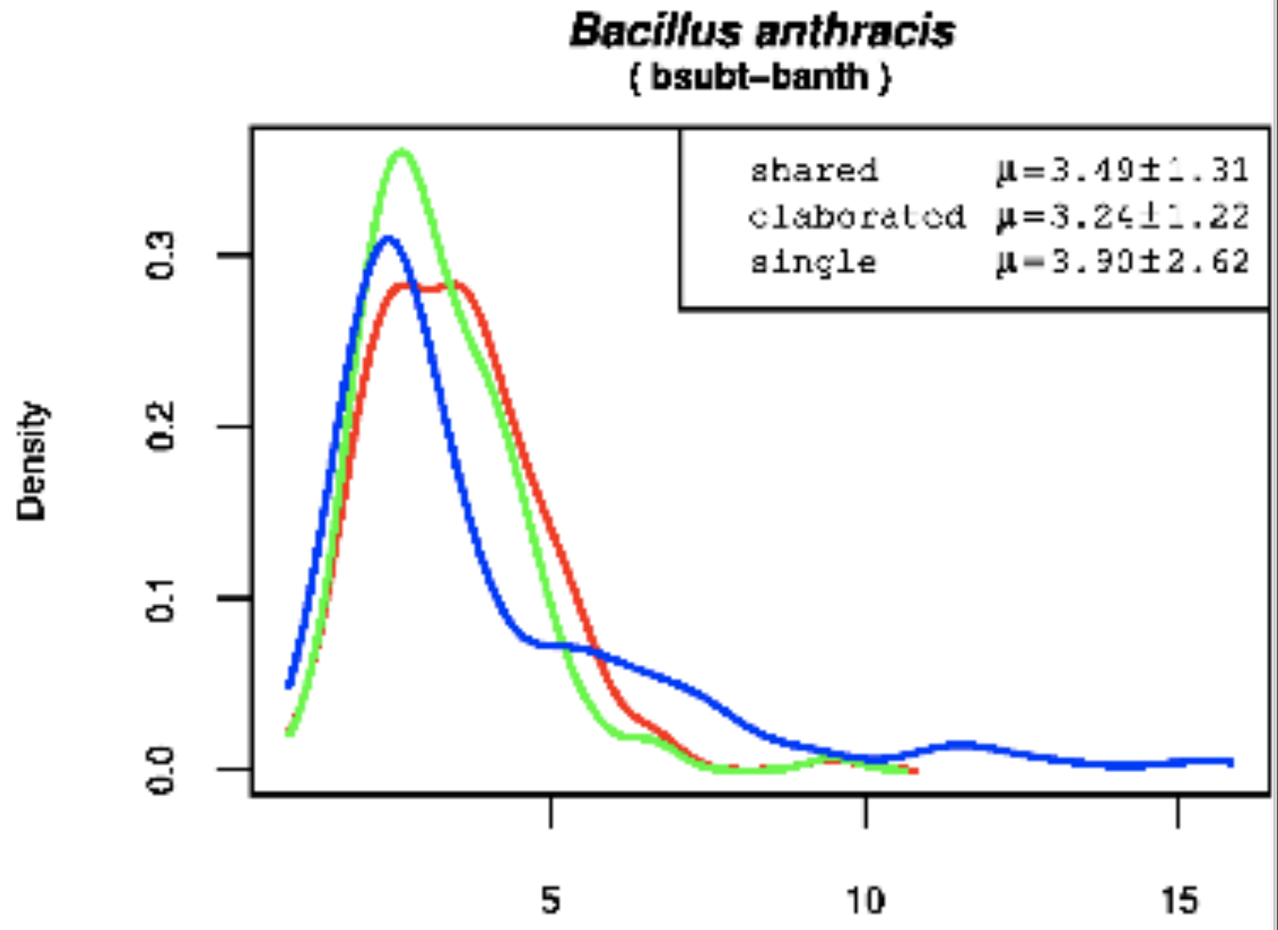
	B. subt.	*L. mono.*
Shared-Single	**0.28**	**0.14**
Elaborated-Single	**0.73**	**1.64E-03**

B. anthracis - L. monocytogenes

	B. anth	*L. mono*
Shared-Single	**0.01**	**0.18**
Elaborated-Single	**0.03**	**0.03**

Sequence p-values (-log10)





Shared

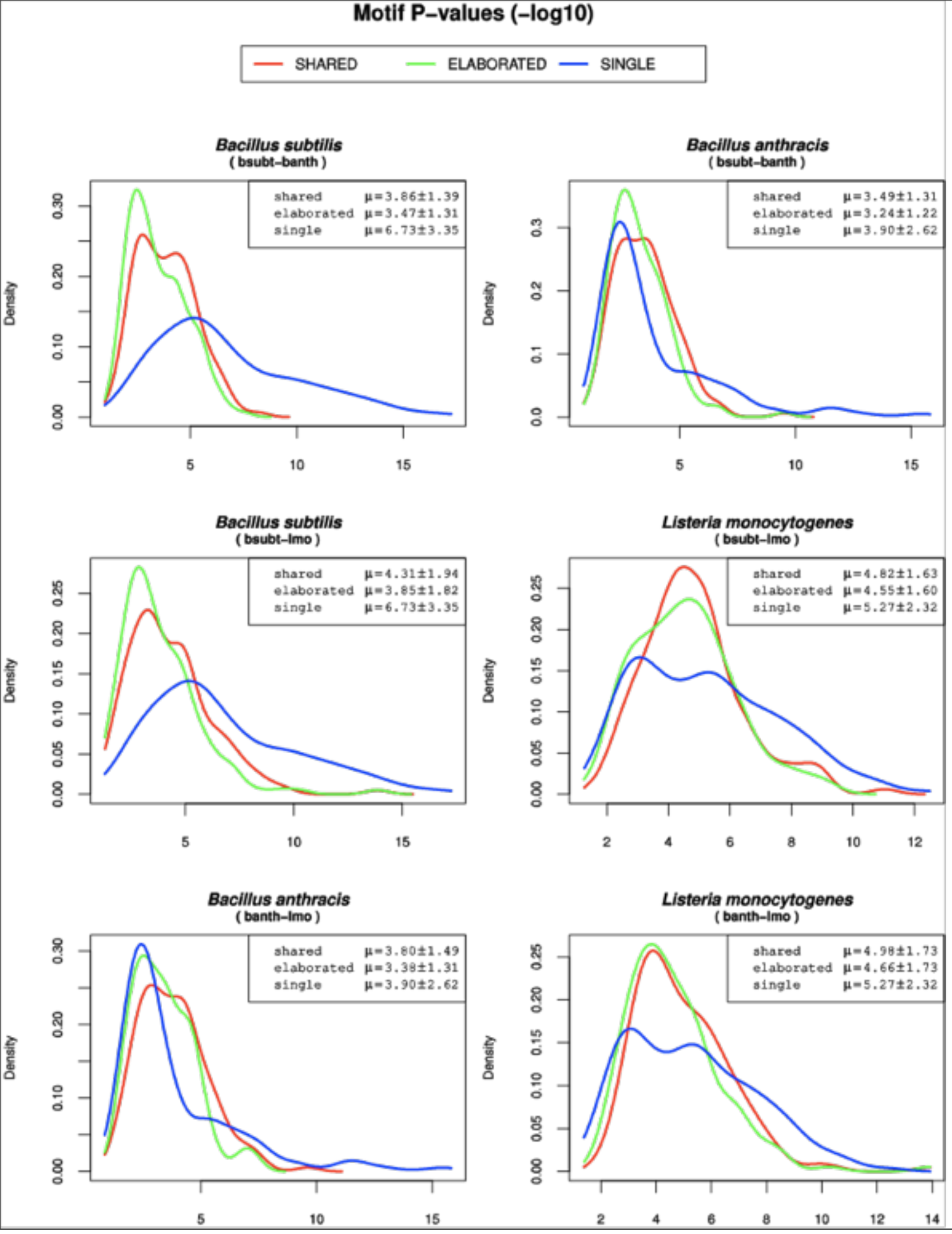
Elaborated

Single

P-values from two-sided Wilcoxen's rank test (α=0.01):

B. subtilis - B. anthracis

	B. subt.	*B. anth.*
Shared-Elaborated	**0.01**	**0.04**
Shared-Single	**1.42E-22**	**0.85**
Elaborated-Single	**1.70E-29**	**0.36**



$$p(seq) = \prod_{m \in motifs} p(seq \mid mod\ el_m)$$

$$p(seq \mid mod\ el) = \prod_{i=1}^{L} q_i$$

P-values from two-sided Wilcoxen's rank test (α=0.01)

B. subtilis - B. anthracis

	B. subt	*B. anth*
Shared-elab	0.01	0.04
Shared-single	1.42E-22	0.85
Elaborated-single	1.70E-29	0.36

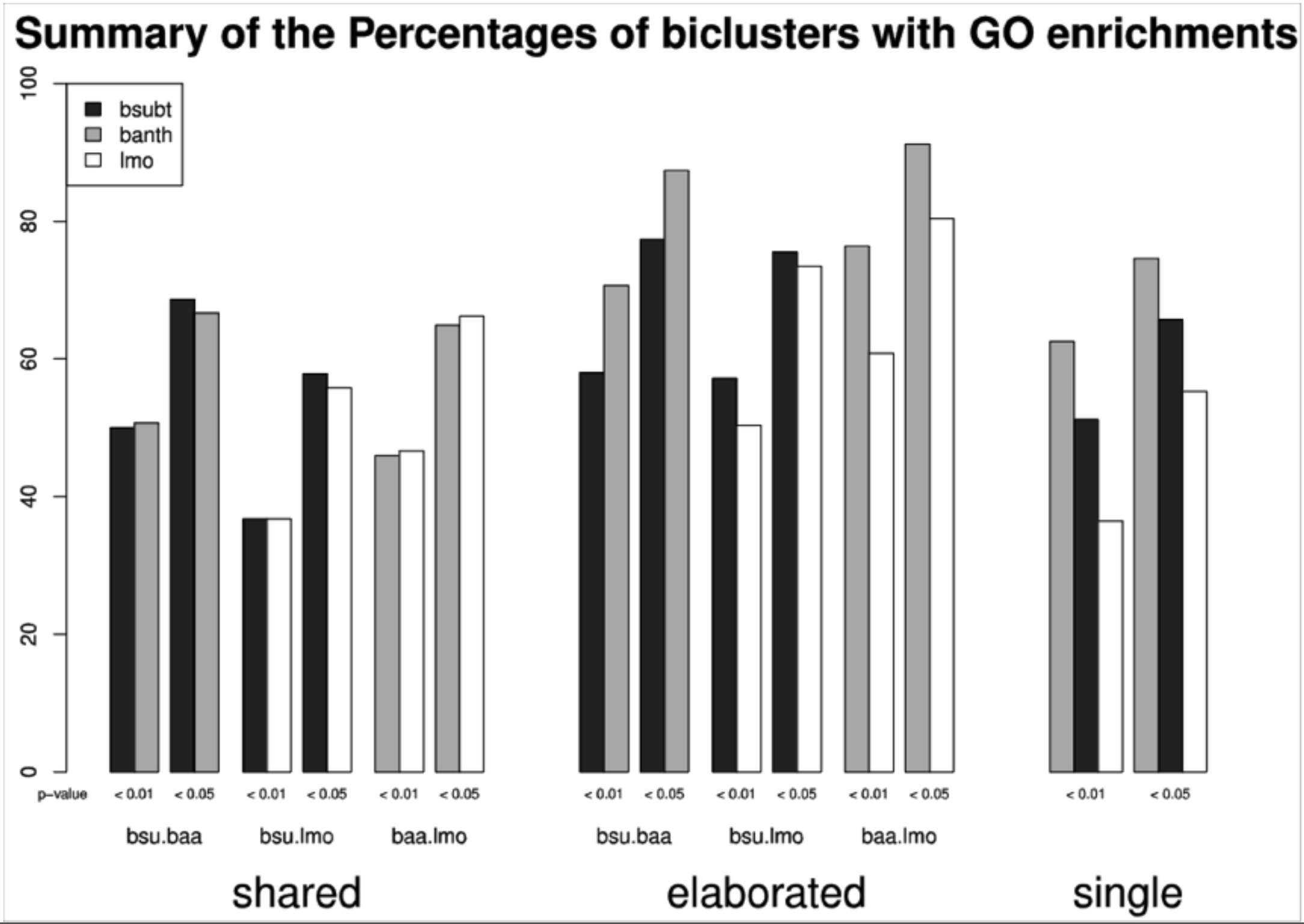
B. subtilis - L. monocytogenes

	B. subt	*L. mono*
Shared-elab	0.01	0.1
Shared-single	3.37E-15	0.07
Elaborated-single	7.36E-23	9.23E-03

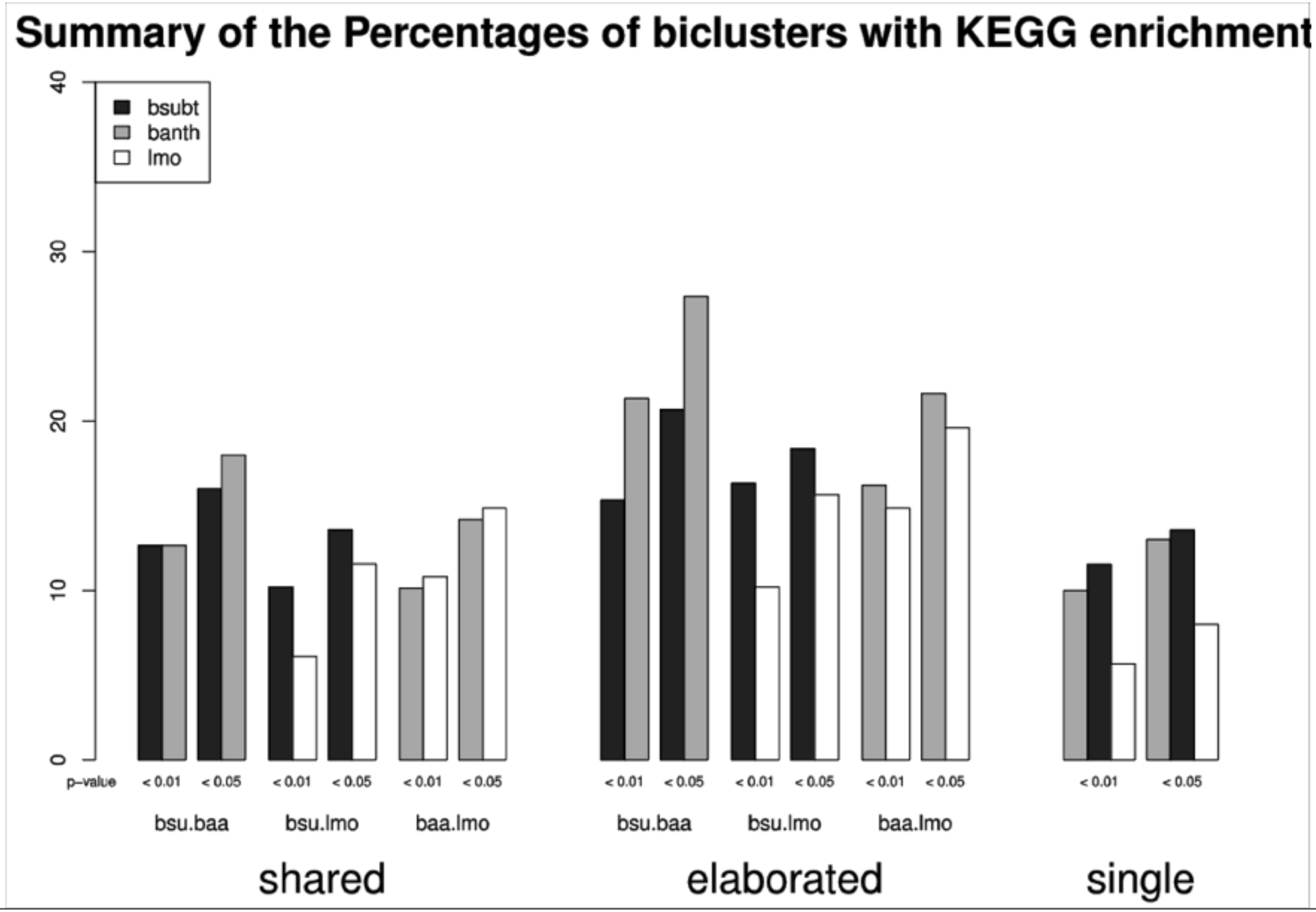
B. anthracis - L. monocytogenes

	B. anth	*L. mono*
Shared-elab	0.01	0.03
Shared-single	0.99	0.24
Elaborated-single	0.55	0.01

Multi-species retrieves more biologically meaningful results



Multi-species retrieves more biologically meaningful results



Conclusions

- Multi-species *cMonkey* improves bicluster quality over conserved modules
  - –Expression (residuals)
  - –Networks (association p-values)
  - –Motifs are area of potential improvement
- Retrieves more
  - –biologically significant results (GO/KEGG)
  - –conserved modules

Proposed Multi-species cMonkey model

- Explore the optimization further:
  - Alternate objective functions for OC_{12} optimization:
    - Bi-variate model: $\pi_{ik} \equiv p\left(y_{ik}^1, y_{ik}^2 \mid g_{ik}^1, g_{ik}^2\right) \propto e^{\left(\beta_0 + \beta_1 g_{ik}^1 + \beta_2 g_{ik}^2\right)}$
    - Co-reference model: $\pi_{ik} = \pi_{ik}^1 \pi_{ik}^2$

    If, we let:

$$\pi_{ik}^1 \equiv p\left(y_{ik}^1 \mid \mathbf{X}_k^1, S_i^1, \mathbf{M}_k^1, \mathbf{N}_{i \circledR k}^1\right) \propto e^{\left(\beta_0^1 + \beta_1^1 g_{ik}^1\right)}$$

$$\pi_{ik}^2 \equiv p\left(y_{ik}^2 \mid \mathbf{X}_k^2, S_i^2, \mathbf{M}_k^2, \mathbf{N}_{i \circledR k}^2\right) \propto e^{\left(\beta_0^2 + \beta_1^2 g_{ik}^2\right)}$$

- Application to different species/data sets
  - Cancer: human-mouse, cancer-normal
  - Additional triplets, i.e. (*E. coli*, *Salmonella*, *Vibrio (already have preliminary results)*)

Acknowledgments

Bonneau lab:
Glenn Butterfoss
Kevin Drew
Aviv Madar
Peter Waltman
Thadeous Kacmarczyk
Shailla Musharof
Devorah Kengmana
Chris Poultny (Shasha)
Irina Nudelman
Alex Pearlman (Ostrer)
Alex Pine

NYU:
Eric Vanden-Eijnden
Harry Ostrer
Mike Purugganan
Patrick Eichenberger
Dennis Shasha

Tacitus-
Howard Coale

- **IBM**
  - Robin Wilner
  - Bill Boverman
  - Viktors Berstis
  - Rick Alther
- ETH Zurich
  - Reudi Aebersold
  - Lars Malmstroem

Mike Boxem
Marc Vidal

Dave Goodlett

Jochen Supper (Zell Lab)

- ISB
- Nitin Baliga (&lab)
- Leroy Hood
- Marc Facciotti
- David Reiss
- Vesteinn Thorsson

- Paul Shannon
- Iliana Avila-Campillo (MERC)

Alan Aderem

Rosetta Commons
Charlie Strauss (los alamos)
David Baker (UW seattle)

DOD-computing and society,
NSF ABI,
NSF Plant genome
NSF DBI,
DOE GTL