Microbial Data Analysis: towards Systems approach

Yixiang Shi

Shanghai Center for Bioinformation Technology

May 11, 2011

Is Microbiology Vanishing?

Why? —Most of the researchable problems are either "having been solved," or "too difficult to be solved."

1. Clinical microbiology:

Pathogen and host: mechanism of pathogenesis and virulence, immune response and escape

2. Agriculture and industry microbiology:

A well developed applied science, great contributions and great potential.

How to further improve the present industry?

How to solve the difficult problems?

3. Basic research in microbiology:

Known、Unknow、Hard to know

Fundamental problems + Key technology platforms

Microbial Genomics Research: An opportunity and challenge to revitalize microbiology

- **Initiation of microbial genomics:**
  - **Model system for human genomics**
- **Advantages of microorganisms as targets for genomic studies:**
  - **Highly diversified physiology**
  - **Extremely long history of evolution**
  - **Close relationship with host/environment**
  - **Relatively small genome size**
  - **Relatively easy for functional analysis**
- **Microbial genomics solved the bottle-neck technology for further development of microbiology:**
  - **It is impossible to establish genetic systems for each of the majority of the organisms being studied**
- **New scientific challenges brought in by microbial genomics research:**
  - **Environment and evolution**
  - **Structure and function (novel genes)**
  - **Non-cultured microorganism**

Historical Recall

- **Whole genome sequencing for bacteriophage and virus chromosome:** φX174 (1978, **5386bp**)
- Microbial genomics research **before the heat wave of HGP:**

  E. coli Kohara library (**cosmid contig**, middle 80's), model systems for HGP (early 90's)

- **The completion of the genomic sequence of *Haemophilus influenzae* Rd** (1995, **1.8Mb,** encoding 1743 hypothetical genes):

  Whole genome shotgun sequencing strategy (**500 bps/read**) and **bioinformatics** (**assembly and annotation/gene assignment**).

Status of Genomic Sequence for Microorganisms

- **June 6, 2000**
  - **Completed, present in public databases [31 genomes]**
    Completed, annotation in progress [17 genomes]
    Sequencing in progress [70 genomes]
- **October 28, 2001**
  - **Completed, present in public databases [58 genomes]**
    Completed, annotation in progress [17 genomes]
    Sequencing in progress [110 genomes]
- **April 19, 2003**
  - **Completed [112 genomes, NCBI]**
- **May, 2011**
  - **Completed [`1676 Bacterial, 89 Archaeal and 305 Eukaryotic genomes, NCBI]**

Some basic questions asked for microbial genomics research

- Are new concepts emerging about how cells work?
  - Yes.
  - Completeness, Comparison and Compact
- Have there been practical benefits in the fields of medicine and agriculture?
  - Perhaps.
  - Encapsulation, Type III secretion system, Pathogenicity islands and symbiotic islands
- Is it feasible to determine the genomic sequence of every bacterial species on Earth?
  - No. But…

How to start? Selection of research objectives

- Significance (scientific and application)
  - Medicine:
    - Disease mechanism, drug resistance, new drug targets, candidate protein for vaccine development
  - Agriculture, industry, ecology and environment:
    - Secondary metabolism, pollution, remediation
  - Science:
    - Evolution, origination of sub-cellular organelles,
    - genetic resources, non-cultured microorganisms
- Technology development
- Reliability of the biology for the strains being sequenced.

Microbial Genomics Research in China (1999, the starter)

Human Genome Project of the *Knowledge Innovation Program*, Chinese Academy of Sciences:

***Thermoanaerobacter tengchongensis* MB4:**

Sequence completed by BGI. Established the technology and the research team for 1% human genome project

Before Sept. 2000: Shotgun sequencing finished. 110,000 reads.

Jan. 2001: Assembly finished. *Sfi*I，*Asc*I and *SgrA*I *in silico* physical map matched with the experimental data.

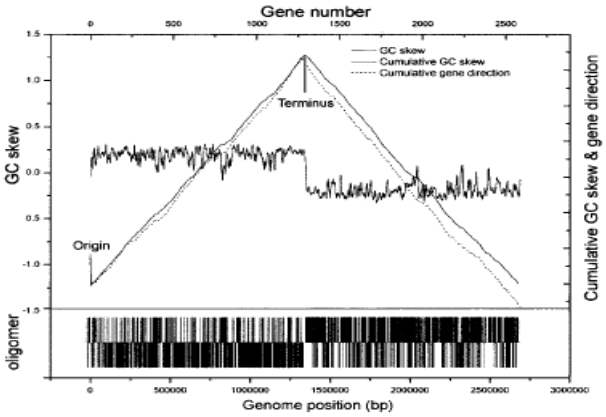
Feb-May, 2001: Annotation employing Glimmer and other software.

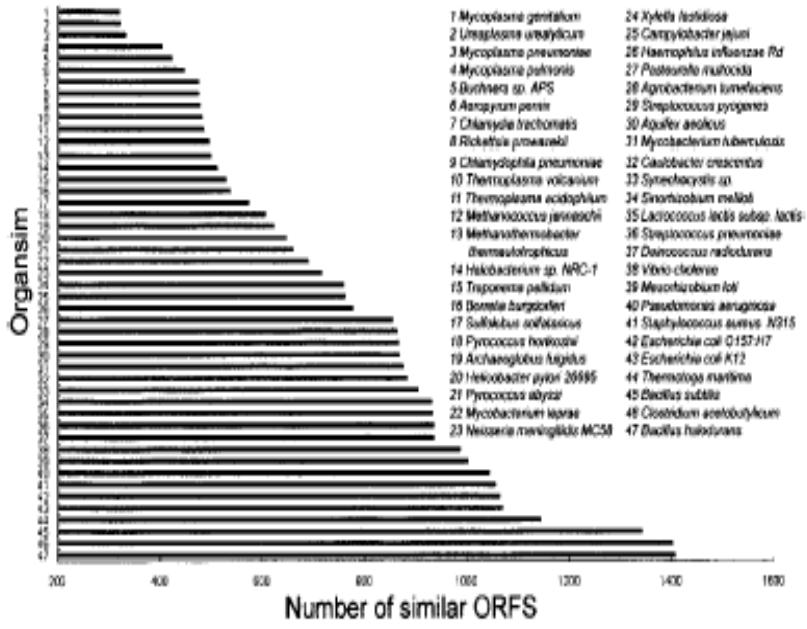
2002: Qiyu BAO *et al*. A complete sequence of the *T. tengcongensis* genome. *Genome Research* **12:** 689-700

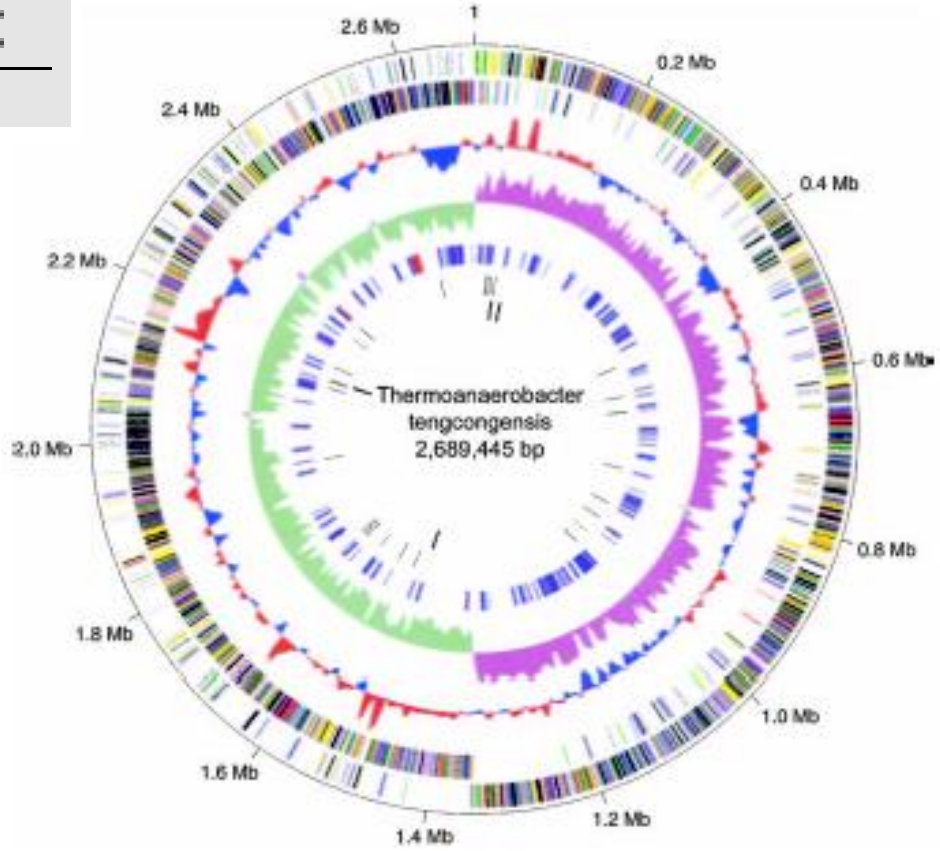
Table 1. General Features of the *Thermoanaerobacter tengcongensis* Genome

Genome size (bp)		2,689,445
G+C content		37.6%
Protein coding		87.1%
Average CDS length (bp)		905
Predicted CDS		2,588
Homologous to	Known proteins	1494 (57.8%)
	Protein domains/motifs	270 (10.4%)
	Hypothetical proteins	301 (11.6%)
No homology		523 (20.2%)
Stable RNAs		0.9%
rRNA operons		4
rRNAs		55
Major repetitive elements		
Short noncoding repeats		0.5%
Long coding repeats		8.6%

CDS, coding sequences.







Microbial Genomics Research in China (2000-, Step Up)

- ***Shigella flexneri* strain 2a:**
  - **The first pathogenic bacterial genomic sequencing project.**
  - **2000：** Sequencing finished within one year.
  - **2002：** Qi JIN *et al.* Genome sequence of *Shigella flexneri* 2a: Insights into pathogenecity through comparison with genomes of *Escherichia coli* K12 and O157 *Nucleic Acids Res.* **30** (20): 4432-4441
- *Leptospira interrogans* serogroup Icterohaemorrhagiae serotype lai
- *Staphylococcus epidermidis* ATCC 12228
- *Xanthomonas campestris* pv. *campestris* strain 8004

Table 1. General features of the Sf301 genome compared with genomes of *E.coli* K12 and 0157, and the virulence plasmid, pWR501, from *S.flexneri* M90T 5a

Chromosome	Sf301	MG1655[a]	EDL933[b]
Total length (bp)	4 607 203	4 639 221	5 528 445
No. of total ORFs	4434	4289	5349
Average length of ORFs (bp)	891	954	905
Percentage of coding sequence (%)	80.4	87.8	87.1
G + C content			
Total genome (%)	50.89	50.79	50.40
Protein coding regions (%)	51.95	51.85	51.51
RNA genes (%)	54.79	54.84	54.88
Intergenic regions (%)	46.07	42.28	42.76
Ribosomal RNA			
No. of 16S	7	7	7
No. of 23S	7	7	7
No. of 5S	8	8	8
No. of transfer RNA	97	92	93
No. of tmRNA	1	1	1
No. of non-classical RNA	9	5	5
Translocations and inversions[c]	13	–	1
IS elements	314	39	40
Of which partial copies	67	7	19
Plasmid	**pCP301**	**pWR501[d]**	
Total length (bp)	221 618	221 851	
No. of total ORFs	267	293	
Average length of ORFs (bp)	658	636	
Percentage of coding sequence	76.24	82.09	
G + C content			
Total (%)	45.77	46.36	
Coding regions (%)	46.13	46.95	
Intergenic regions (%)	44.59	43.69	
IS elements	88	92	
Of which partial copies	62	69	

[a]Data are from Blattner *et al.* (10).
[b]Data are from Perna *et al.* (11).
[c]Only those with DNA segments >5 kb are listed.
[d]Data are from Venkatesan *et al.* (8).

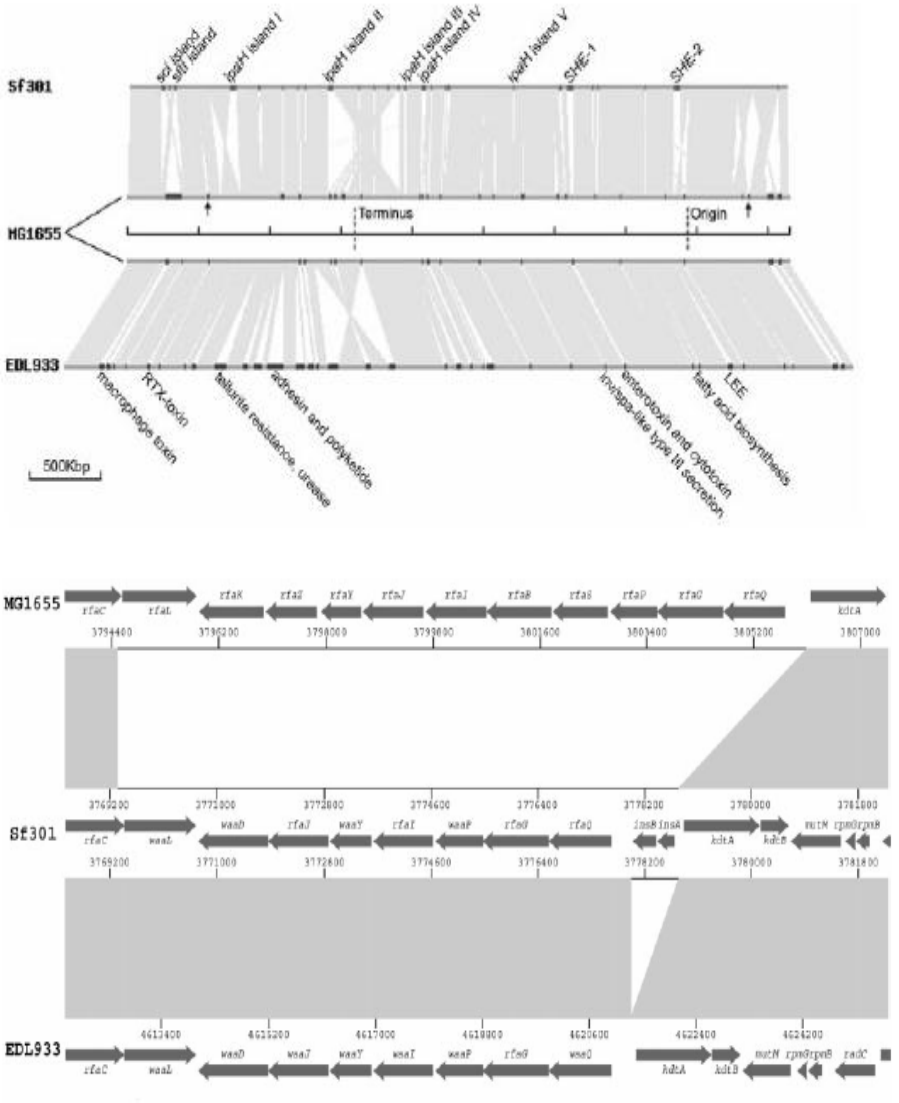
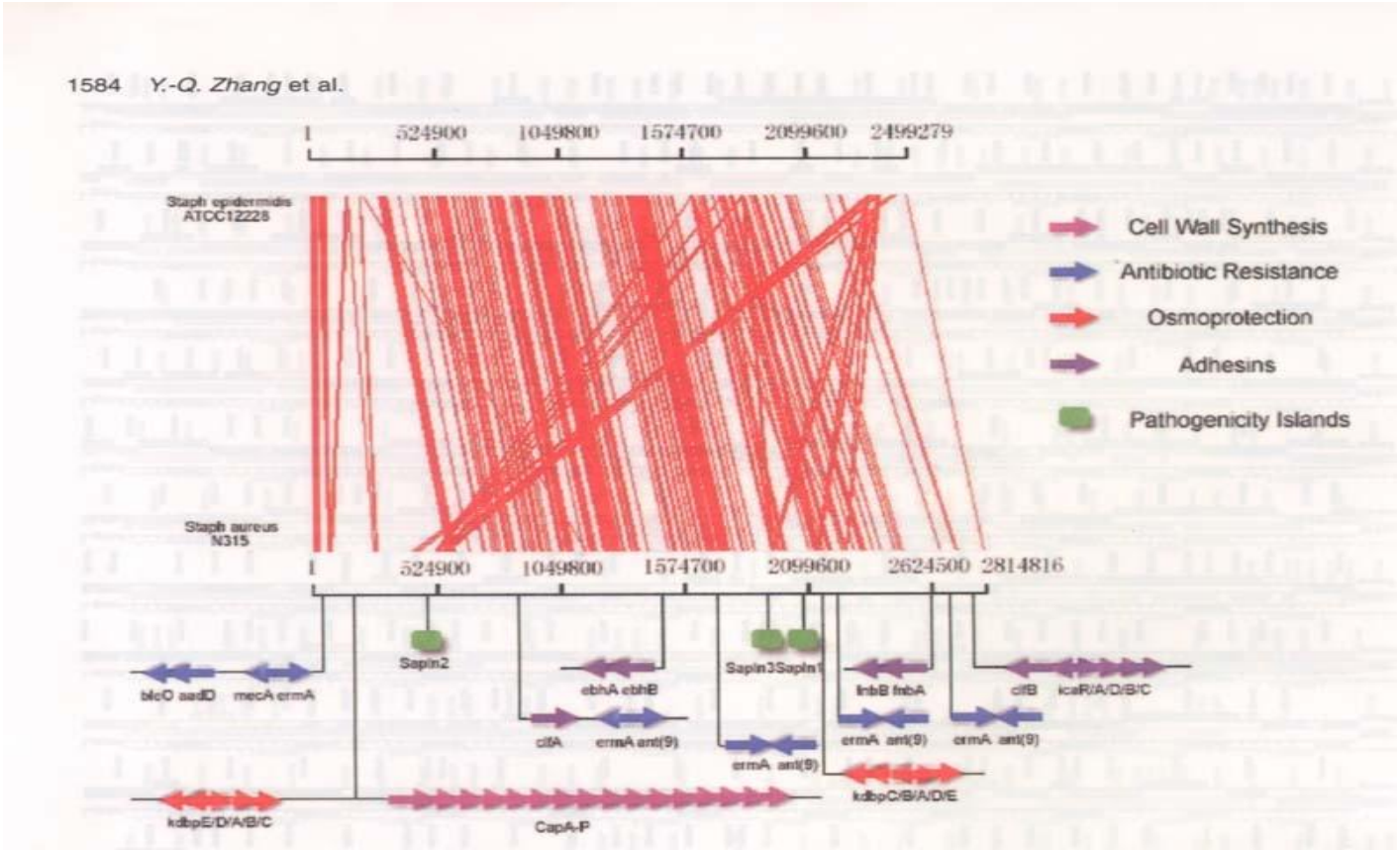
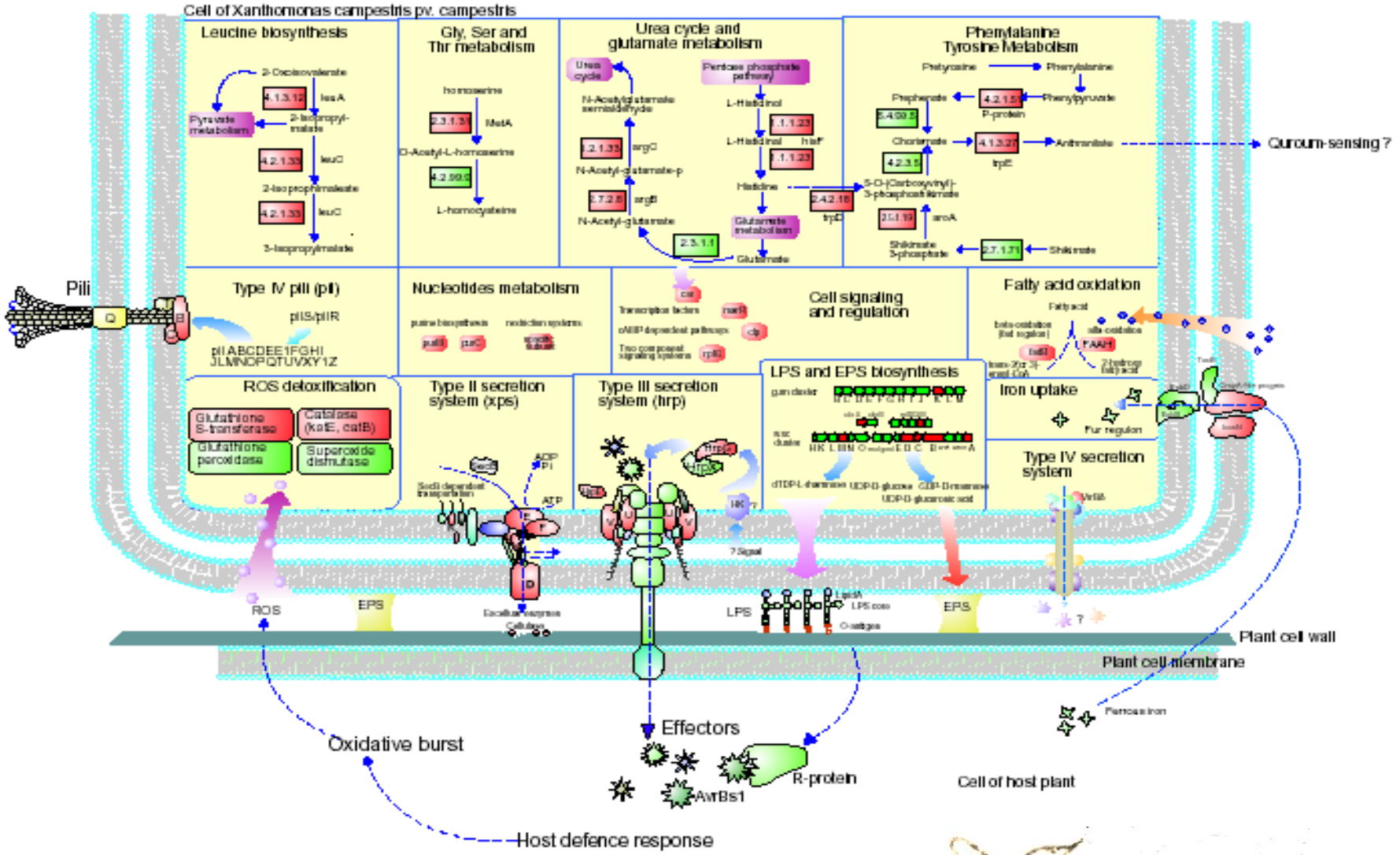


Figure 4. Comparison of the *rfa/waa* region (to scale). Arrows indicate predicted ORFs in both strands. Regions in gray indicate identical sequences among strains and the non-filling areas indicate sequences with non or low homology.

Genome-based analysis of virulence genes in a non-biofilm-forming *Staphylococcus epidermidis* strain (ATCC 12228)

Yue-Qing ZHAO *et al. Mol. Microbiol.* 49 (6): 1577-1593 (2003)



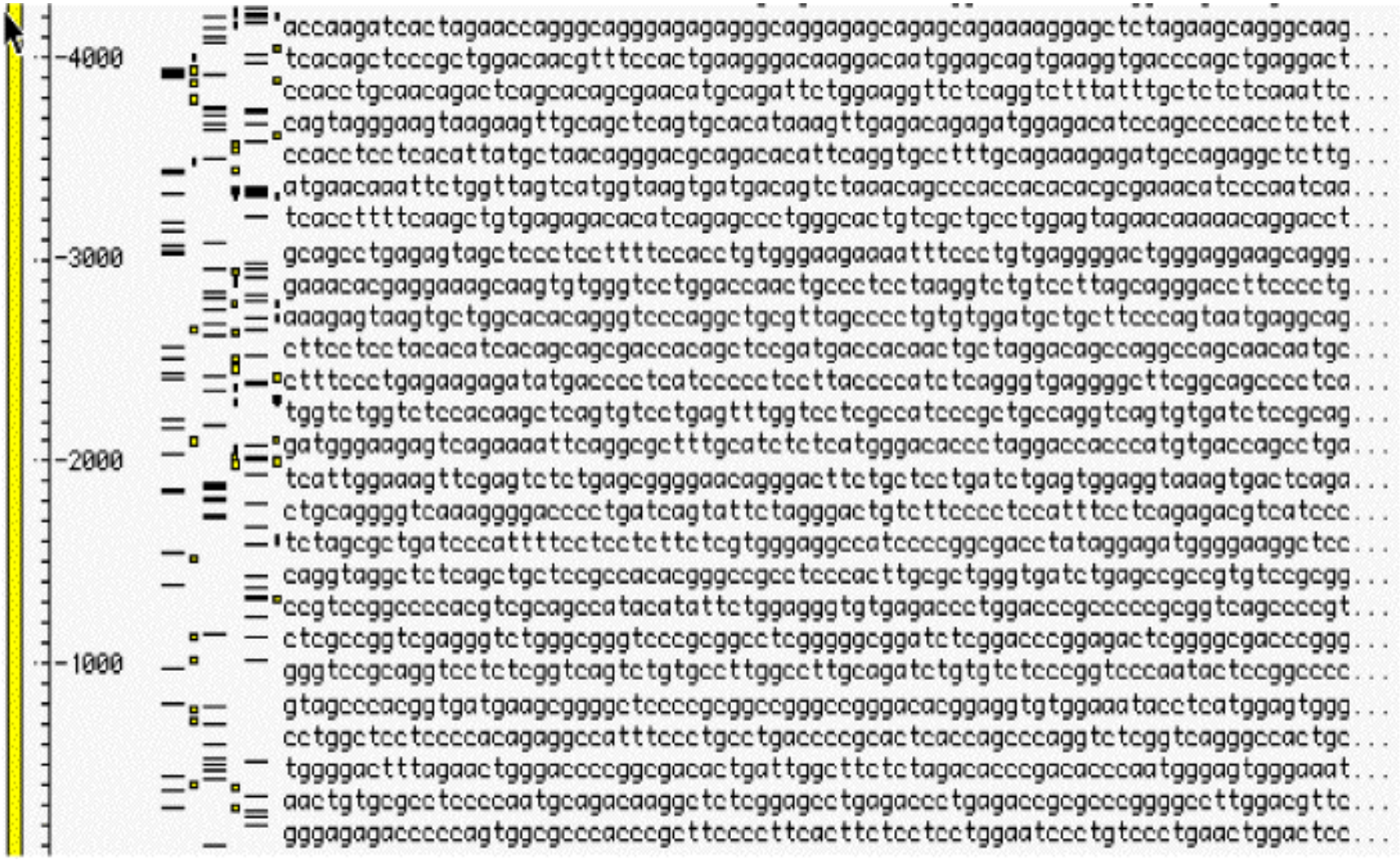


***Xanthomonas campestris* pv. *campestris* Strain 8004**

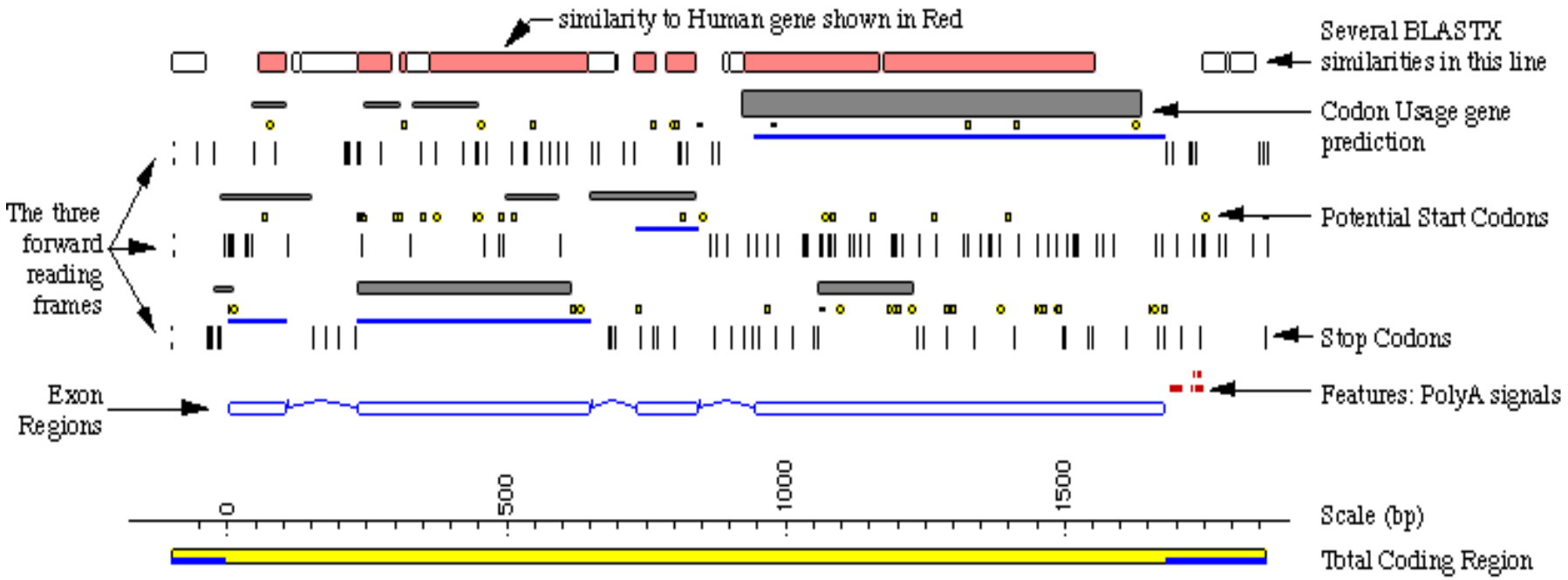
The value of genome sequences lies in their annotation

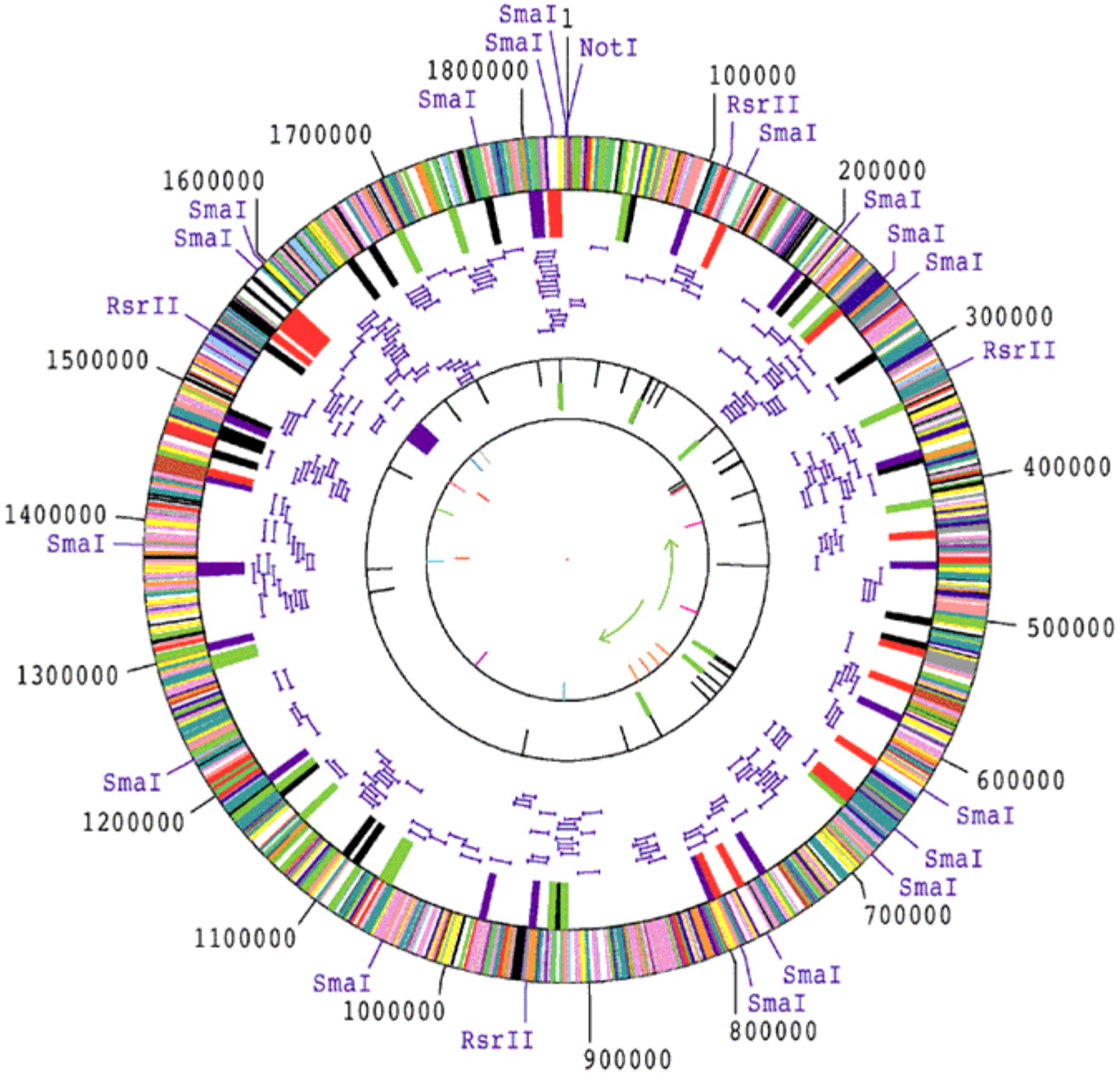
- Annotation - Characterizing genomic features using computational and experimental methods
- Genes: Four levels of annotation
  - Gene Prediction - Where are genes?
  - What do they look like?
  - Domains - What do the proteins do?
  - Role - What pathway(s) involved in?

How do we get from here …



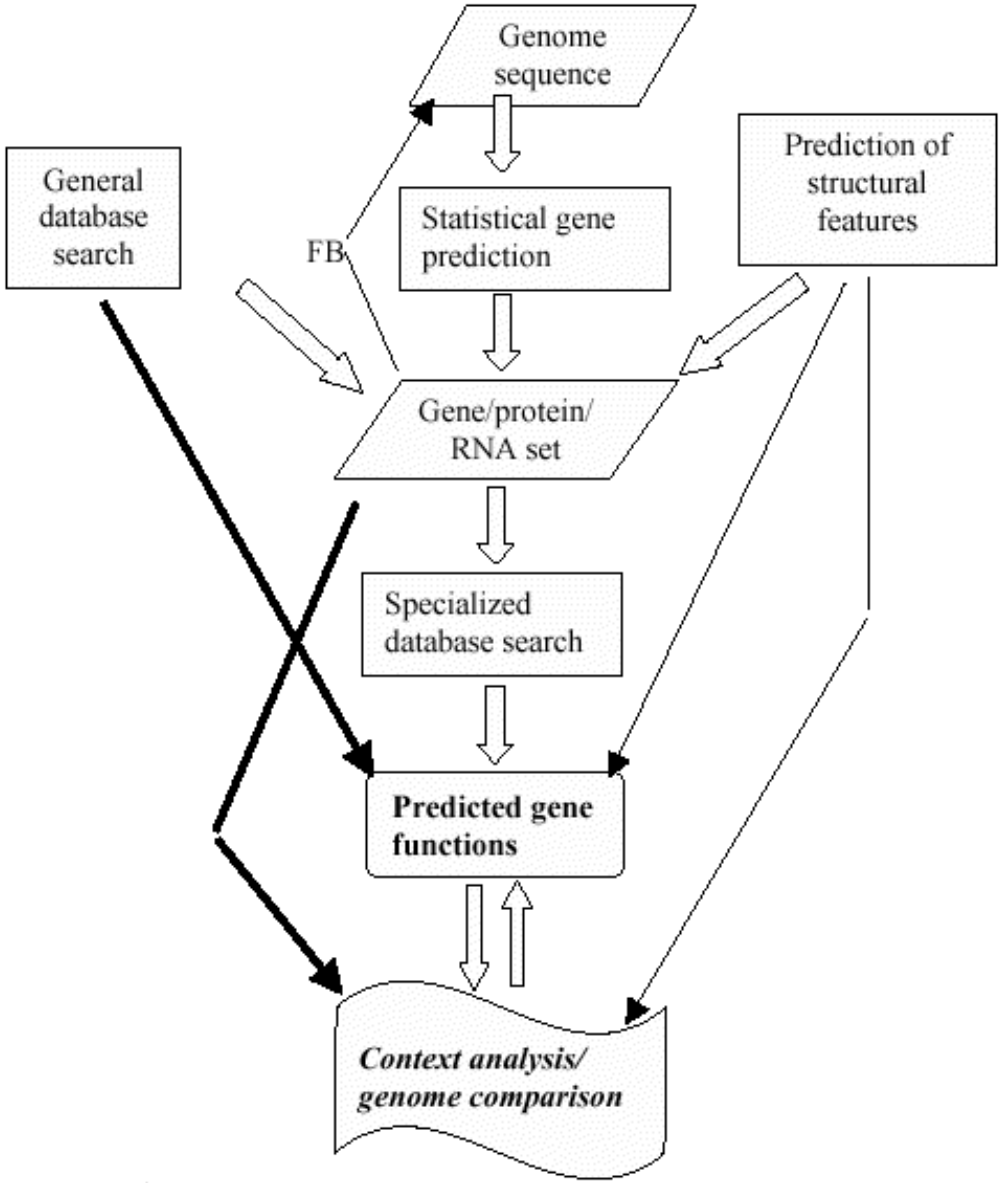
to here,


similarity to Human gene shown in Red
Several BLASTX similarities in this line
Codon Usage gene prediction
The three forward reading frames
Potential Start Codons
Stop Codons
Exon Regions
Features: PolyA signals
0
500
1000
1500
Scale (bp)
Total Coding Region



Fleischmann et.al (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269(5223): 496-512.

Generalized Genome Annotation Flowchart



Genome Annotation Explained

- **FB:** feedback from gene identification for correction of sequencing errors, primarily frameshifts
- **General database search:** searching sequence databases (typically, NCBI NR) for sequence similarity, usually using BLAST.
- **Specialized database search:** searching domain databases, such as Pfam, SMART, and CDD, for conserved domains, genome-oriented databases, such as COGs, for identification of orthologous relationship and refined functional prediction, metabolic databases, such as KEGG for metabolic pathway reconstruction, and possibly, other database searches.
- **Statistical gene prediction:** use of methods like GeneMark or Glimmer to predict protein-coding genes.
- **Prediction of structural features:** prediction of signal peptide, transmembrane segments, coiled domain and other features in putative protein functions.

What are genes?

- Complete DNA segments responsible for making functional products
- Products
  - Proteins
  - Functional RNA molecules
    - RNAi (interfering RNA)
    - rRNA (ribosomal RNA)
    - snRNA (small nuclear)
    - snoRNA (small nucleolar)
    - tRNA (transfer RNA)

Expansions and Clarifications

- ORFs
  - Start – triplets – stop
  - Prokaryotes: gene = ORF
  - Eukaryotes: spliced genes **or** ORF genes

- Exons
  - Remain after introns have been removed
  - Flanking parts contain non-coding sequence (5'- and 3'-UTRs)

Gene identification

- **Homology-based gene prediction**
  - Similarity Searches (*e.g.* BLAST, BLAT)
  - RNA evidence (ESTs....)
- ***Ab initio gene prediction***
  - Prokaryotes
    - ORF identification
  - Eukaryotes
    - Promoter prediction
    - PolyA-signal prediction
    - Splice site, start/stop-codon predictions

Approaches to Gene Finding

- Direct
  - Exact or near-exact matches of EST, cDNA, or Proteins from the same, or closely related organism
- Indirect
  1. Look for something that looks like one gene (***homology***)
  2. Look for something that looks like all genes (***ab initio***)

Homology-based gene prediction

- Similarity Searches (*e.g.* BLAST)
- Dependent genomes closely

Ab initio gene prediction

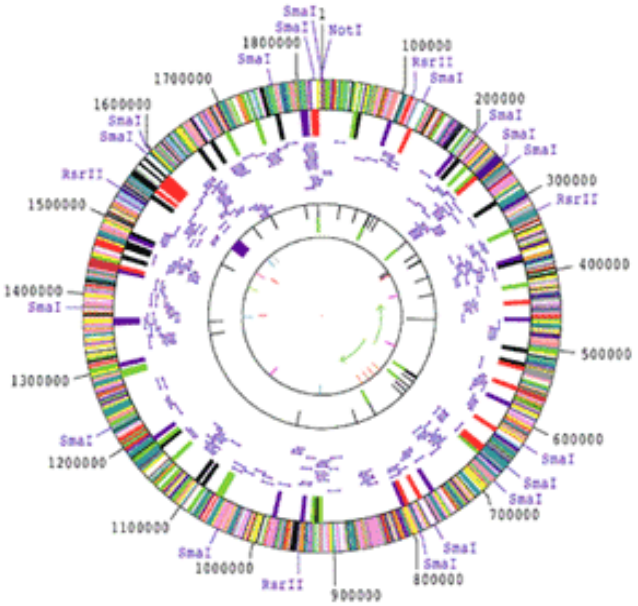
- Prokaryotes
  - ORF-Detectors
- Eukaryotes
  - Position, extent & direction: through promoter and polyA-signal predictors
  - Structure: through splice site predictors
  - Exact location of coding sequences: through determination of relationships between potential start codons, splice sites, ORFs, and stop codons

Bioinformatics as *Extrapolation*

- Computational gene finding is a process of:
  - Identifying common phenomena in known genes
  - Building a computational framework/model that can accurately describe the common phenomena
  - Using the model to scan uncharacterized sequence to identify regions that match the model, which become putative genes
  - Test and validate the predictions

Prokaryotic gene model: ORF-genes

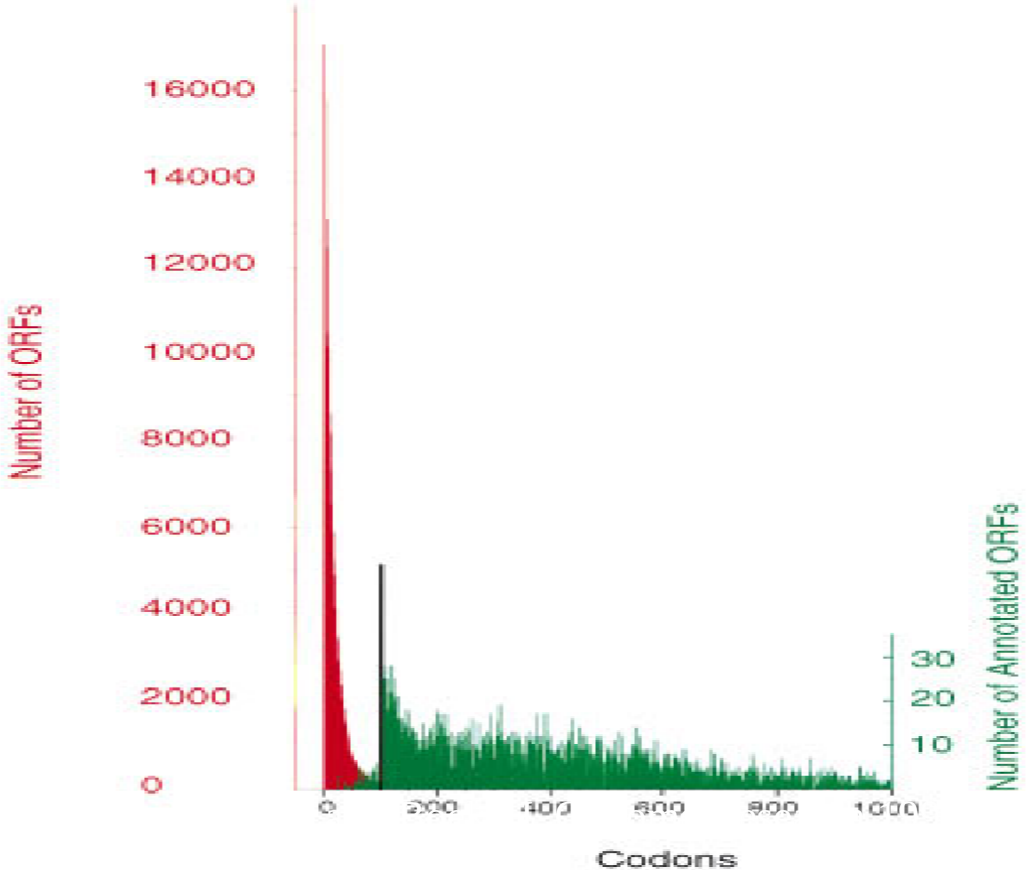
- "Small" genomes, high gene density
  - *Haemophilus influenza* genome 85% genic
- Operons
  - One transcript, many genes
- No introns.
  - One gene, one protein
- Open reading frames
  - One ORF per gene
  - ORFs begin with start, end with stop codon (def.)



What is it about genes that we can measure (and model)?

- Most of our knowledge is biased towards protein-coding characteristics
  - ORF (Open Reading Frame): a sequence defined by in-frame AUG and stop codon, which in turn defines a putative amino acid sequence.
  - Codon Usage: most frequently measured by CAI (Codon Adaptation Index)
- Other phenomena
  - Nucleotide frequencies and correlations:
    - value and structure
  - Functional sites:
    - splice sites, promoters, UTRs, polyadenylation sites

A simple measure: ORF length Comparison of Annotation and Spurious ORFs in *S. cerevisiae*



Codon Adaptation Index (CAI)

$$CAI = \prod_{i=codons} \left[\frac{f_{codon_i}}{f_{(codon_i)_{\max}}} \right]$$

- Parameters are empirically determined by examining a “large” set of example genes
- This is not perfect
  - Genes sometimes have unusual codons for a reason
  - The predictive power is dependent on length of sequence

Genomic sequence features

- **Repeats ("Junk DNA")**
  - Transposable elements, simple repeats
  - RepeatMasker
- **Genes**
  - Vary in density, length, structure
  - Identification depends on evidence and methods and may require concerted application of bioinformatics methods and lab research
- **Pseudo genes**
  - Look-a-likes of genes, obstruct gene finding efforts.
- **Non-coding RNAs (ncRNA)**
  - tRNA, rRNA, snRNA, snoRNA, miRNA
  - tRNASCAN-SE, COVE

Prokaryotic Gene Prediction

- Glimmer
- GeneMark
- Critica
- ORNL Annotation Pipeline

Non-protein Coding Gene Tools and Information

- tRNA
  - tRNA-ScanSE
    - http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
  - FAStRNA
    - http://bioweb.pasteur.fr/seqanal/interfaces/fastrna.html
- snoRNA
  - snoRNA database
    - http://rna.wustl.edu/snoRNAdb/
- microRNA
  - Sfold
    - http://www.bioinfo.rpi.edu/applications/sfold/index.pl
  - SIRNA
    - http://bioweb.pasteur.fr/seqanal/interfaces/sirna.html

The annotation pipeline

- Mask repeats using RepeatMasker.
- Run sequence through several programs.
- Take predicted genes and do similarity search against ESTs and genes from other organisms.
- Do similarity search for non-coding sequences to find ncRNA.

General Things to Remember about (Protein-coding) Gene Prediction Software

- It is, in general, organism-specific
- It works best on genes that are *reasonably* similar to something seen previously
- It finds protein coding regions far better than non-coding regions
- In the absence of external (direct) information, alternative forms will not be identified
- It is imperfect! (It's biology, after all…)

Open Challenges in Predicting Prokaryotic (Protein-Coding) Genes

- Start site prediction
  - Most algorithms are greedy, taking the largest ORF
- Overlapping Genes
  - This can be very problematic, esp. with use of Viterbi-like algorithms
- Non-canonical coding

After Gene Finding…

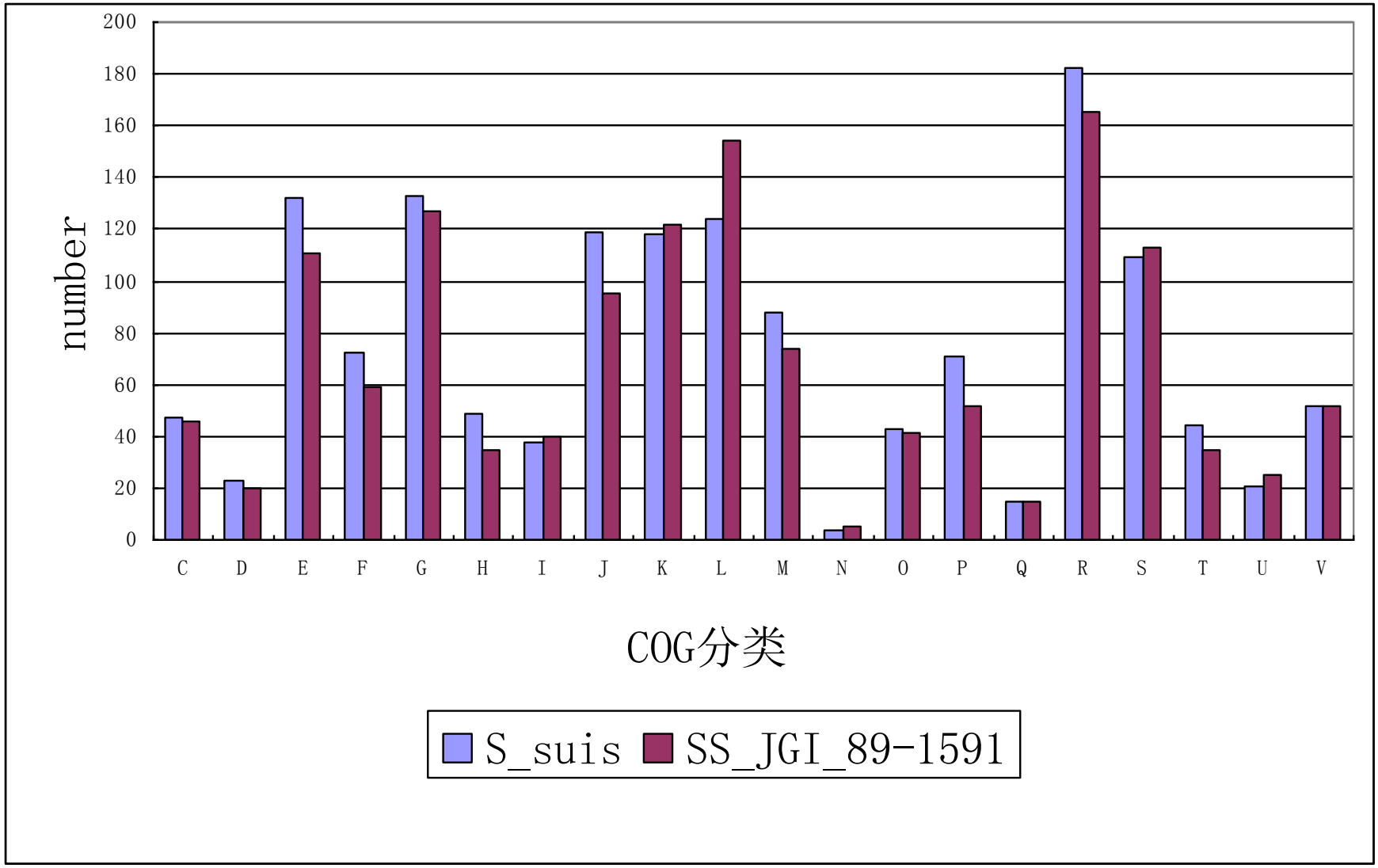
- Genome annotation
  - Gene function, including domain analysis
  - Gene functional group
  - Pathway analysis
  - Specific functional group you interesting….
    - Virulence genes
    - Pseudogenes
    - TCS
- Other Genome characteristic
  - GC content……
  - Genome islands
  - IS, transposons

Genome Annotation

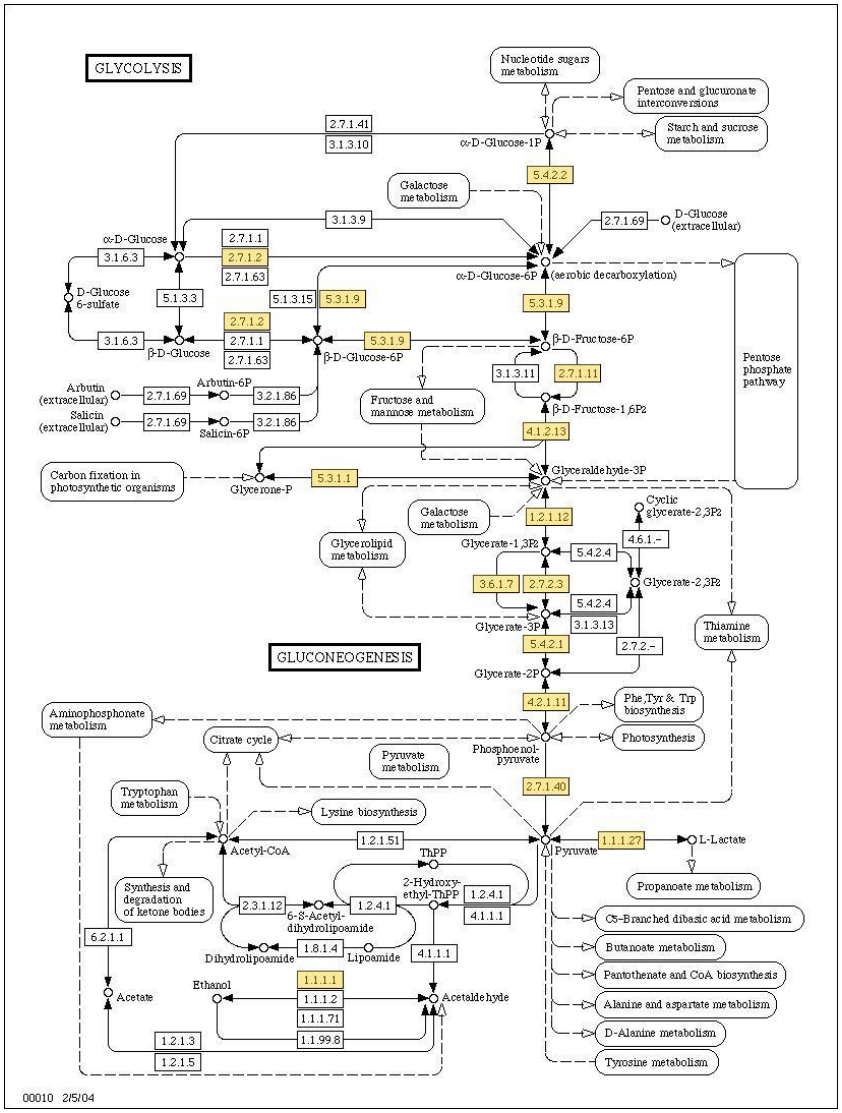
- Gene Function
  - Blast to NR database
  - Domain analysis
    - Pfam (http://www.sanger.ac.uk/Software/Pfam/)
    - InterPro (http://www.ebi.ac.uk/interpro/)
- Gene Cluster
  - COG （http://www.ncbi.nlm.nih.gov/COG/)
- Pathway
  - KEGG pathway (http://www.genome.jp/kegg/pathway.html)

Annotation nomenclature

- **Known Gene -** Predicted gene matches the entire length of a known gene.
- **Putative Gene -** Predicted gene contains region conserved with known gene. Also referred to as “like” or “similar to”.
- **Unknown Gene -** Predicted gene matches a gene or EST of which the function is not known.
- **Hypothetical Gene -** Predicted gene that does not contain significant similarity to any known gene or EST.


200
180
160
140
120
100
80
60
40
20
0
number
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
COG分类
S_suis
SS_JGI_89-1591

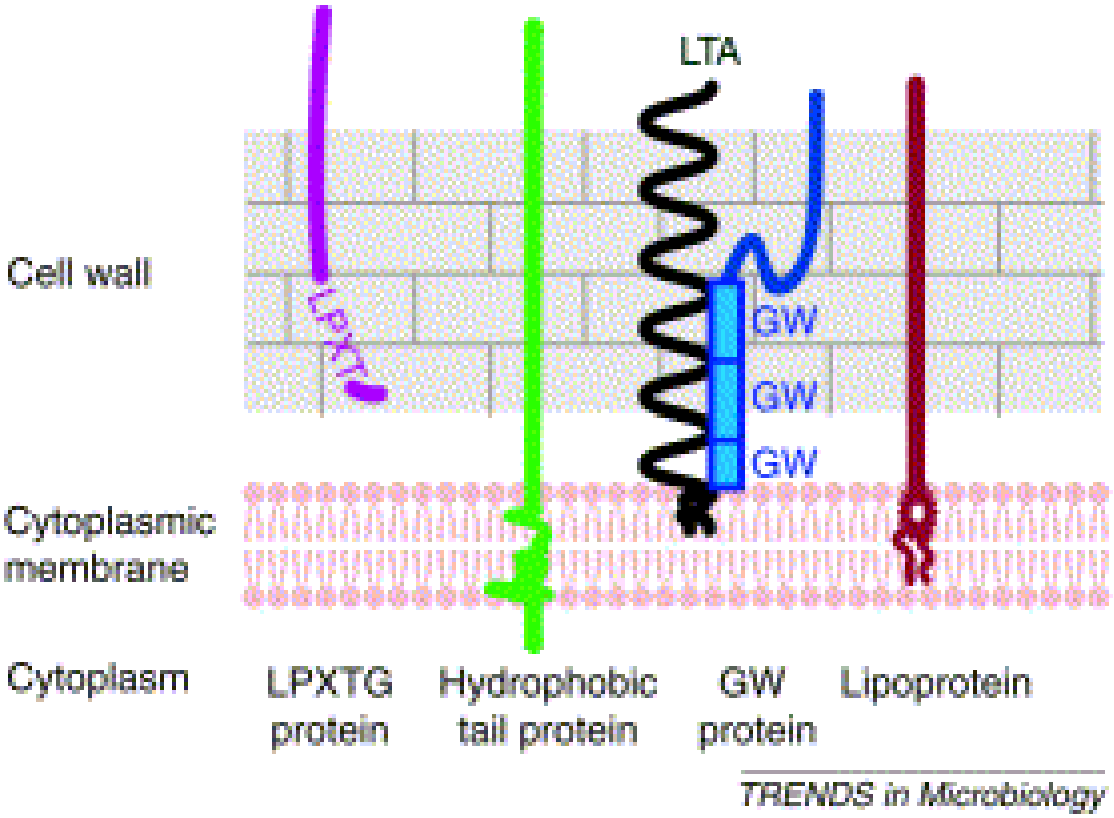
Pathway Analysis



Virulence factors

- **Virulence factors** refers to the properties (i.e., gene products) that enable a microorganism to establish itself on or within a host of a particular species and enhance its potential to cause disease.
- **Virulence factors** include bacterial toxins, cell surface proteins that mediate bacterial attachment, cell surface carbohydrates and proteins that protect a bacterium, and hydrolytic enzymes that may contribute to the pathogenicity of the bacterium.

Identify potential virulence factors



Trends Microbiol. 2002 May;10(5):238-45.

Surface proteins and the pathogenic potential of Listeria monocytogenes.

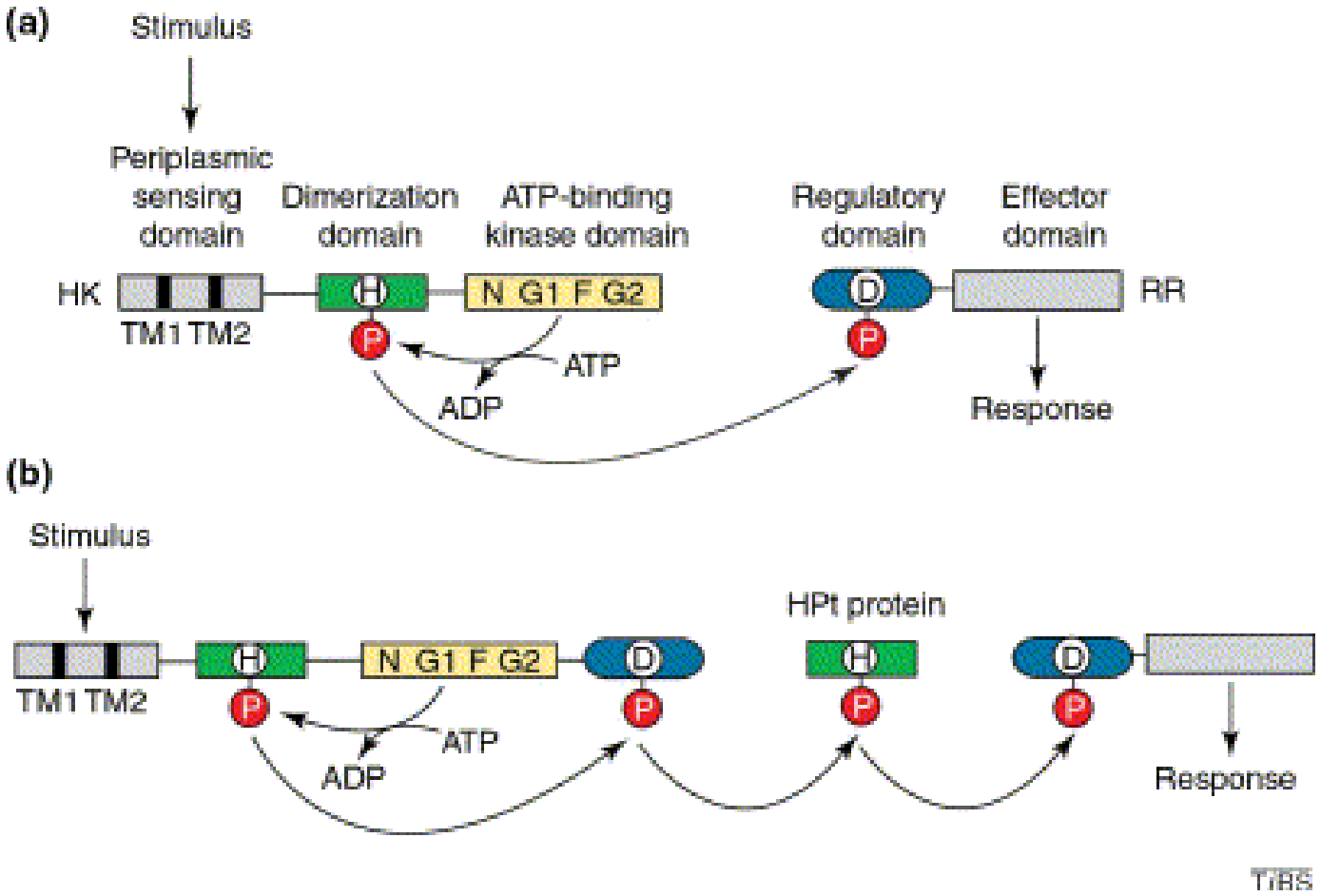
Common mechanisms of antimicrobial resistance in microbes and viruses

Alterations in drug target or activating enzyme[a,b]
Inactivation by enzymes[a]
Changes in cellular permeability towards the drug[a]
Active efflux[a]
Overproduction of target enzymes
Bypass of drug action
Intercellular cooperation??
[a]These represent the major mechanisms of bacterial resistance
[b]This is the only mechanism of relevance to viruses

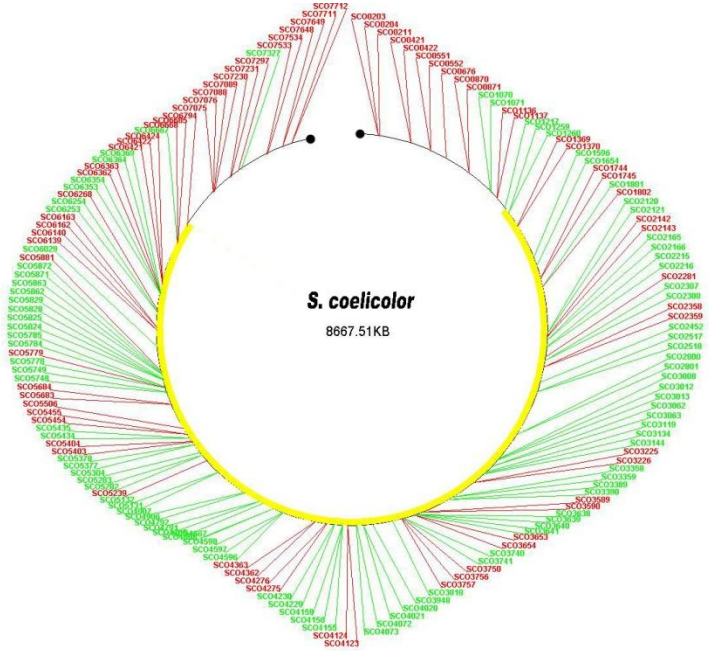
Microbial and viral drug resistance mechanisms

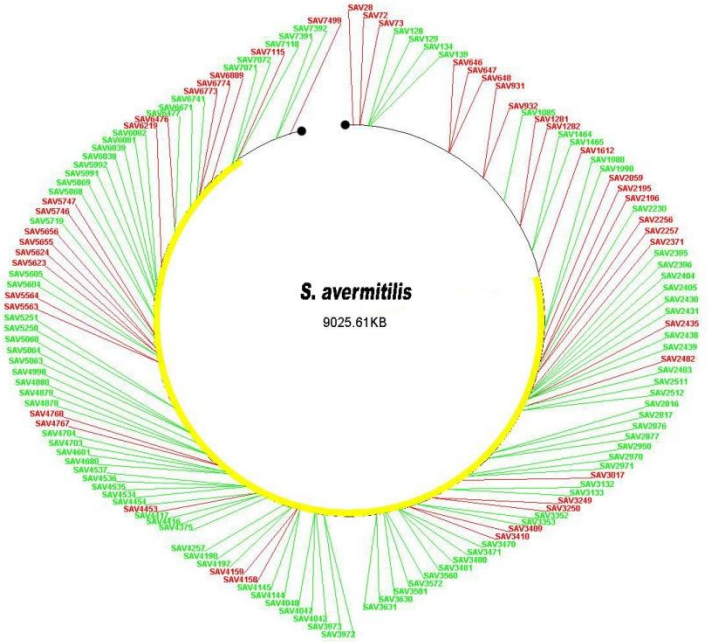
Trends Microbiol. 2002;10(10 Suppl):S8-14.

Two-component signaling systems



Arrangement of TCS genes in two *Streptomyces* strains





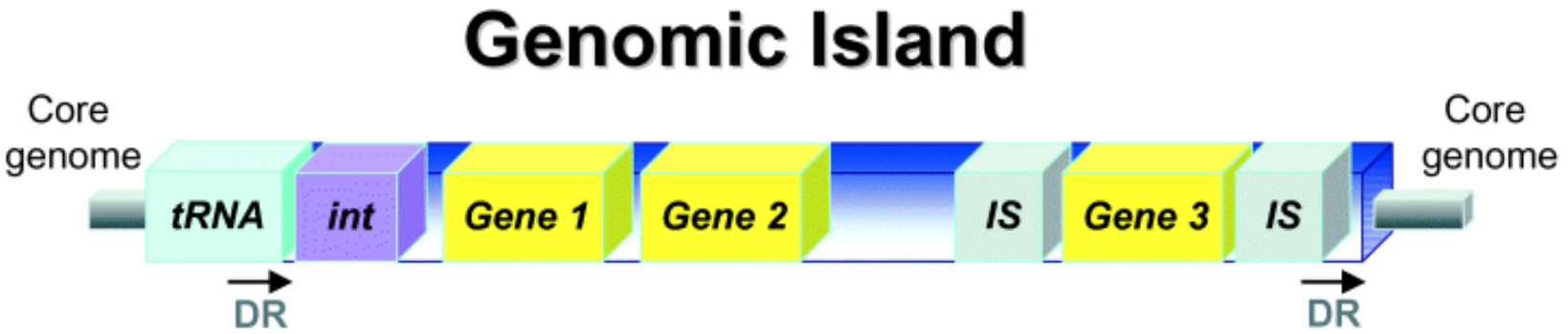
Genomic islands

TABLE 1. *Methods to analyse bacterial strains for the presence of GEI/PAI-related genes/sequences*

Method	**Aim**	**Useful for routine diagnostics**
GEI/PAI-specific PCR	Detection of GEI/PAI genes	Yes
DNA-chip analysis	Simultaneous detection of GEI/PAI-associated genes and their expression	Yes
tRNA screening	Detection of integration in tRNA genes	Limited
Subtractive hybridization	Detection of genomic differences	No
Island probing	Analysis of GEI instability	No

Impact of pathogenicity islands in bacterial diagnostics.
Apmis **112** (11-12), 930-936.

Characteristics of genomic islands



Functions encoded by Genomic Islands:

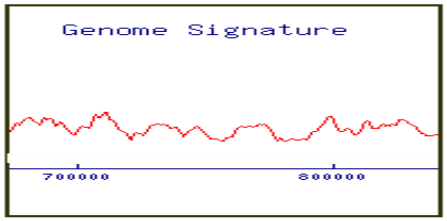
Pathogenicity, Iron Uptake, Secondary Metabolism,
Antibiotic Resistance, Secretion, Degradation of Xenobiotics, Symbiosis

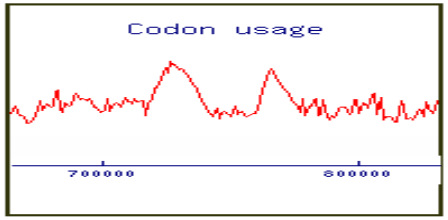
Impact of pathogenicity islands in bacterial diagnostics. *Apmis* **112** (11-12), 930-936.

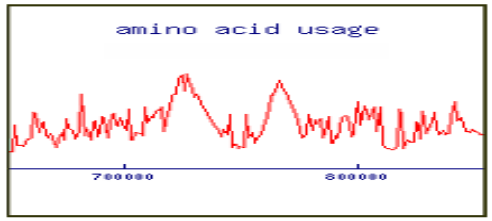
Characterization of anomalous Genomic islands

- Genomic characterization:
  - **Compositional contrasts (standard method):** compare G+C frequency within *W* to the average genomic G+C frequency.
  - **Genome signature contrasts:** compare δ* differences of each window segment to the average genomic signature.
  - **Codon usage contrasts:** compare codon biases of the gene set of each window to the average gene codon usages.
  - **Amino acid contrasts:** compare amino acid biases of proteins in each window relative to the average proteome amino acid frequencies.
  - **Putative alien (pA) gene clusters:** compare differences in codon usages from the RP, TP and CH gene classes, and from the average gene.
- IS sequences, transposonase, tRNA

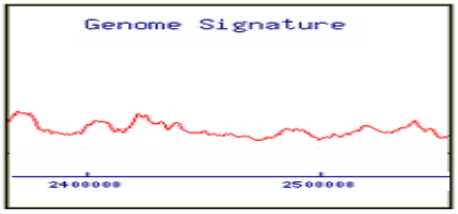
element I

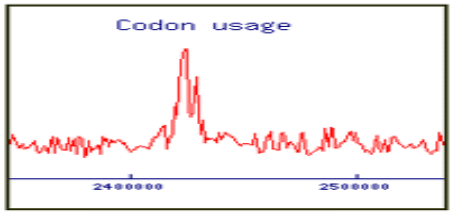


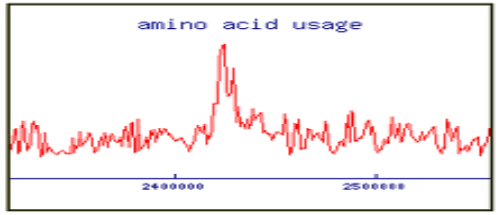




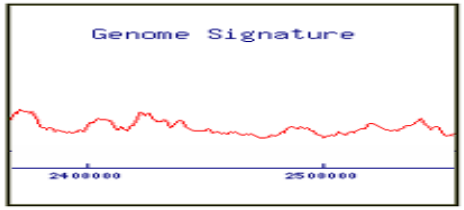
element II

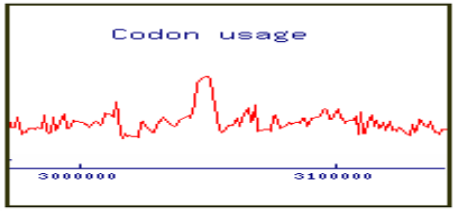


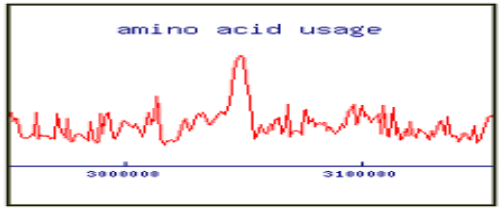




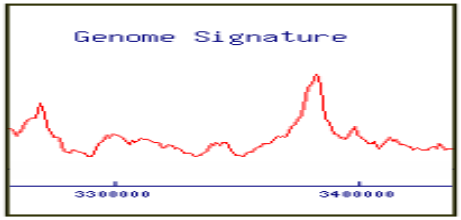
element III

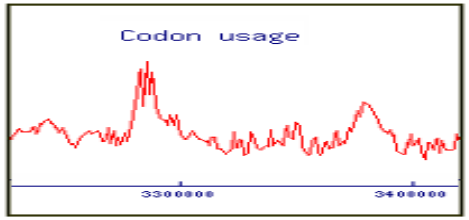


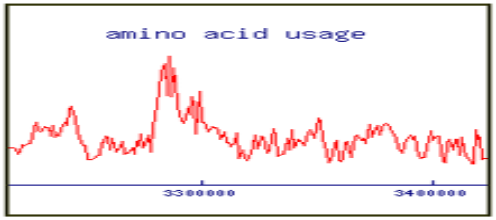




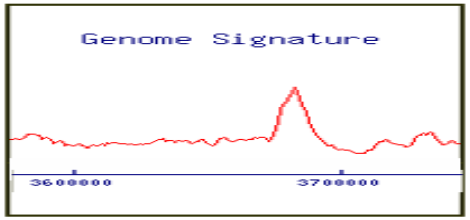
element IV

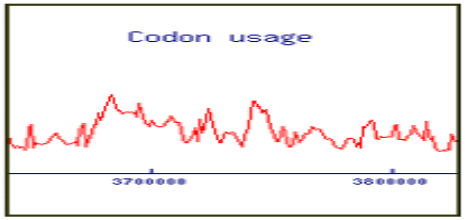


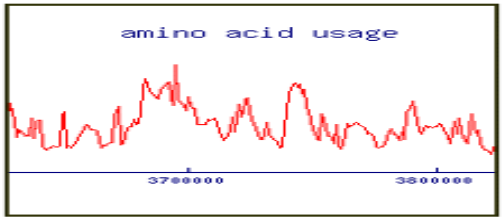


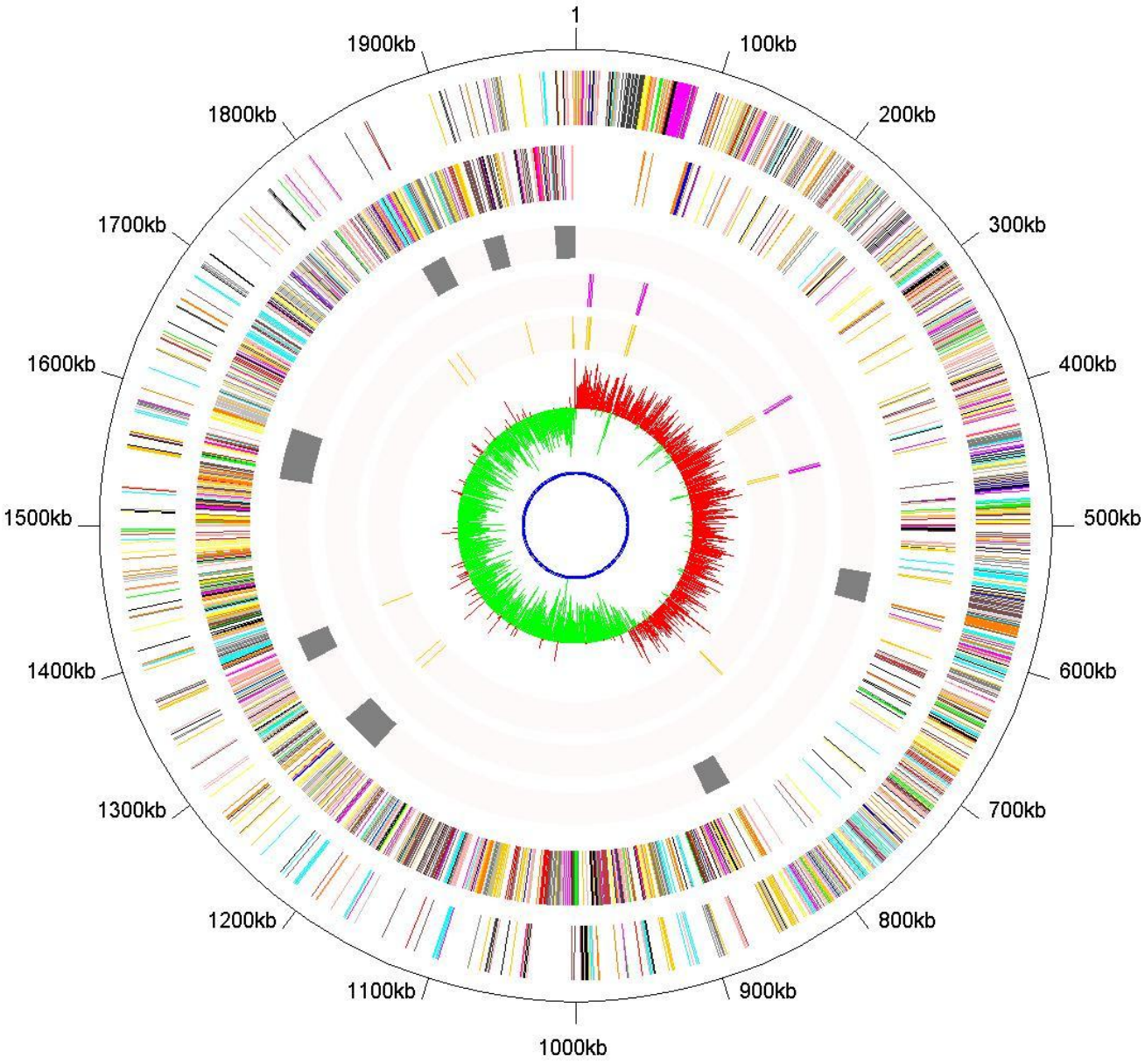


element V








1
100kb
200kb
300kb
400kb
500kb
600kb
700kb
800kb
900kb
1000kb
1100kb
1200kb
1300kb
1400kb
1500kb
1600kb
1700kb
1800kb
1900kb

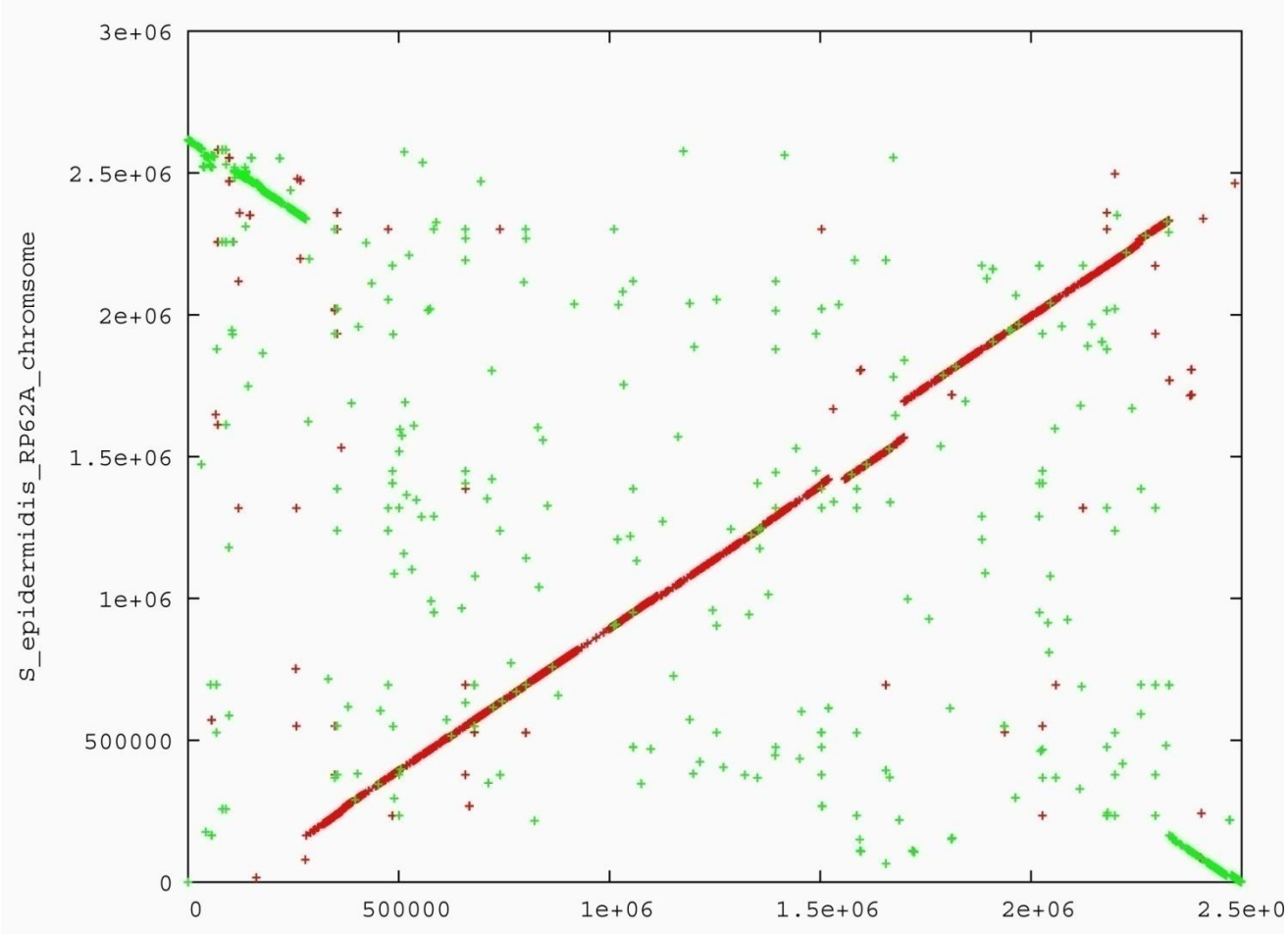
How to do Comparative Genomics

- Genome sequences alignment
  - BlastN, MUMmer
- Gene content comparison
  - BlastP

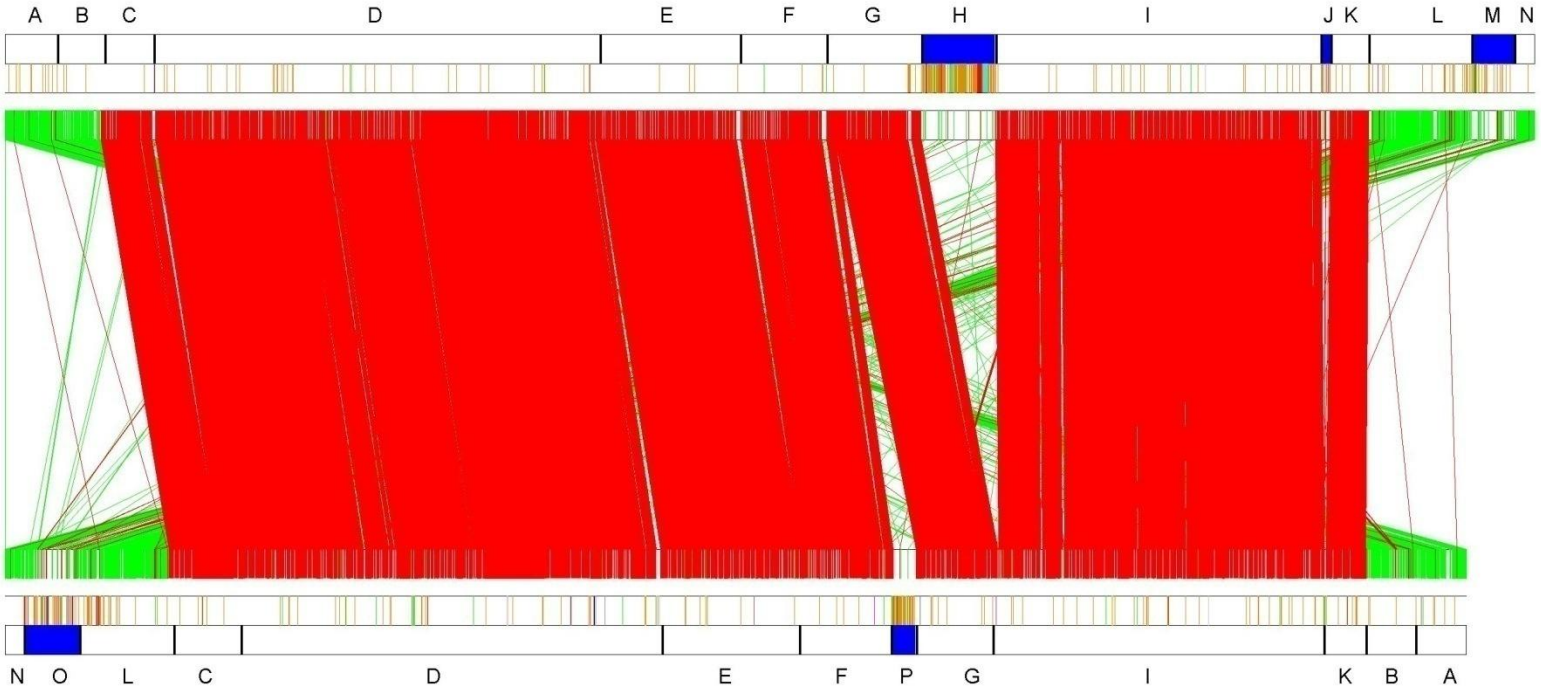
Genome sequences alignment

- Genome islands (a cluster of genes)
- Indels
  - Transpons
  - Tandem repeats
- SNPs
  - Sense-Nonsense substitutions
  - Synonymous/non-synonymous substitutions

Genome Sequence alignment



Genomic comparison



Gene content comparison

- Homolog
  - Ortholog
  - Paralog
- Specific genes

General Information

	streptococcus_suis_89_1591	*SS_Sanger_1-7*
Total CDS number	1918	1969
CDSs categorized by sequence variations		
All homologs	1351	1341
Total ortholog	1306	1306
Orthologs with identical nucleic acid sequences	31	31
Orthologs with SNPs and the same length	848	848
Orthologs with insertoins/deletions	427	427
Paralogs	45	35
Unique compare to each other	567	628

Towards System Level

- Basic Terms of Transcriptional Regulation in Microbe
- A Quick Overview of Transcriptional Regulation Investigation
- Methodology of Gene Network Inference
  - How to infer gene networks from expression profiles
  - Classic Case Studies
- Other Branches
  - Comparative Analysis in the Light of Evolution
  - Learning Biological Networks from Modules to Dynamics
  - Prof. Palsson's Strategy: Integration of Kinds of Networks

Basic Terms of Transcriptional Regulation in Bacteria

- **A Transcription Unit (TU)**
  - a regulatory region
  - a transcription start site
  - one or more ORFs
  - and a transcription termination site.
- **An Operon**
  - The collection of overlapping TUs constitutes an operon
- **Cis Elements/Transcription Factor Binding Sites (TFBS)**
  - The regulatory region contains *cis* elements such as the promoter
  - 400 base pairs

Balleza, E., et al., *Regulation by transcription factors in bacteria: beyond description*. FEMS Microbiol Rev, 2009. **33**(1): p. 133-51.

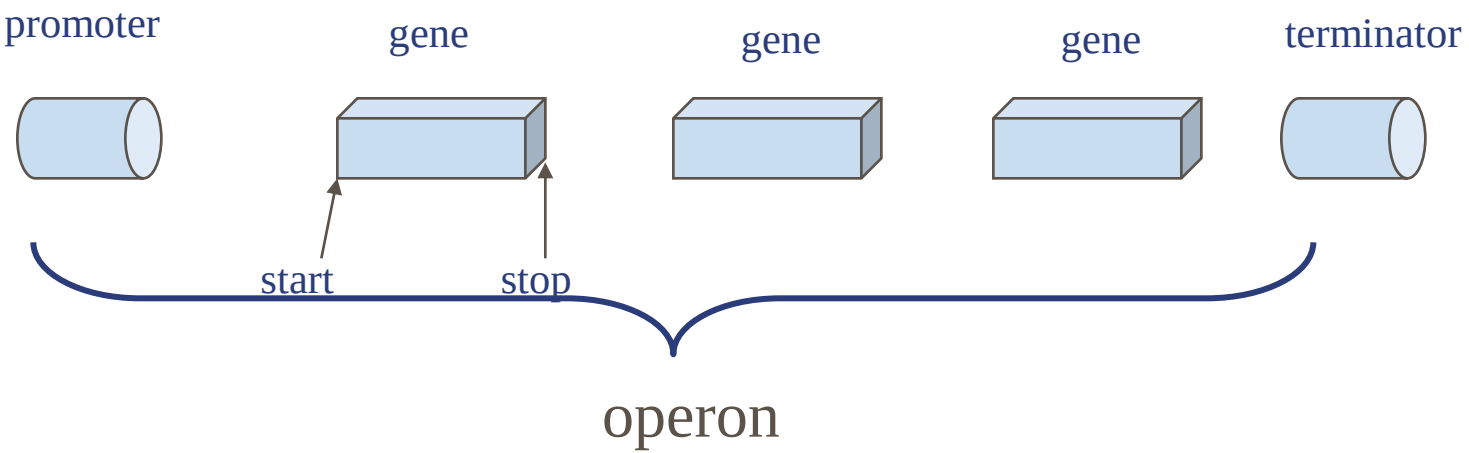
Basic Terms of Transcriptional Regulation in Microbe

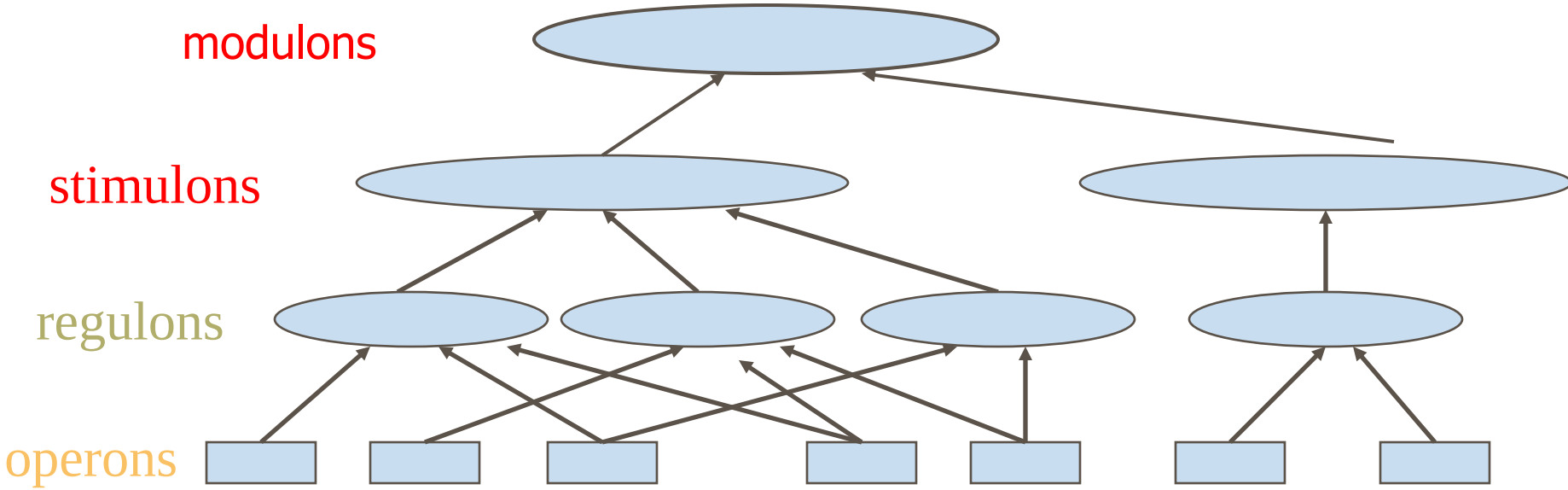
- Transcription initiation in bacteria requires proteins known as **sigma factors (σ)** that essential for proper promoter recognition by RNA polymerase.
  - σ^{70} and σ^{54}.

- **TFs are classified in several families based on at least two domains:**
  - a signal sensor and
  - a responsive element that directly interacts with a target DNA, helix–turn–helix domain is the most common
  - two-component systems
- **Regulon**
  - a set of TGs coregulated by the same set of TFs
  - Regulons are divided into **simple** or **complex(**majority**)** if regulated by a single or by multiple TFs, correspondingly.
  - **RegulonDB**

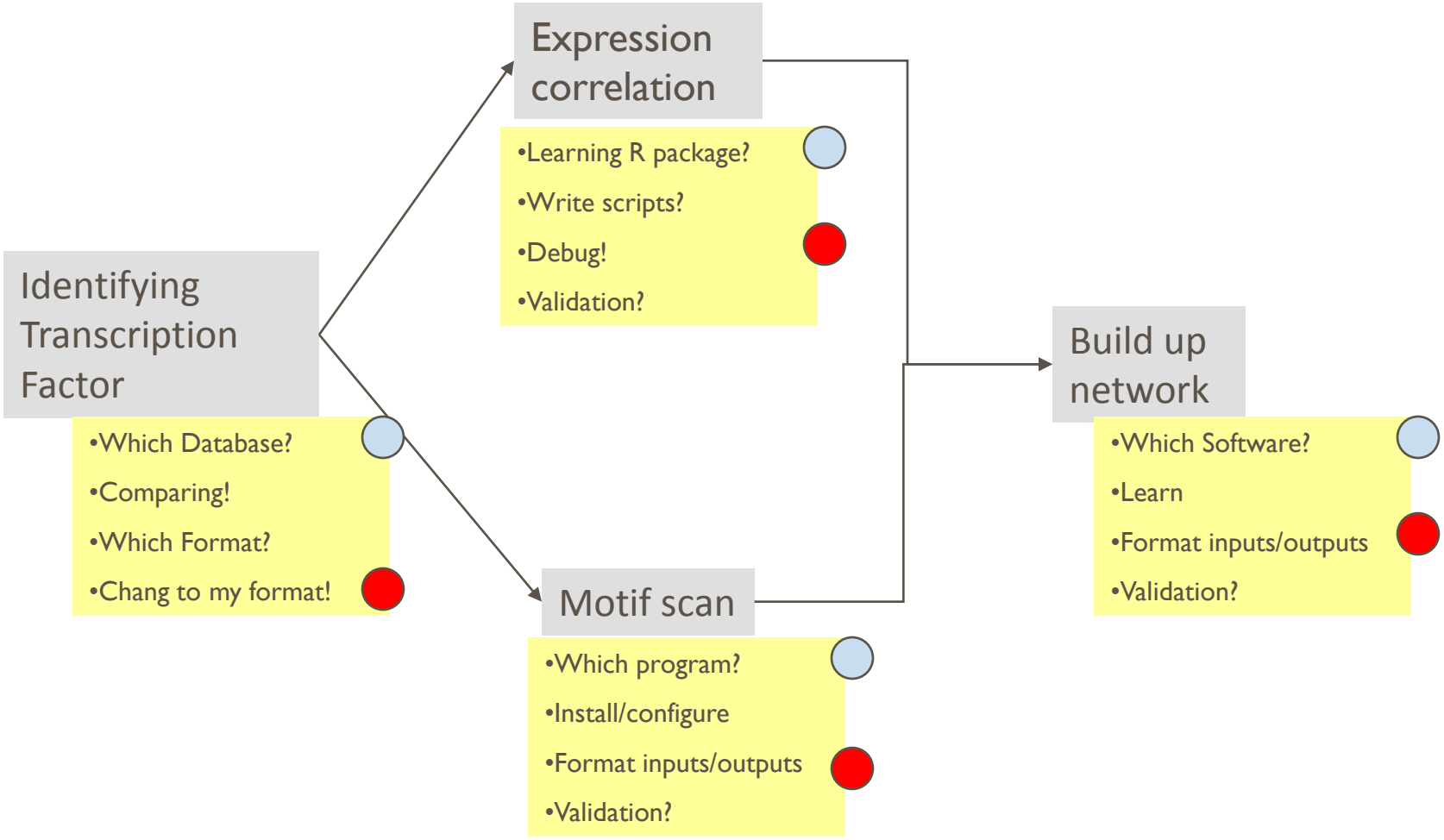
Operon prediction

- Genes are grouped into operons (transcriptional units)




modulons
stimulons
regulons
operons

A Quick Overview of Transcriptional Regulation Investigation



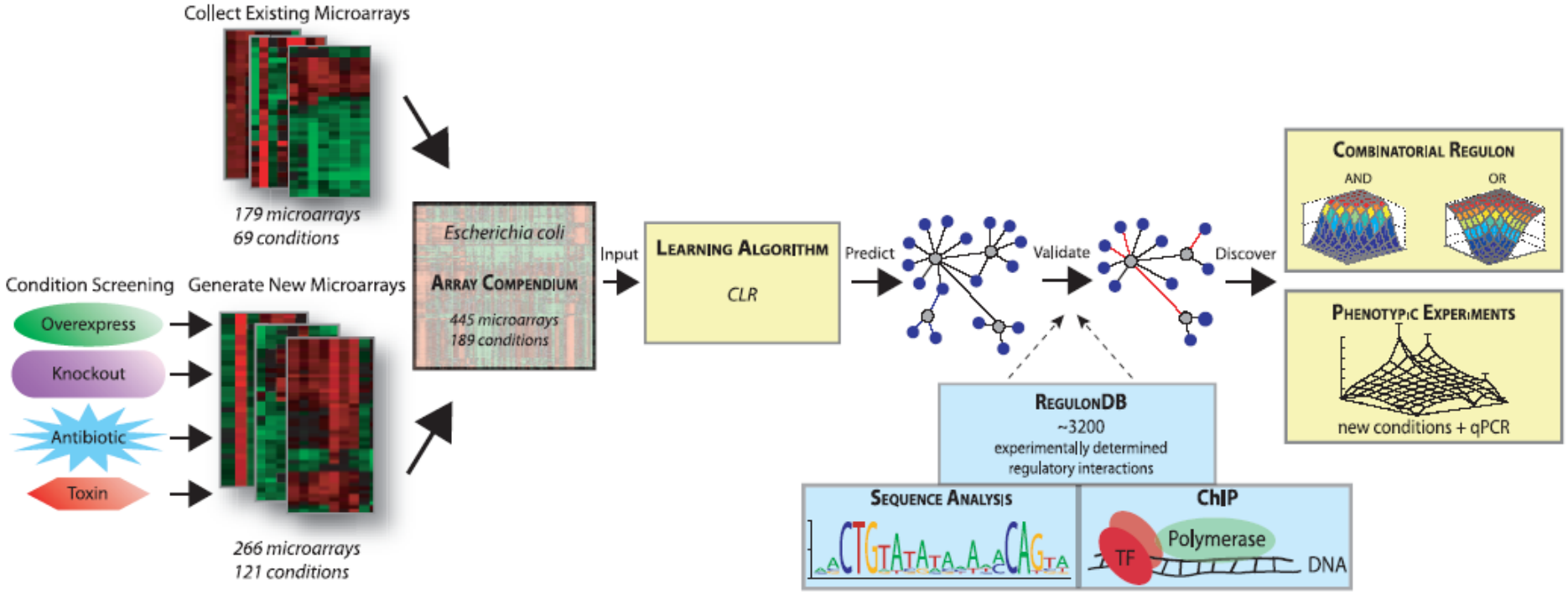


Figure 1. Overview of Our Approach for Mapping the *E. coli* Transcriptional Regulatory Network

Microarray expression profiles were obtained from several investigators. Our laboratory profiled additional conditions, focusing on DNA damage, stress responses, and persistence. These two data sources were combined into one uniformly normalized *E. coli* microarray compendium that was analyzed with the CLR network inference algorithm. The predicted regulatory network was validated using RegulonDB, sequence analysis, and ChIP. The validated network was then examined for cases of combinatorial regulation, one of which was explored with follow-up real-time quantitative PCR experiments.

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles**. *PLoS Biol* 2007, **5**:e8.

Computational approaches to investigate transcriptional regulatory networks

- Template-based method
  - Coexpression networks and clustering algorithms
  - Network Alignment
- Reverse engineering using gene expression data (network inference)
  - DREAM is a Dialogue for Reverse Engineering Assessments and Methods with its main objective to catalyse the interaction between experiment and theory in the area of cellular network inference (http://wiki.c2b2.columbia.edu/dream/).

Babu MM: **Computational approaches to study transcriptional regulation**. *Biochem Soc Trans* 2008, **36**:758-765.

Gene network inference algorithms

- **Bayesian networks**
- **Information-theoretic approaches**
- **Ordinary differential equations**
- **Choose the most suitable network inference algorithms according to the problem to be addressed.**

Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D: **How to infer gene networks from expression profiles**. *Mol Syst Biol* 2007, **3**:78

Bayesian networks

- A Bayesian network is a graphical model for probabilistic relationships among a set of random variables X_i, where i=1 ... n.
- In order to reverse-engineer a Bayesian network model of a gene network, we must find the directed acyclic graph G (i.e. the regulators of each transcript) that *best* describes the gene expression data D, where D is assumed to be a steady-state data set.

$$P(G/D) = \frac{P(D/G) * P(G)}{P(D)}$$

- Choosing the G with the maximum Bayesian score is an NP-hard problem. Therefore, a heuristic search method is used, like the greedy-hill climbing approach, the Markov Chain Monte Carlo method or simulated annealing.
- Bayesian networks cannot contain cycles(i.e. no feedback loops). Dynamic Bayesian networks are an extension of Bayesian networks able to infer interactions from a data set D consisting of time-series rather than steady-state data.

Information-theoretic approaches

- Information-theoretic approaches use Mutual Information (MI), to compare expression profiles from a set of microarrays.

Mutual information, MI_{ij}, between gene i and gene j is computed as:

$$MI_{i,j} = H_i + H_j - H_{ij} \quad (3)$$

where H, the entropy, is defined as:

$$H_i = -\sum_{k=1}^{n} p(x_k) \log(p(x_k)) \quad (4)$$

- The higher the entropy, the more randomly distributed are gene expression levels across the experiments. A higher MI indicates that the two genes are non-randomly associated to each other.
- MI is symmetric, $M_{ij}=M_{ji}$, therefore the network is described by an undirected graph G, thus differing from Bayesian networks (directed acyclic graph).
- Deal with steady-state gene expression data set, or with time-series data as long as the sampling time is long enough to assume that each point is independent of the previous points.

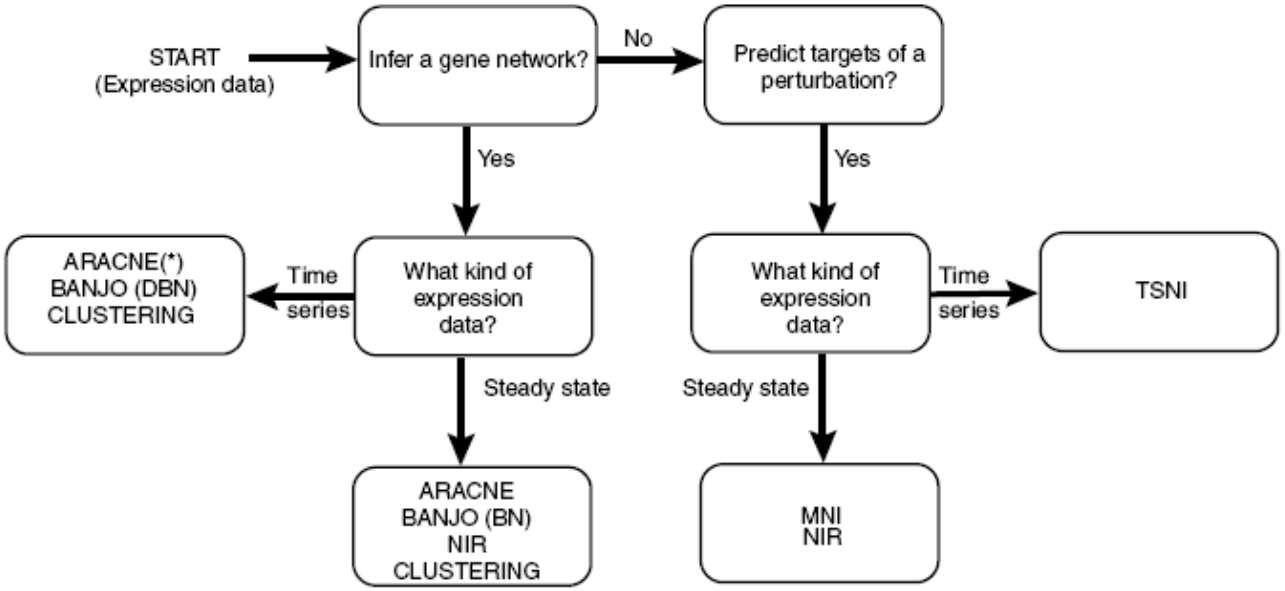


Figure 1 Flowchart to choose the most suitable network inference algorithms according to the problem to be addressed. (*): check for independence of time points (see text for details); (BN): Bayesian networks; (DBN): Dynamic Bayesian Networks.

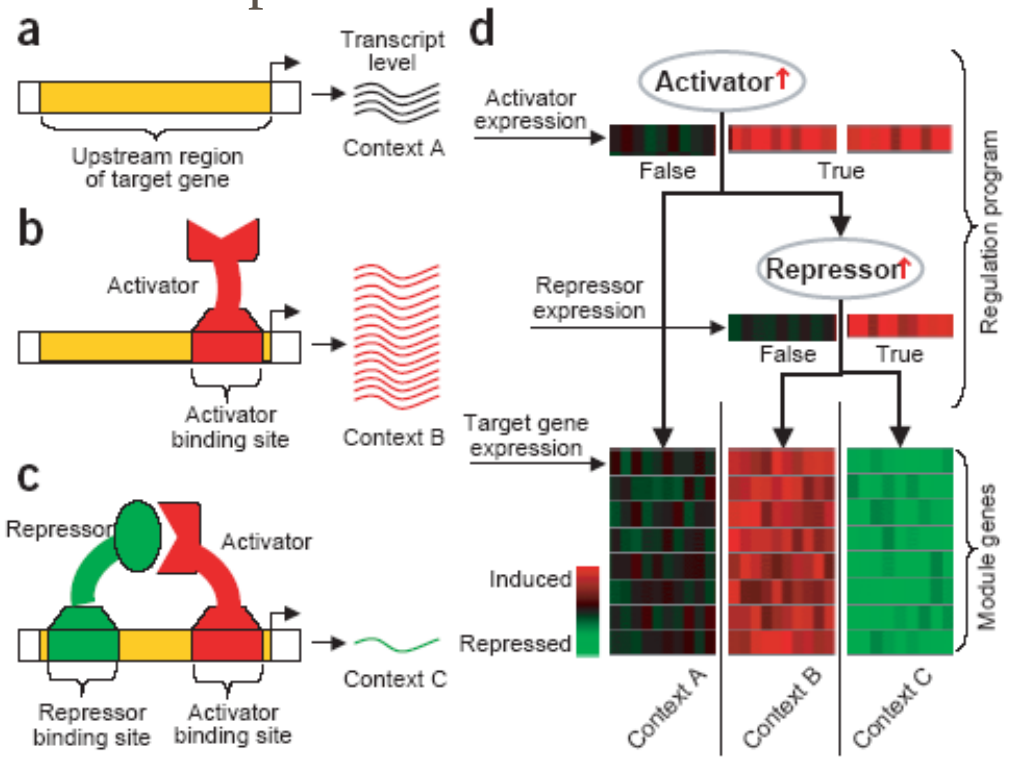
Table 1 Features of the network inference algorithms reviewed in this tutorial

Software	Download	Data type	Command line	Notes
BANJO	www.cs.duke.edu/~amink/software/banjo	S/D	java-jar banjo.jar setting-File=mysettings.txt	Good performance if large datasets is available ($M \gg N$)
ARACNE	amdec-bioinfo.cu-genome.org/html/caWork-Bench/upload/arcane.zip	S/D	arance-i inputfile-o outputfile [options]	Good performance even for $M \leqslant N$. Not useful for short time series
NIR/MNI[a]	tgardner@bu.edu	S	MATLAB	NIR: good performance but requires knowledge of perturbed genes/MNI: good performance for inferring targets of a perturbation
Hierarchical clustering	http://bonsai.ims.u-tokyo.ac.jp/mde-hoon/software/cluster	S/D	GUI	Useful for finding coexpressed genes, but not for network inference

Abbreviations: D: dynamic time-series; N: number of genes; M: number of experiments; S: steay-state.
[a]Predicts only targets of a perturbation (see text for details).

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**:166-176.

- A regulatory module is a set of genes that are regulated in concert by a shared regulation program that governs their behavior.
- A regulation program specifies the behavior of the genes in the module as a function of the expression level of a small set of regulators.



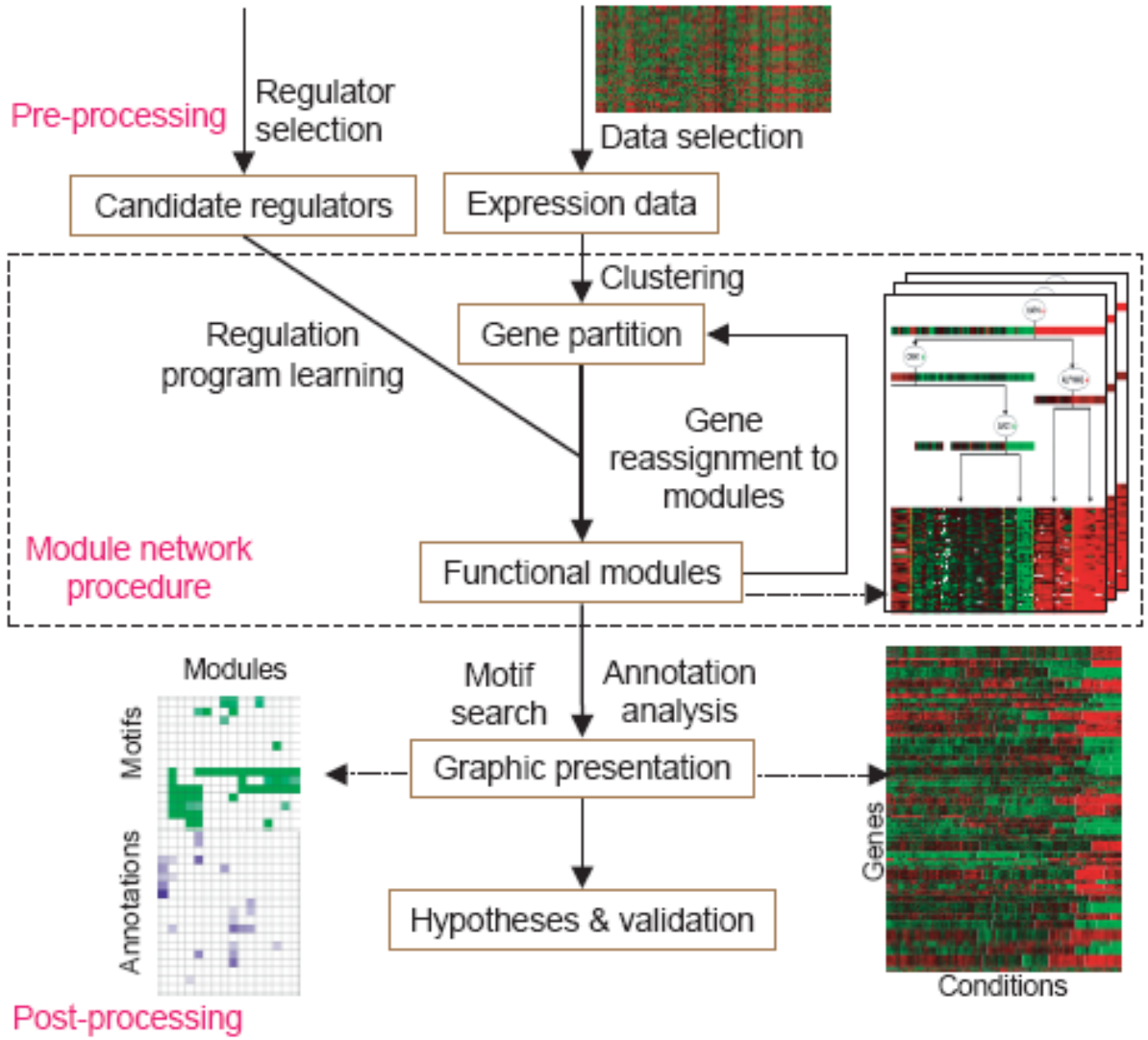
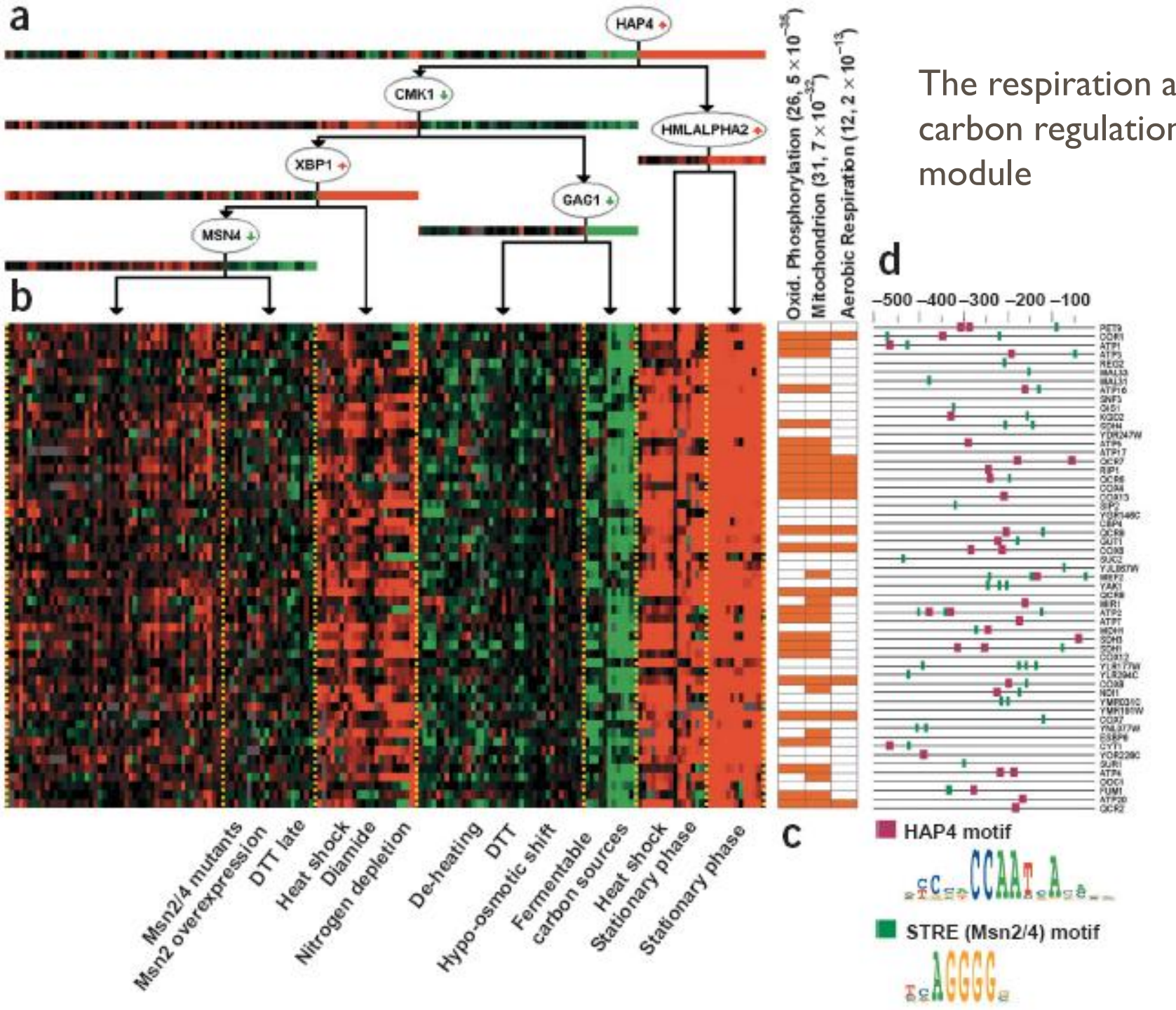


Figure 1 Overview of the module networks algorithm and evaluation procedure. The procedure takes as input a data set of gene expression profiles and a large precompiled set of candidate control genes. The method itself (dotted box) is an iterative procedure that determines both the partition of genes to modules and the regulation program (right icon in dotted box) for each module. In a post-processing phase, modules are tested for enrichment of gene annotations and *cis*-regulatory binding site motifs.

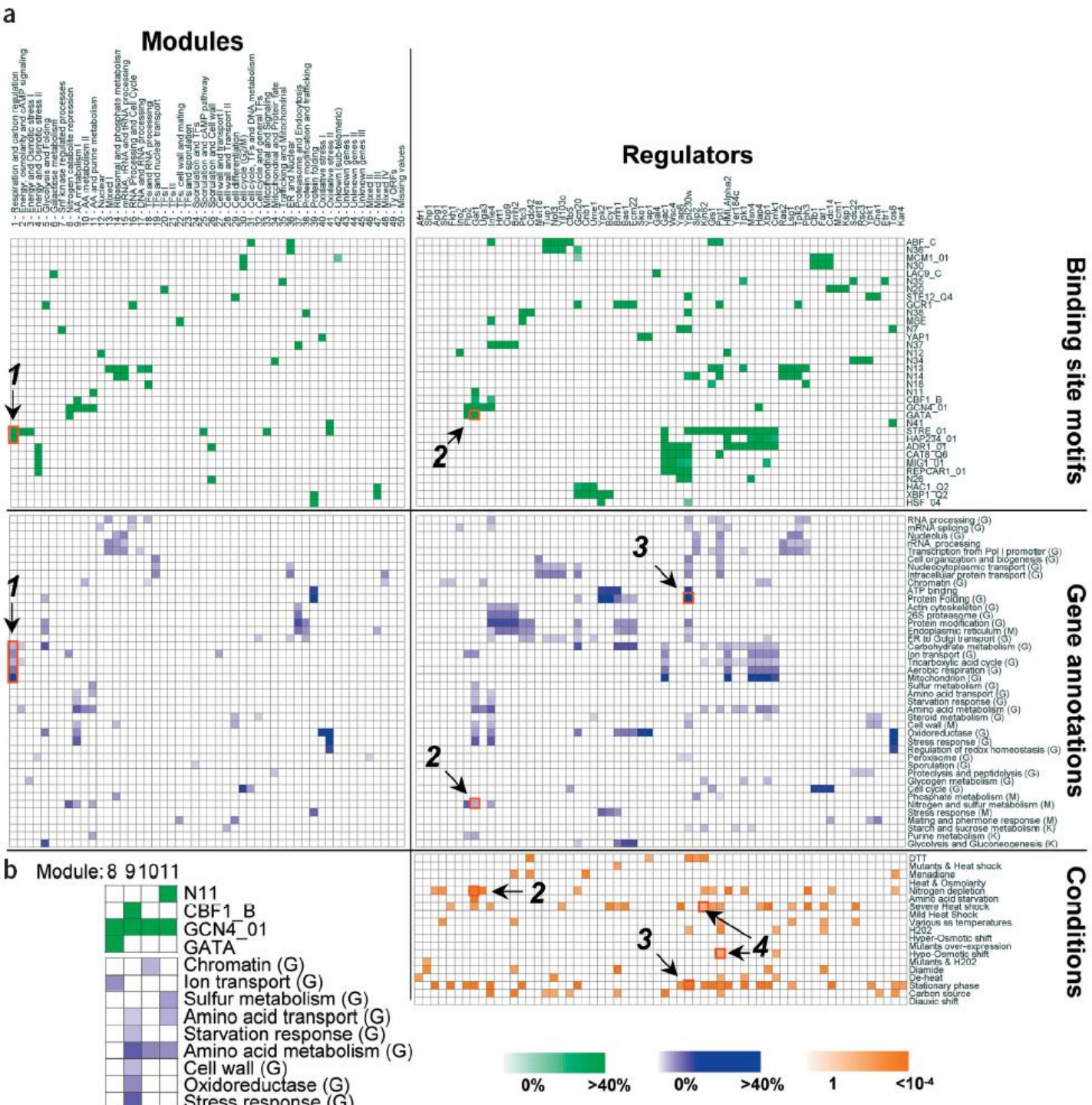
Input a gene expression data set and a large precompiled set of candidate regulatory genes, containing both known and putative transcription factors and signal transduction molecules.

The algorithm searches simultaneously for a partition of genes into modules and for a regulation program (Fig. 2) for each module that explains the expression behavior of genes in the module.

The procedure gives as output a list of modules and associated regulation programs, generating testable hypotheses in the form 'regulator X regulates module Y under conditions W'



The respiration and carbon regulation module


a
Modules
Regulators
Binding site motifs
Gene annotations
Conditions
b
Module: 8 9 10 11
N11
CBF1_B
GCN4_01
GATA
Chromatin (G)
Ion transport (G)
Sulfur metabolism (G)
Amino acid transport (G)
Starvation response (G)
Amino acid metabolism (G)
Cell wall (G)
Oxidoreductase (G)
Stress response (G)
0%
>40%
0%
>40%
1
<10-4

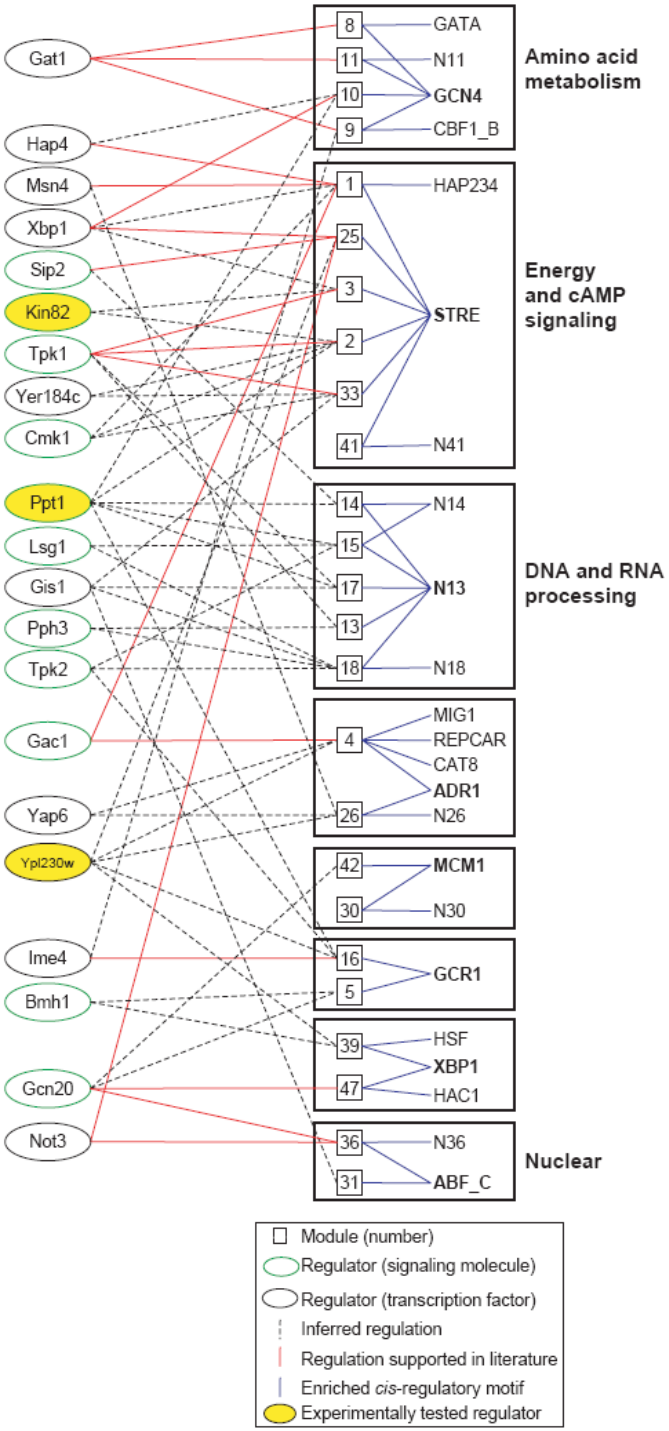


Figure 5 Global view and higher order organization of modules. The graph depicts inferred modules (middle; numbered squares), their significantly enriched *cis*-regulatory motifs (right; significant motifs from **Fig. 4a**) and their associated regulators (left; ovals with black border for transcription factors or with green border for signal transduction molecules). Modules are connected to their significantly enriched motifs by solid blue lines. Module groups, consisting of sets of modules that share a common motif, and their associated motifs are enclosed in bold boxes. Only connected components that include two or more modules are shown. Motifs connected to all modules of their component are marked in bold. Modules are also connected to their predicted regulators. Red edges between a regulator and a module are supported in the literature: either the module contains genes that are known targets of the regulator (**Table 1**, G column) or upstream regions of genes in the module are enriched for the *cis*-regulatory motif known to be bound by the regulator (**Table 1**, M column). Regulators that we tested experimentally are marked in yellow. Module groups are defined as sets of modules that share a single significant *cis*-regulatory motif. Module groups whose modules are functionally related are labeled (right). Modules belonging to the same module group seem to share regulators and motifs, with individual modules having different combinations of these regulatory elements.

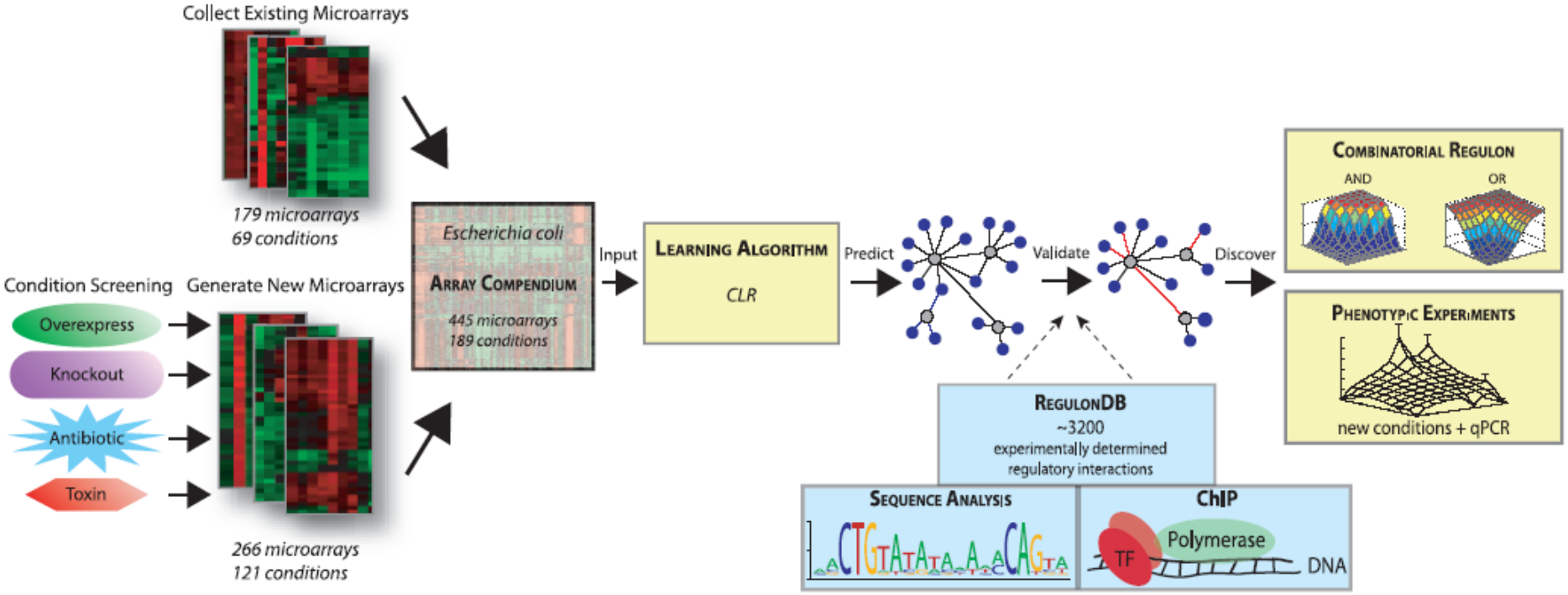


Figure 1. Overview of Our Approach for Mapping the *E. coli* Transcriptional Regulatory Network

Microarray expression profiles were obtained from several investigators. Our laboratory profiled additional conditions, focusing on DNA damage, stress responses, and persistence. These two data sources were combined into one uniformly normalized *E. coli* microarray compendium that was analyzed with the CLR network inference algorithm. The predicted regulatory network was validated using RegulonDB, sequence analysis, and ChIP. The validated network was then examined for cases of combinatorial regulation, one of which was explored with follow-up real-time quantitative PCR experiments.

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles**. *PLoS Biol* 2007, **5**:e8.

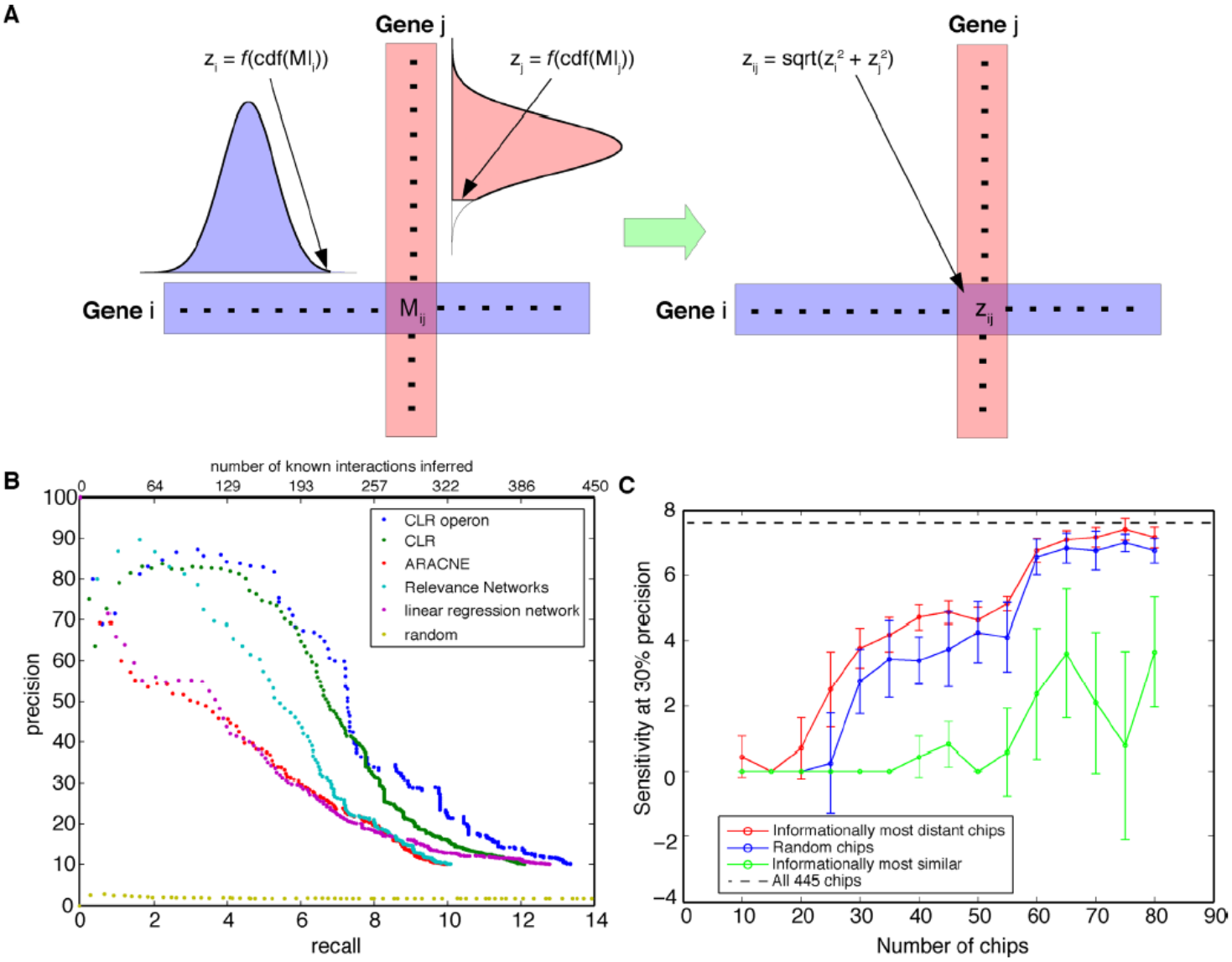
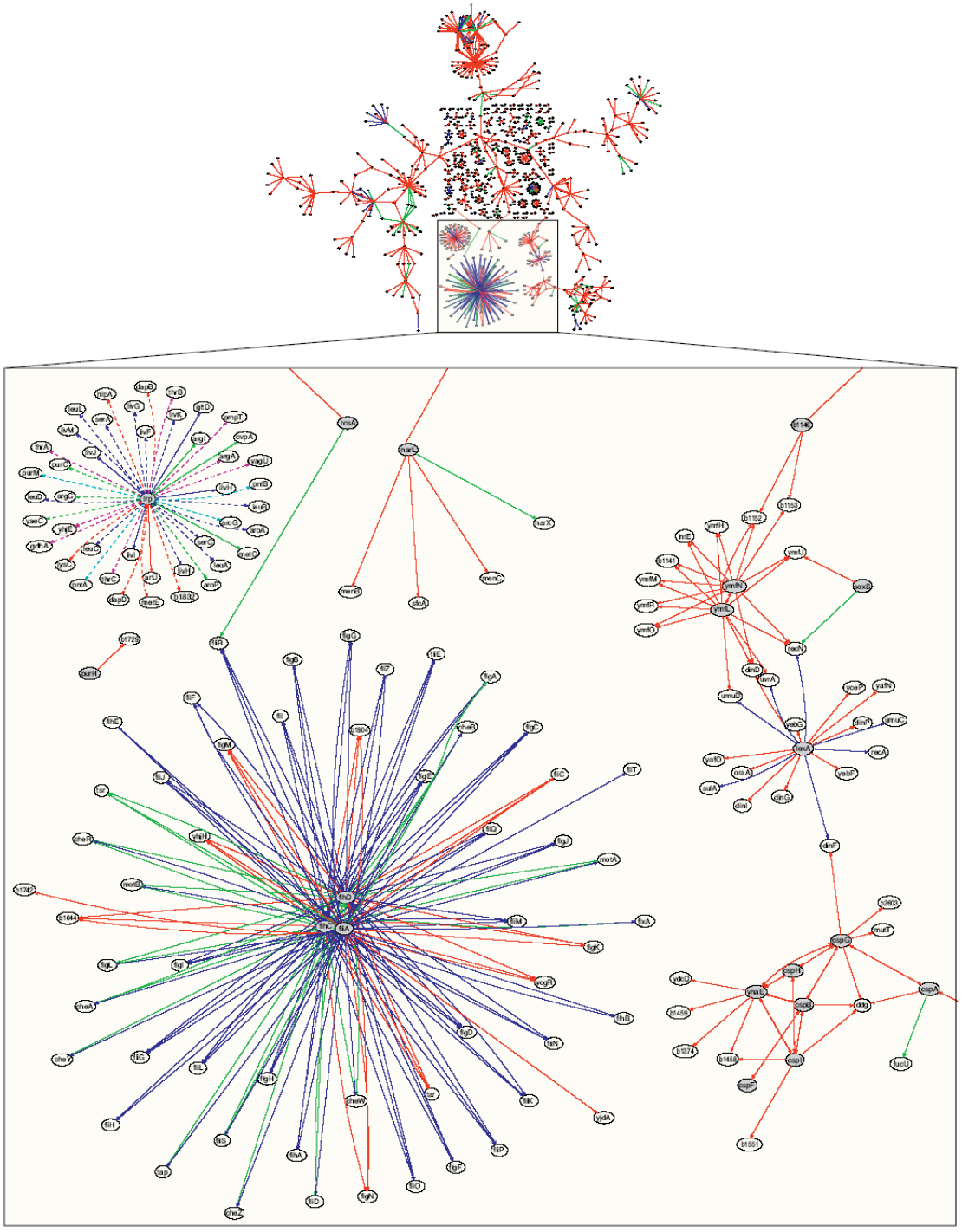


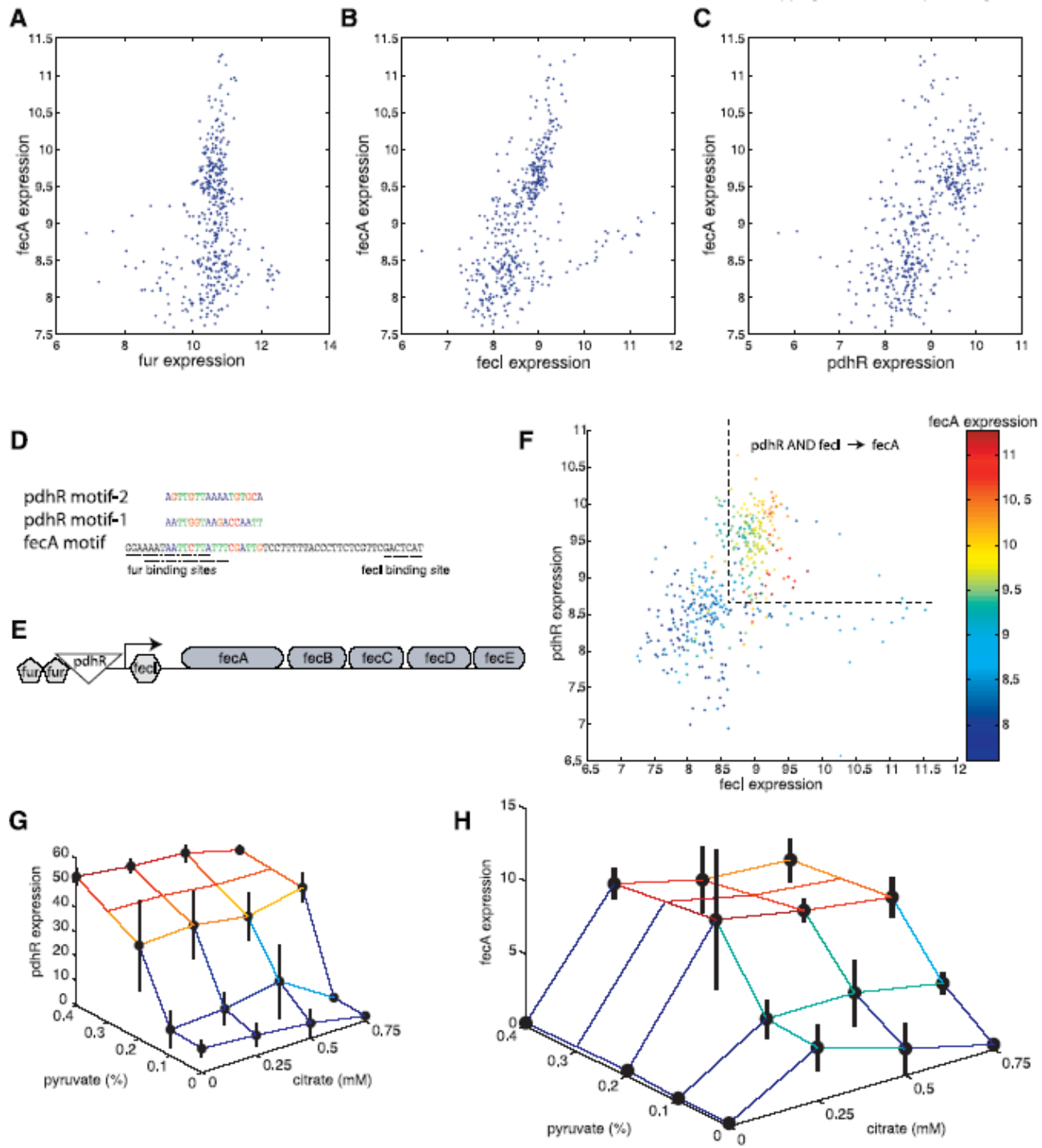
Figure 2. The CLR Algorithm: Methods and Comparison to Other Approaches

(A) A schema of the CLR algorithm. The *z*-score of each regulatory interaction depends on the distribution of MI scores for all possible regulators of the target gene (z_i) and on the distribution of MI scores for all possible targets of the regulator gene (z_j).

(B) Precision and recall for several different network inference methods applied to all genes in the *E. coli* microarray compendium were calculated using RegulonDB. The number of correctly inferred interactions (within RegulonDB) for each recall value is labeled on the top of the chart. All algorithms performed far better than the random method. Both CLR and relevance networks reach high precisions, but CLR attains almost twice the recall of relevance networks at some levels of precision.

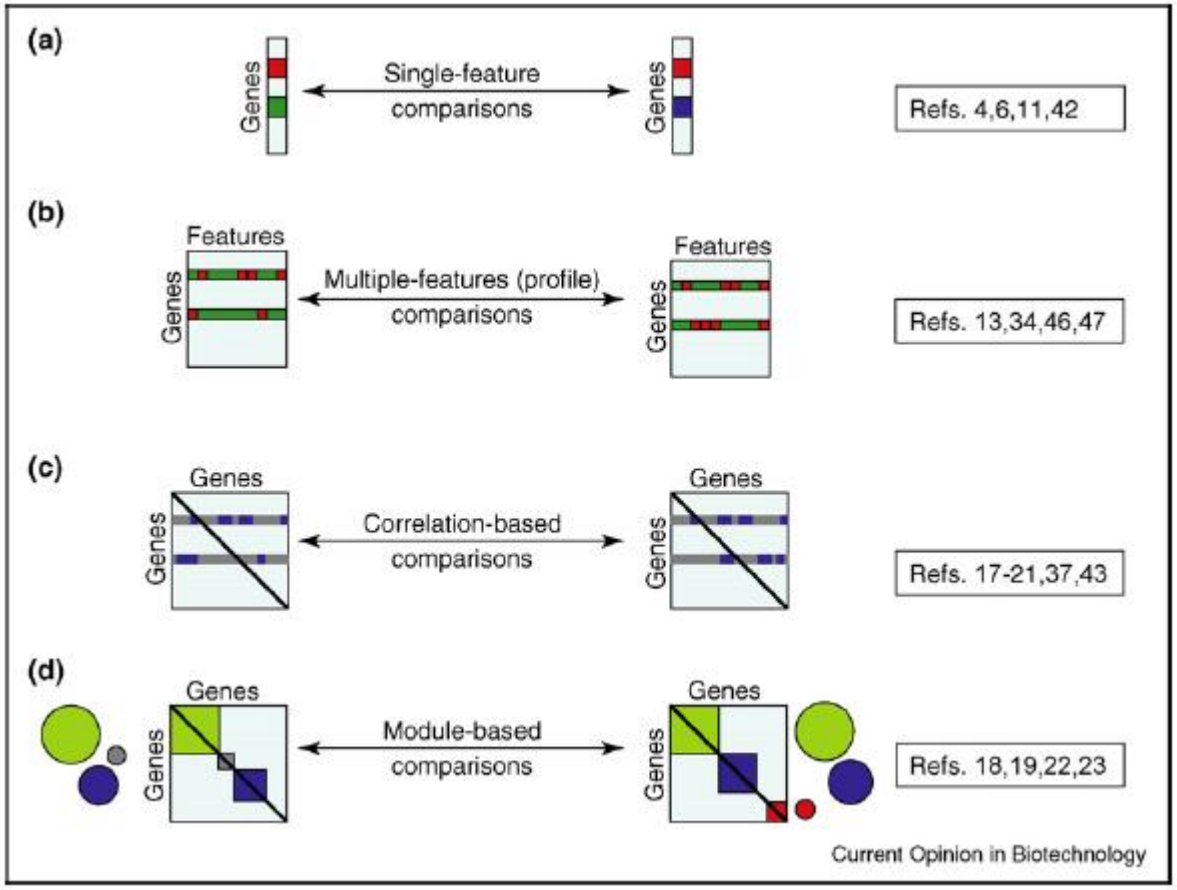
(C) Using 60 well-chosen arrays, we can infer a network, nearly equivalent in recall and precision to the network inferred using all 445 microarrays in the compendium (dotted horizontal line), reflecting the redundancy of the compendium and the potential for improvement in choosing subsequent perturbations to profile.



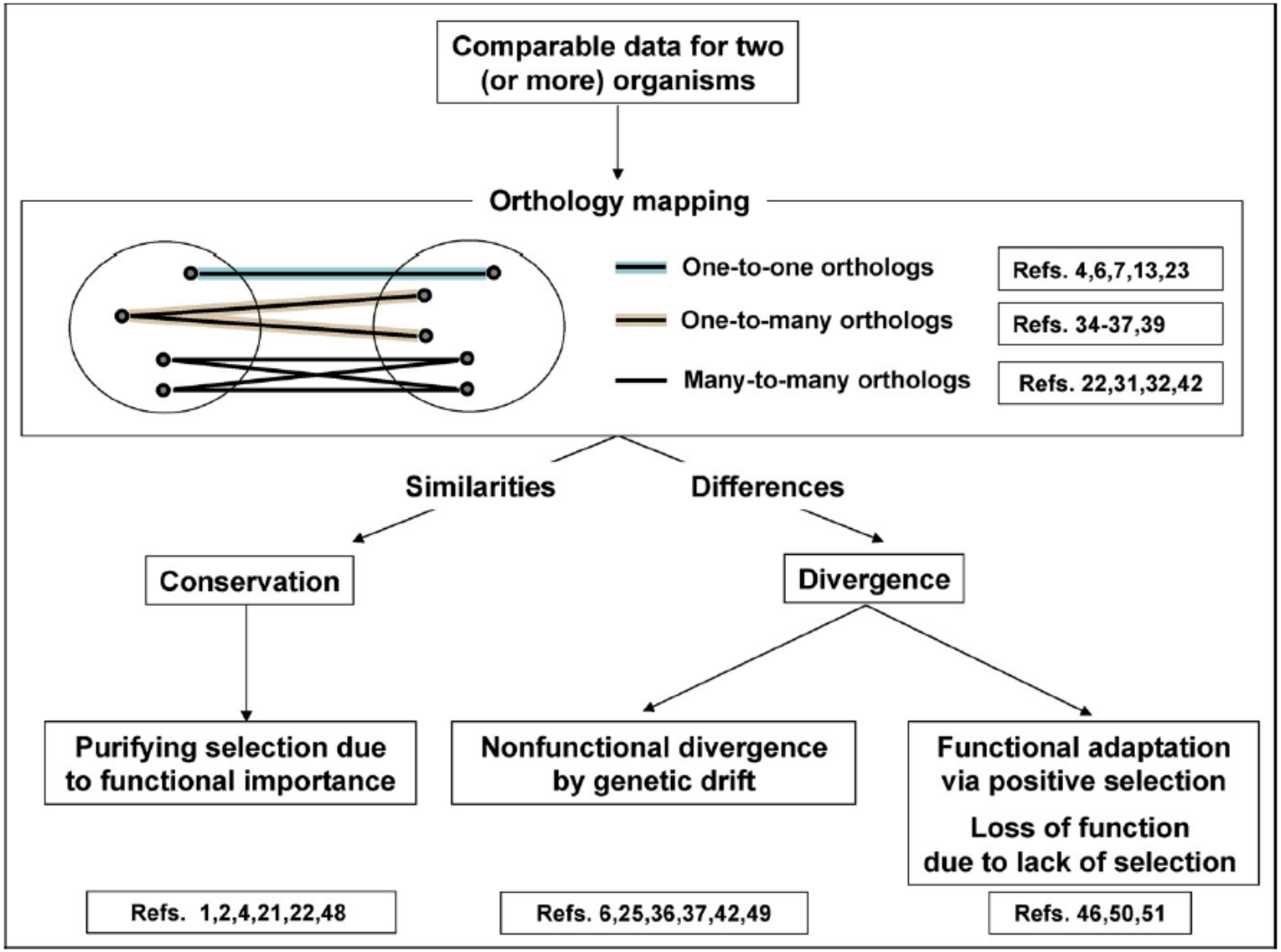


fecABCDE is an operon that encodes a ferric citrate transporter and plays a central role in the import of cellular iron. Existing literature described only two regulators of *fecABCDE*—FecI and Fur. The Fur regulation is not apparent in the compendium (Figure 6A), while the FecI regulation is clear (Figure 6B). However, the bifurcation of the plot suggests a more complex combinatorial regulation for *fecABCDE*. The CLR algorithm identified PdhR, a pyruvate-sensing repressor and necessary component of the energy transduction cascade, as a possible additional regulator of the *fecA* operon (Figure 6C). We also identified a potential PdhR binding motif in the promoter region of the operon (Figure 6D and 6E). Moreover, in undefined, rich media (Luria-Bertani [LB] with 0.2% glucose), our ChIP results showed a significant enrichment for PdhR-*fecA* binding when judged by a *t*-test (p value = 0.004) and a modest enrichment using a nonparametric rank-sum test (p value = 0.1).

Comparative Analysis in the Light of Evolution

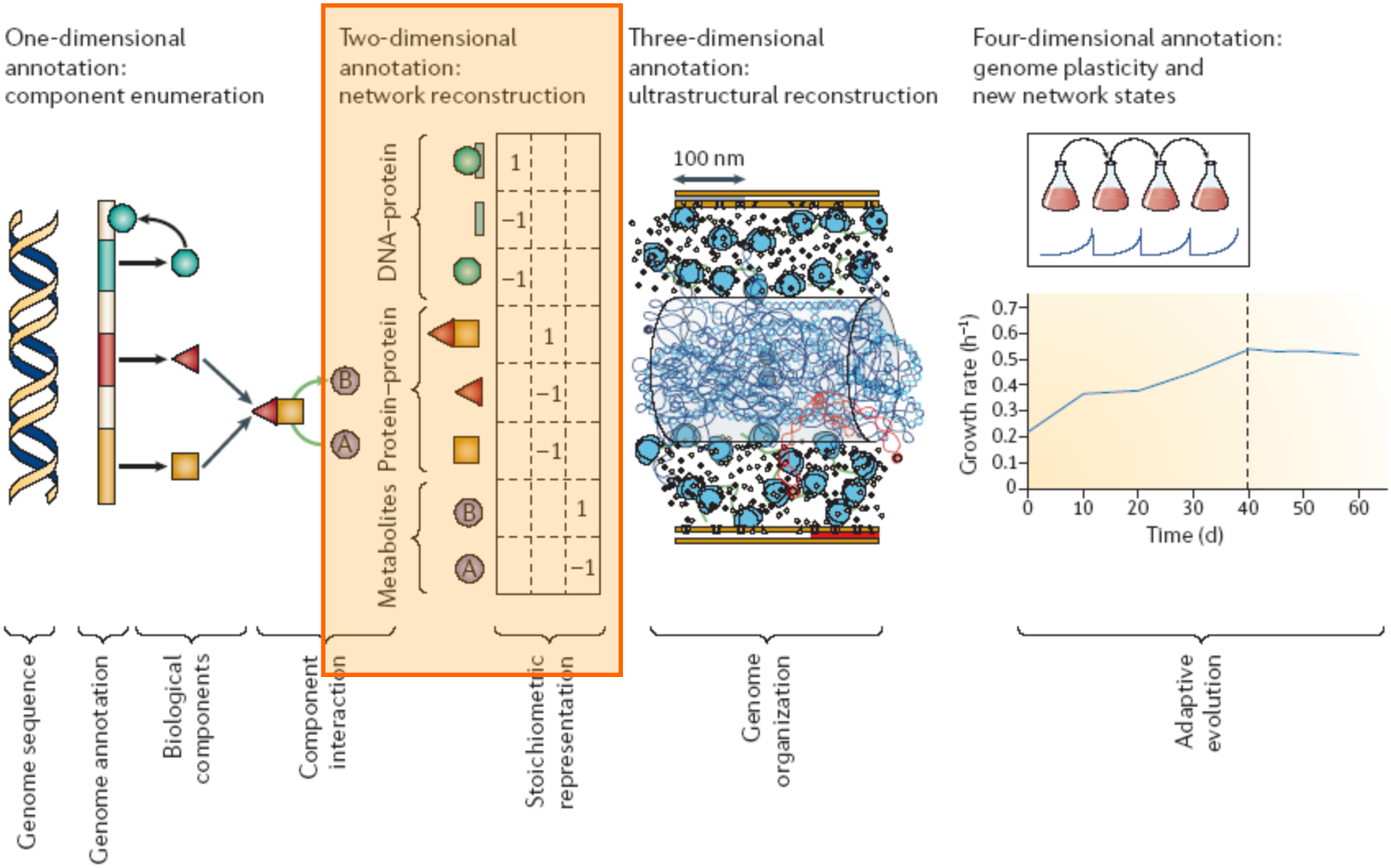


Tirosh I, Bilu Y, Barkai N: **Comparative biology: beyond sequence analysis**. *Curr Opin Biotechnol* 2007, **18**:371-377.



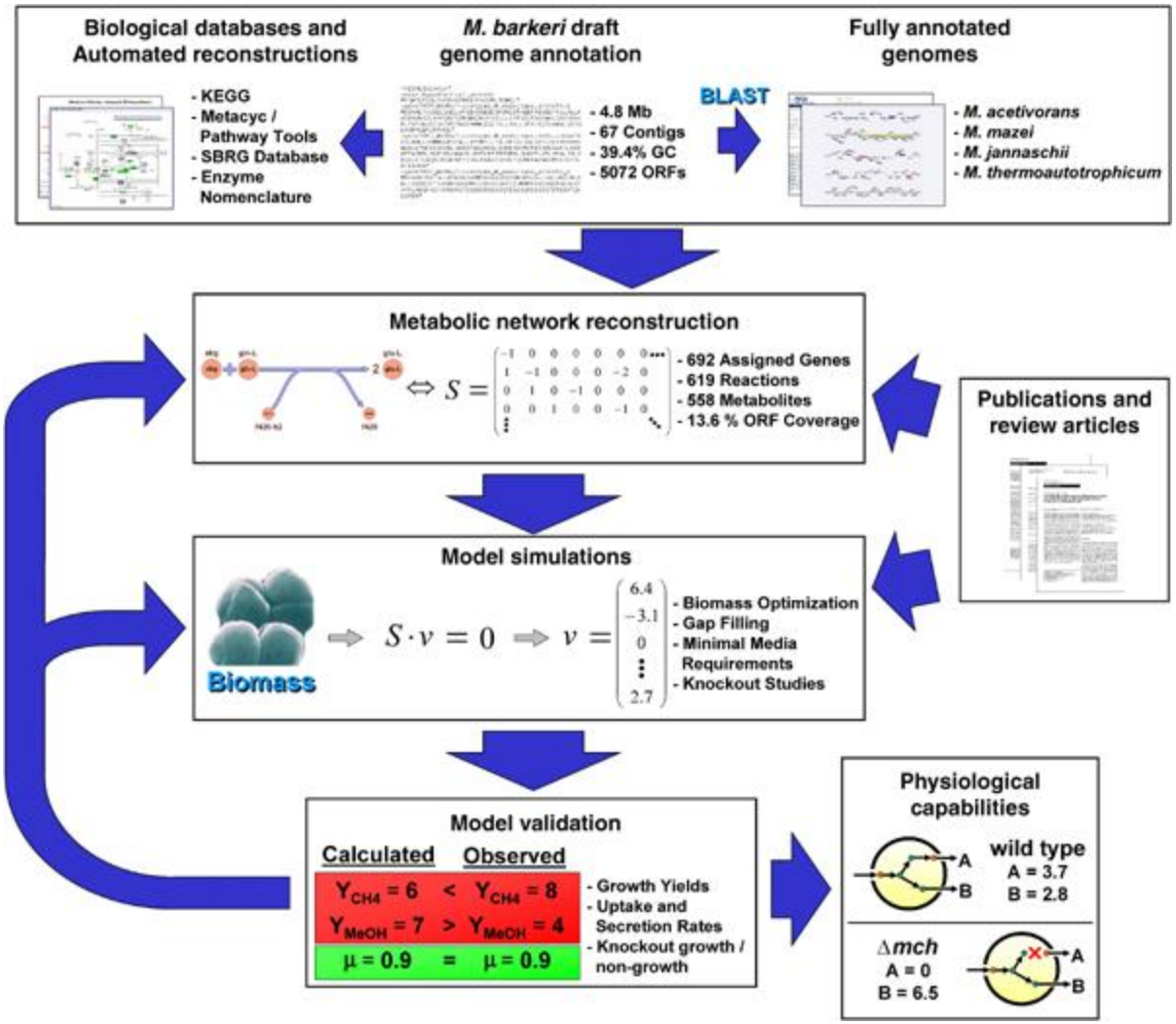
Principles of comparative analysis. Comparative analysis typically starts by collecting comparable data for two or more organisms. To compare the datasets, an orthology mapping and the type of orthology comparison have to be determined. Three types of comparisons are shown: (i) many-to-many, which considers all potential orthology relationships [22•,31,32,42•]; (ii) one-to-one, which considers only the best match of each gene and in some cases excludes ambiguities (i.e. the middle gene in the right circle) [4••,6•,7,13••,23•]; (iii) one-to-many, which focuses on gene duplication and losses [34–37,39]. After the compared objects have been determined, their inter-species similarities are examined. Objects with significant similarity can be identified, which suggests that they were evolutionarily conserved.This conservation may be interpreted as the result of purifying selection and therefore as an indication for functional importance [1,2,4••,21,22•,48•]. Conversely, objects with significant differences are likely to be evolutionarily divergent. This divergence may be associated with either a functional change, being the result of positive selection or lack of selection [6•,25••,36,37,42•,49], or functionally neutral, being the result of random drift [46,50,51]. Several references are given as examples for each scenario.

reconstruct a genome-wide metabolic network



Towards Multidimensional Genome Annotation
Jennifer L. Reed, Iman Famili, Ines Thiele and Bernhard O. Palsson
Nature Reviews Genetics, Vol. 7, No. 2. pp. 130-141

A Workflow

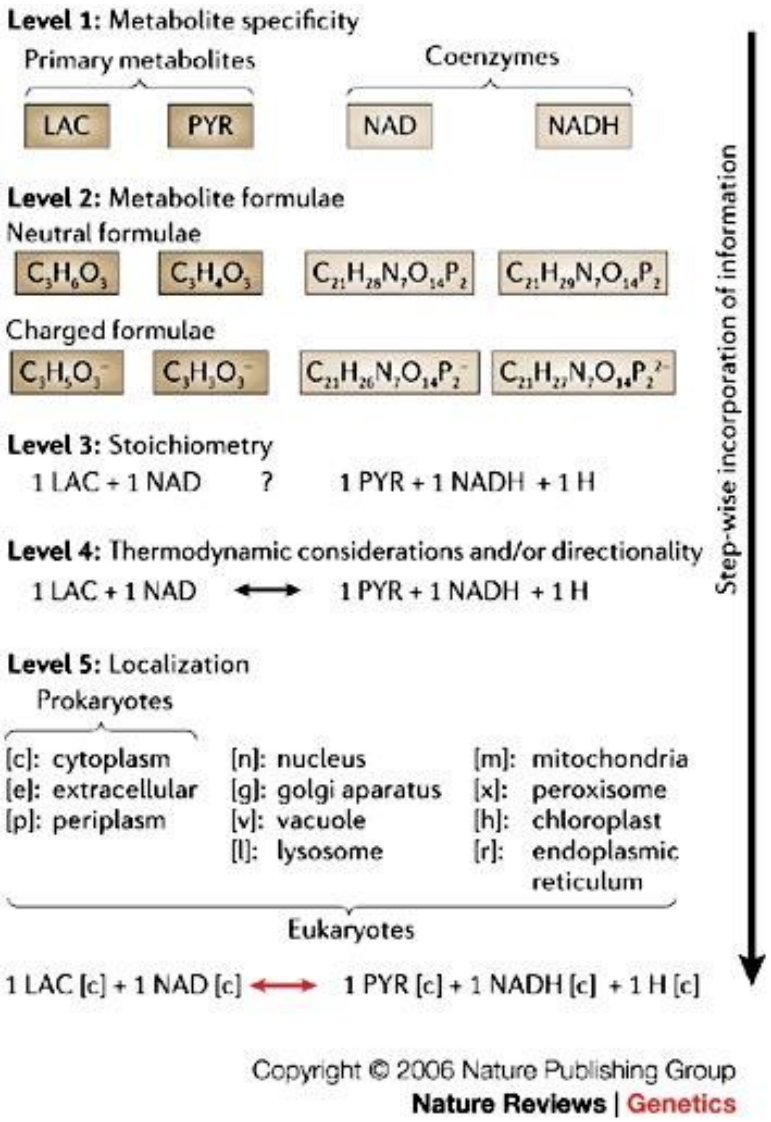


The iterative model building procedure used to generate *i*AF692. The draft genome annotation was used as a scaffold, on which GPR assignments were made. The reactions added to the model were taken from both biochemical databases and published data. Once a reaction was found to be in the network, it was manually curated and either associated to a potential ORF or added with no gene assignment. A biomass objective function was formulated to perform model simulations based on cellular composition. Modeling simulations were run under steady-state conditions to determine the reaction flux distribution in the network. The results from the simulations were interpreted and compared to experimental data. From the comparison, physiological capabilities of the cell were confirmed or the network was further refined or updated.

Molecular Systems Biology ***2*** *Article number: 2006.0004 doi:10.1038/msb4100046 .*

Metabolic network reconstruction mainly involves:

1. **Metabolite specificity of an enzyme**
2. **Molecular formulae**
3. **Stoichiometry**
4. **Directionality or reversibility from biochemical study and thermodynamic properties**
5. **Localization of reactions and proteins to specific cellular compartments.**



Potential Problems

- *Incorrect substrate specificity*
- *Reaction reversibility is not defined.*
- *Enzyme subunits are shown as catalyzing a reaction independently although they are active only in a complex.*
- *Cofactor requirements are often specific for the given organism and have to be found elsewhere.*
- *Several reactions that are necessary for making a functional cell have not been assigned a corresponding ORF.*

Genome Res., Vol. 15, No. 6. (1 June 2005), pp. 820-829

Reed *et al. Nature Reviews Genetics* **7**, 130–141 (February 2006) | doi:10.1038/nrg1769

Sources of information and involved data types:

	KEGG	BRENDA	UniProtKB	Entrez Gene	PubChem	MetaCyc	Transport DB	TIGR	PSORTdb
Information about the definition of metabolic reactions									
Substrate specificity	✓	✓	✓			✓	✓		
Metabolite formulae	✓	✓			✓	✓	✓		
Stoichiometry	✓	✓	✓			✓			
Reaction directionality	✓	✓				✓	✓		
Subcellular localization				✓		✓			✓
Other information about metabolic-reaction properties									
Genome sequence and annotation	✓		✓	✓				✓	
GPR associations	✓	✓				✓	✓		
Literature	✓	✓	✓	✓		✓	✓		

GPR associations, gene–protein–reaction associations.

Textbooks and literature:
Biochemical data, such as characterization of enzymes, essentiality of enzymes or genes.
Physiological data, such as minimal medium requirements and favorable growth environments.
Phylogenetic data are useful when a particular organism is not well studied but a close relative is; in these cases information can be inferred from a close relative.

Reed *et al. Nature Reviews Genetics* **7**, 130–141 (February 2006) | doi:10.1038/nrg1769

Representation of a reconstructed metabolic network:

1.Textually

2. Graphically - A map of nodes and edges can be useful for analyzing topological features of a network.

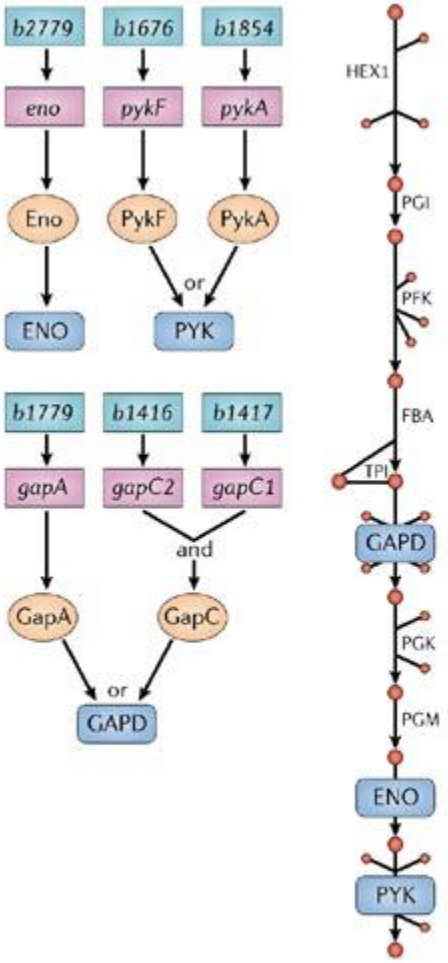
GPR** (gene–protein–reaction) **association and GPR scheme:

- *indicate which genes encode which proteins and which enzymatic reactions these proteins catalyze.*
- *relate various data types, including genomic, transcriptomic, proteomic and flux data.*
- *distinguish between isozymes, enzyme complexes, enzyme subunits*

3. Mathematically as a matrix

Abbreviation	Glycolytic reactions	Genes
HEX1	[c]GLC + ATP → G6P + ADP + H	*glk*
PGI	[c]G6P ↔ F6P	*pgi*
PFK	[c]ATP + F6P → ADP + FDP + H	*pfkA, pfkB*
FBA	[c]FDP ↔ DHAP + G3P	*fbaA, fbaB*
TPI	[c]DHAP ↔ G3P	*tpiA*
GAPD	[c]G3P + NAD + PI ↔ 13DPG + H + NADH	*gapA, gapC1, gapC2*
PGK	[c]13DPG + ADP ↔ 3PG + ATP	*pgk*
PGM	[c]3PG ↔ 2PG	*gpmA, gpmB*
ENO	[c]2PG ↔ H_2O + PEP	*eno*
PYK	[c]ADP + H + PEP → ATP + PYR	*pykA, pykF*

	HEX1	PGI	PFK	FBA	TPI	GAPD	PGK	PGM	ENO	PYK
ATP	−1	0	−1	0	0	0	1	0	0	1
GLC	−1	0	0	0	0	0	0	0	0	0
ADP	1	0	1	0	0	0	−1	0	0	−1
G6P	1	−1	0	0	0	0	0	0	0	0
H	1	0	1	0	0	1	0	0	0	−1
F6P	0	1	−1	0	0	0	0	0	0	0
FDP	0	0	1	−1	0	0	0	0	0	0
DHAP	0	0	0	1	−1	0	0	0	0	0
G3P	0	0	0	1	1	−1	0	0	0	0
NAD	0	0	0	0	0	−1	0	0	0	0
PI	0	0	0	0	0	−1	0	0	0	0
13DPG	0	0	0	0	0	1	−1	0	0	0
NADH	0	0	0	0	0	1	0	0	0	0
3PG	0	0	0	0	0	0	1	−1	0	0
2PG	0	0	0	0	0	0	0	1	−1	0
PEP	0	0	0	0	0	0	0	0	1	−1
H_2O	0	0	0	0	0	0	0	0	1	0
PYR	0	0	0	0	0	0	0	0	0	1



Network Evaluation

1. I

2. C

exp

Collec

molec

cellul

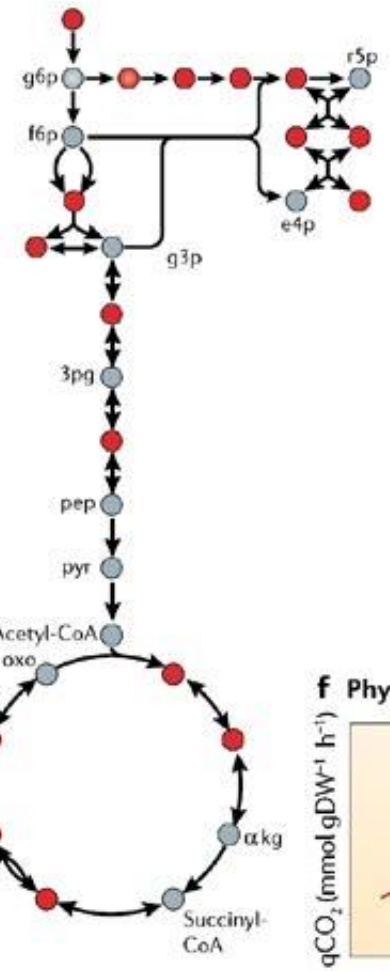
3. C

exp

kno

on

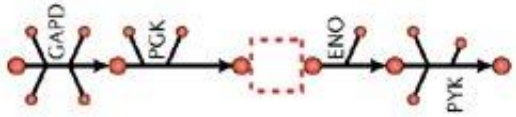
a Precursor metabolite formation



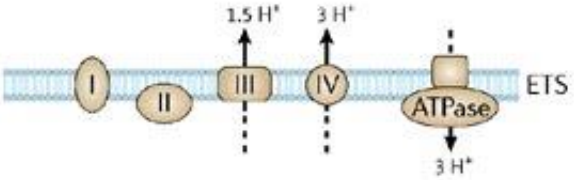
b Incorporating biomass composition

D = 0.1	% (w/w)
Proteins	
Amino acids	45.0
Free amino acids	1.1
Carbohydrates	
Monosaccharides	–
Disaccharides	
Trehalose	0.8
Oligosaccharides	–
Polysaccharides	
Glycogen	8.4
Mannan	13.1
Other carbohydrates	18.4
Nucleotides	
RNA	6.3
DNA	0.4
Lipids	2.9
Ash	5.0
Total	**101.4**

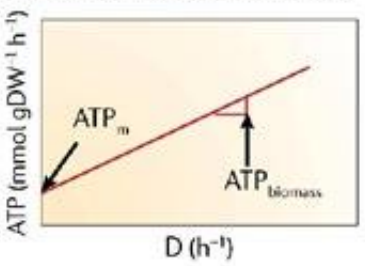
c Filling network gaps



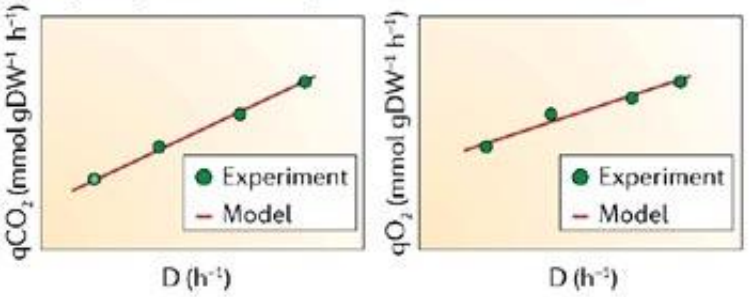
d P/O ratio calculation



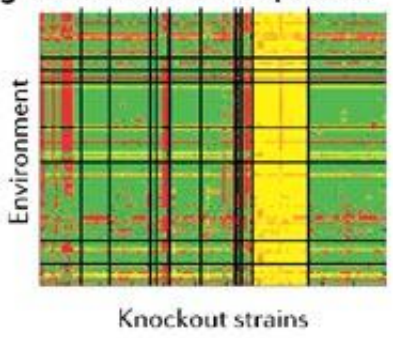
e ATP-maintenance calculation



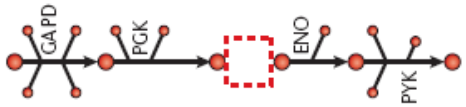
f Physiological-data comparison



g Knockout-data comparison



Gap Finding and Filling



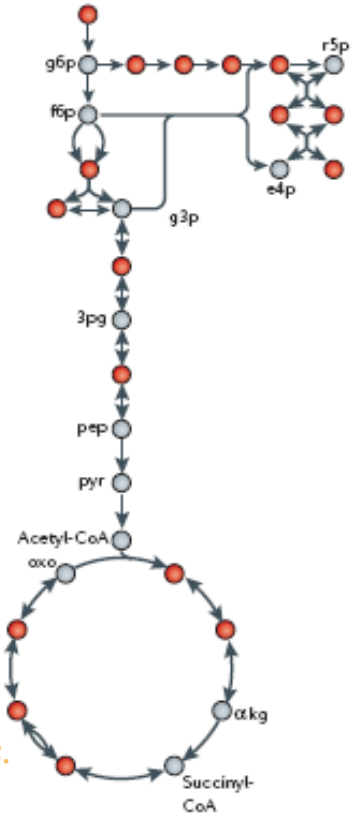
D = 0.1	% (w/w)
Proteins	
Amino acids	45.0
Free amino acids	1.1
Carbohydrates	
Monosaccharides	–
Disaccharides	
Trehalose	0.8
Oligosaccharides	–
Polysaccharides	
Glycogen	8.4
Mannan	13.1
Other carbohydrates	18.4
Nucleotides	
RNA	6.3
DNA	0.4
Lipids	2.9
Ash	5.0
Total	**101.4**

List the biomass components that the organism is known to generate and a complete set of anabolic pathways to synthesize the biomass components. And compare these with what the reconstructed network is able to generate.

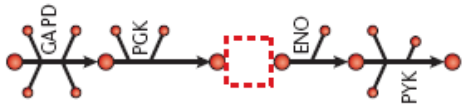
Biomass components: proteins, carbohydrates, lipids and nucleotides, the individual monomers such as amino acids, vitamins, cofactors, metals and minerals that make up a cell.

List the precursor metabolites that are required for the synthesis of biomass components, and compare these with what the network is able to generate.

Precursor metabolites: Metabolites that are generated by catabolic pathways and used by anabolic pathways to generate biomass components. Precursor metabolites link catabolic pathways to anabolic pathways in the cell and are the intermediate molecules used to form macromolecular subunits and biomass components.



Gap Finding and Filling



List other special metabolites that the organism is known to produce or degrade, and compare these with what the network is able to produce or degrade.

Collect biochemical data, such as essential enzymes; physiological data, such as the growth capabilities, minimal medium requirements and favorable growth environments. This information can be used to identify missing reactions.

Analytical tools can also be used to identify network gaps that involve reactions (blocked reactions or pathway holes) or metabolites (dead-end metabolites) that are isolated from the rest of the network.

Flux balance analysis (FBA) can be used to calculate steady-state flux distributions through the metabolic network and is a useful tool for evaluating reconstructed networks.

A stoichiometric matrix, $\boldsymbol{S}$ ($m \times n$), was constructed for the metabolic network where m is the number of metabolites and n is the number of reactions. The corresponding entry in the stoichiometric matrix, S_{ij}, represents the stoichiometric coefficient for the participation of the ith metabolite in the jth reaction.

The linear steady-state problem can be represented by the equation:

$$\boldsymbol{S} \cdot \boldsymbol{v} = 0$$

where $\boldsymbol{v}(n \times 1)$ is a vector of reaction fluxes.

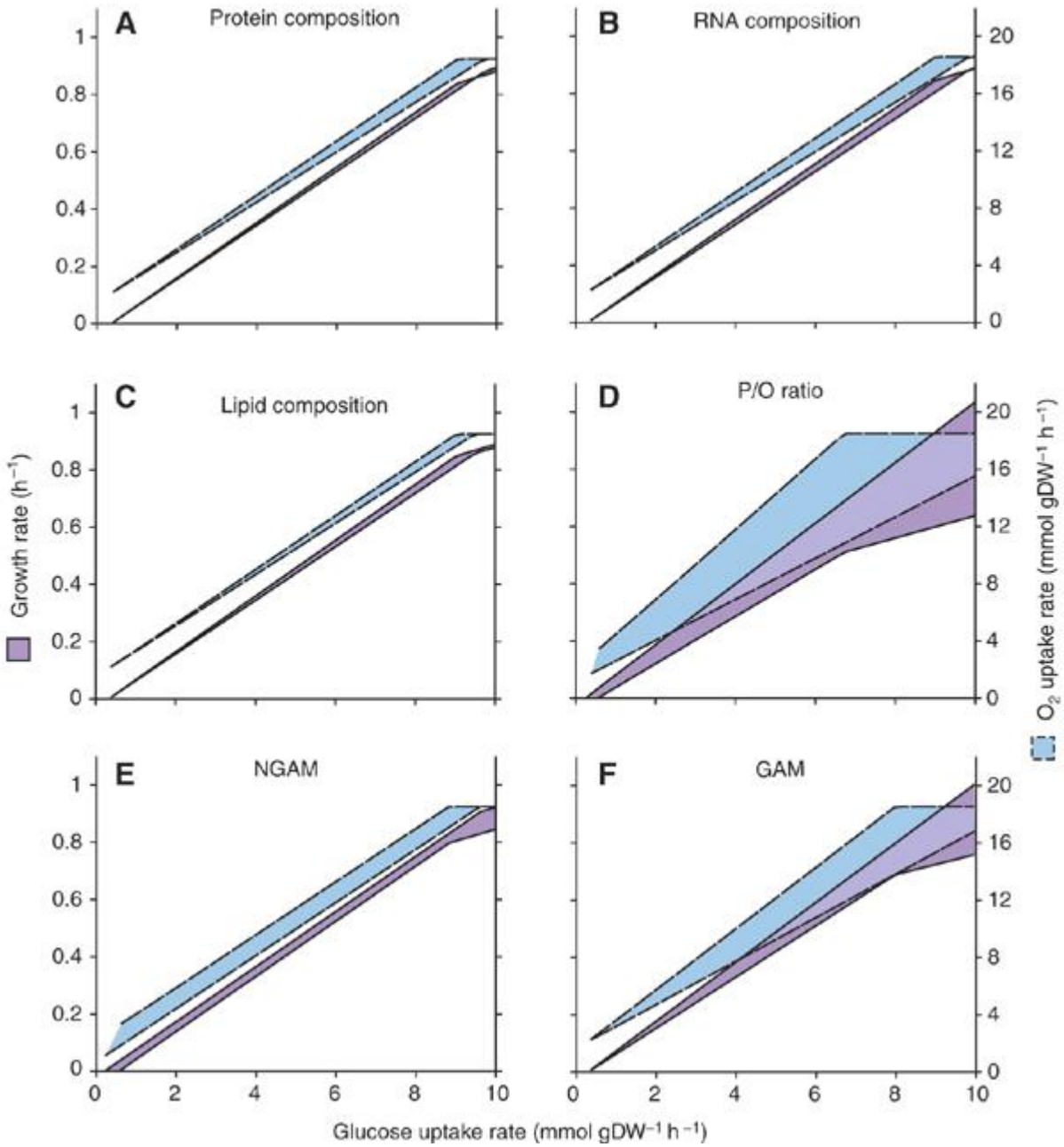
To find a solution for v, the cellular **objective** of producing the maximal amount of biomass constituents, represented by the ratio of metabolites in the BOF, is optimized in the linear system.

constraints that are imposed on the system are in the form of:

$$\alpha_i \leqslant v_i \leqslant \beta_i \qquad (2)$$

where α_i and β_i are the lower and upper limits placed on each reaction flux, v_i, respectively. For reversible reactions, $-\infty \leqslant v_i \leqslant \infty$, and for irreversible reactions, $0 \leqslant v_i \leqslant \infty$.

Sensitivity Analysis



Sensitivity analysis on the modeling parameters used in analyzing *i*AF1260. The relationship between the GUR (mmol gDW^{-1} h^{-1}) (bottom axes, the dependant variable) and the resulting (1) GR (h^{-1}) (left axes) and (2) OUR (mmol gDW^{-1} h^{-1}) (right axes) produced during the sensitivity analysis using *i*AF1260. Using FBA and *i*AF1260, optimal growth was simulated under glucose aerobic conditions while varying (**A**) the dry weight percentage of protein (50–80%), (**B**) RNA (10–25%) and (**C**) lipid (7–15%) in the BOF_{CORE} using physiologically measured values. Also analyzed was (**D**) potential P/O ratios (1.0–2.7) in the network, as well as the (**E**) NGAM (±50%) and (**F**) GAM (±50%) that were determined for these conditions.

Adam M Feist *et.al* ***M****olecular* ***S****ystems* ***B****iology* ***3*** *Article number: 121 doi:10.1038/msb4100155 .*

Gene Essentiality Analysis

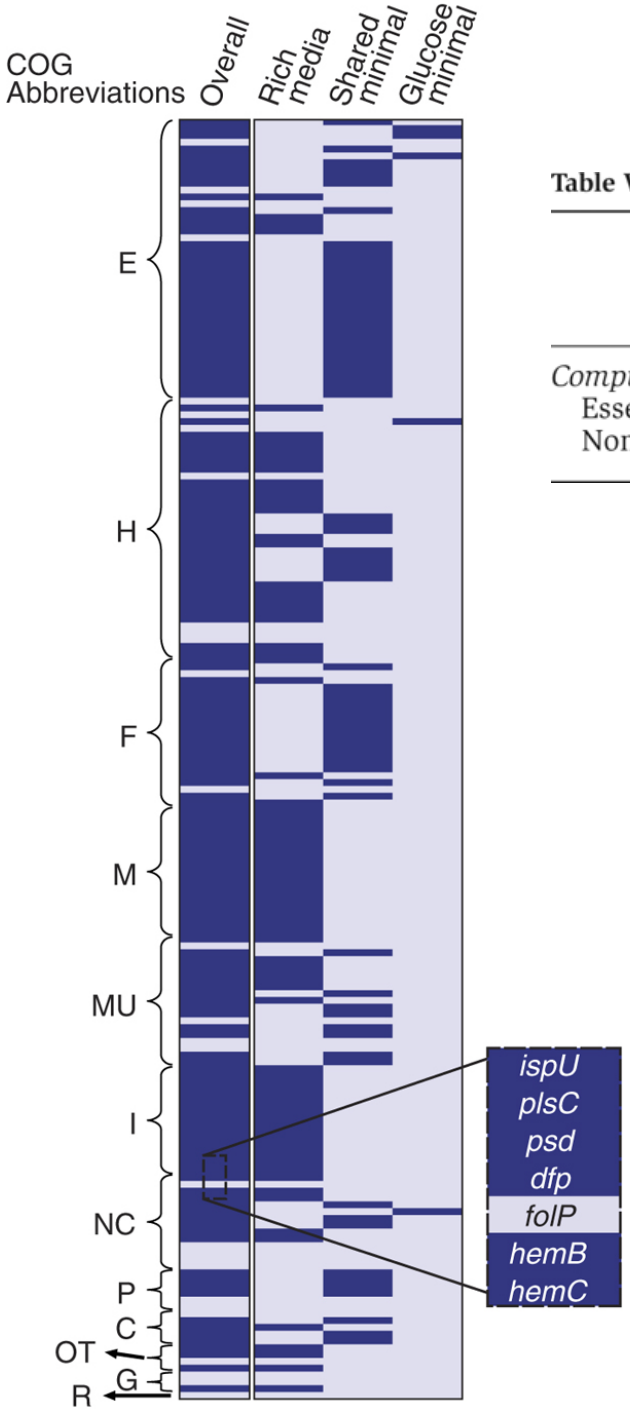


Table V Computational essentiality predictions

	Experimental	
	Essential	Non-essential
Computational		
Essential	159 (13%)	29 (2%)
Non-essential	79 (6%)	993 (79%)

To determine the effect of a gene deletion, the reaction(s) associated with each gene in *i*AF1260 were individually deleted from *S* and FBA was used to predict the mutation growth phenotype. The simulations were performed using glucose minimal medium conditions with a GUR of 10 mmol gDW^{-1} h^{-1}, an OUR of 20 mmol gDW^{-1} h^{-1}, the BOF_{CORE}, an NGAM of 8.39 mmol ATP gDW^{-1} h^{-1}, a GAM of 59.81 mmol ATP gDW^{-1} and zero flux through the 152 reactions regulated under glucose aerobic conditions. The flux through the BOF_{CORE} was optimized in the mutated network, *S*', and a positive flux through the BOF ($vBOF_{core}$>0) was considered non-essential. Experimental criteria for gene essentiality are described in detail in Joyce *et al* (2006).

Adam M Feist *et.al* ***M****olecular* ***S****ystems* ***B****iology* ***3*** *Article number: 121 doi:10.1038/msb4100155 .*

Growth Condition Analysis

Table IV Growth condition analysis

Source	Computational		Experimental	Agreement (*i*AF1260/*i*JR904)			Disagreement (*i*AF1260/*i*JR904)		
	Potential substrates	Support growth[a]	Total possible comparisons	E-G C-G	E-NG C-NG	% Total	E-NG C-G	E-G C-NG	% Total
Carbon	262	174/90	87	54/46	11/15	75%/70%	22/18	0/8	25%/30%
Nitrogen	163	78/34	51	28/24	8/12	71%/71%	8/4	7/11	29%/29%
Phosphorous	63	49/4	20	20/3	0/0	100%/15%	0/0	0/17	0%/85%
Sulfur	25	11/2	12	8/2	0/0	67%/17%	0/0	4/10	33%/83%

[a]Results using the *i*AF1260/*i*JR904 computational model; G, growth; NG, no growth; E, experimental; C, computational.

To determine the carbon, nitrogen, phosphorus and sulfur sources that could support simulated growth, we screened all of the metabolites that could be exchanged with the environment (i.e., exchange reactions) in the *i*AF1260 and *i*JR904 models. The identified metabolites formed the **potential substrate sets**.

If a positive flux could be generated through the BOF_{CORE} reaction (*v*BOFcore>0), then the substrate was considered a viable source. Experimental data used in the comparison were provided by Biolog (http://www.biolog.com)

Adam M Feist *et.al* ***M****olecular* ***S****ystems* ***B****iology* ***3*** *Article number: 121 doi:10.1038/msb4100155 .*

Notes for evaluation of network reconstruction:

- Network evaluation is highly dependent on the availability of data, especially physiological data, which can often be the most limiting factor.

- Network reconstruction is an iterative process, involving network evaluation, genome re-annotation and the availability of new experimental data.

Significance of such work

A knowledge database for an organism with information ranging from genome to metabolism.

Analyzing high-throughput data such as transcriptomic, proteomic and metabolomic data in the context of network reconstruction (that is, ‘putting content into context’) provides the means to improve the accuracy of the network reconstruction, to evaluate the consistency of various heterogeneous data sets within the context of functional roles, to generate testable hypotheses that drive experimental discovery.

……

Thanks

Q & A