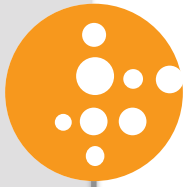


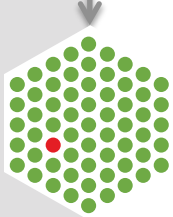
Toolchains for integrating heterogeneous algorithms and data: an anecdote

Aaron Brooks

Nitin Baliga

Dave Reiss

Antoine Allard

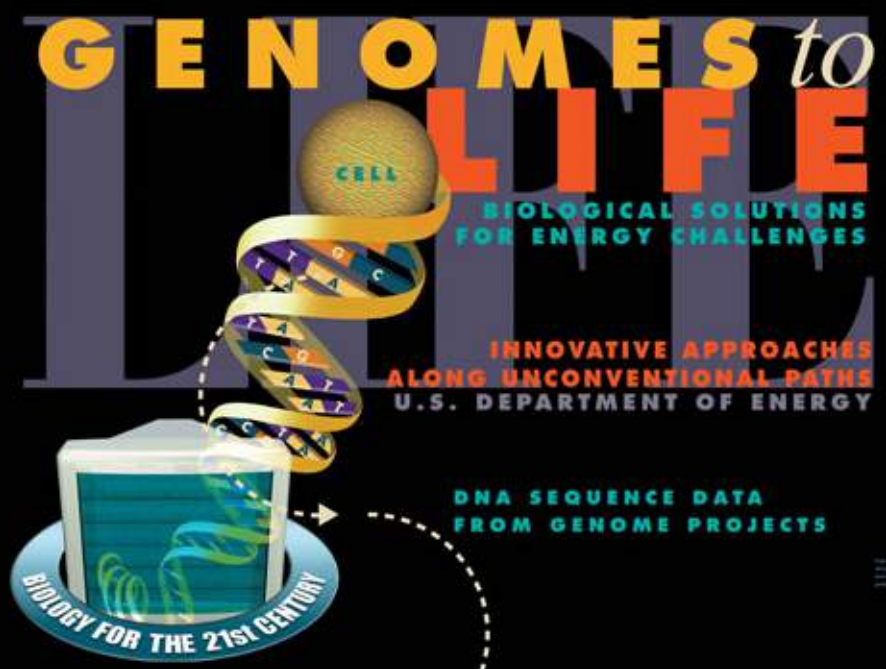


GENOMES to LIFE

BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

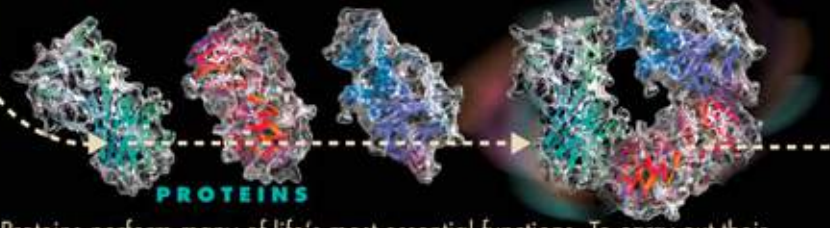
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

U.S. DEPARTMENT OF ENERGY



DNA SEQUENCE DATA FROM GENOME PROJECTS

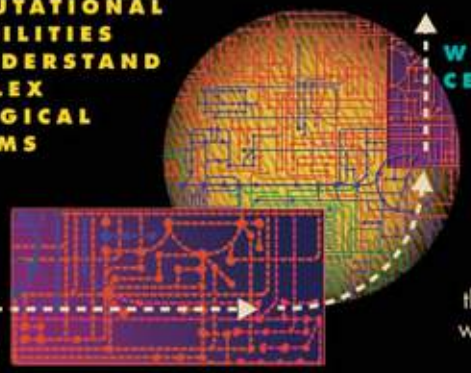
Genes and other DNA sequences contain instructions on how and when to build proteins



Proteins perform many of life's most essential functions. To carry out their specific roles, they often work together in the cell as protein machines.

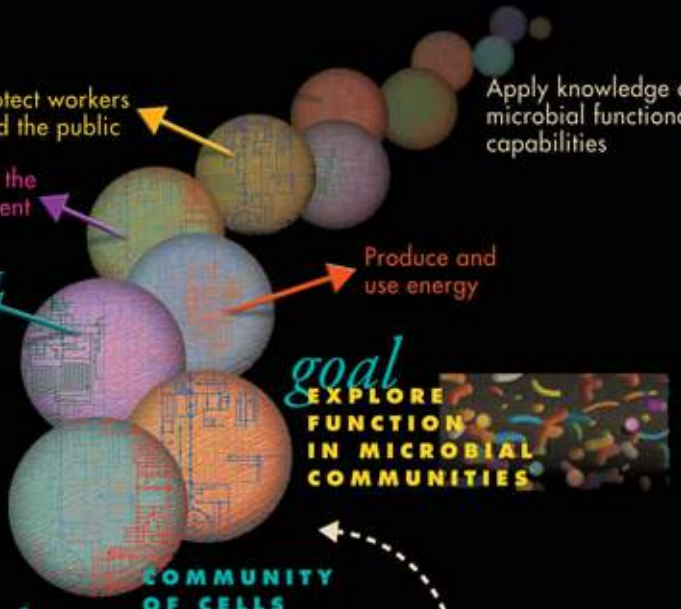
goal IDENTIFY PROTEIN MACHINES

goal DEVELOP COMPUTATIONAL CAPABILITIES TO UNDERSTAND COMPLEX BIOLOGICAL SYSTEMS



Many protein machines interact through complex, interconnected pathways. Analyzing these dynamic processes will lead to models of life processes.

goal CHARACTERIZE GENE REGULATORY NETWORKS

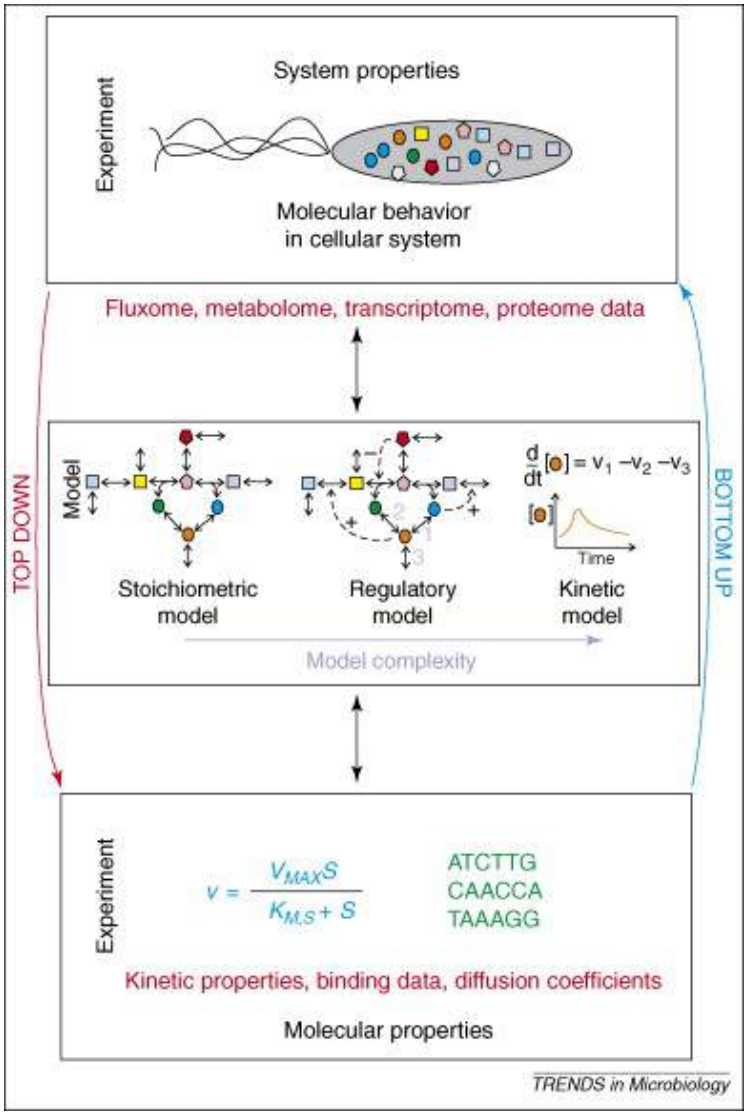


goal EXPLORE FUNCTION IN MICROBIAL COMMUNITIES

URL DOEGenomesToLife.org

10/02

Two conceptual approaches in systems biology

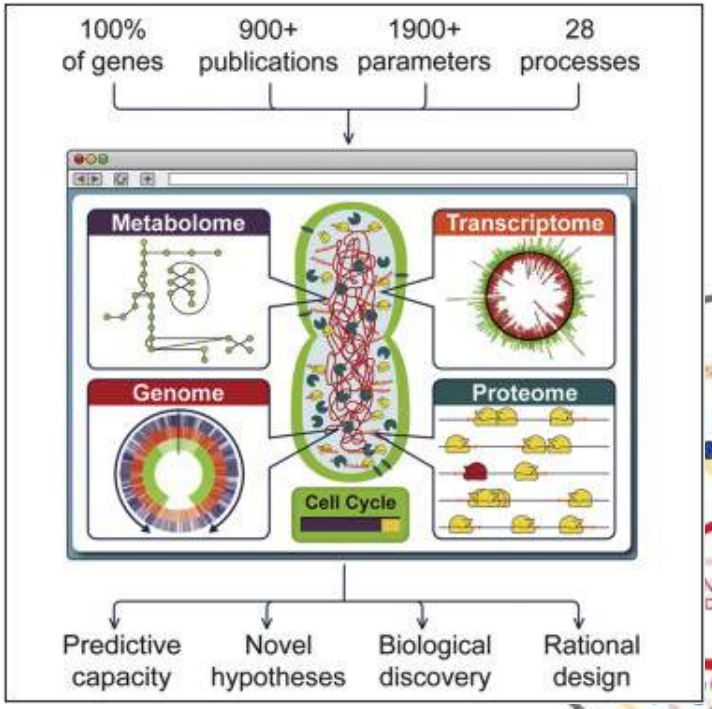


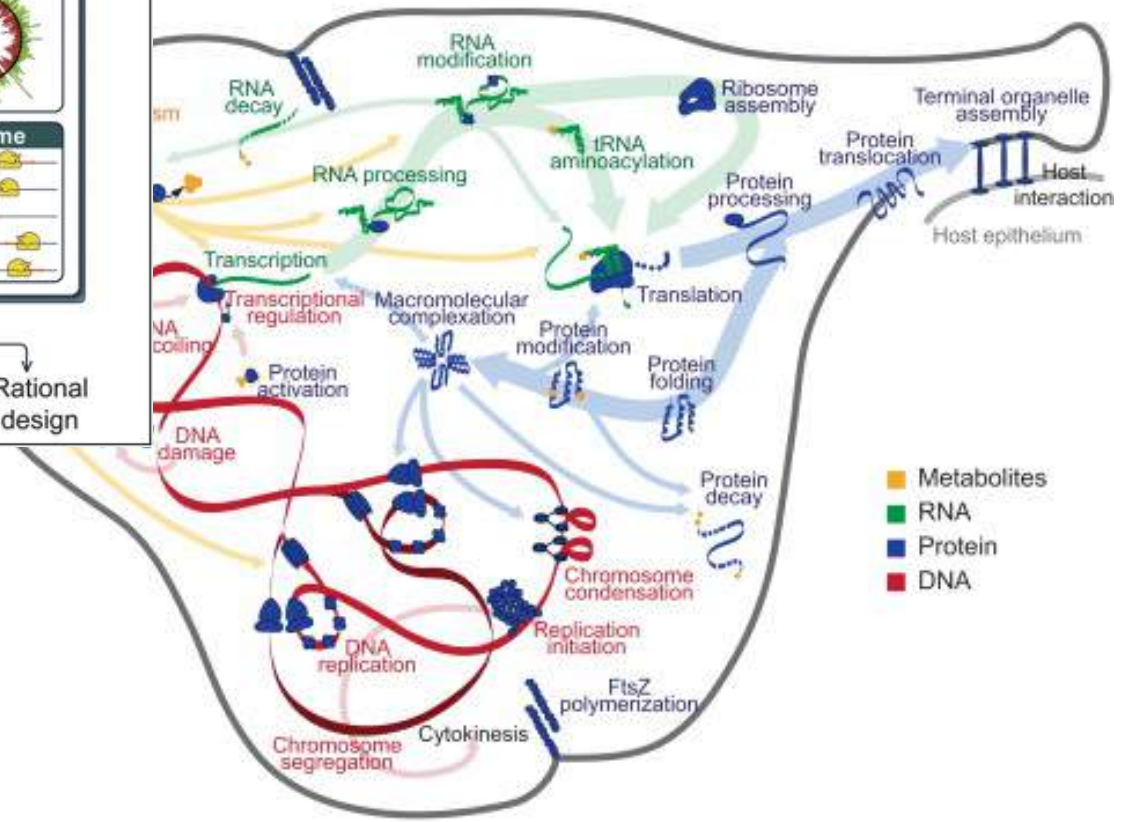
Top down

Bottom up

Bruggeman and Westerhoff 2007

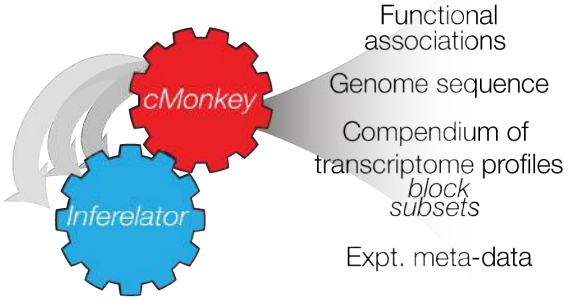
Bottom up

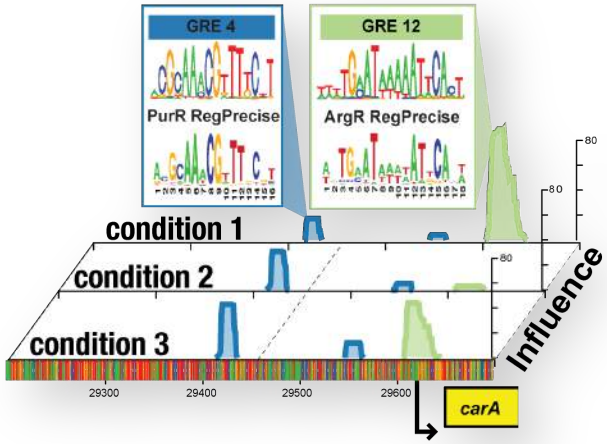


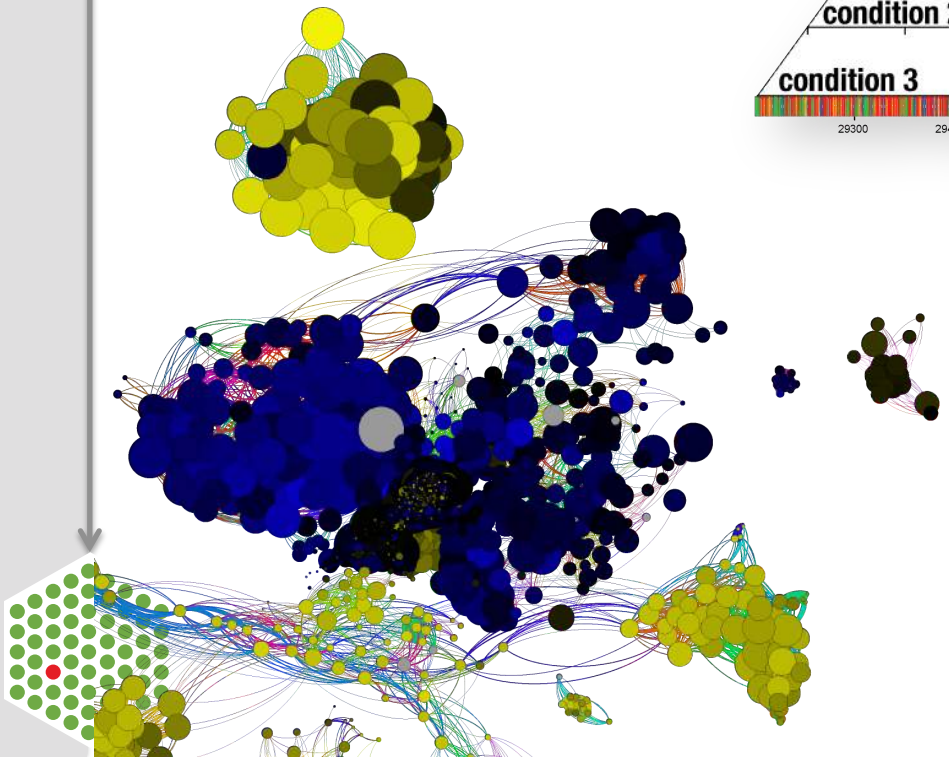


Karr et al 2012

Top down

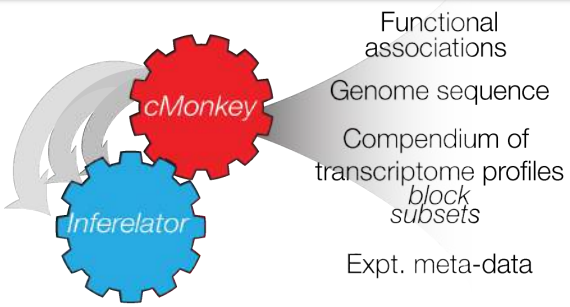


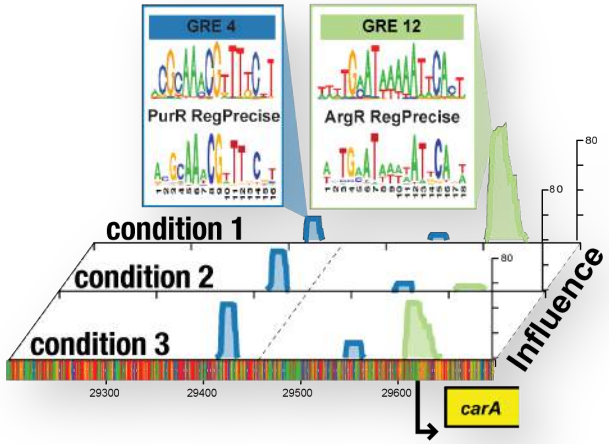


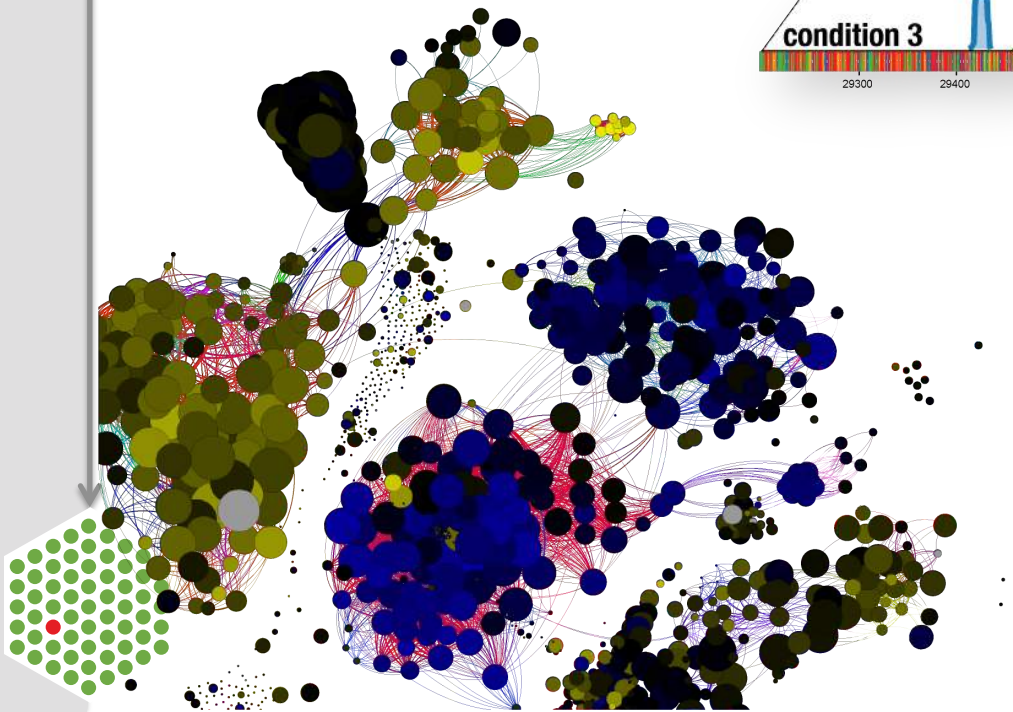


EGRIN2.0
ensemble network inference

Top down



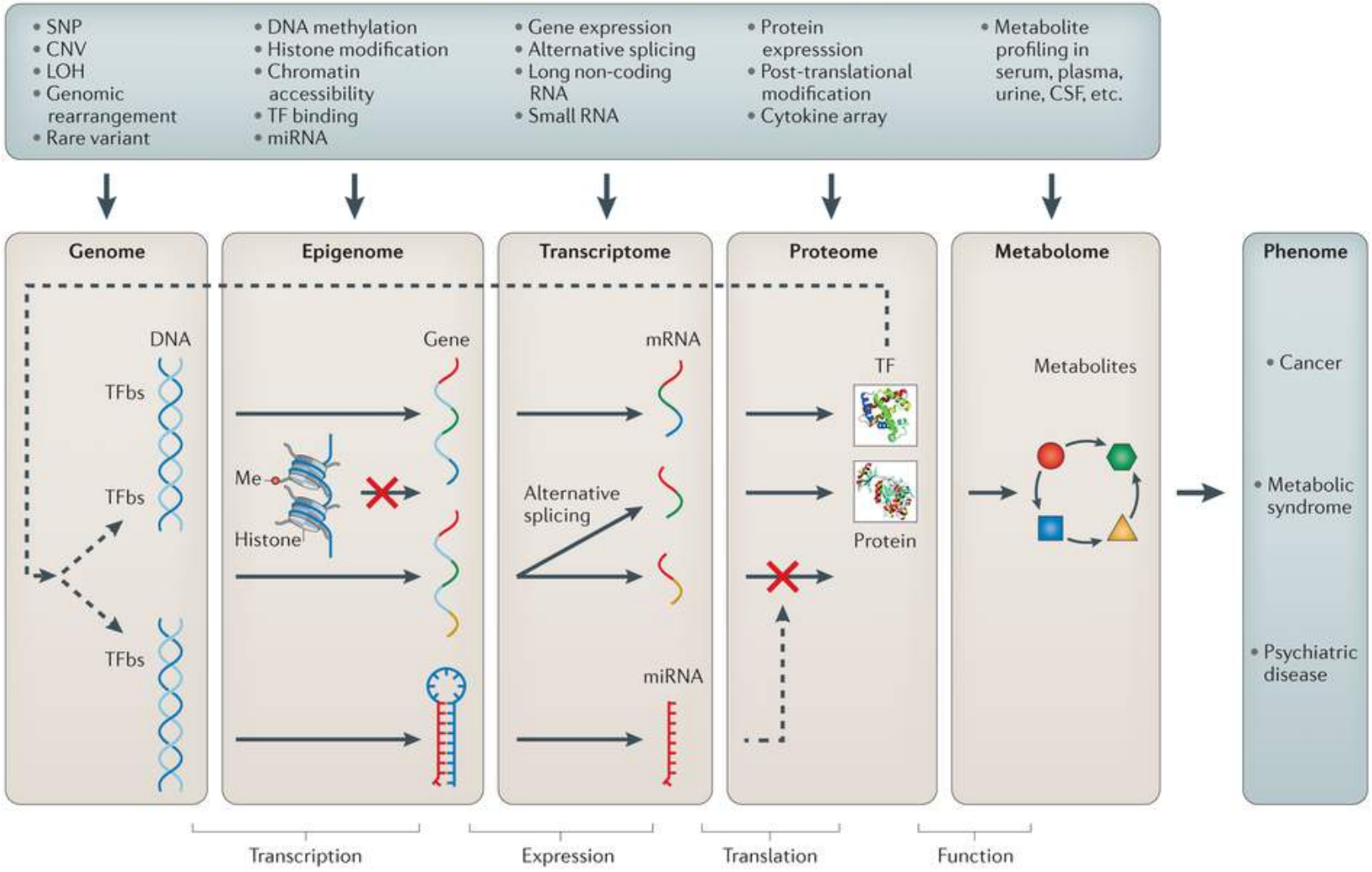




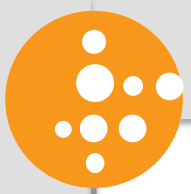
EGRIN2.0
ensemble network inference

Data is heterogeneous

Challenge: Heterogeneous data



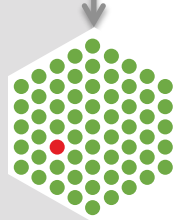
Ritchie et al 2015



Modeling challenges in systems biology

Data is heterogeneous

Algorithms are numerous and diverse



Challenge: Many Algorithms

Table 1 | Network inference methods

ID	Synopsis	Reference
	Regression: transcription factors are selected by target gene–specific (i) sparse linear-regression and (ii) data-resampling approaches.	
1	Trustful Inference of Gene REgulation using Stability Selection (TIGRESS): (i) Lasso; (ii) the regularization parameter selects five transcription factors per target gene in each bootstrap sample.	33[a]
2	(i) Steady-state and time-series data are combined by group Lasso; (ii) bootstrapping.	34[a]
3	Combination of Lasso and Bayesian linear regression models learned using reversible-jump Markov chain Monte Carlo simulations.	35[a]
4	(i) Lasso; (i) bootstrapping.	36
5	(i) Lasso; (ii) area under the stability selection curve.	36
6	Application of the Lasso toolbox GENLAB using standard parameters.	37
7	Lasso models are combined by the maximum regularization parameter selecting a given edge for the first time.	36[a]
8	Linear regression determines the contribution of transcription factors to the expression of target genes.	—[a,b]
	Mutual information: edges are (i) ranked based on variants of mutual information and (ii) filtered for causal relationships.	
1	Context likelihood of relatedness (CLR): (i) spline estimation of mutual information; (ii) the likelihood of each mutual information score is computed based on its local network context.	11[a,b]
2	(i) Mutual information is computed from discretized expression values.	38[a,b]
3	Algorithm for the reconstruction of accurate cellular networks (ARACNE): (i) kernel estimation of mutual information; (ii) the data processing inequality is used to identify direct interactions.	9[a,b]
4	(i) Fast kernel-based estimation of mutual information; (ii) Bayesian local causal discovery (BLCD) and Markov blanket (HITON-PC) algorithm to identify direct interactions.	39[a]
5	(i) Mutual information and Pearson's correlation are combined; (ii) BLCD and HITON-PC algorithm.	39[a]
	Correlation: edges are ranked based on variants of correlation.	
1	Absolute value of Pearson's correlation coefficient.	38
2	Signed value of Pearson's correlation coefficient.	38[a,b]
3	Signed value of Spearman's correlation coefficient.	38[a,b]
	Bayesian networks: optimize posterior probabilities by different heuristic searches.	
1	Simulated annealing (catnet R package, http://cran.r-project.org/web/packages/catnet/), aggregation of three runs.	—
2	Simulated annealing (catnet R package, hyperlink above).	—
3	Max-min parent and children algorithm (MMPC), bootstrapped data sets.	40
4	Markov blanket algorithm (HITON-PC), bootstrapped data sets.	41
5	Markov boundary induction algorithm (TIE*), bootstrapped data sets.	42
6	Models transcription factor perturbation data and time series using dynamic Bayesian networks (Infer.NET toolbox, http://research.microsoft.com/infernet/).	—[a]
	Other approaches: network inference by heterogeneous and novel methods.	
1	GENIE3: a Random Forest is trained to predict target gene expression. Putative transcription factors are selected as tree nodes if they consistently reduce the variance of the target.	19[a]
2	Codependencies between transcription factors and target genes are detected by the nonlinear correlation coefficient η^2 (two-way ANOVA). Transcription-factor perturbation data are up-weighted.	20[a]
3	Transcription factors are selected by maximizing the conditional entropy for target genes, which are represented as Boolean vectors with probabilities to avoid discretization.	43[a]
4	Transcription factors are preselected from transcription-factor perturbation data or by Pearson's correlation and then tested by iterative Bayesian model averaging (BMA).	44
5	A Gaussian noise model is used to estimate whether [illegible] in transcription-factor perturbation measurements.	45
6	After scaling, target genes are clustered by Pearson's correlation. A neural network is trained (genetic algorithm) and parameterized (back-propagation).	46[a]

Marbach et al 2012

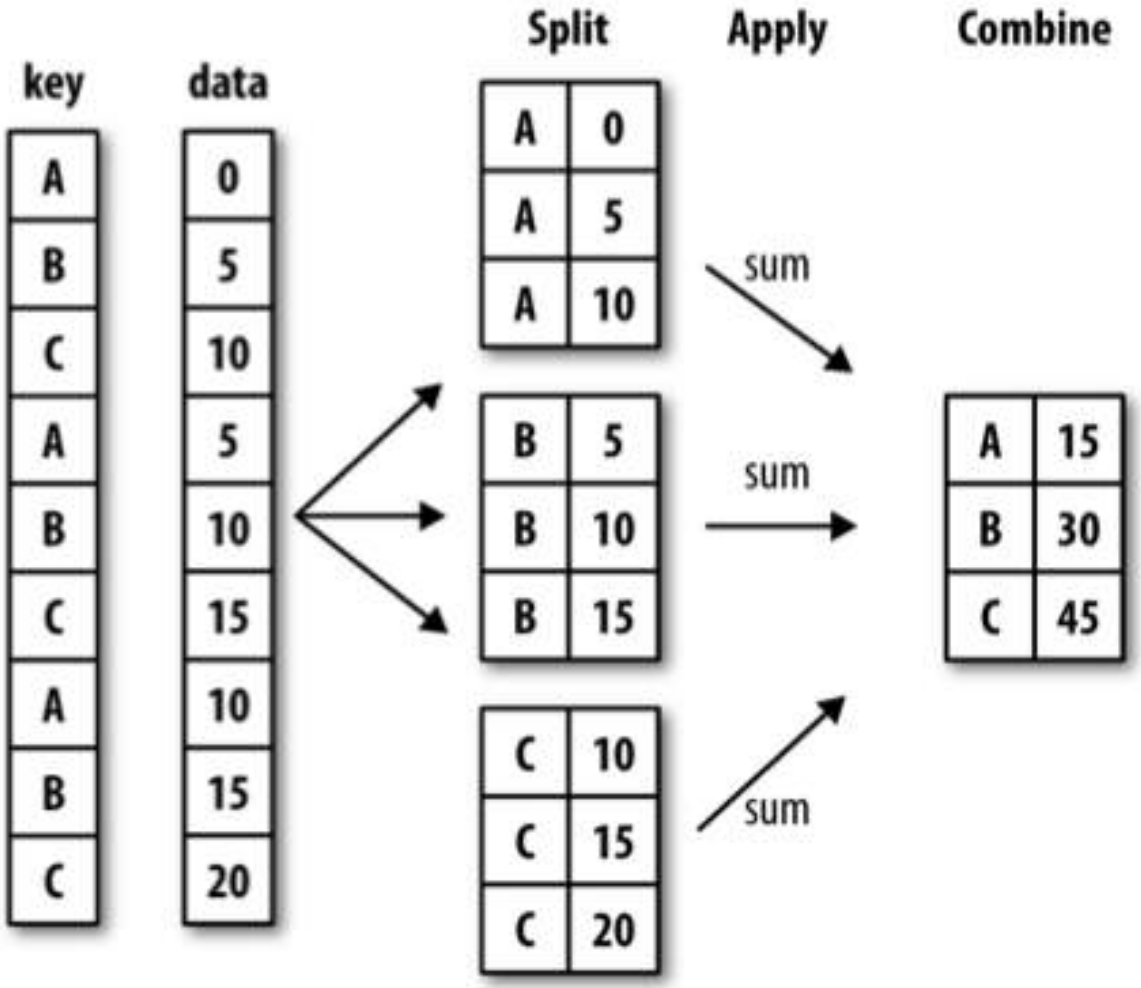
Goals for this talk

Introduce a generalized strategy for building multistep, data-driven inference toolchains, using:

- Large, heterogeneous data sets
- Multiple algorithms combined in an ensemble learning framework
- Graph-based methods to reconcile predictions

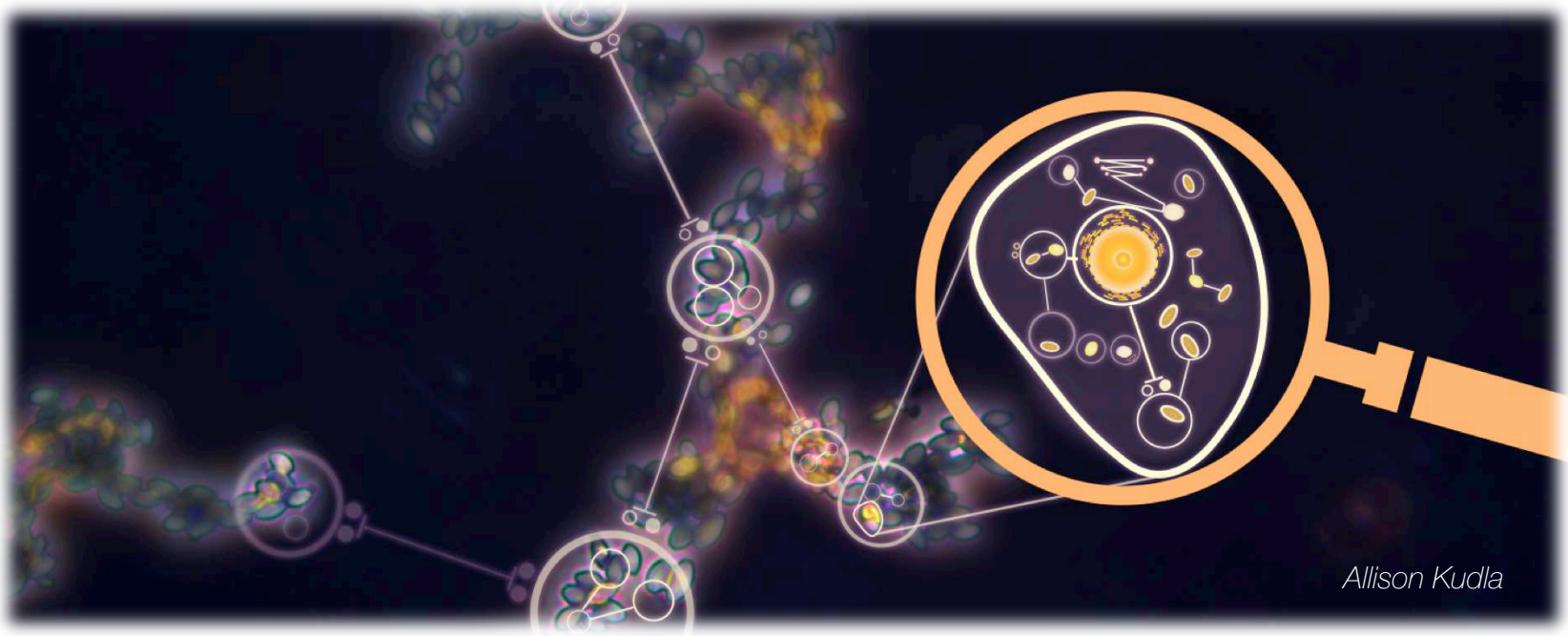
Extension: a suggestion for integrating models across data types and biological processes

Split-Apply-Combine Approach to Biological Inference



SPLIT → APPLY → COMBINE

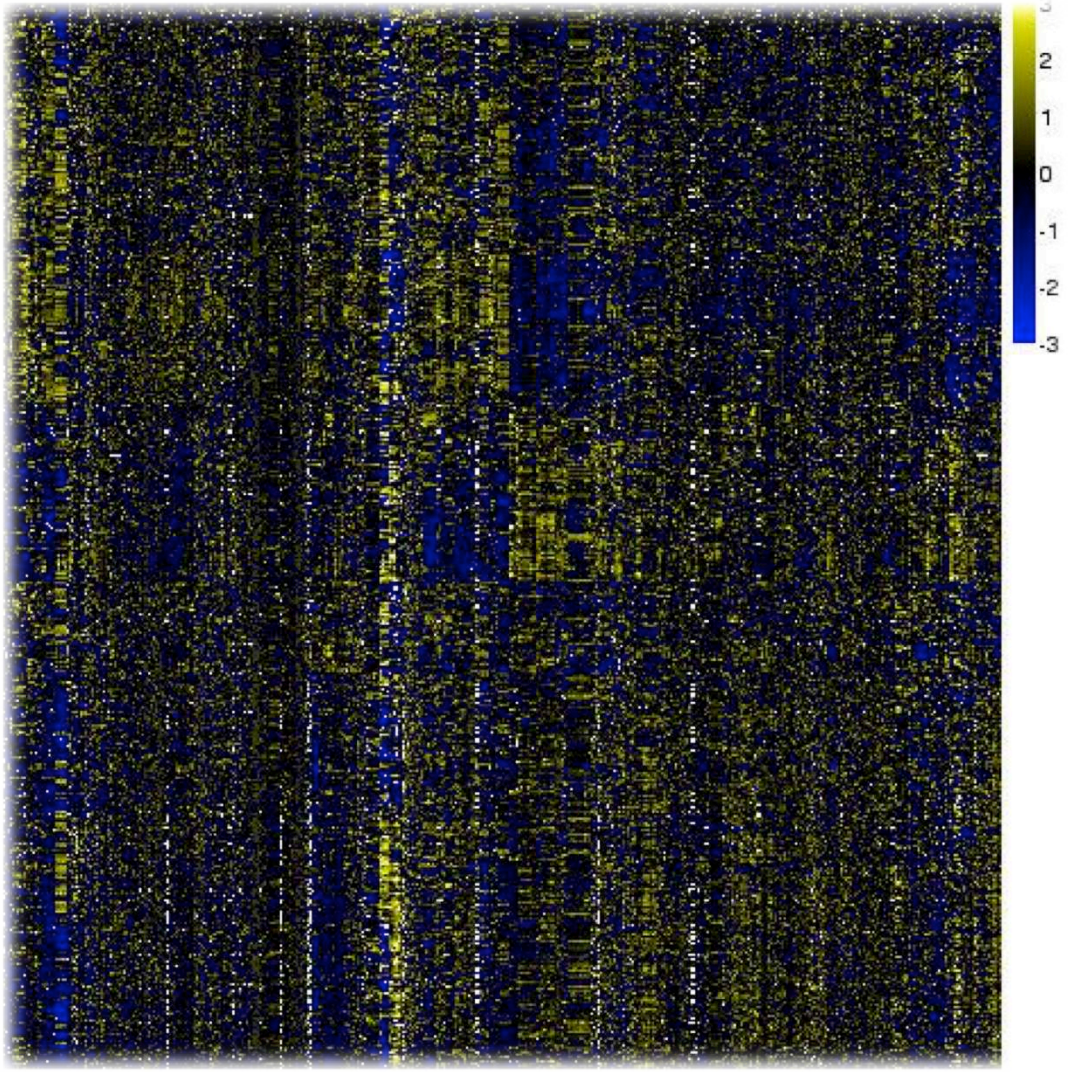
How do microbes control expression of their genomes?




TARA
OCEANS
agnès b.

From gene expression to regulatory networks

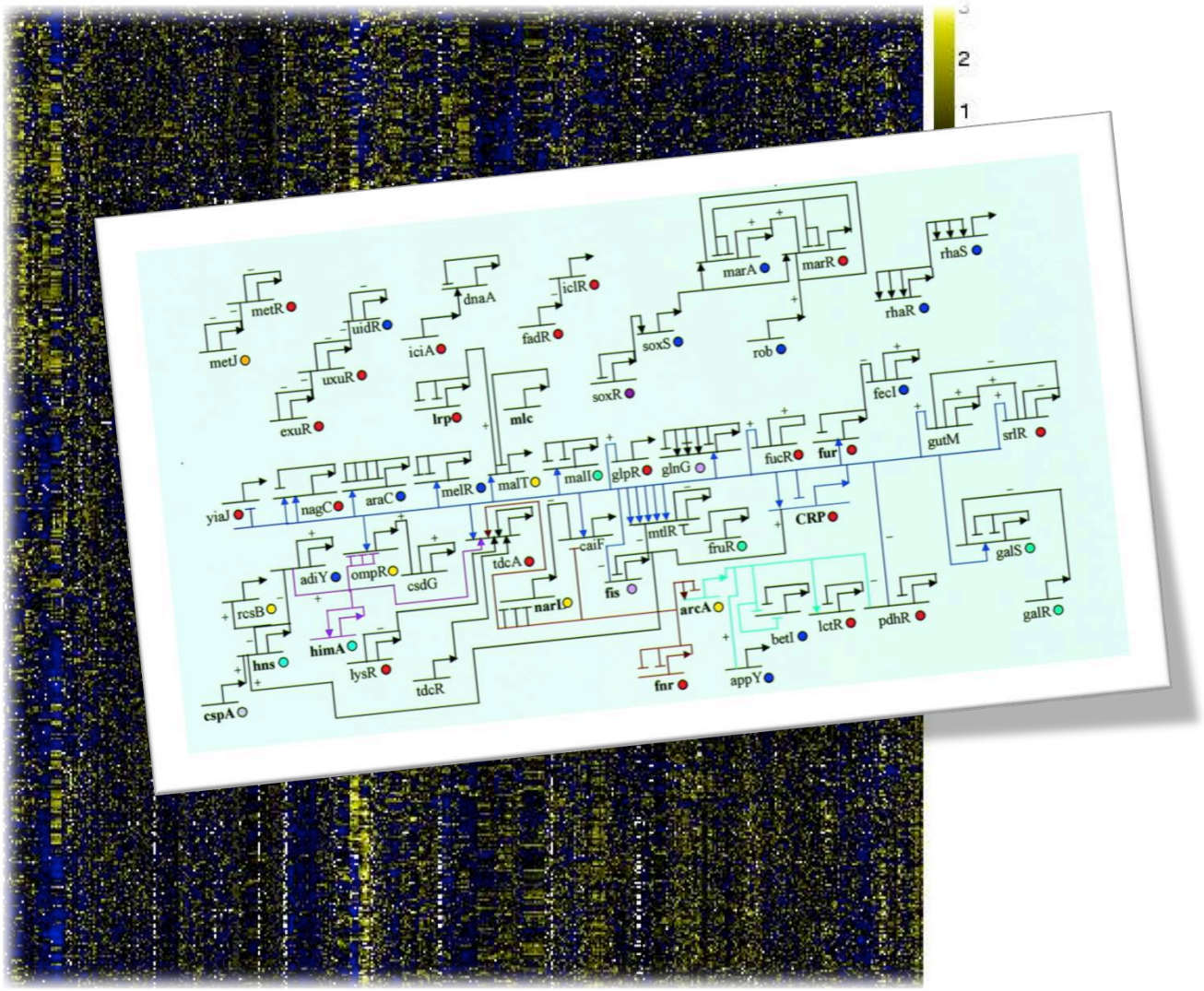
4,213 genes



868 conditions

From gene expression to regulatory networks

4,213 genes

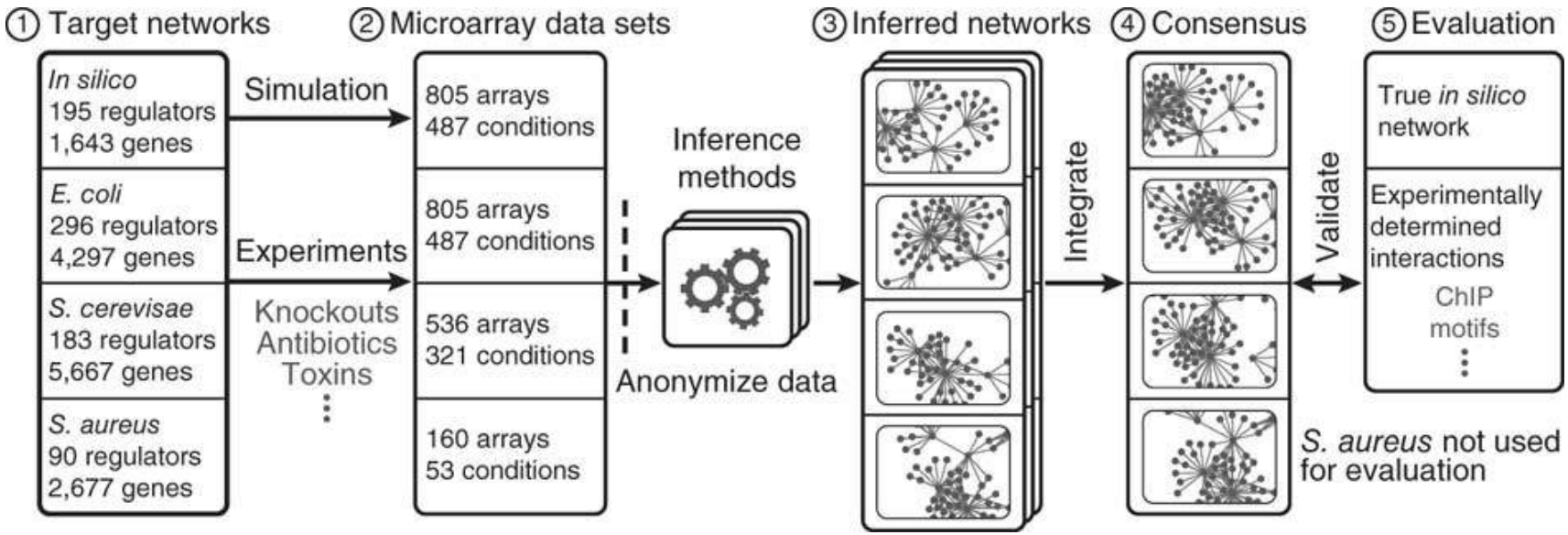


868 conditions

Wisdom of crowds for robust gene network inference

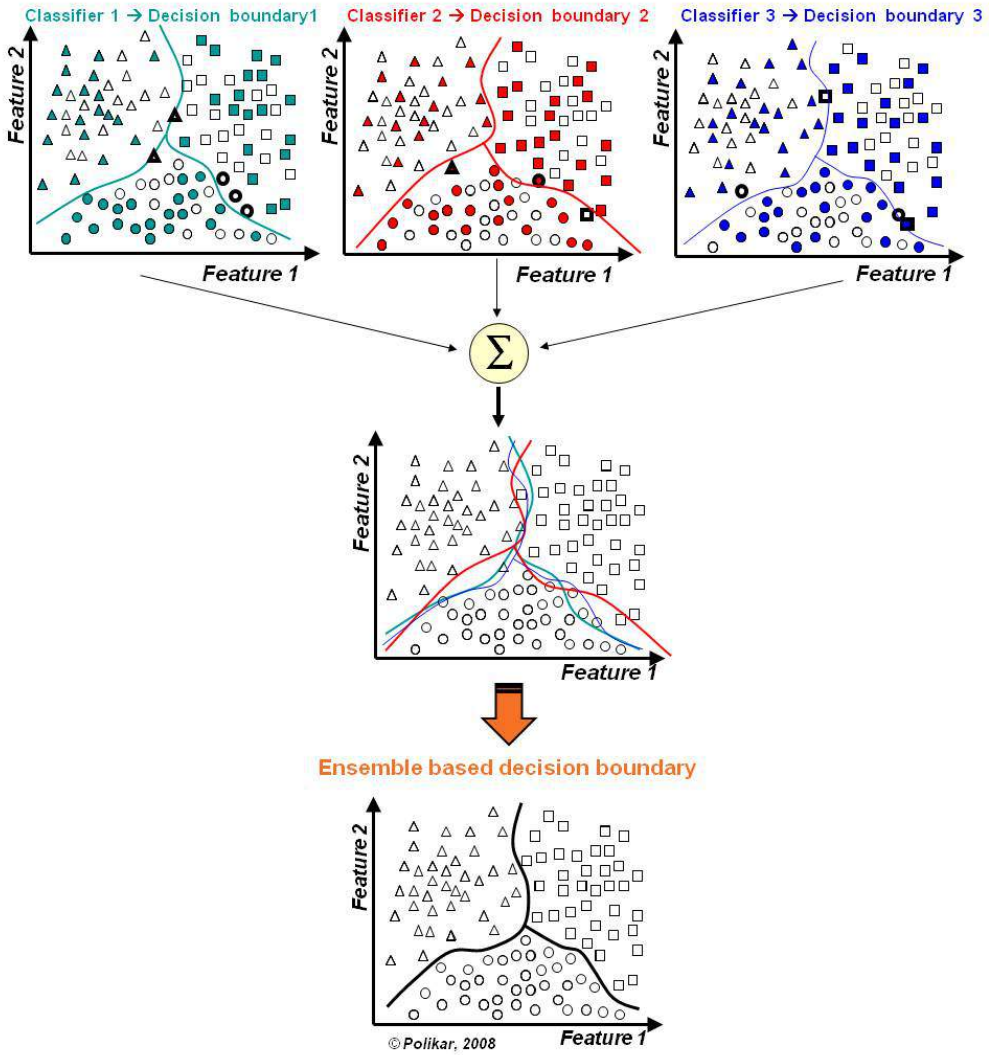
Marbach et al 2012

Combine algorithms to increase performance



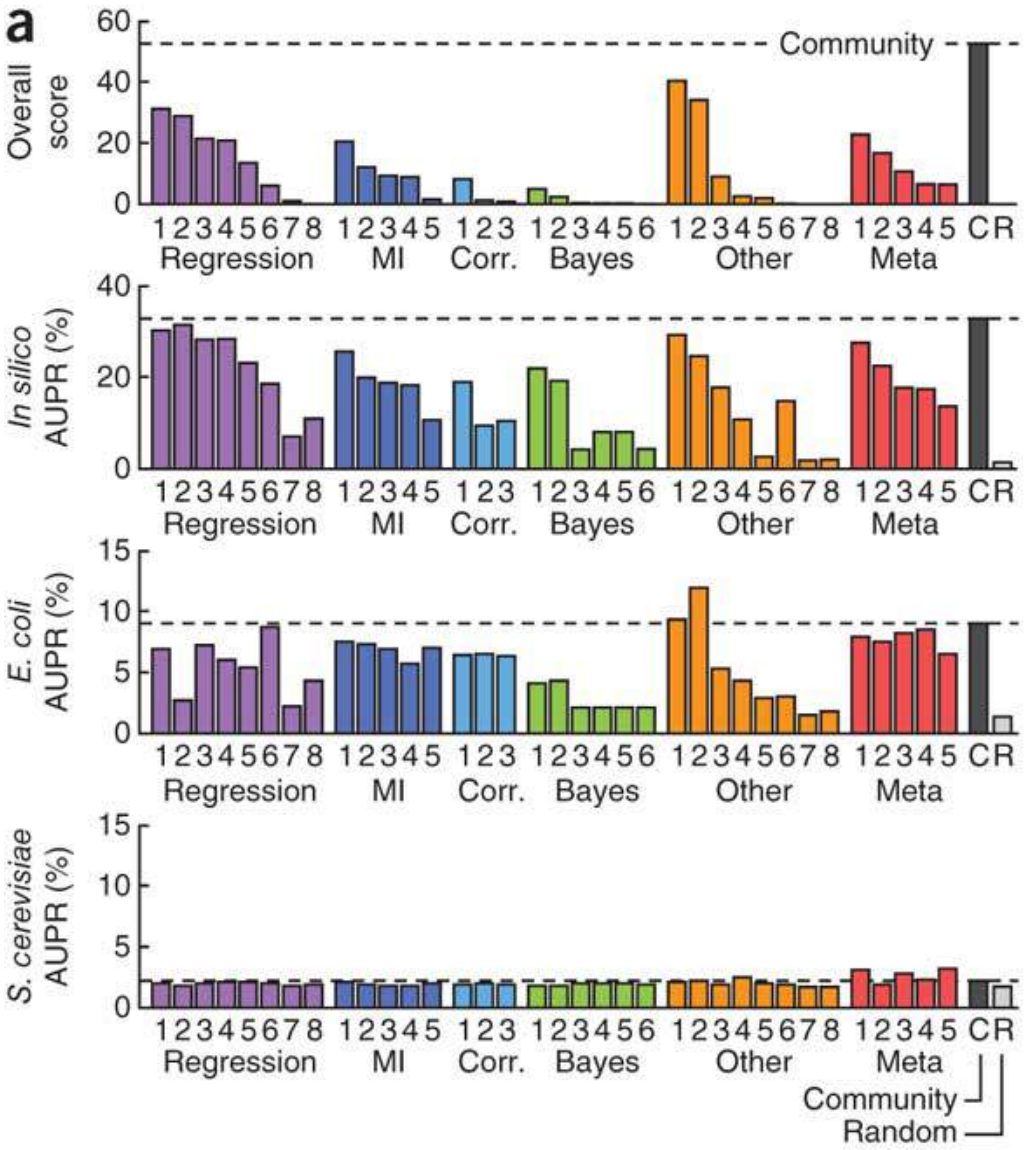
Marbach et al 2012

Ensemble learning

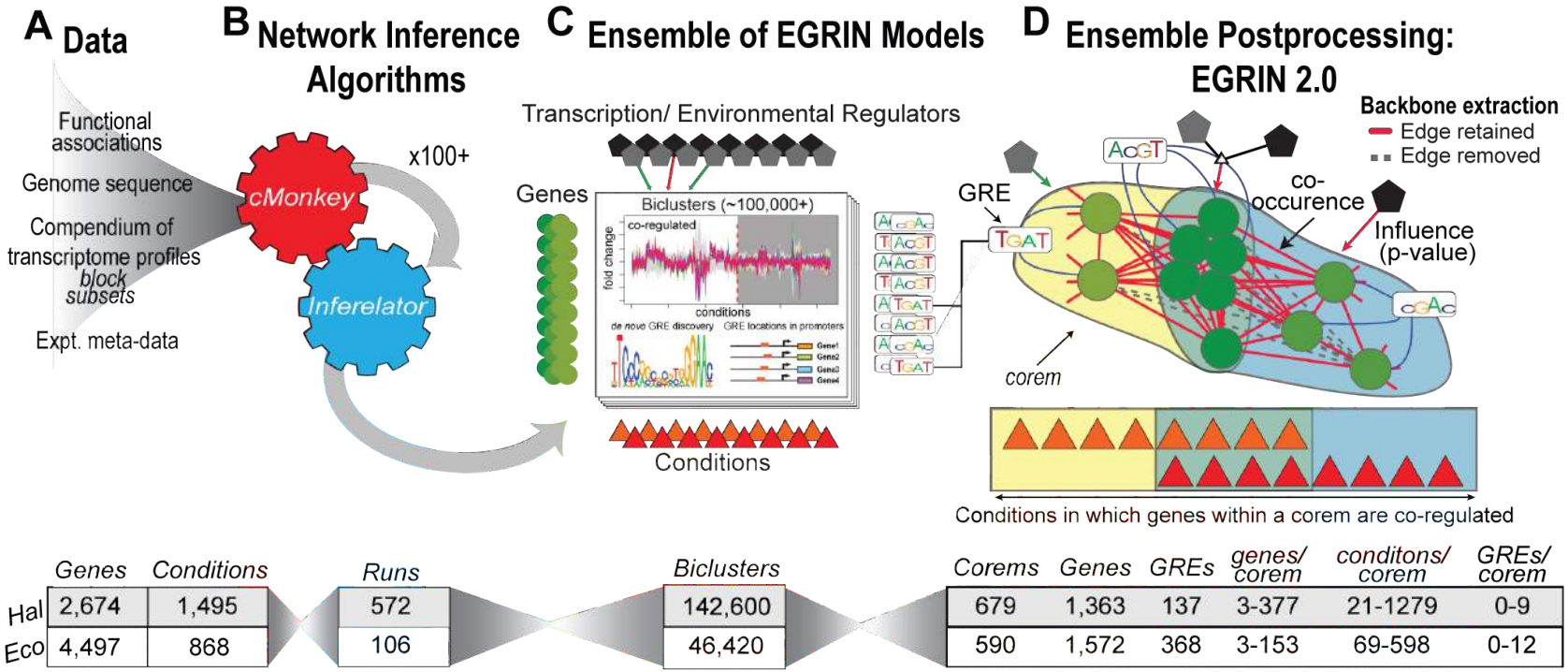


Pollikar 2008

Combining learners increases performance



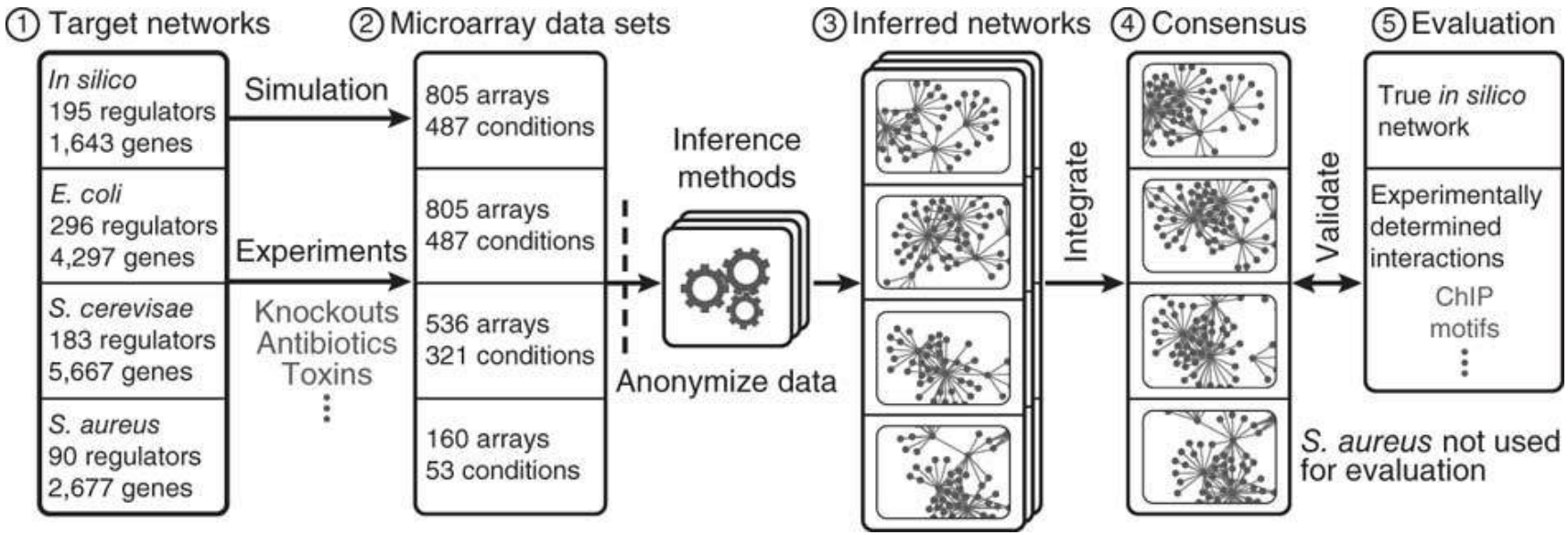
Marbach et al 2012



	Genes	Conditions	Runs	Biclusters	Corems	Genes	GREs	genes/ corem	conditons/ corem	GREs/ corem
Hal	2,674	1,495	572	142,600	679	1,363	137	3-377	21-1279	0-9
Eco	4,497	868	106	46,420	590	1,572	368	3-153	69-598	0-12

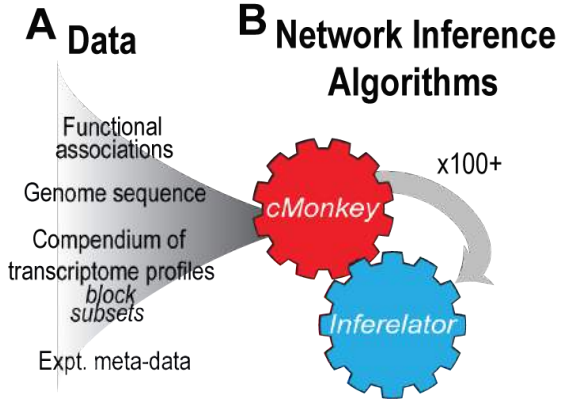
Brooks and Reiss et al 2014

Combine algorithms to increase performance



Marbach et al 2012

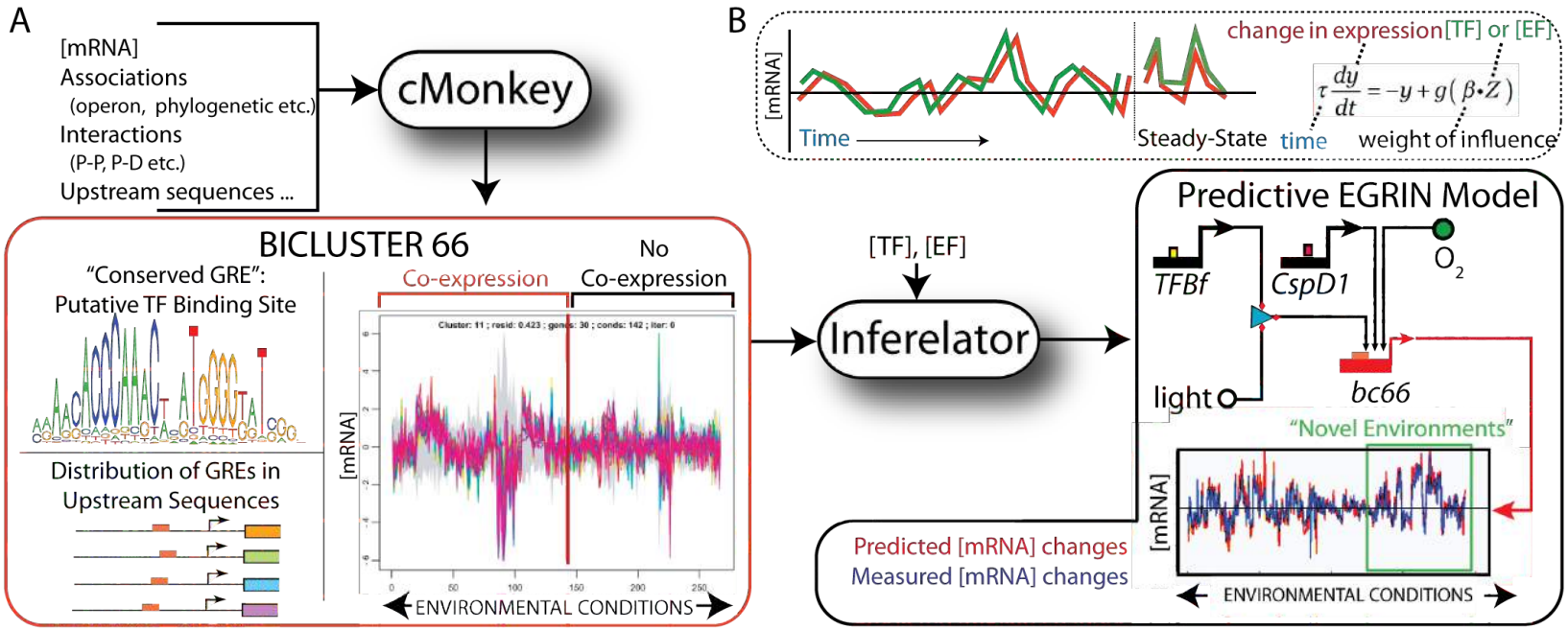
SPLIT: Bagging



	Genes	Conditions	Runs
Hal	2,674	1,495	572
Eco	4,497	868	106

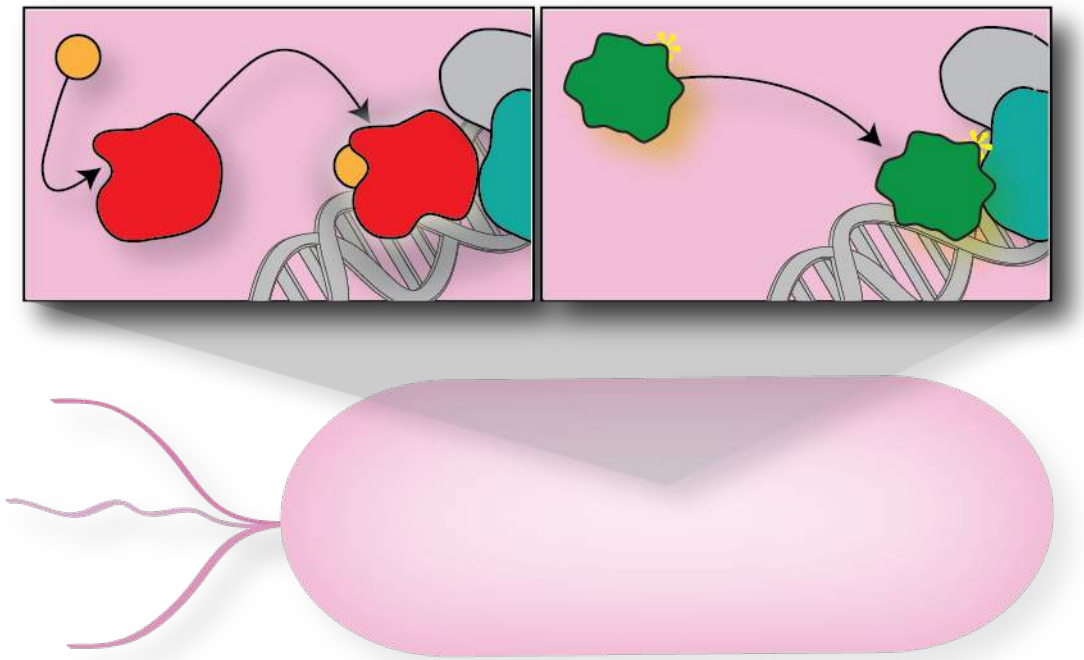
Brooks and Reiss et al 2014

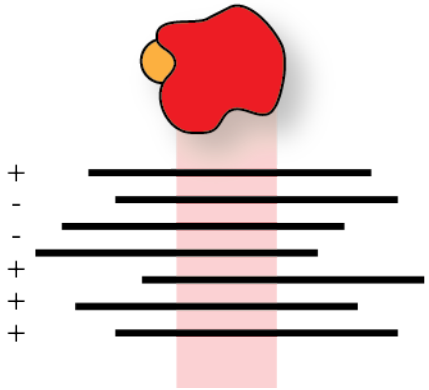
APPLY: cMonkey and Inferelator



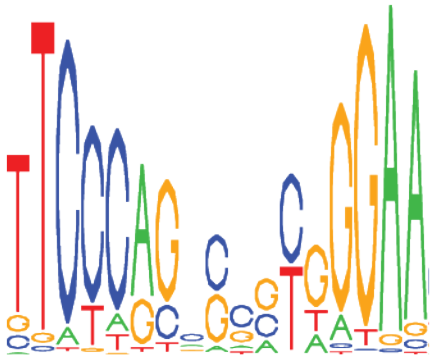
Reiss et al 2006

cMonkey clusters and detects motifs de novo



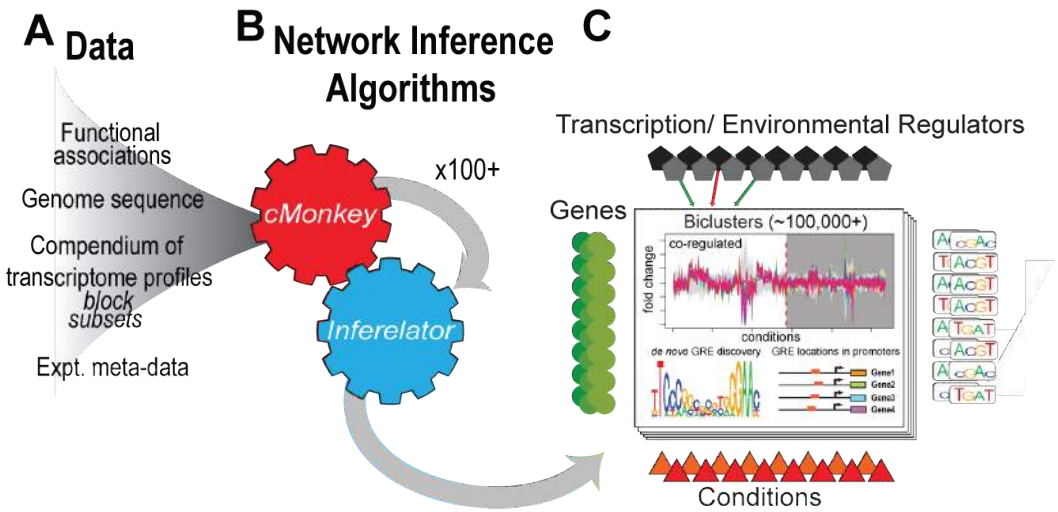


	A	C	G	T
1	0.03	0.08	0.10	0.79
2	0.01	0.03	0.04	0.93
3	0.06	0.90	0.01	0.02
4	0.01	0.78	0.03	0.18
5	0.09	0.81	0.02	0.08
6	0.70	0.01	0.25	0.04
7	0.02	0.21	0.70	0.07
8	0.22	0.44	0.19	0.15
9	0.08	0.48	0.42	0.02
10	0.12	0.49	0.28	0.12



Brooks 2014

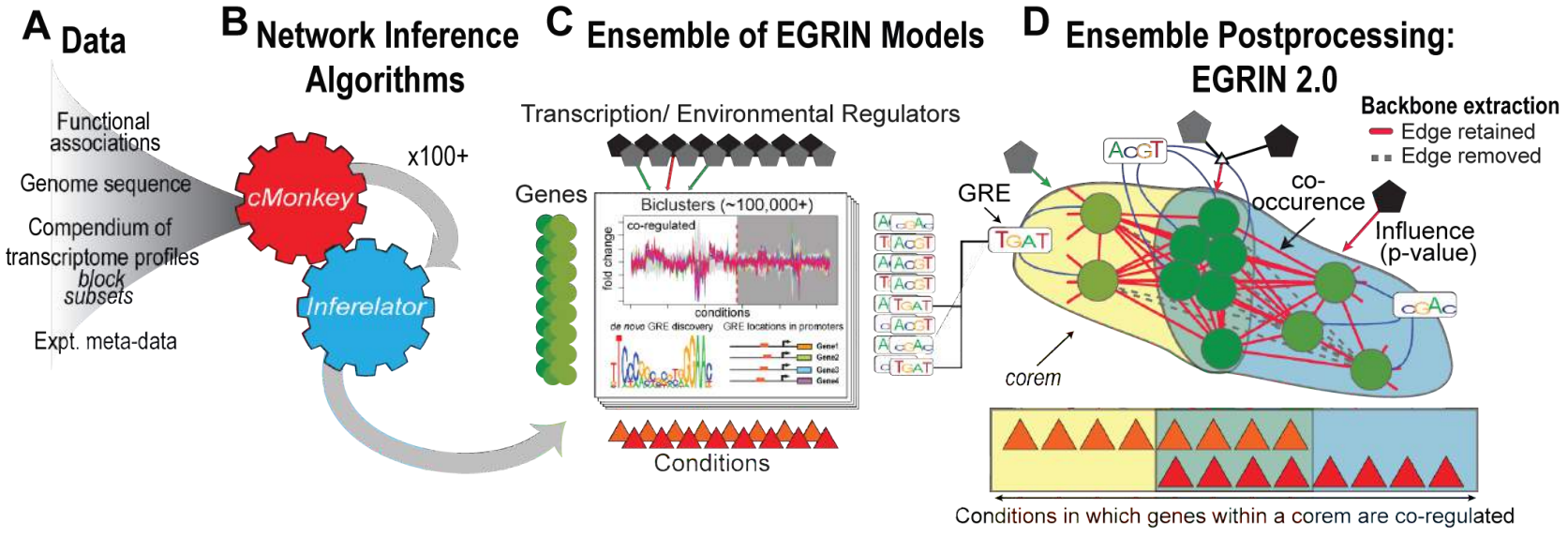
Full EGRIN 2.0 toolchain



	Genes	Conditions	Runs	Biclusters
Hal	2,674	1,495	572	142,600
Eco	4,497	868	106	46,420

Brooks and Reiss et al 2014

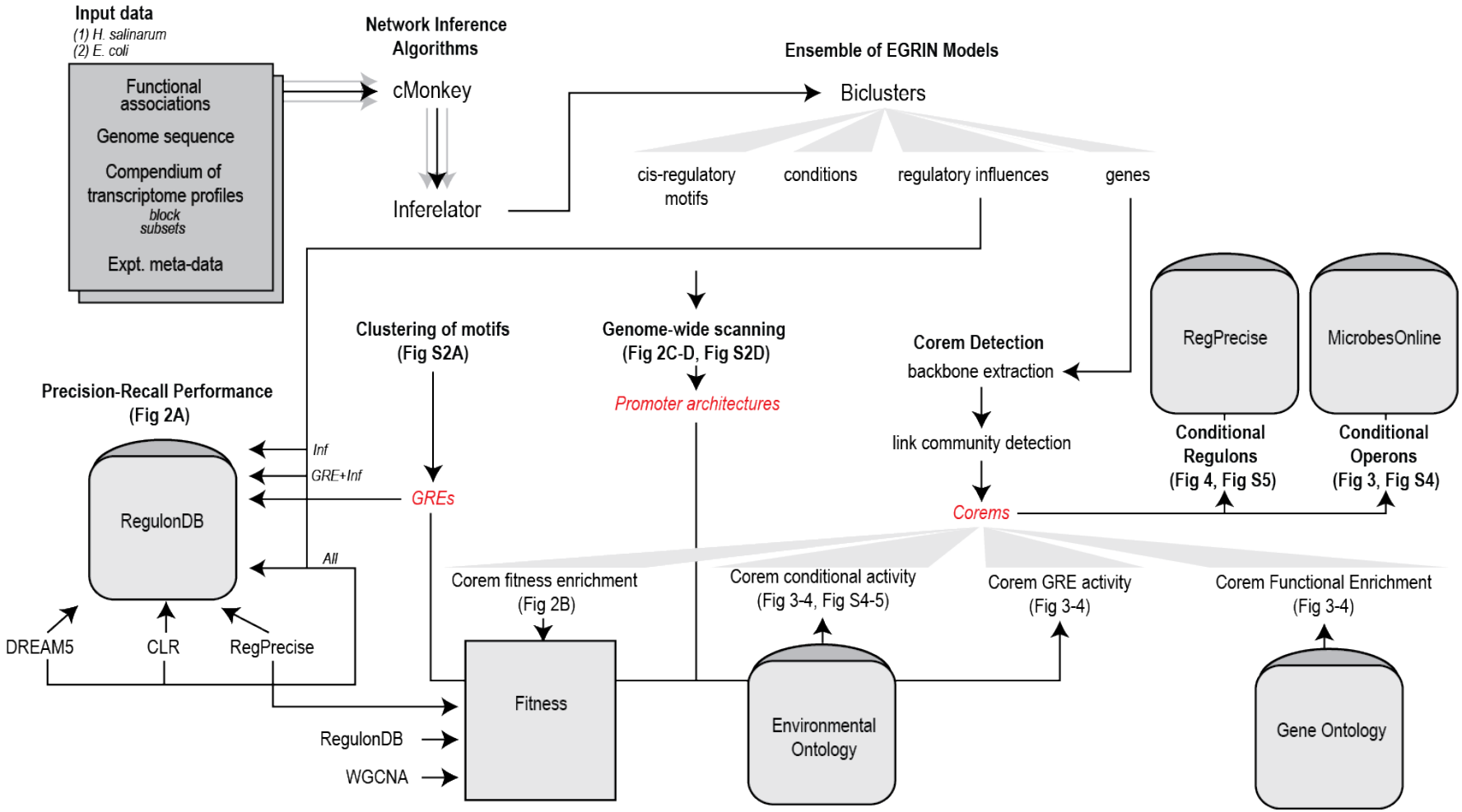
Full EGRIN 2.0 toolchain



	Genes	Conditions	Runs	Biclusters	Corems	Genes	GREs	genes/ corem	conditons/ corem	GREs/ corem
Hal	2,674	1,495	572	142,600	679	1,363	137	3-377	21-1279	0-9
Eco	4,497	868	106	46,420	590	1,572	368	3-153	69-598	0-12

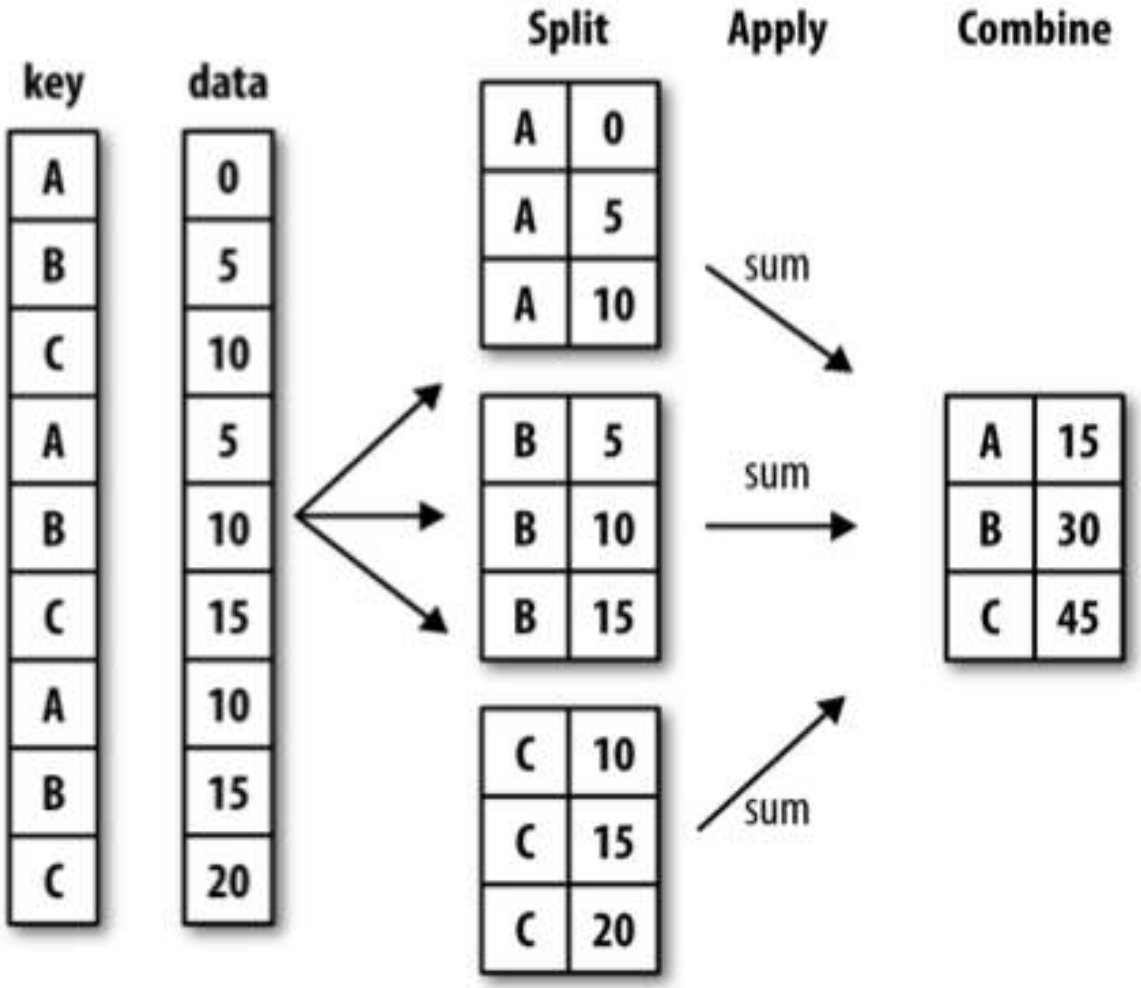
Brooks and Reiss et al 2014

Full EGRIN 2.0 toolchain



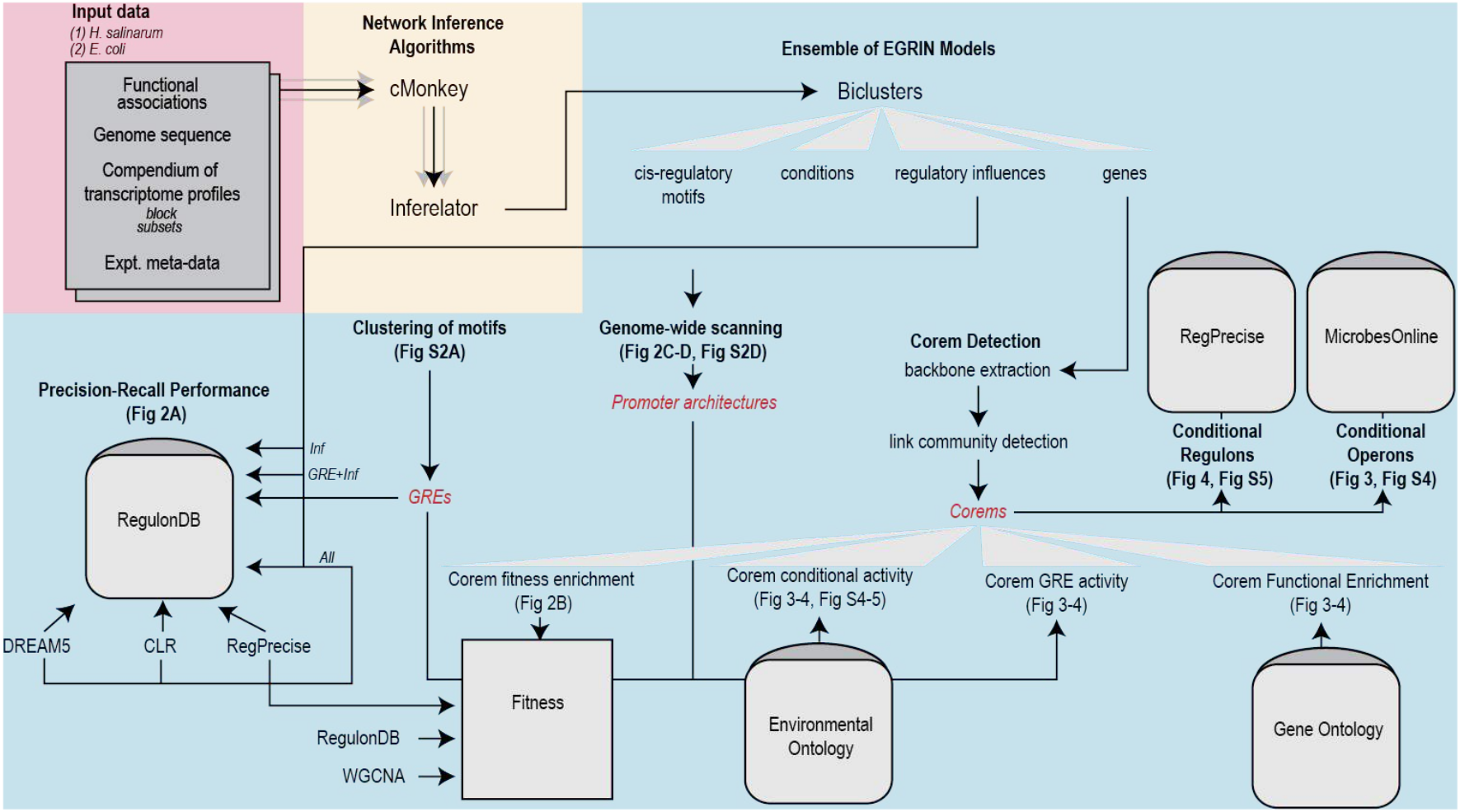
Brooks and Reiss et al 2014

Split-Apply-Combine Approach to Biological Inference



SPLIT → APPLY → COMBINE

Full EGRIN 2.0 toolchain



SPLIT → **APPLY** → **COMBINE**

Brooks and Reiss et al 2014

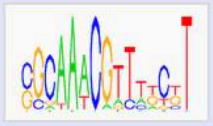
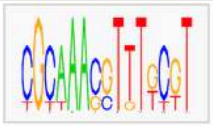
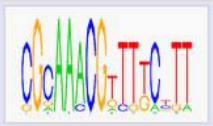
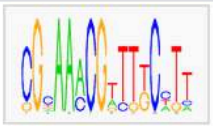
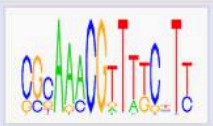
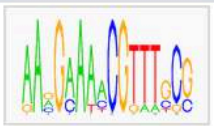
Generalized pattern for reconciling predictions

- *Combine predictions in a graph (clean as needed)*

- *Interpret with graph-based methods (clustering)*

Determine consensus GRE

GRE = gene regulatory element (putative TF binding site)

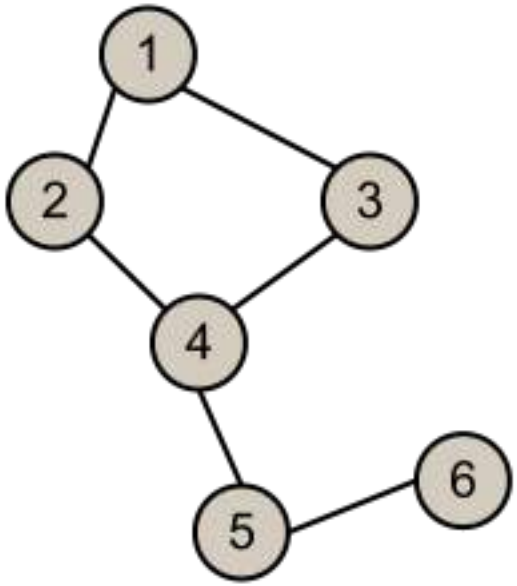
CRM ID	Bicluster	PSSM	eval
eco_2716_1	eco_2716		0
eco_1863_1	eco_1863		0
eco_2223_1	eco_2223		0
eco_2666_1	eco_2666		0
eco_859_1	eco_859		0
eco_74_1	eco_74		0
eco_357_1	eco_357		0
eco_2076_1	eco_2076		0
eco_3077_1	eco_3077		0
eco_2741_1	eco_2741		0

Showing 1 to 10 of 331 entries

egrin2.systemsbiology.net

Combine: network representation for GREs

Undirected Graph & Adjacency Matrix

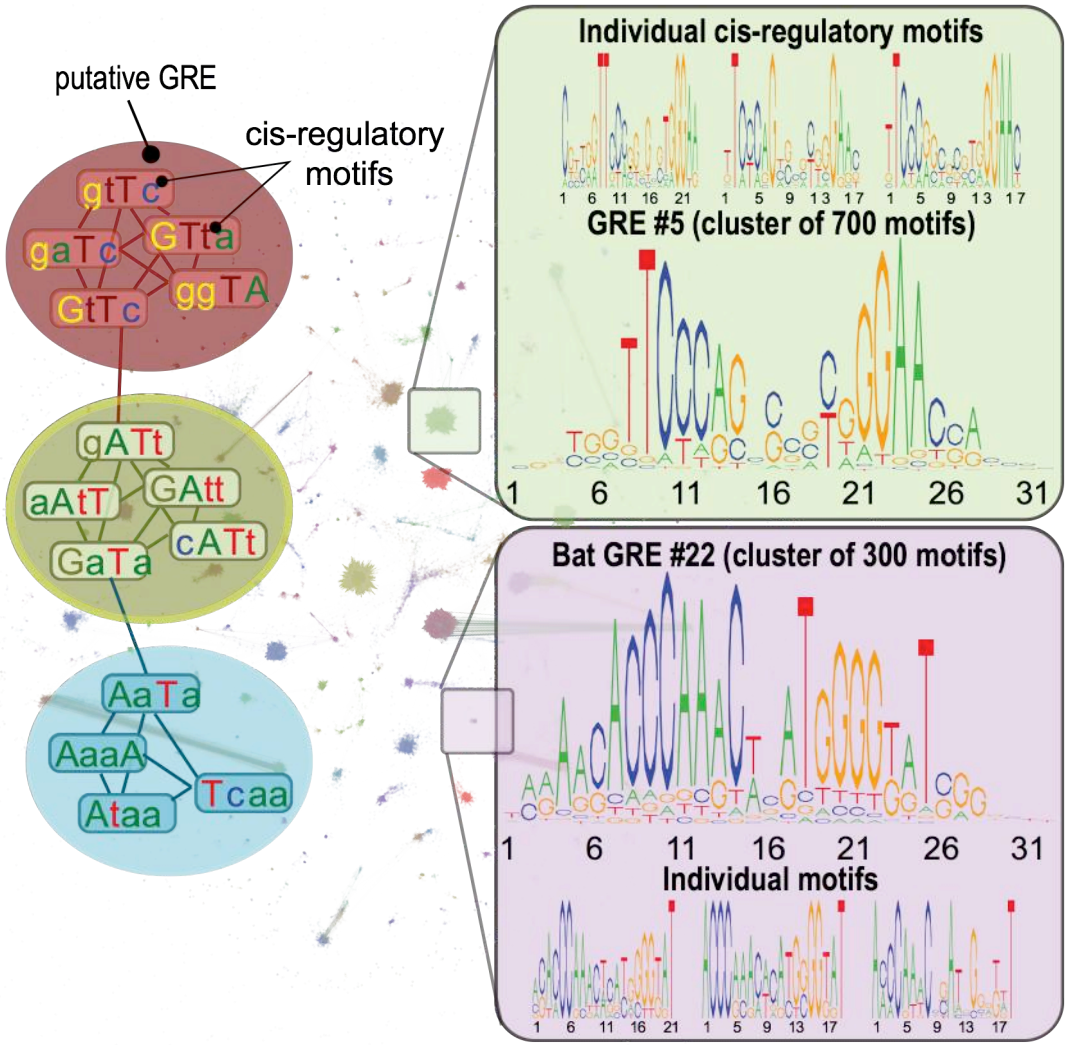


Undirected Graph

	1	2	3	4	5	6
1	0	1	1	0	0	0
2	1	0	0	1	0	0
3	1	0	0	1	0	0
4	0	1	1	0	1	0
5	0	0	0	1	0	1
6	0	0	0	0	1	0

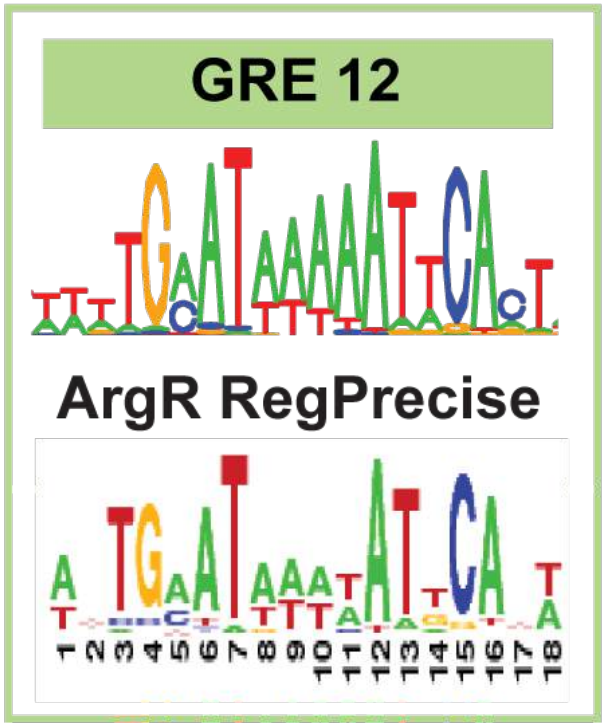
Adjacency Matrix

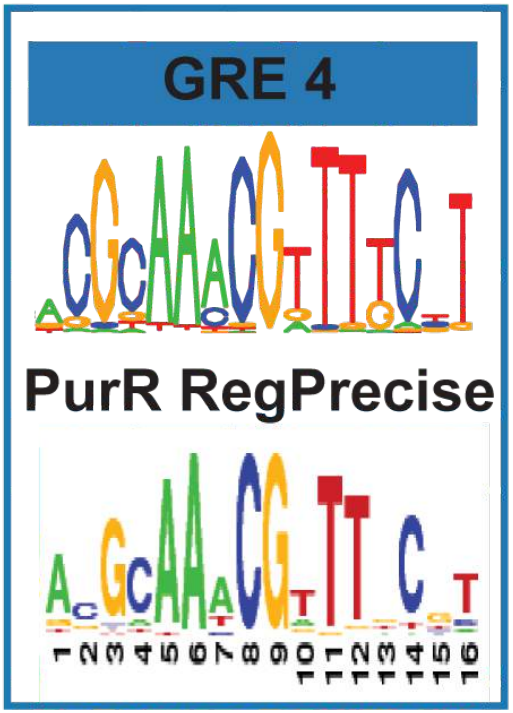
Combine: GRE clustering



Brooks and Reiss et al 2014

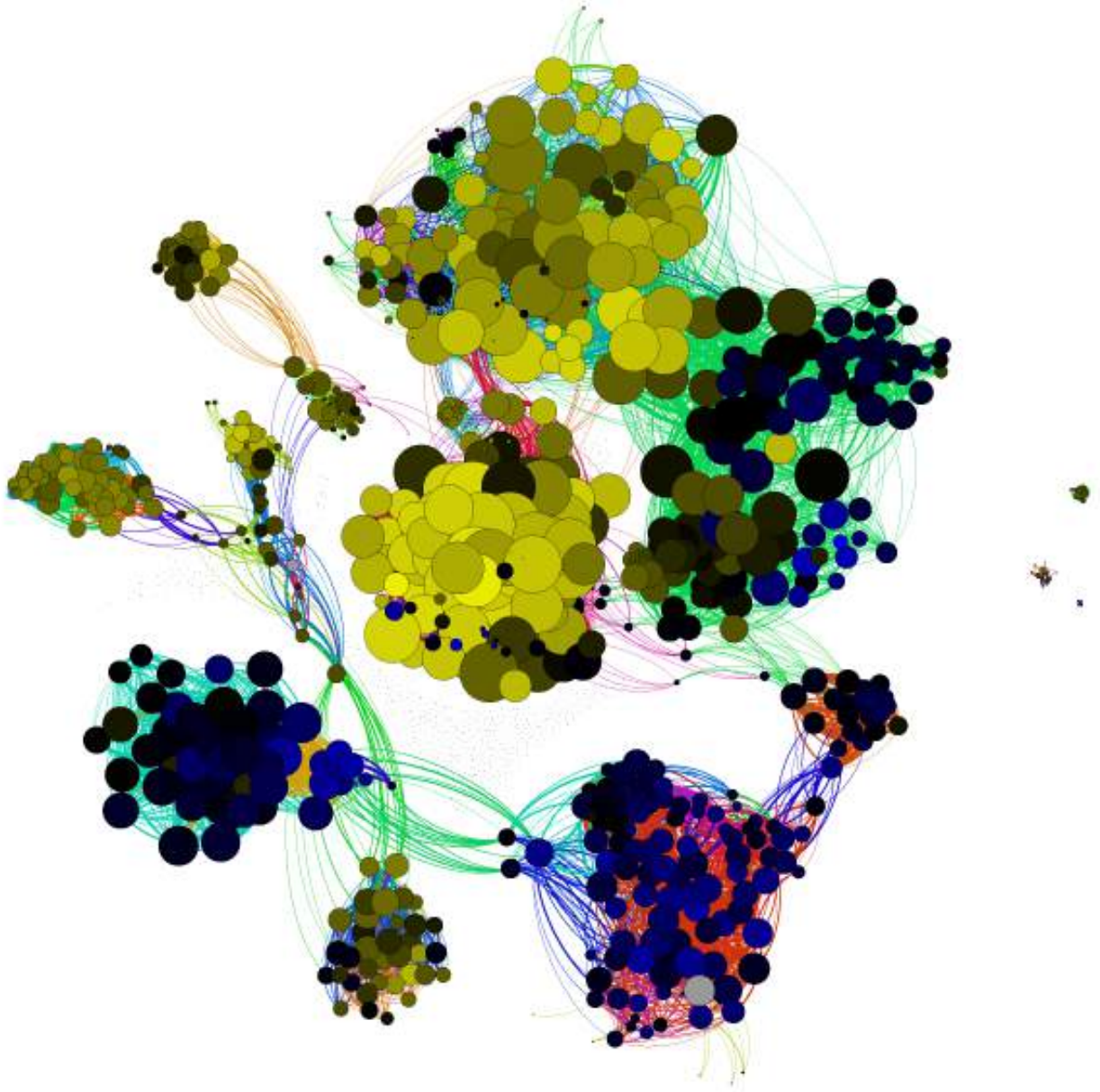
Combine: Inferred GREs resemble known TFBs





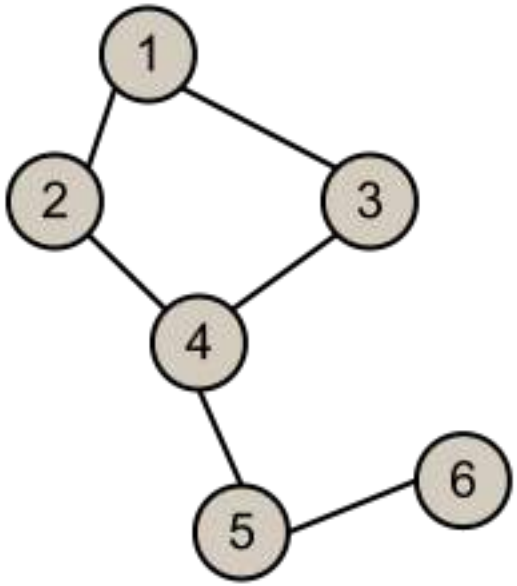
Brooks and Reiss et al 2014

Discover co-regulatory modules



Combine: Network representation for gene-gene co-regulation

Undirected Graph & Adjacency Matrix



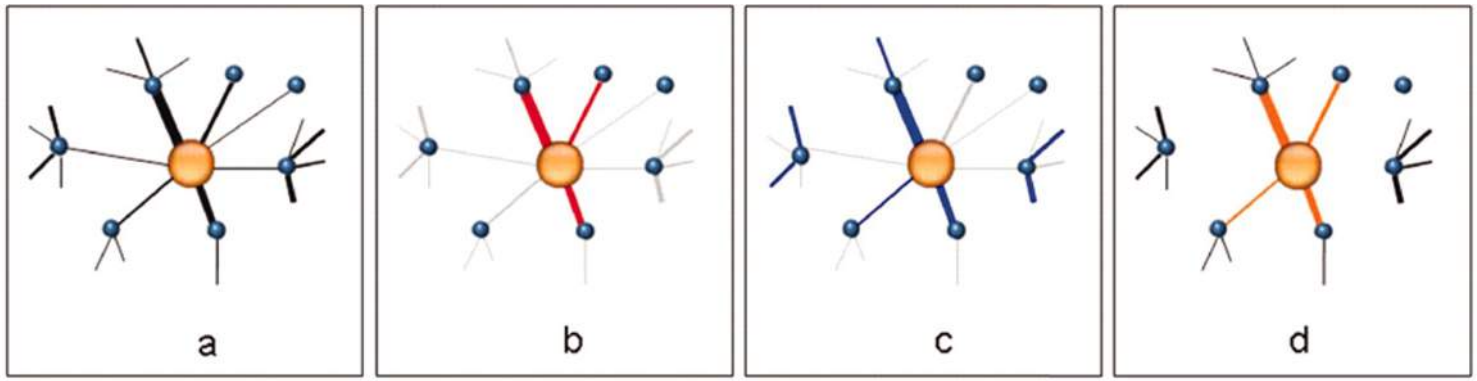
Undirected Graph

	1	2	3	4	5	6
1	0	1	1	0	0	0
2	1	0	0	1	0	0
3	1	0	0	1	0	0
4	0	1	1	0	1	0
5	0	0	0	1	0	1
6	0	0	0	0	1	0

Adjacency Matrix

Combine: network backbone extraction

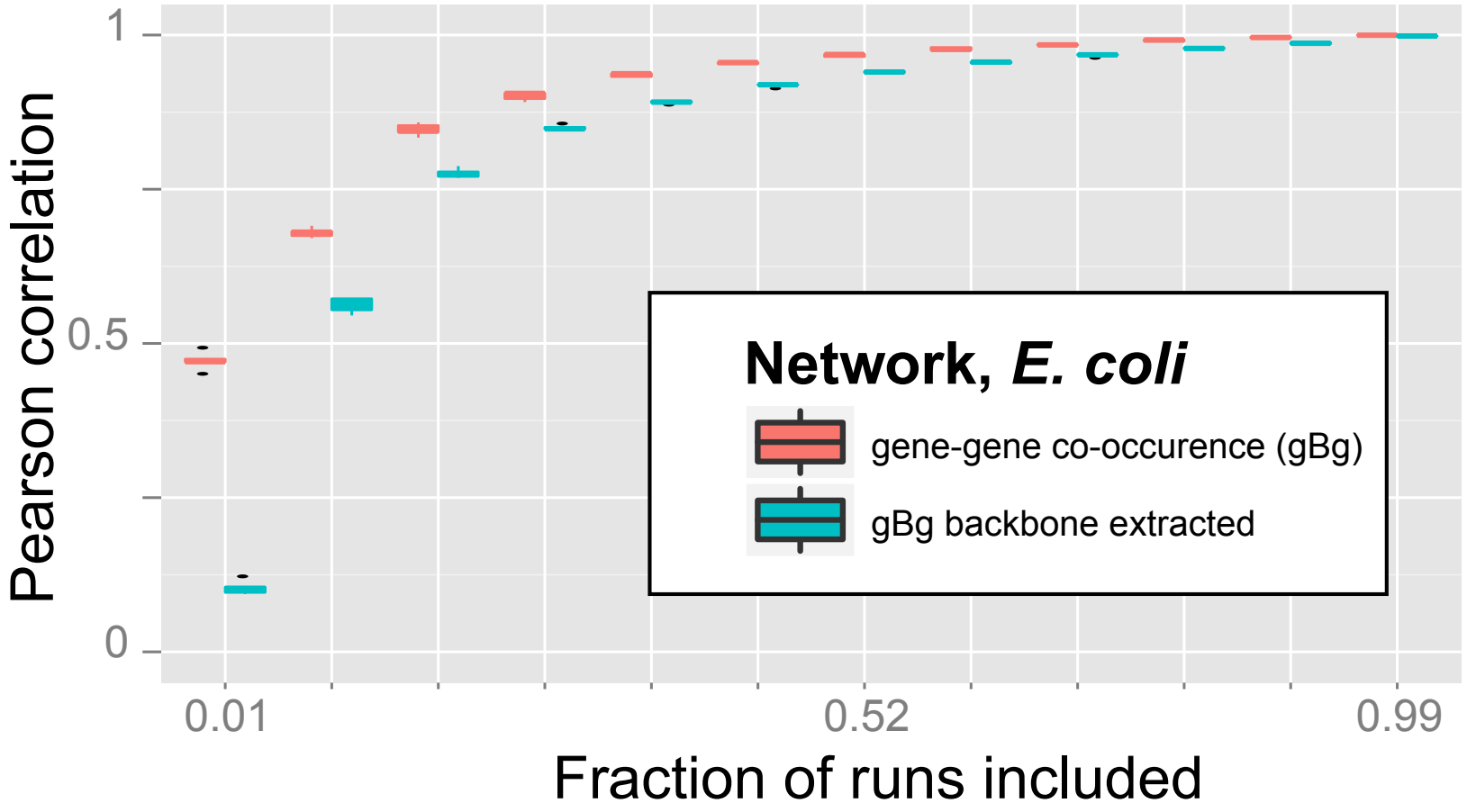
Original → Final



$$\alpha_{ij} = 1 - (k-1)\int_0^{w_{ij}} (1-x)^{k-2} dx \leq 0.05,$$

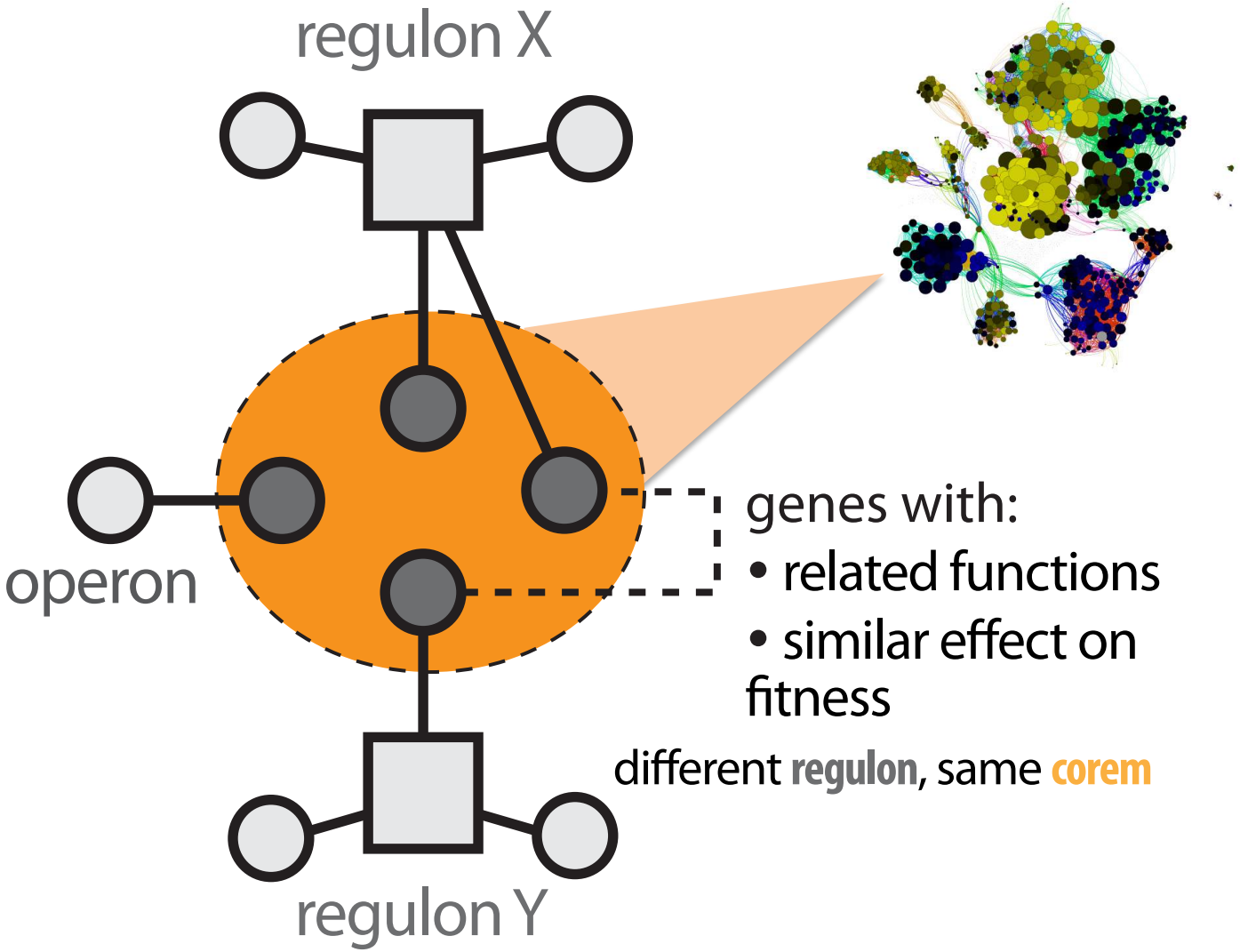
Serrano et al 2009

Network structure converges



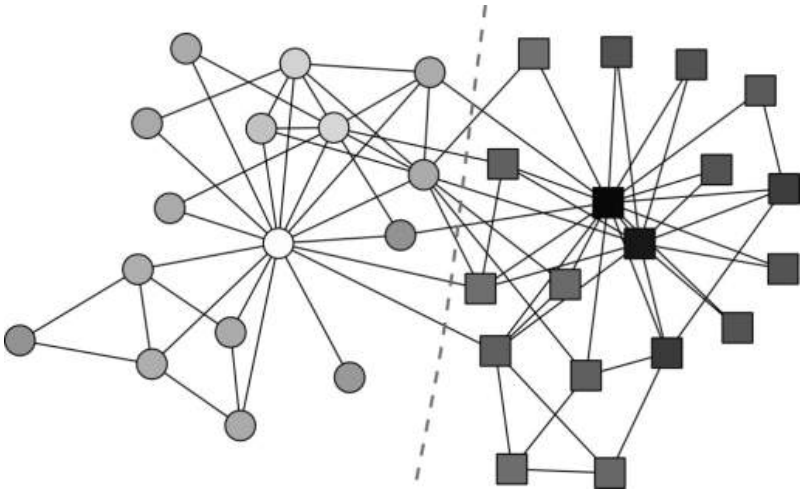
Brooks and Reiss et al 2014

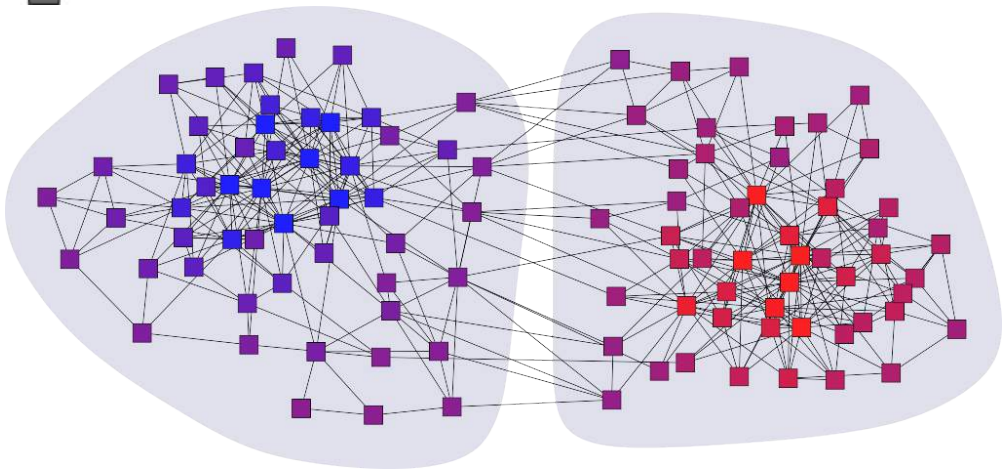
Corems: conditionally co-regulated modules



Brooks and Reiss et al 2014

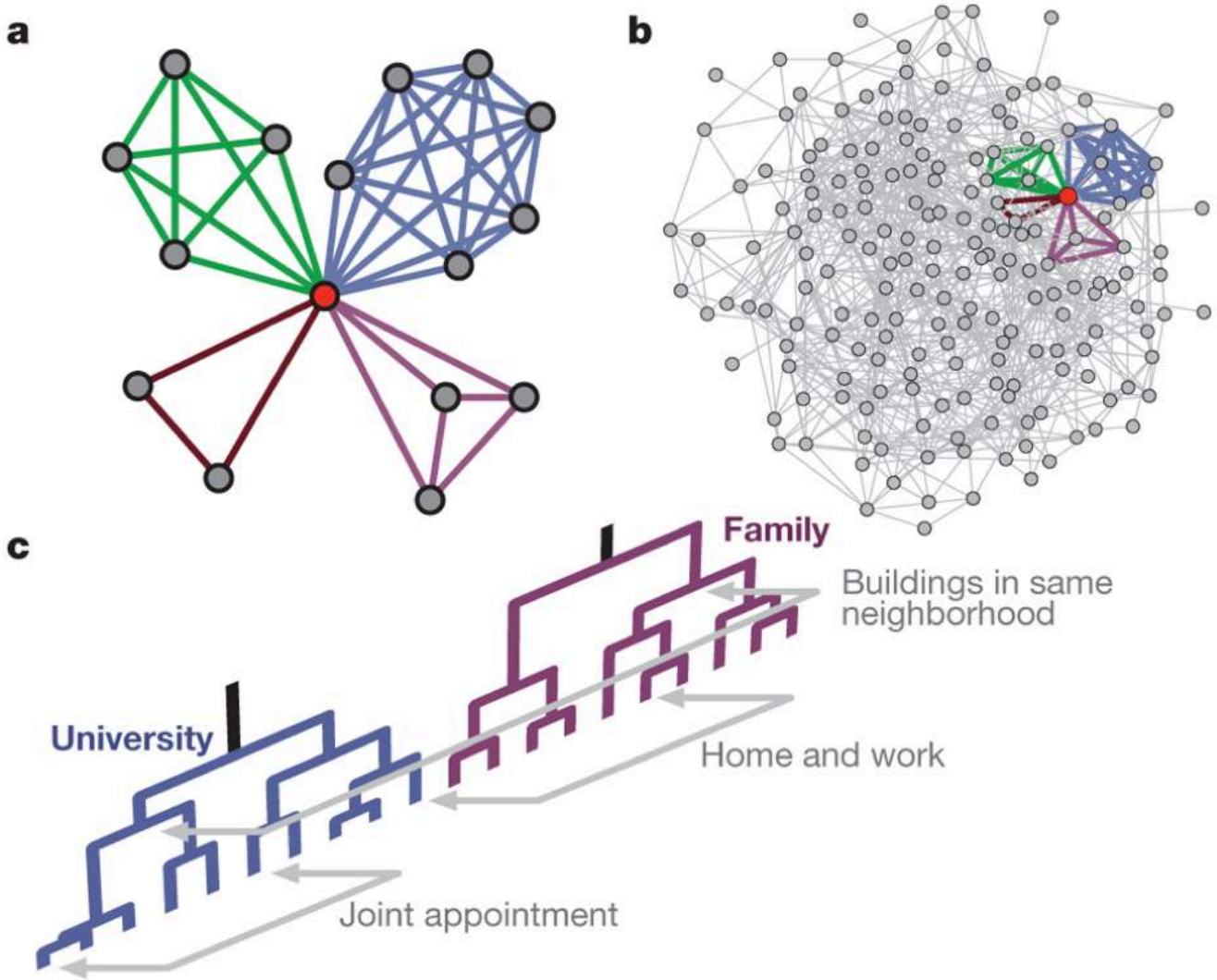
Community Detection





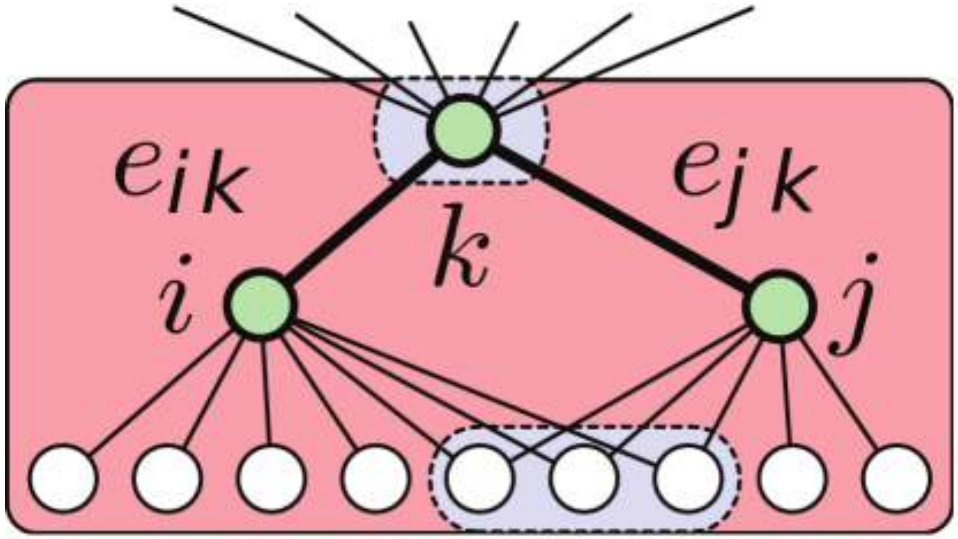
Newman 2006

Link Community Detection



Ahn et al 2010

Compute weighted similarity for every pair of edges

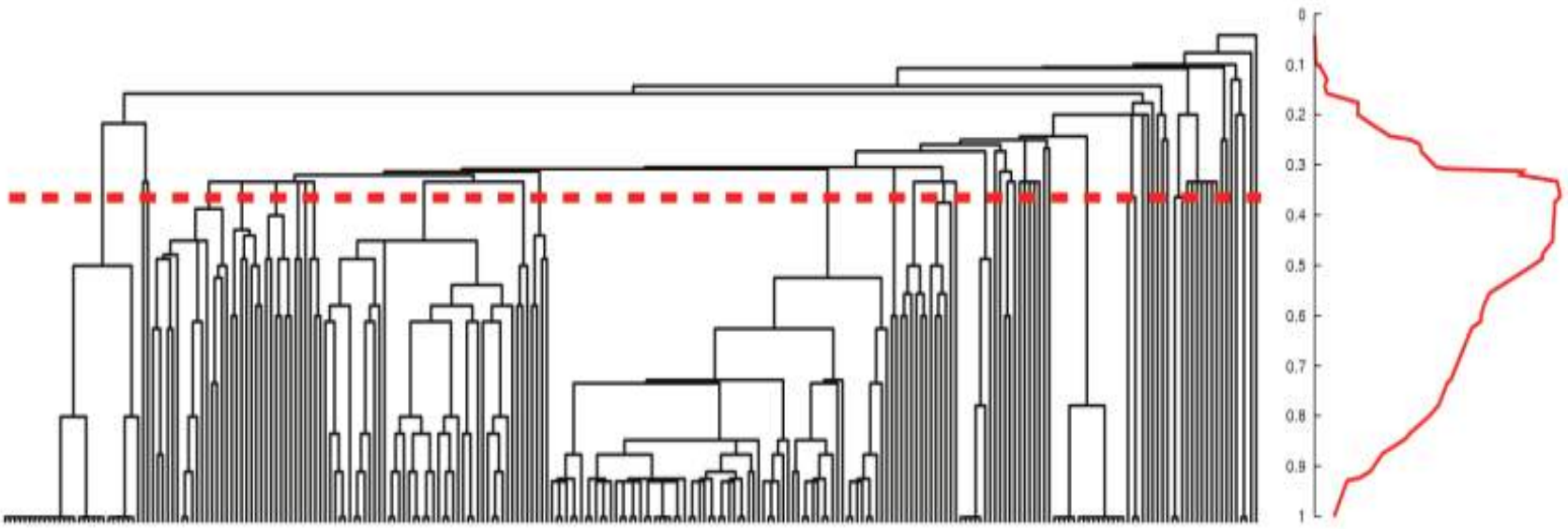


$$\widetilde{A}_{ij} = \frac{1}{k_i} \sum_{i' \in n(i)} w_{ii'} \delta_{ij} + w_{ij}$$

$$S(e_{ik}, e_{jk}) = \frac{\mathbf{a}_i \cdot \mathbf{a}_j}{|\mathbf{a}_i|^2 + |\mathbf{a}_j|^2 - \mathbf{a}_i \cdot \mathbf{a}_j}$$

Ahn et al 2010

Cluster edges to maximize intra-cluster density

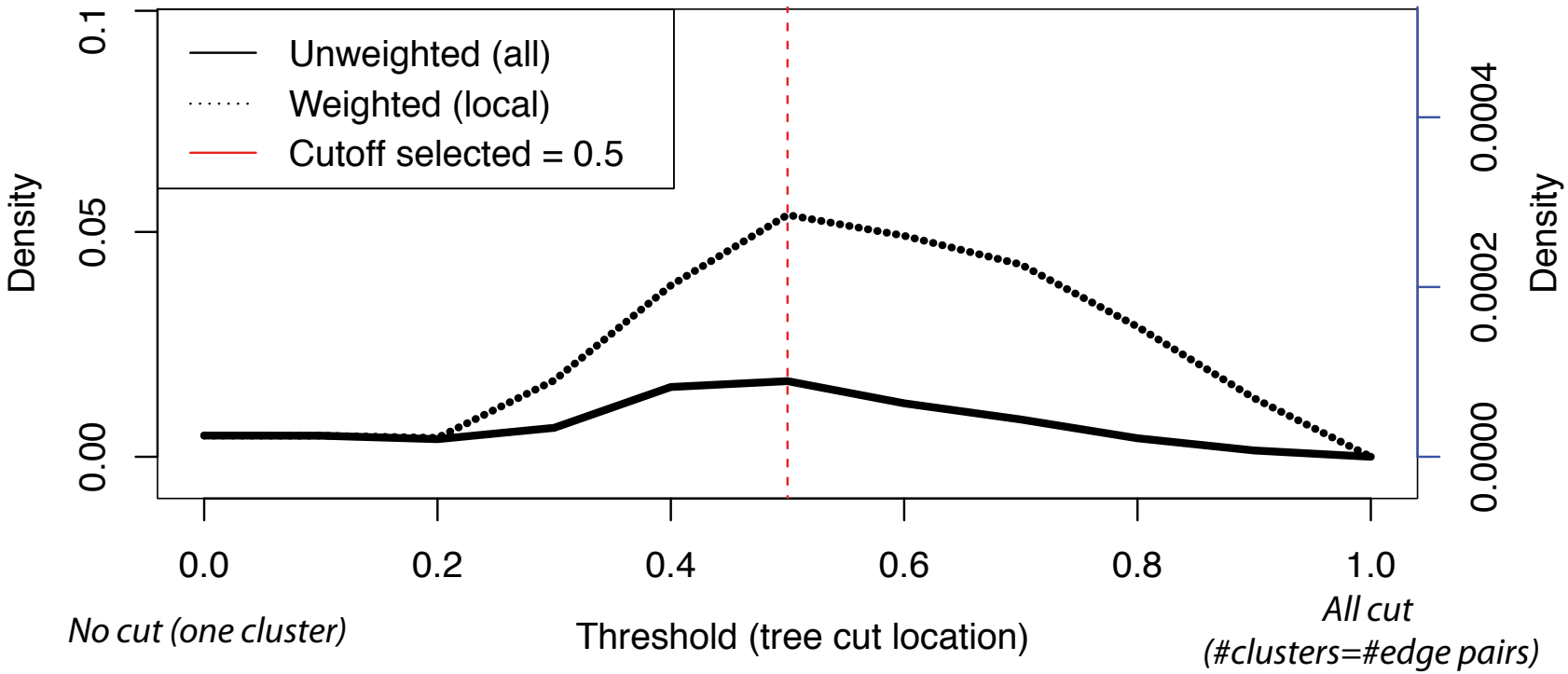


N	Total number of nodes in the network
M	Total number of edges in the network
$\langle w \rangle$	Average weight of edges of the network
$\mathcal{C}$	Partition of the network into edge communities
$\mathcal{E}$	Set of all edges in the network ($\lvert\mathcal{E}\rvert = M$)
n_c	Number of nodes in community c
m_c	Number of edges in community c
$\langle w \rangle_c$	Average weight of edges in community c

$$D^{(5)} = \frac{1}{M\langle w \rangle} \sum_{c \in \mathcal{C}} m_c \langle w \rangle_c \left(\langle w \rangle_c \left[\frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)} \right] \right)$$

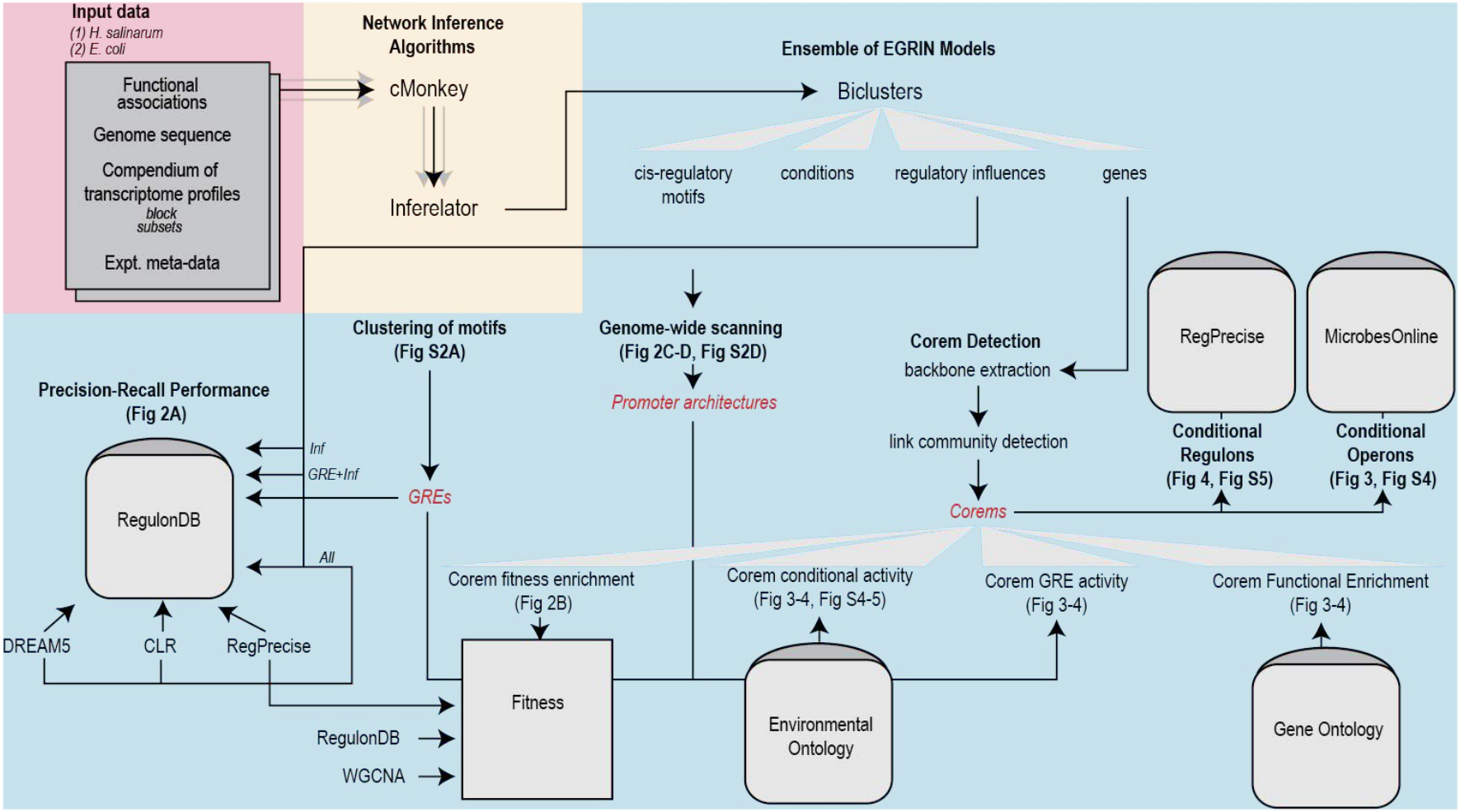
Ahn et al 2010

Network density given threshold selection for corem selection, *E. coli*



Brooks and Reiss et al 2014

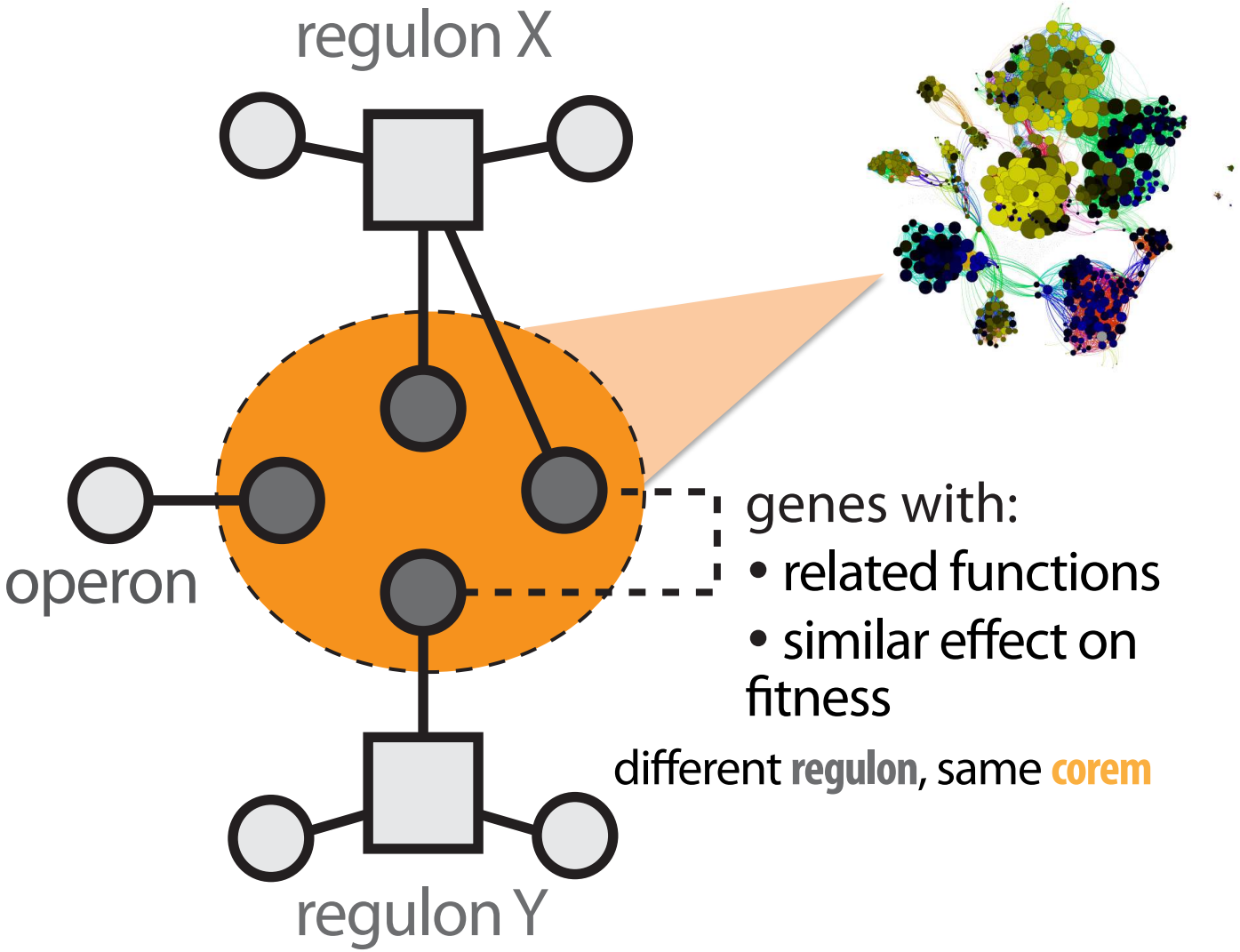
Full EGRIN 2.0 toolchain



SPLIT → **APPLY** → **COMBINE**

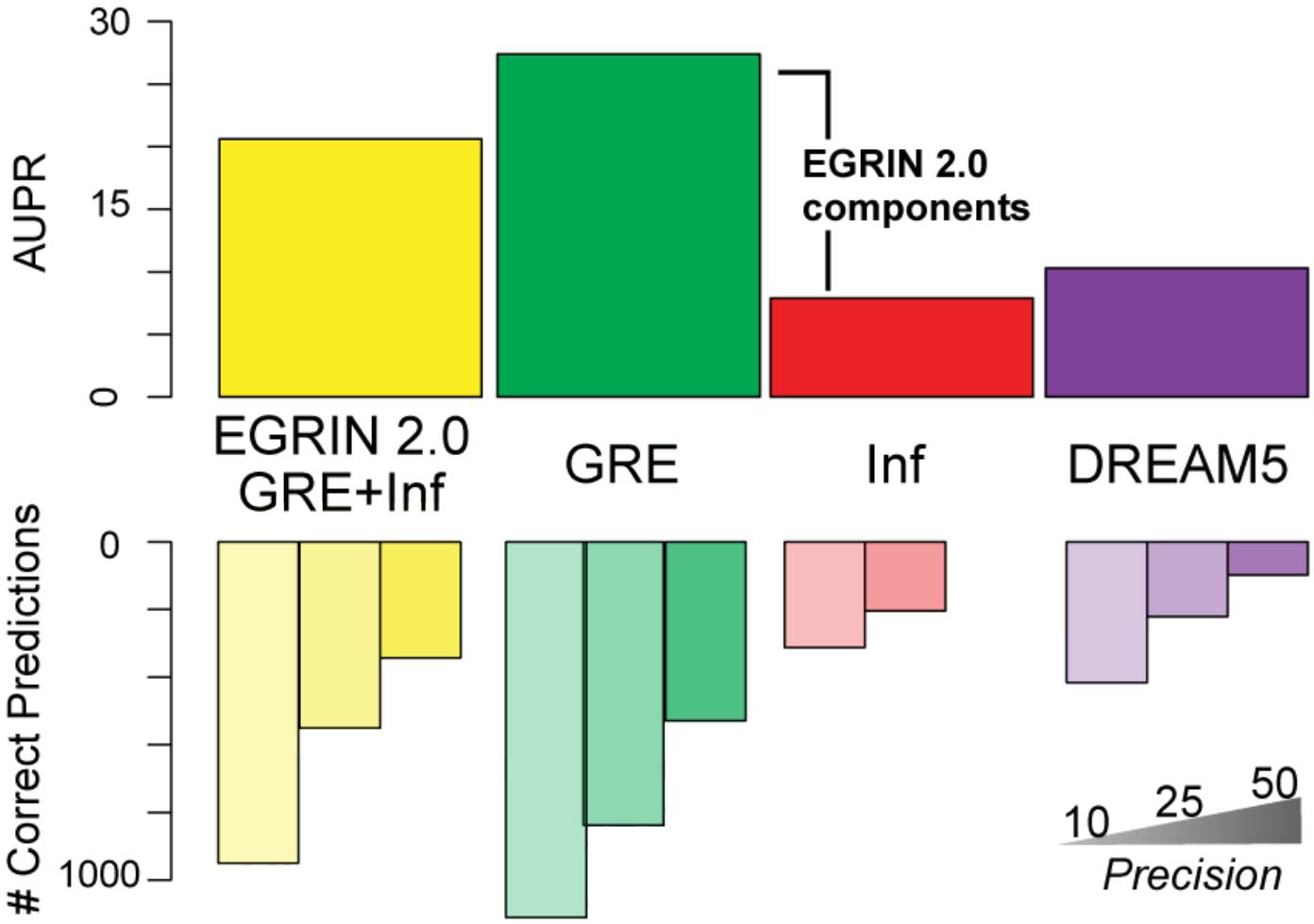
Brooks and Reiss et al 2014

Corems: conditionally co-regulated modules



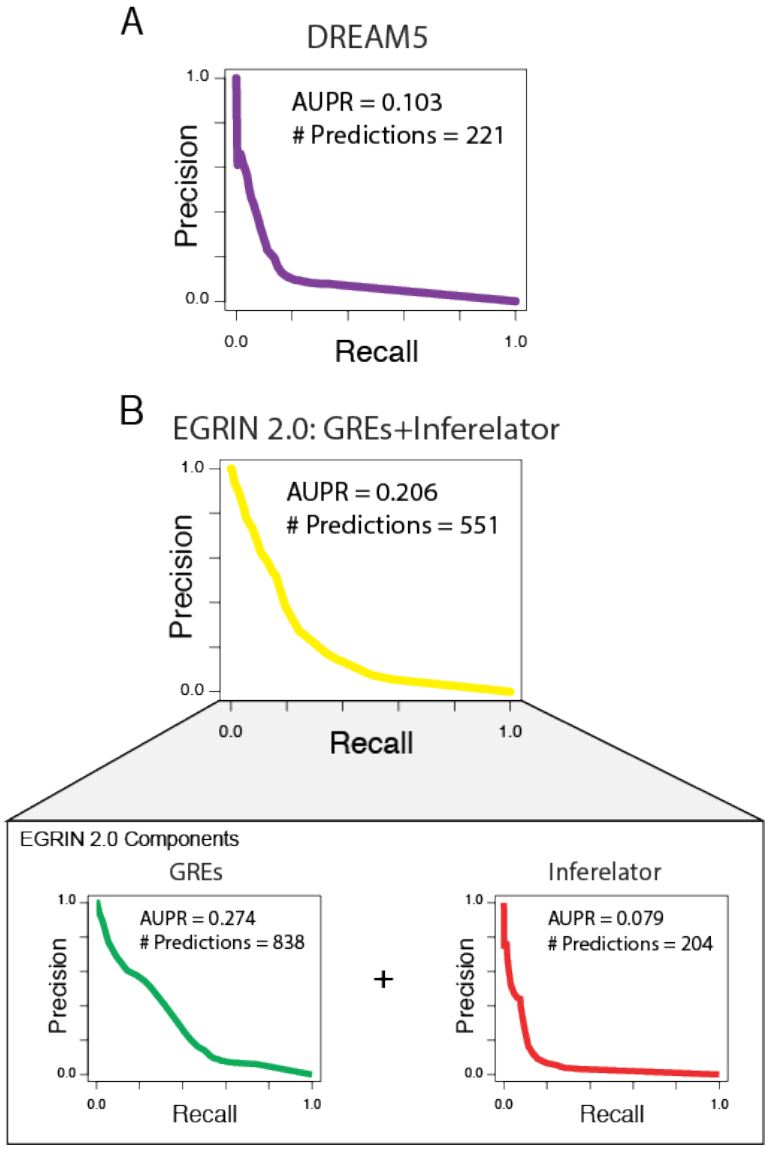
Brooks and Reiss et al 2014

EGRIN 2.0 has ~3-fold increased performance

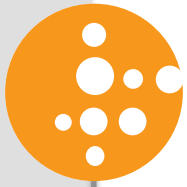


Brooks and Reiss et al 2014

Performance increase due to motif detection

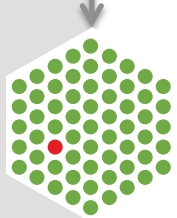


Brooks and Reiss et al 2014

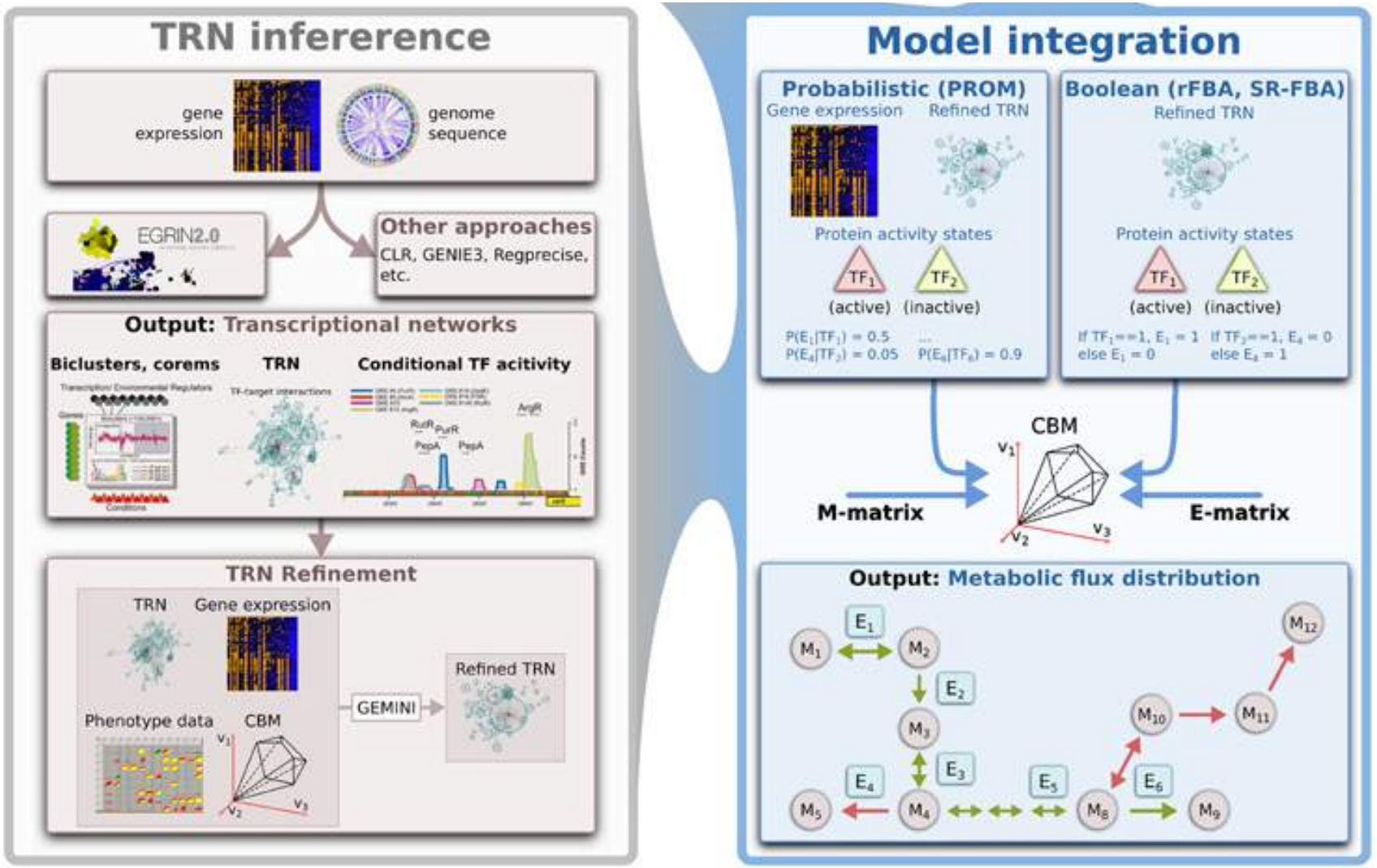


Extensions

Integration with metabolic models

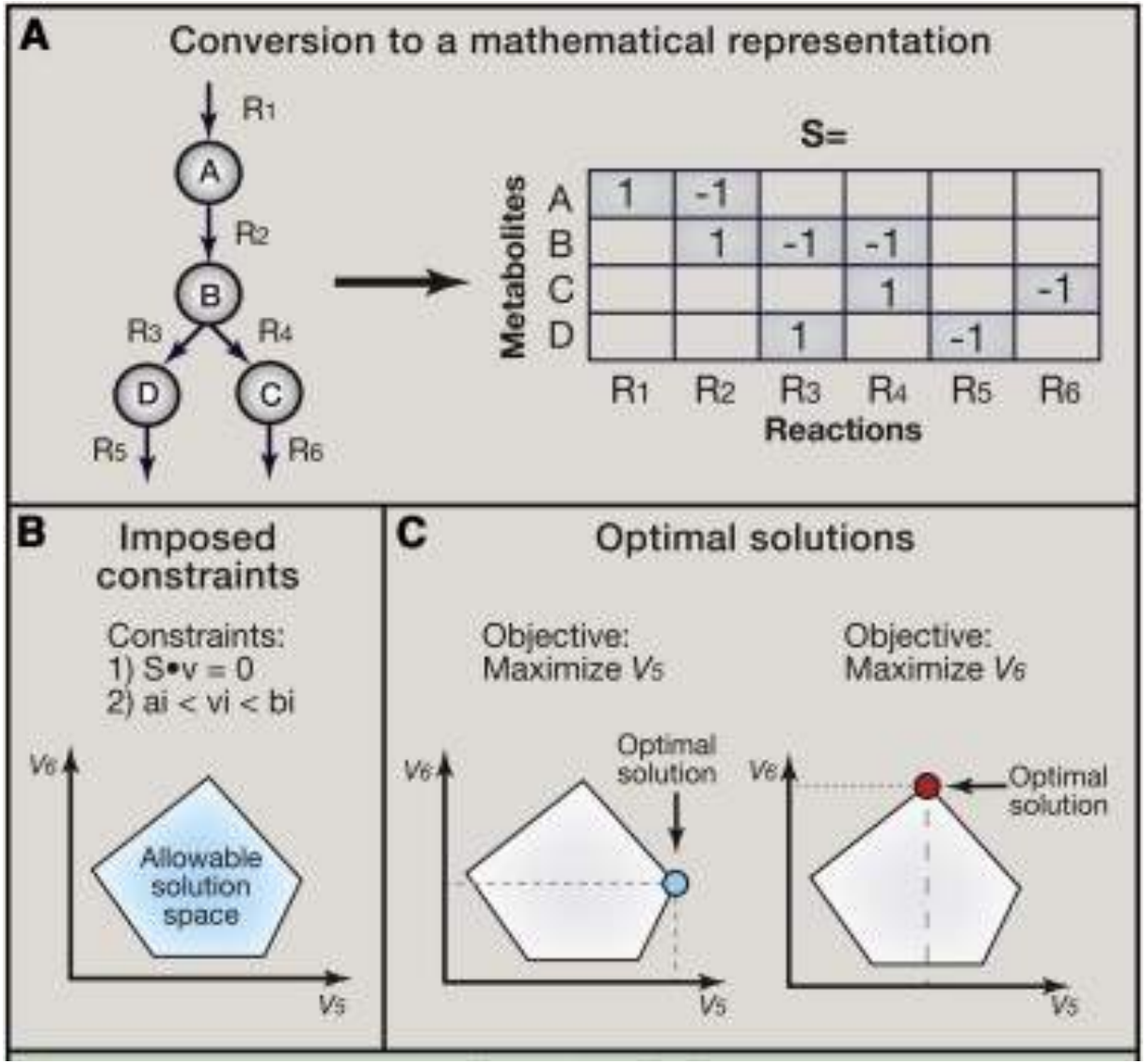


Integration with models of metabolism



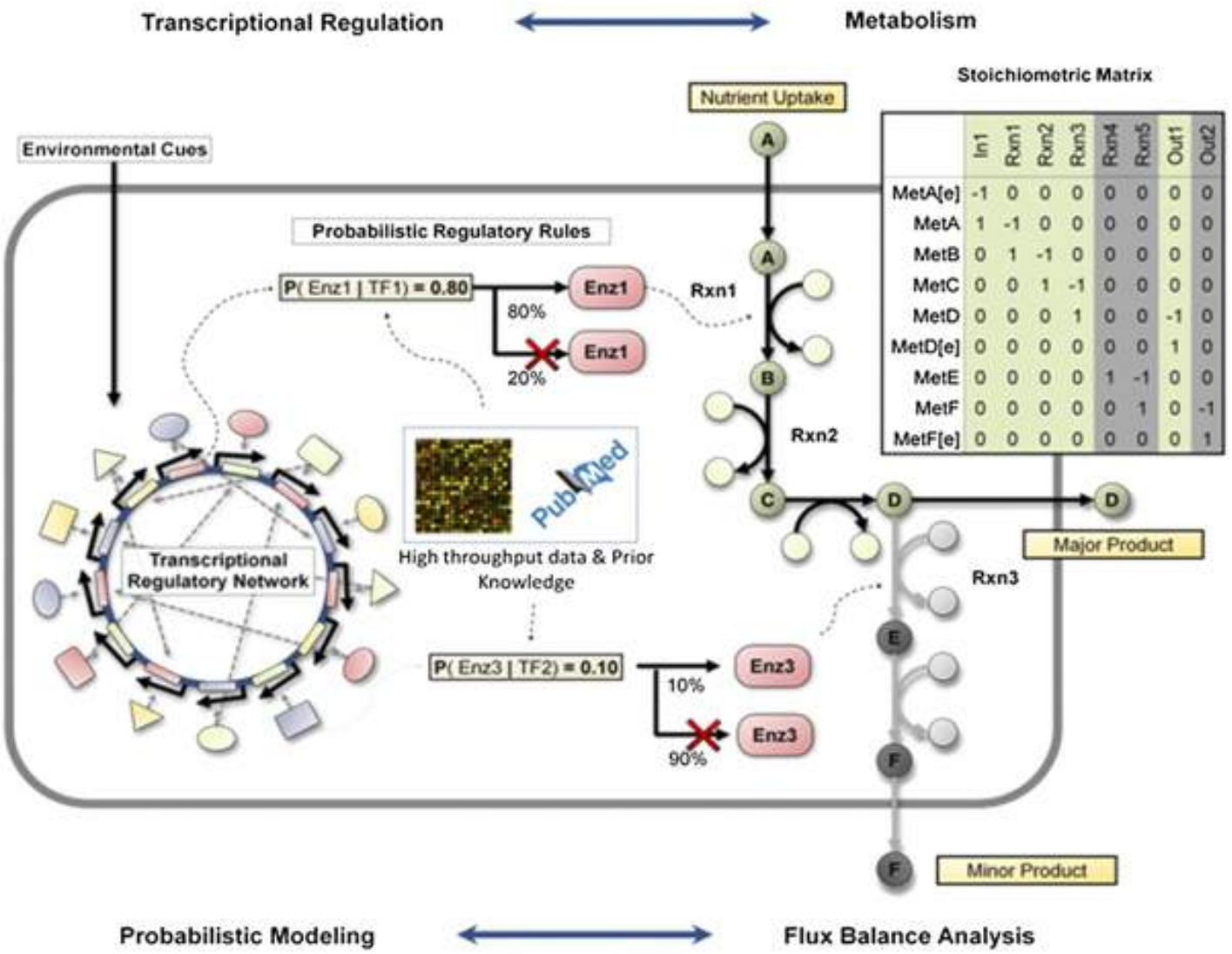
Imam et al 2015

Genome-scale models of metabolism



O'Brien et al 2015

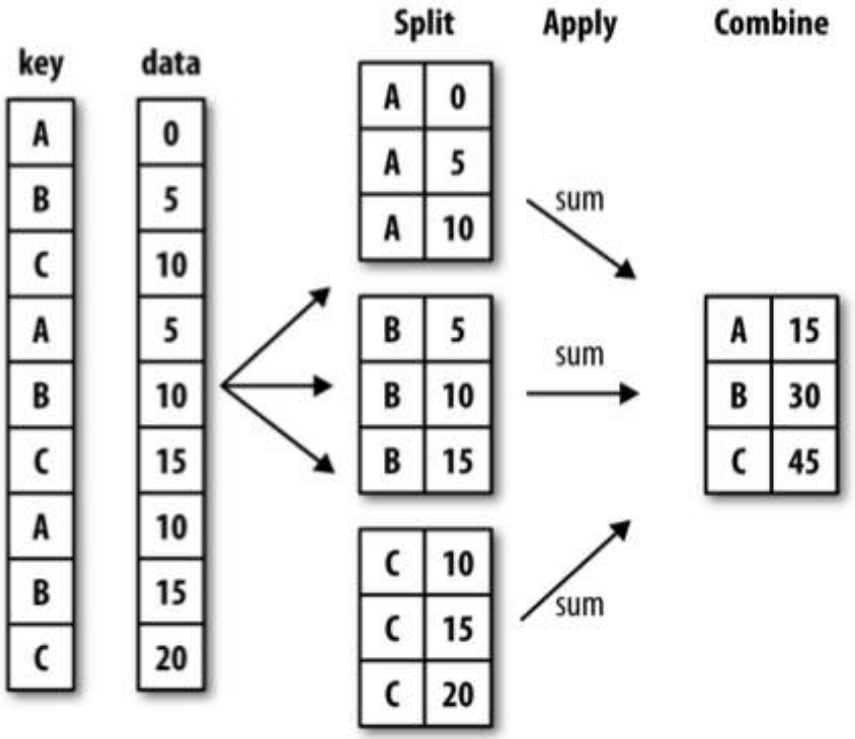
Probabilistic Regulation of Metabolism (PROM)



	In1	Rxn1	Rxn2	Rxn3	Rxn4	Rxn5	Out1	Out2
MetA[e]	-1	0	0	0	0	0	0	0
MetA	1	-1	0	0	0	0	0	0
MetB	0	1	-1	0	0	0	0	0
MetC	0	0	1	-1	0	0	0	0
MetD	0	0	0	1	0	0	-1	0
MetD[e]	0	0	0	0	0	0	1	0
MetE	0	0	0	0	1	-1	0	0
MetF	0	0	0	0	0	1	0	-1
MetF[e]	0	0	0	0	0	0	0	1

Chandrasekaran and Price 2010

Split-Apply-Combine-like strategy for building multistep, data-driven inference toolchains



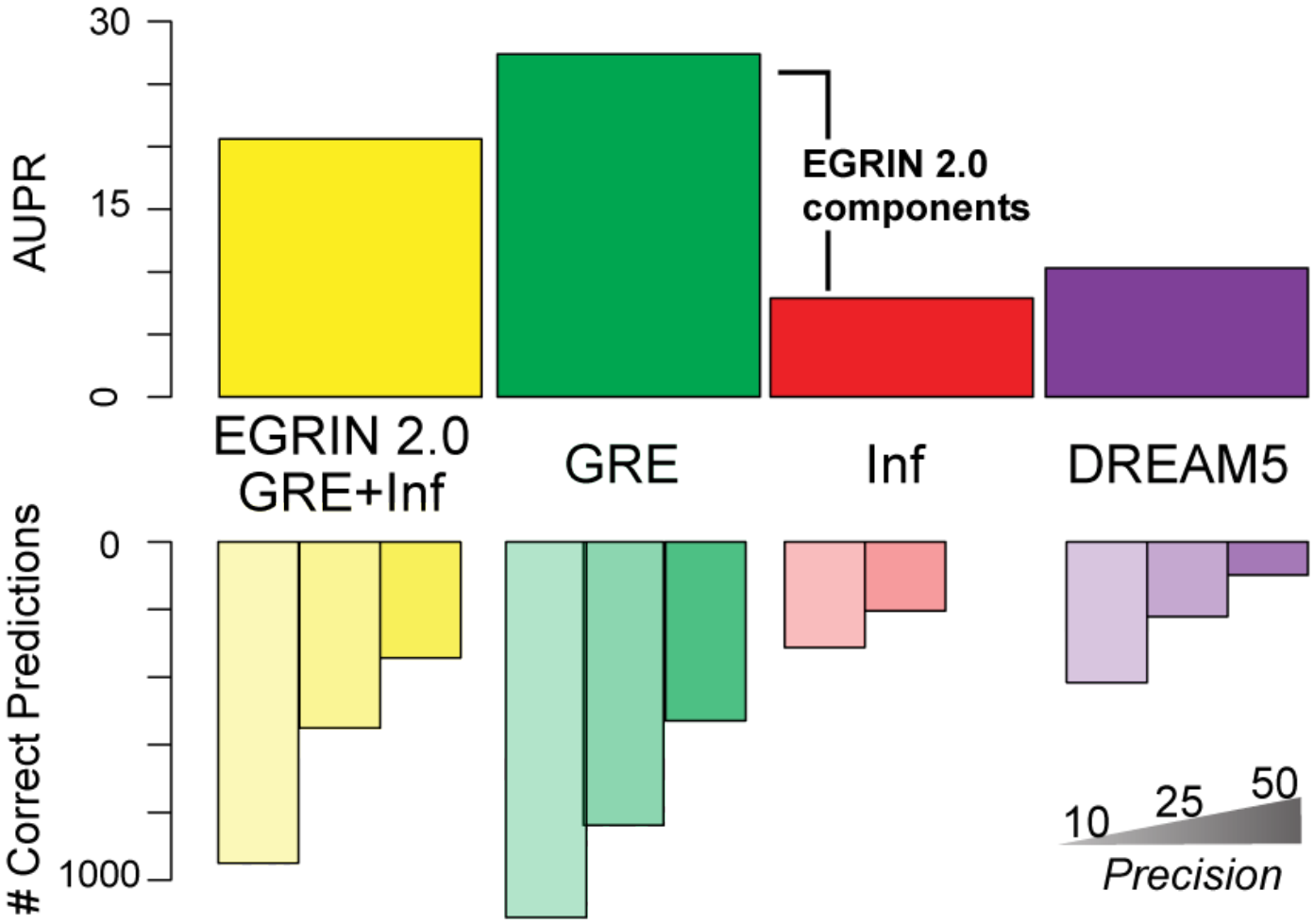
SPLIT → APPLY → COMBINE

Split-Apply-Combine-like strategy for building multistep, data-driven inference toolchains

- Heterogeneous data can (and should) be integrated

  Increased predictive performance by simultaneous motif detection

EGRIN 2.0 has ~3-fold increased performance



Brooks and Reiss et al 2014

Split-Apply-Combine-like strategy for building multistep, data-driven inference toolchains

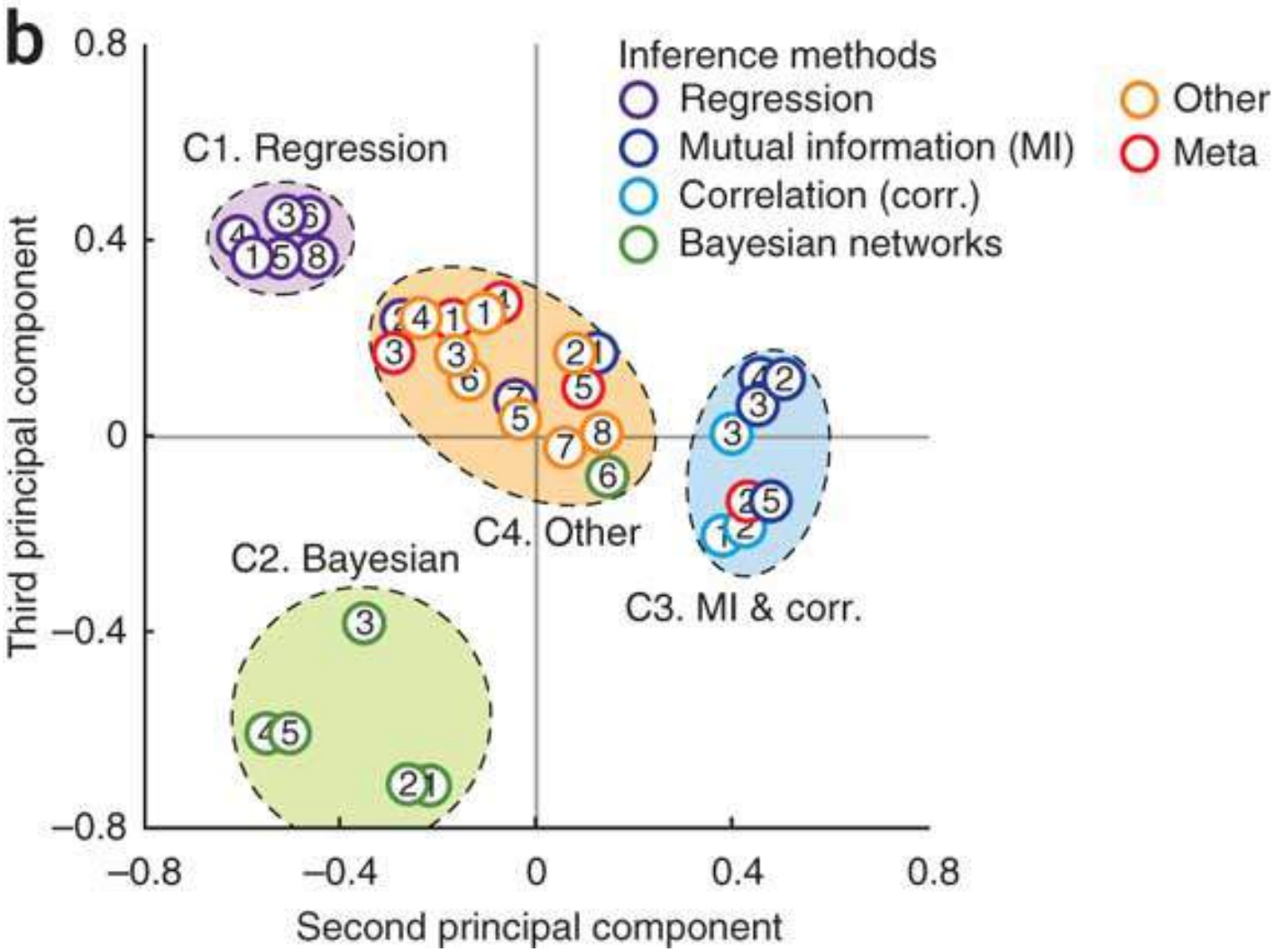
- Heterogeneous data can (and should) be integrated

  Increased predictive performance by simultaneous motif detection

- Ensemble learning frameworks can harness model diversity

  Combining models increases overall model performance

Algorithms learn different network features



Marbach et al 2012

Summary

Split-Apply-Combine-like strategy for building multistep, data-driven inference toolchains

- Heterogeneous data can (and should) be integrated

  Increased predictive performance by simultaneous motif detection

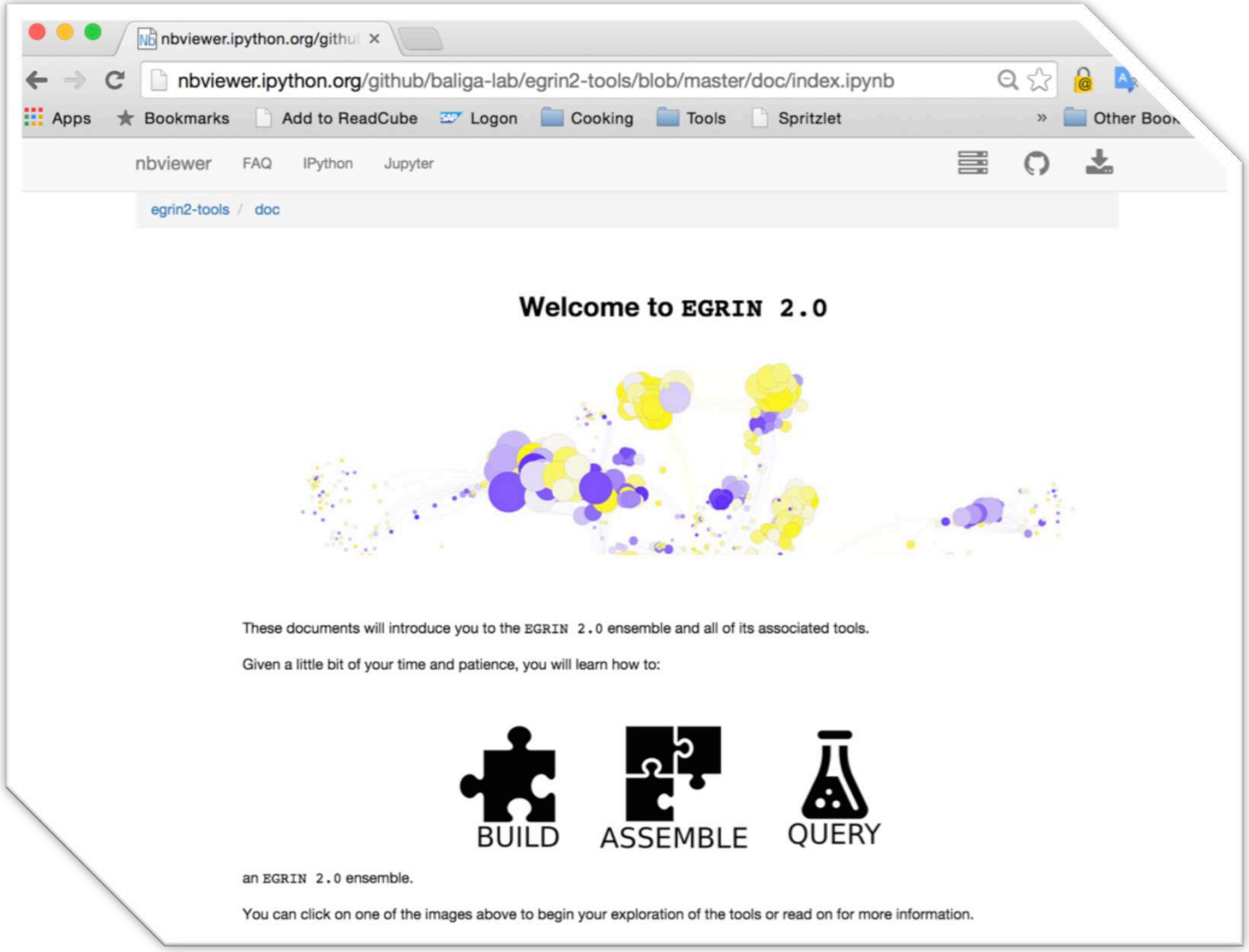
- Ensemble learning frameworks can harness model diversity

  Combining models increases overall model performance

- Graph-based methods can flexibly encode and reconcile predictions

  Examples of model reconciliation with graphs: GRE and corem detection

New tools for SAC-biology



github.com/baliga-lab/egrin2-tools

ISB

Sriram Chandrasekaran
Amardeep Kaur
Min Pan
Chris Plaisier
Dave Reiss
Diego Salvanha
Wei-ju Wu
Nitin Baliga

Université Laval

Antoine Allard

Committee

Nitin Baliga
Elhanan Borenstein
Caroline Harwood
Herbert Sauro
Ilya Shmulevich

Funding

U.S. DEPARTMENT OF ENERGY | Office of Science
NIH
NSF
MOLECULAR AND CELLULAR BIOLOGY
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.