V7 – Gene Regulation

- transcription factors
- binding motifs
- gene-regulatory networks

Fri., Nov 18, 2016

Coming from PPI networks "Assembly in time"

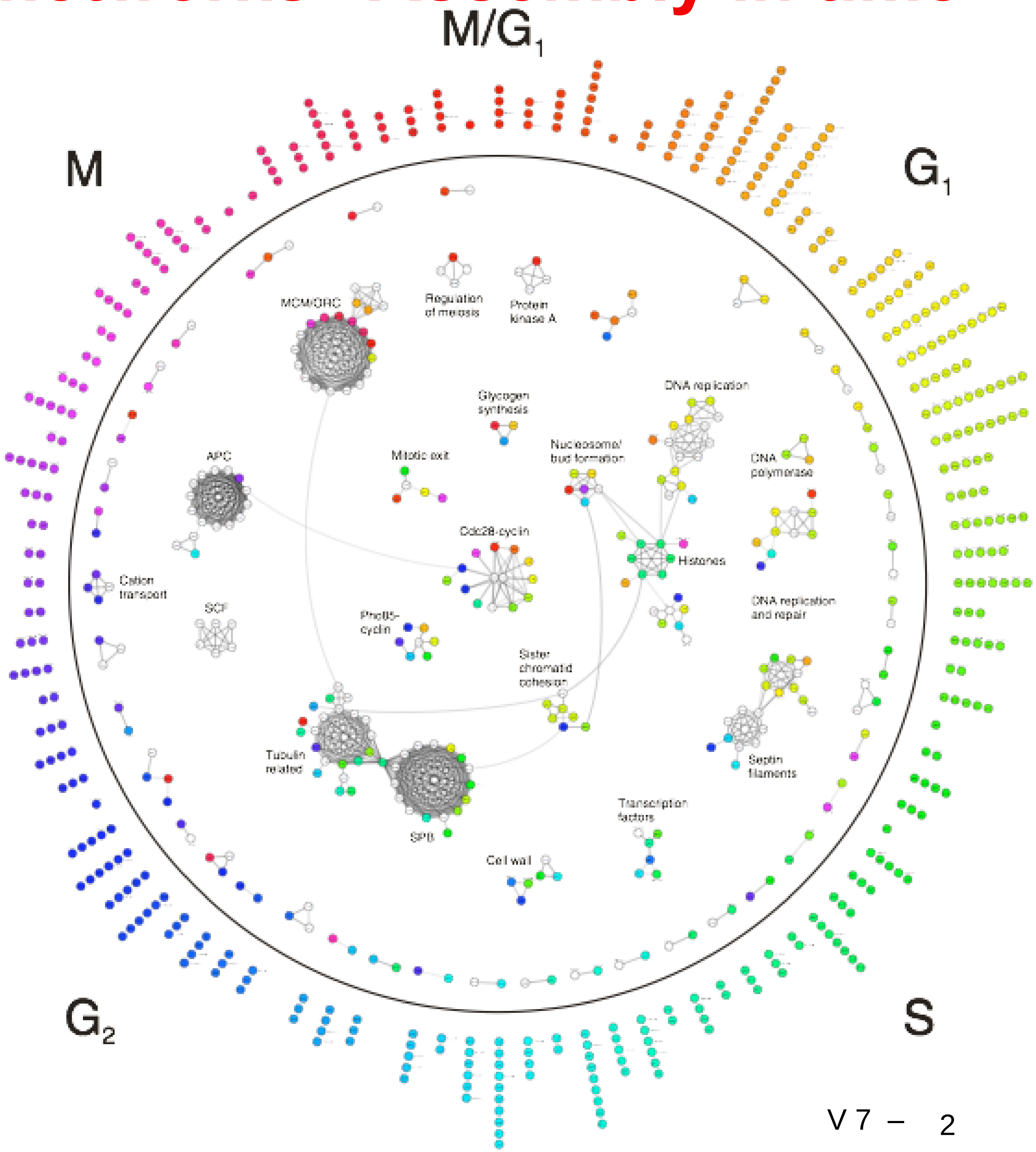
From Lichtenberg et al,
Science 307 (2005) 724:

The wheel represents the 4 stages of a cell cycle in *S. cerevisiae*.

Colored proteins are components of protein complexes that are (only) expressed at certain stages.

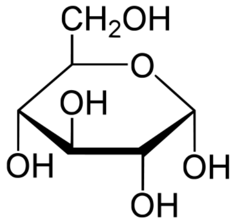
Other parts of these complexes have constant expression rates (white).

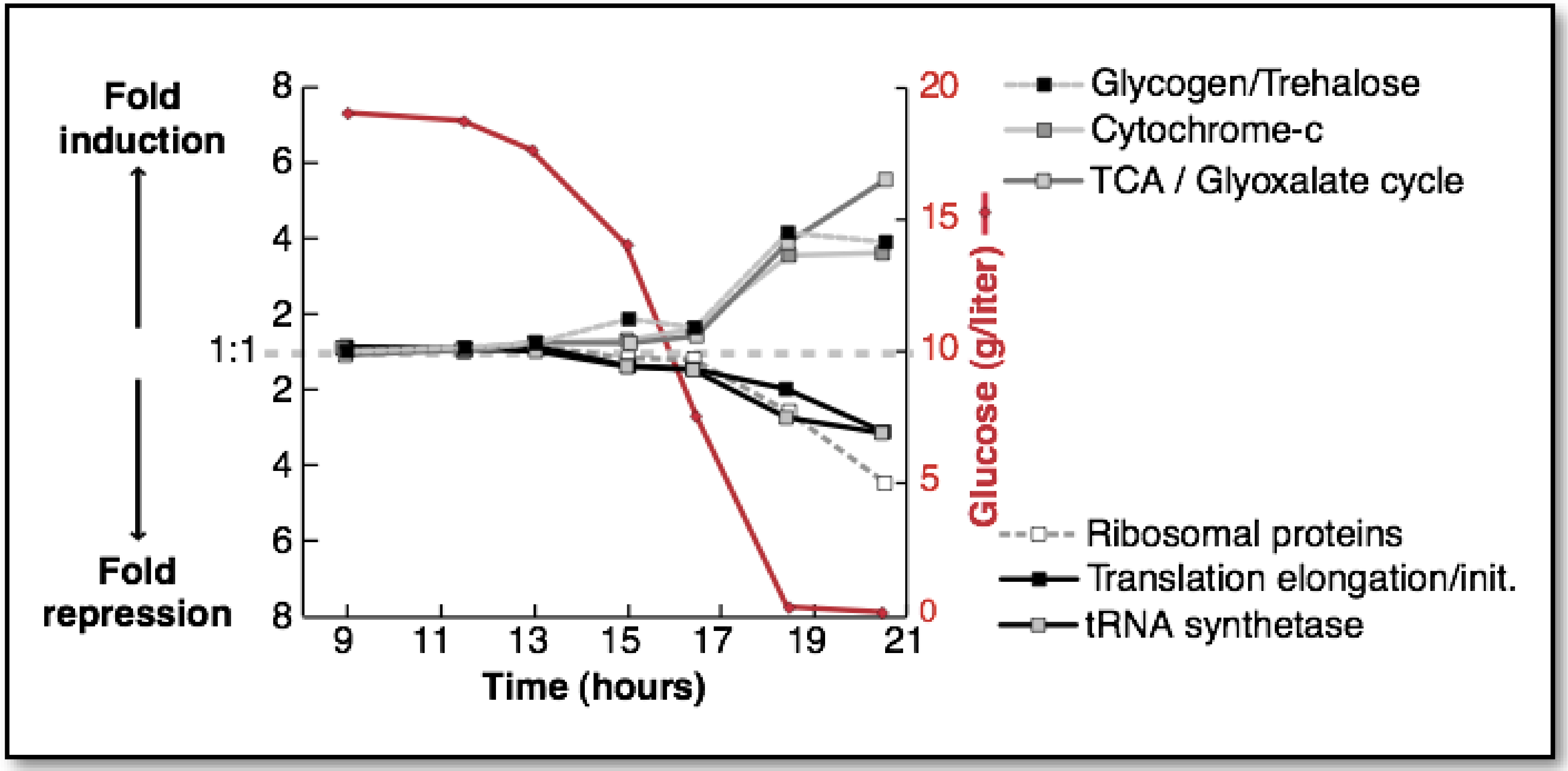
→ "assembly in time"

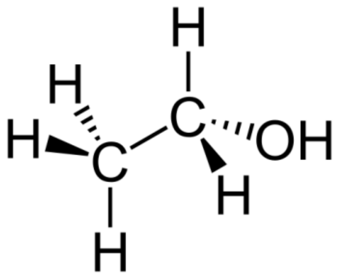


Classic: External triggers affect transcriptome

Re-routing of metabolic fluxes during the "diauxic shift" in *S. cerevisiae*
→ changes in mRNA levels (leads to changes of protein abundance)







anaerobic fermentation:
fast growth on glucose → ethanol

Diauxic shift →

aerobic respiration:
ethanol as carbon source,
cytochrome *c* as electron carrier in respiration and enzymes of TCA cycle (in mitochondrial matrix) and glyoxalate cycles upregulated

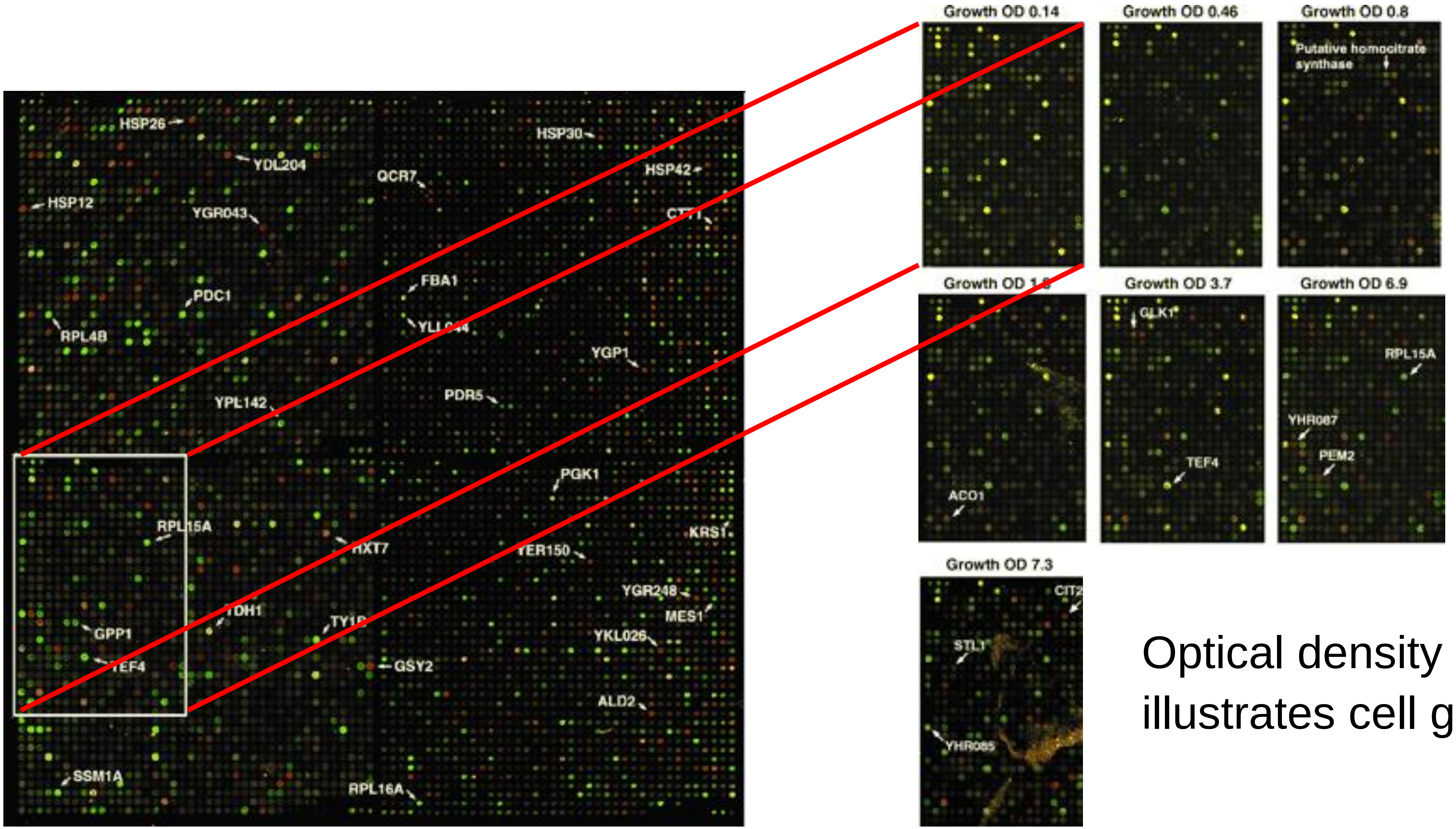
DeRisi et al., *Science* **278** (1997) 680

Diauxic shift affects hundreds of genes

Cy3/Cy5 labels (these are 2 dye molecules for the 2-color microarray), comparison of 2 probes at 9.5 hours distance; w and w/o glucose

Red: genes induced by diauxic shift (710 genes > 2-fold)

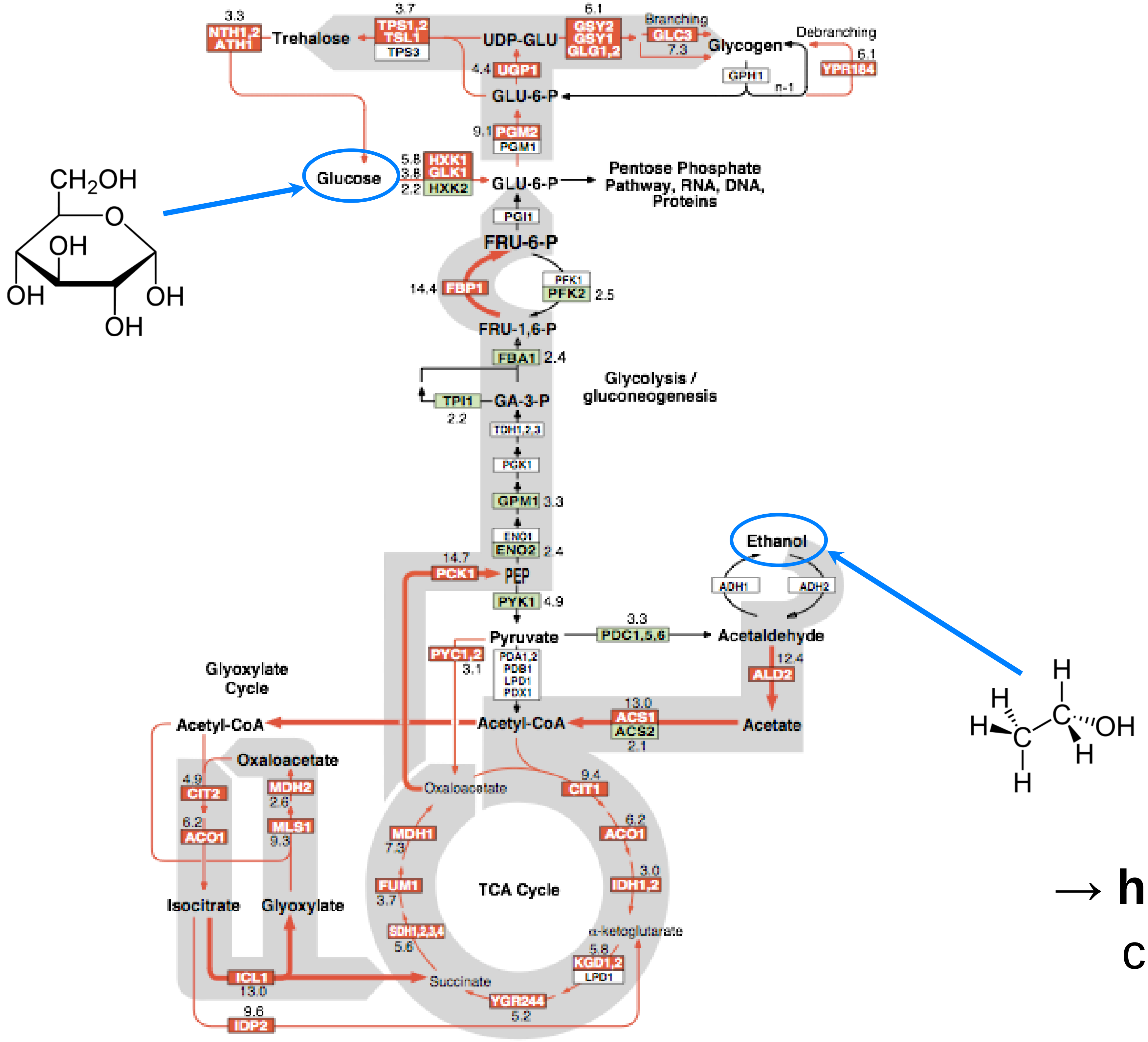
Green: genes repressed by diauxic shift (1030 genes change > 2-fold)



Optical density (OD) illustrates cell growth;

DeRisi et al., *Science* **278** (1997) 680

Flux Re-Routing during diauxic shift



fold change

🟥 expression increases

□ expression unchanged

🟩 expression diminishes

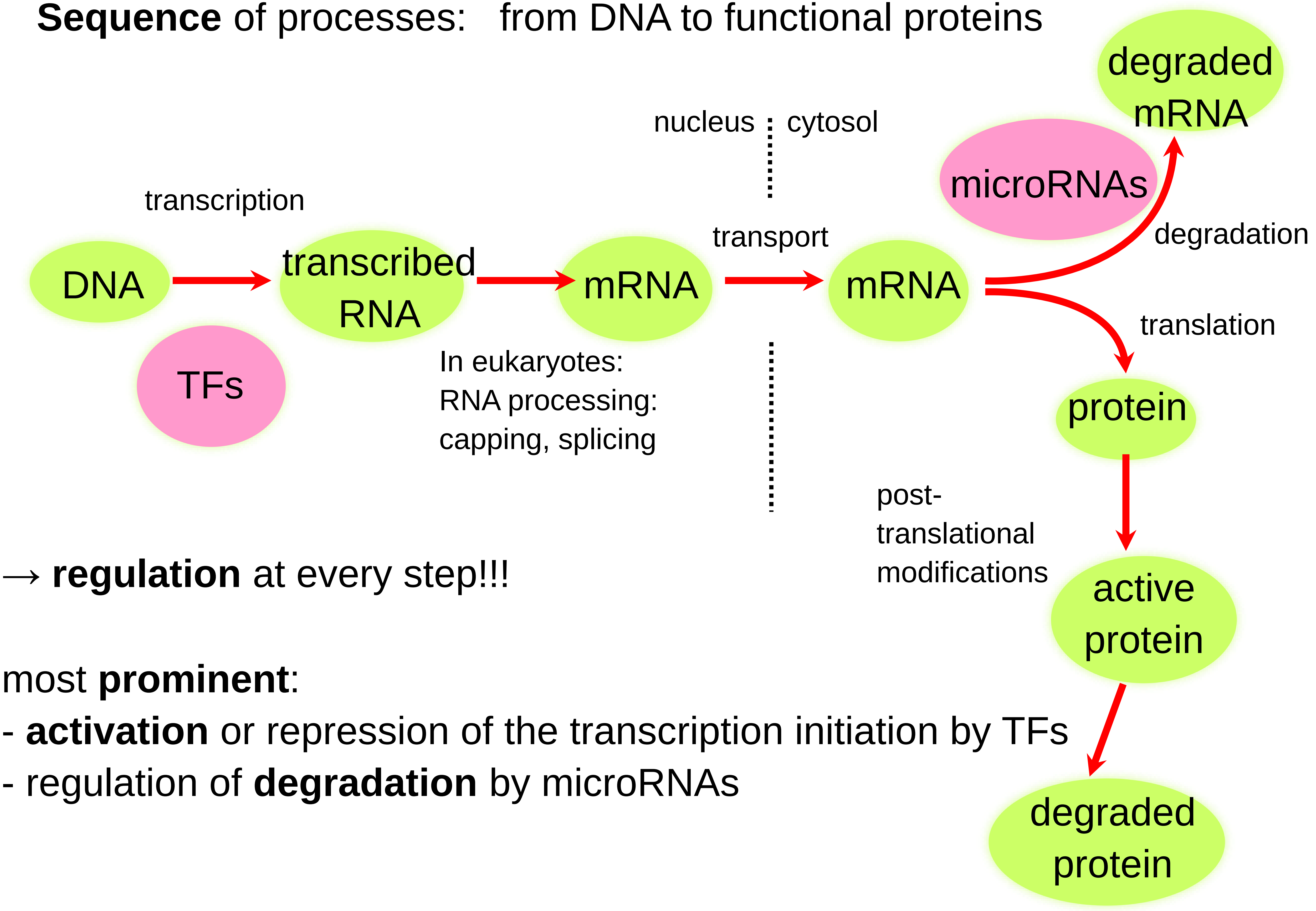
metabolic flux increases

→ **how** are these changes **coordinated?**

DeRisi et al., *Science* **278** (1997) 680

Gene Expression

Sequence of processes: from DNA to functional proteins



→ **regulation** at every step!!!

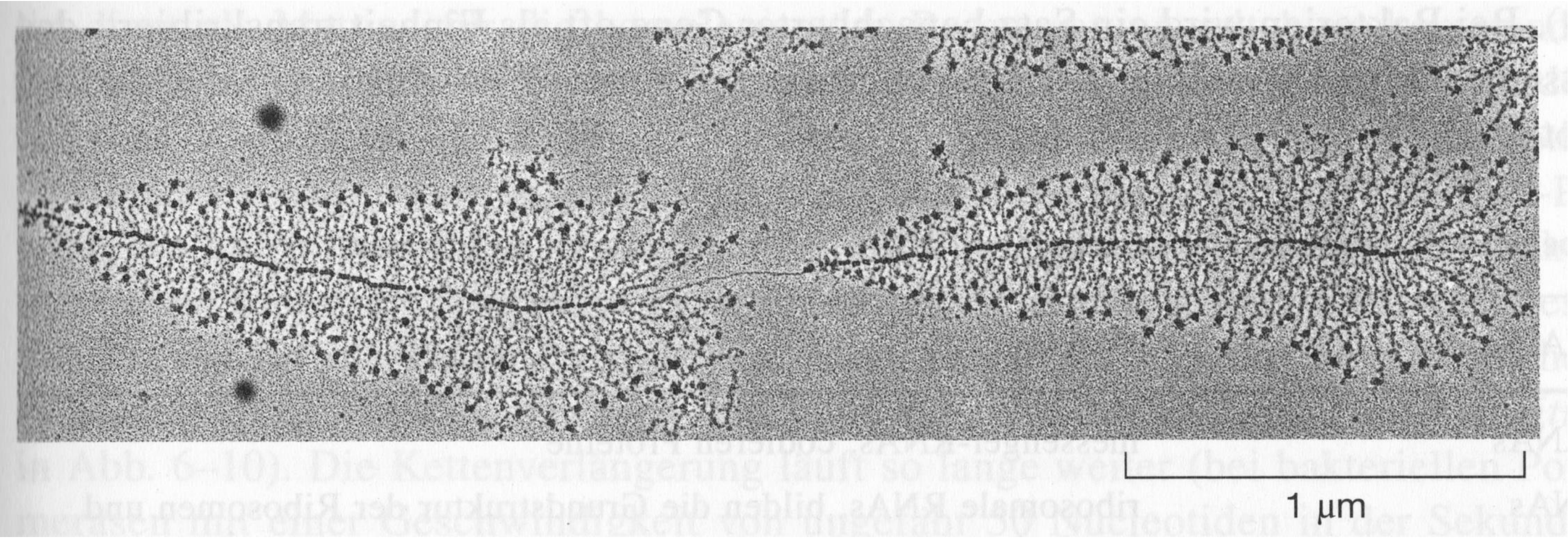
most **prominent**:

- **activation** or repression of the transcription initiation by TFs
- regulation of **degradation** by microRNAs

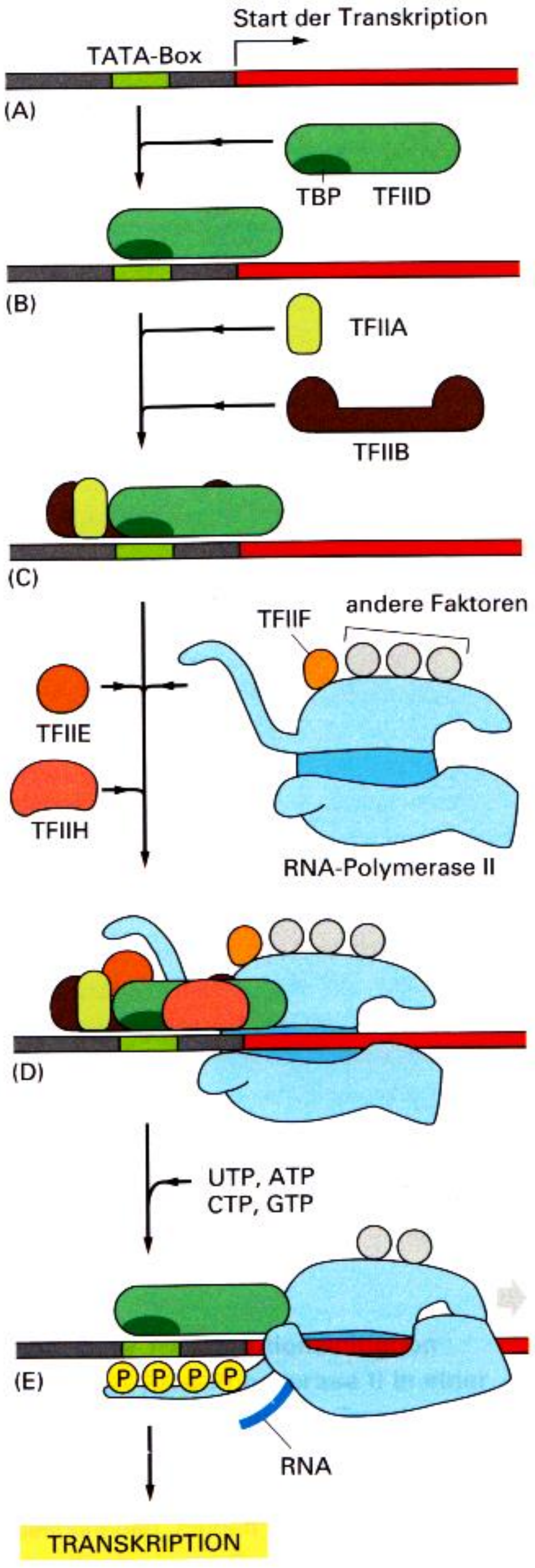
Transcription Initiation

In eukaryotes:

- several **general** transcription factors **have** to bind to gene promoter
- **specific** enhancers or repressors **may** bind
- then the RNA polymerase binds
- and starts transcription



Shown here: many RNA polymerases read central DNA at different positions and produce ribosomal rRNAs (perpendicular arms). The large particles at their ends are likely ribosomes being assembled.



Alberts et al.
"Molekularbiologie der Zelle", 4. Aufl.

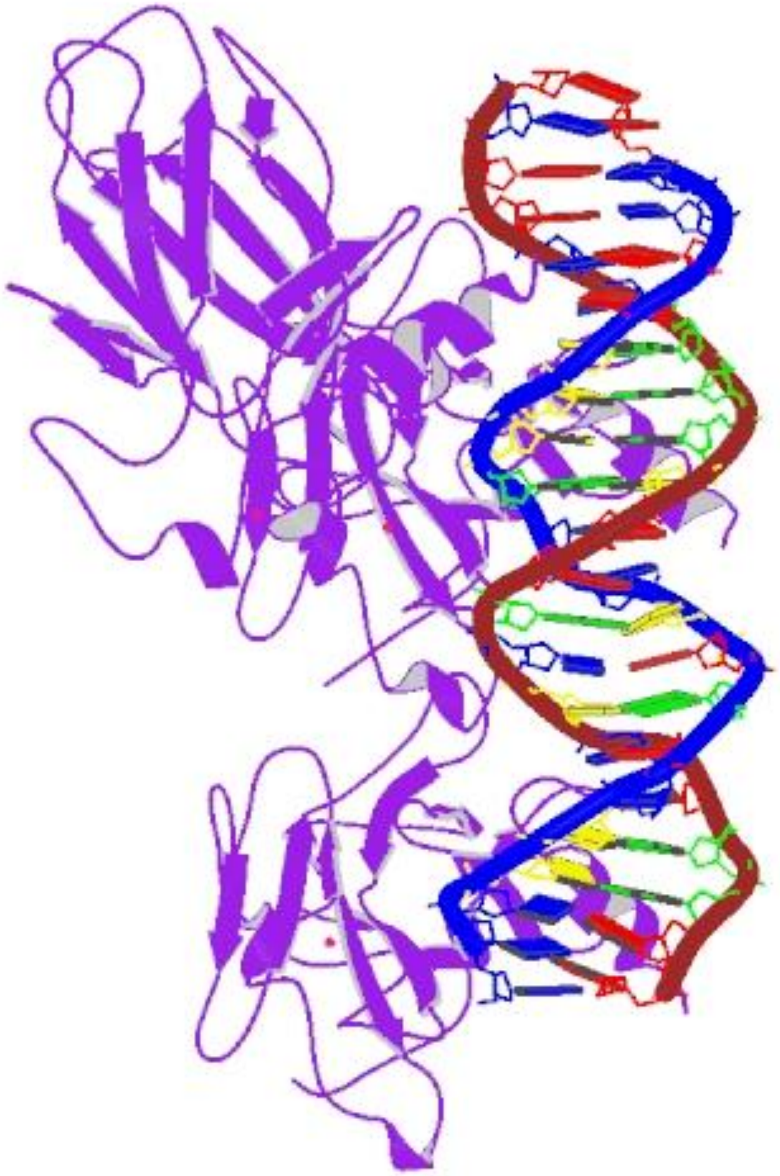
p53: example of a Protein-DNA-complex

PDB-Structure 1TUP: tumor suppressor **p53**

Determined by X-ray crystallography

Purple (left): p53-protein

Blue/red DNA double strand (right)



www.rcsb.org

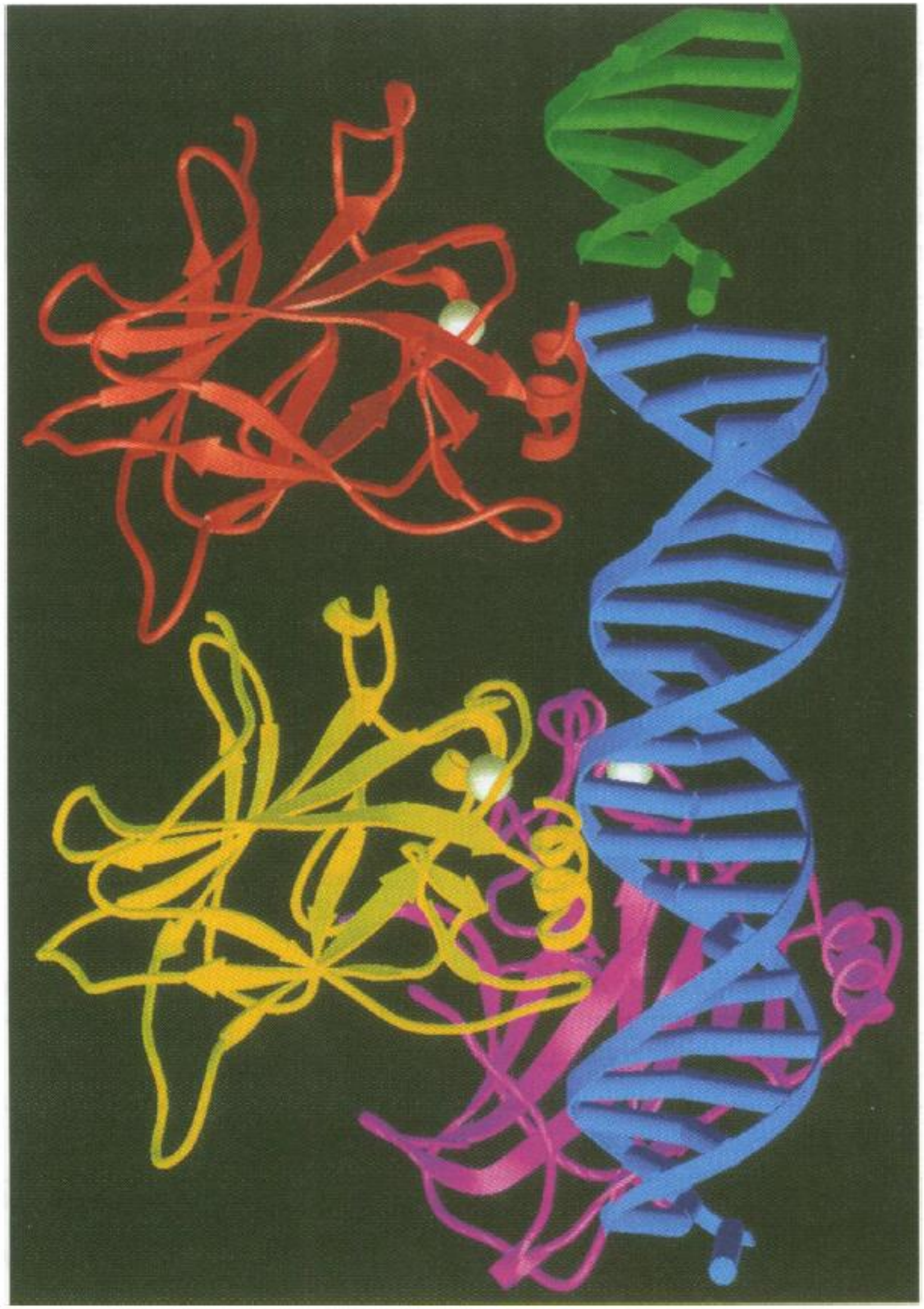
The protective action of the wild-type *p53* gene helps to suppress tumors in humans. The *p53* gene is the most commonly mutated gene in human cancer, and these mutations may actively promote tumor growth.

www.sciencemag.org (1993)

Contacts establish specific binding mode

Nikola Pavletich,
Sloan Kettering
Cancer Center

Fig. 3. Schematic ribbon drawing of the asymmetric unit, which contains three p53 core domain molecules and one DNA duplex. Two of the core domains bind DNA (blue); one (yellow) interacts extensively with a consensus binding site, and the other (red) binds at a nonconsensus site at the interface of DNA fragments related by crystallographic symmetry (a portion of the symmetry-related DNA fragment is shown in green). The third core domain molecule (purple) does not bind DNA, but makes protein-protein contacts stabilizing crystal packing. The zinc atoms are shown as white spheres.



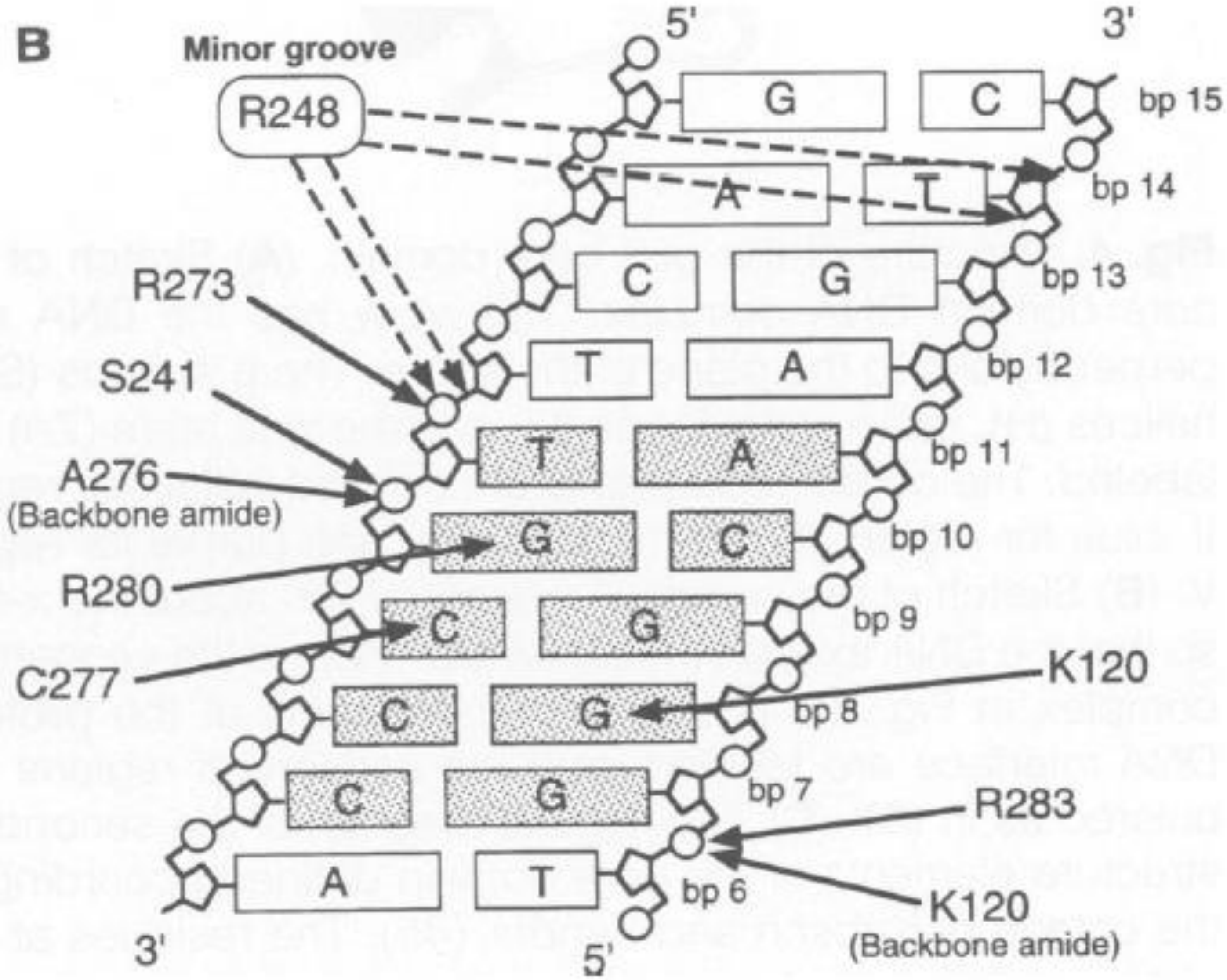
RESEARCH ARTICLE

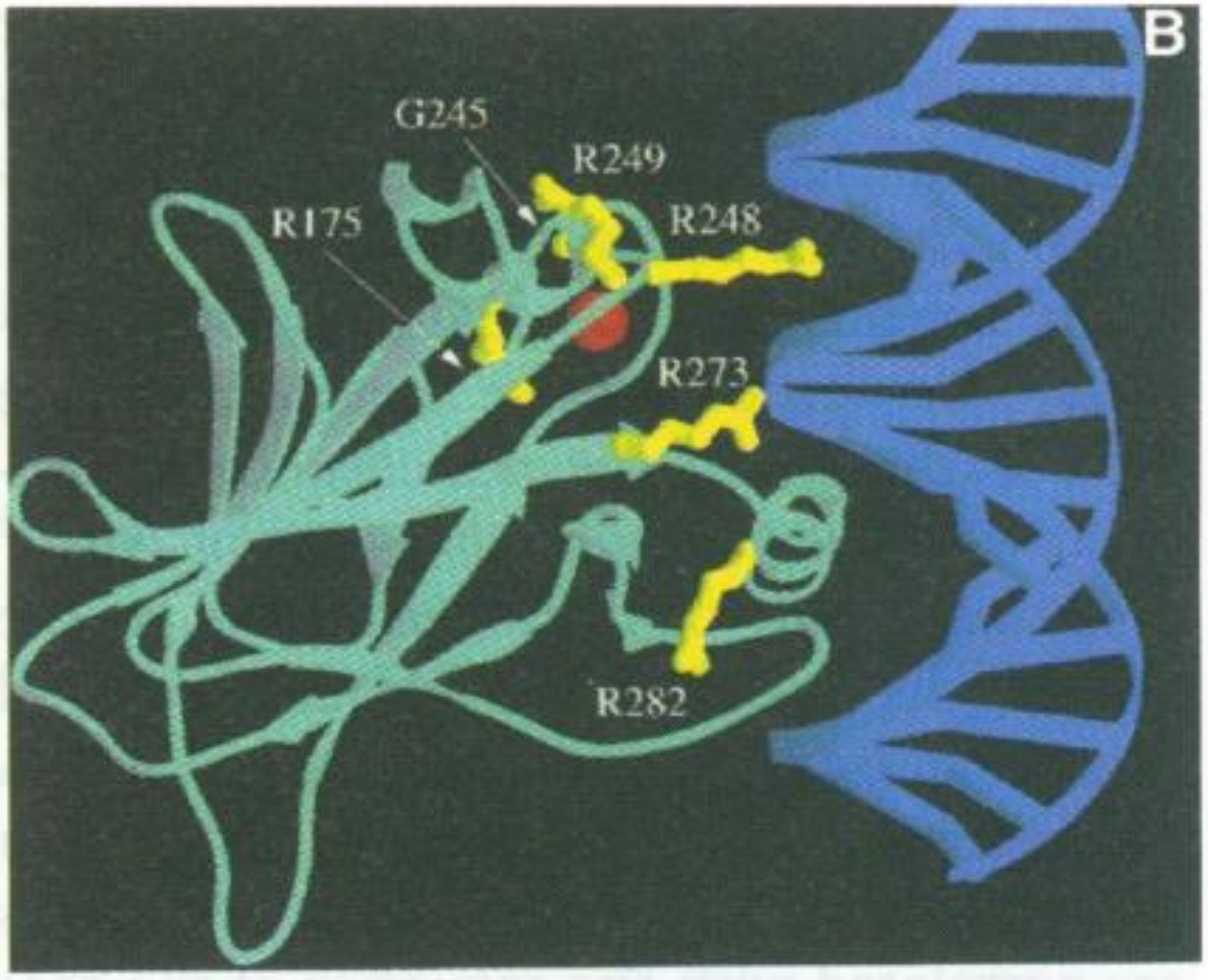
Crystal Structure of a p53 Tumor Suppressor–DNA Complex: Understanding Tumorigenic Mutations

Yunje Cho, Svetlana Gorina, Philip D. Jeffrey, Nikola P. Pavletich

Science 265, 346-355 (1994)

Contact residues





Left: Protein – DNA contacts involve many arginine (R) and lysine (K) residues

Right: the 6 most frequently mutated amino acids (yellow) in cancer.
5 of them are Arginines.

In p53 all 6 residues are located at the binding interface for DNA!

Science 265, 346-355 (1994)

What is a GRN?

Gene regulatory networks (GRN) are model representations of how genes regulate the expression levels of each other.

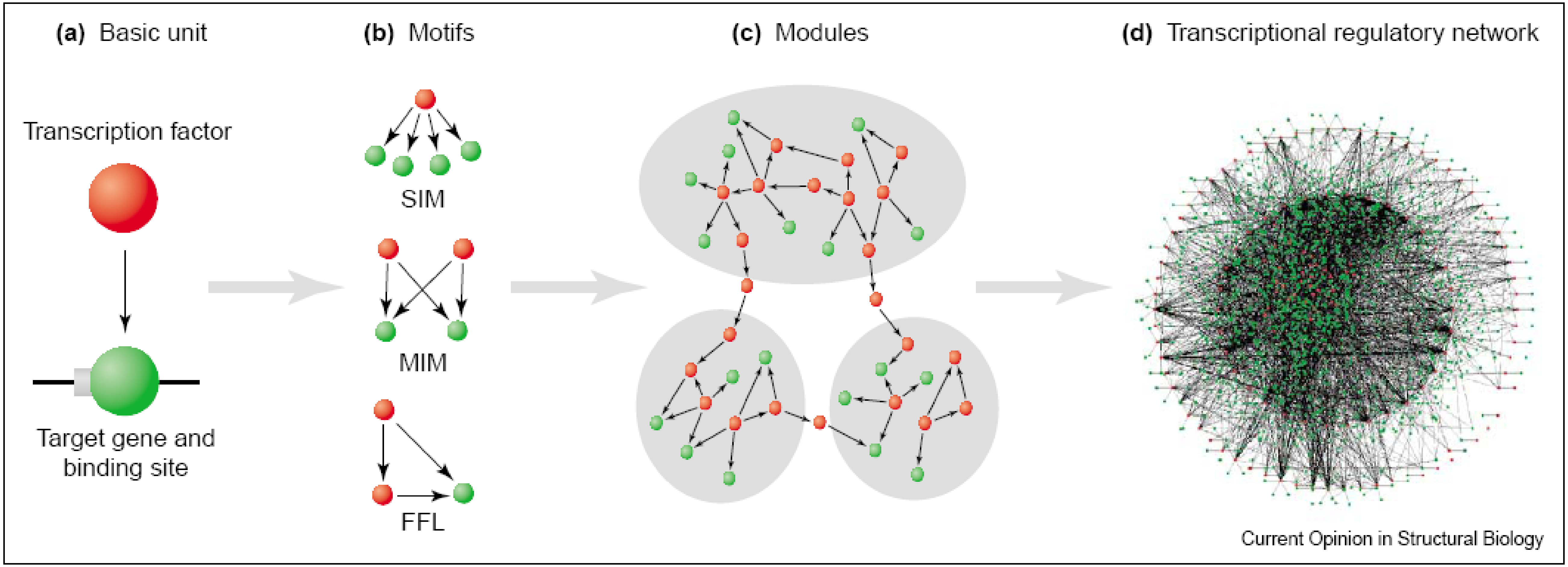
In **transcriptional regulation**, proteins called **transcription factors (TFs)** regulate the transcription of their **target genes** to produce messenger RNA (mRNA).

In **post-transcriptional regulation microRNAs** (miRNAs) cause **degradation** and repression of target mRNAs.

These interactions are represented in a GRN by adding edges linking TF or miRNA genes to their target mRNAs.

Narang et al. (2015). PLoS Comput Biol 11(9): e1004504

Structural organization of transcription/regulatory networks



(a) Basic unit
(b) Motifs
(c) Modules
(d) Transcriptional regulatory network
Transcription factor
Target gene and binding site
SIM
MIM
FFL
Current Opinion in Structural Biology

Regulatory networks are highly interconnected,
very few modules can be entirely separated from the rest of the network.

We will discuss motifs in GRNs in a subsequent lecture.

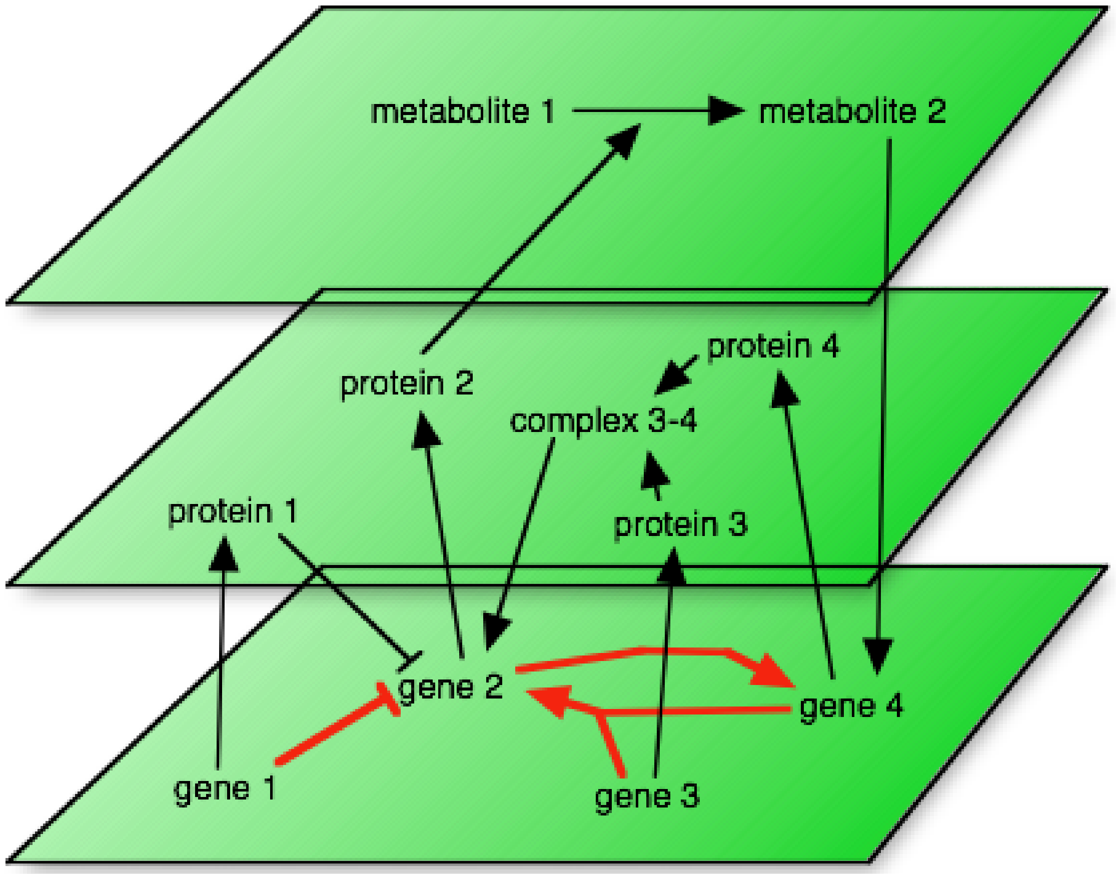
Babu et al. Curr Opin Struct Biol. 14, 283 (2004)

Layers upon Layers

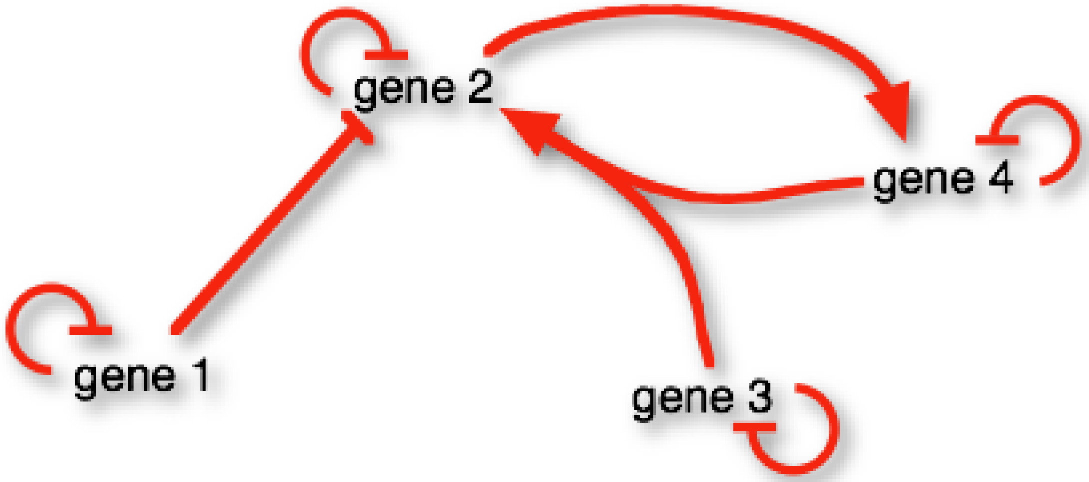
Biological regulation
via proteins and metabolites

<=>

Projected regulatory network



<=>



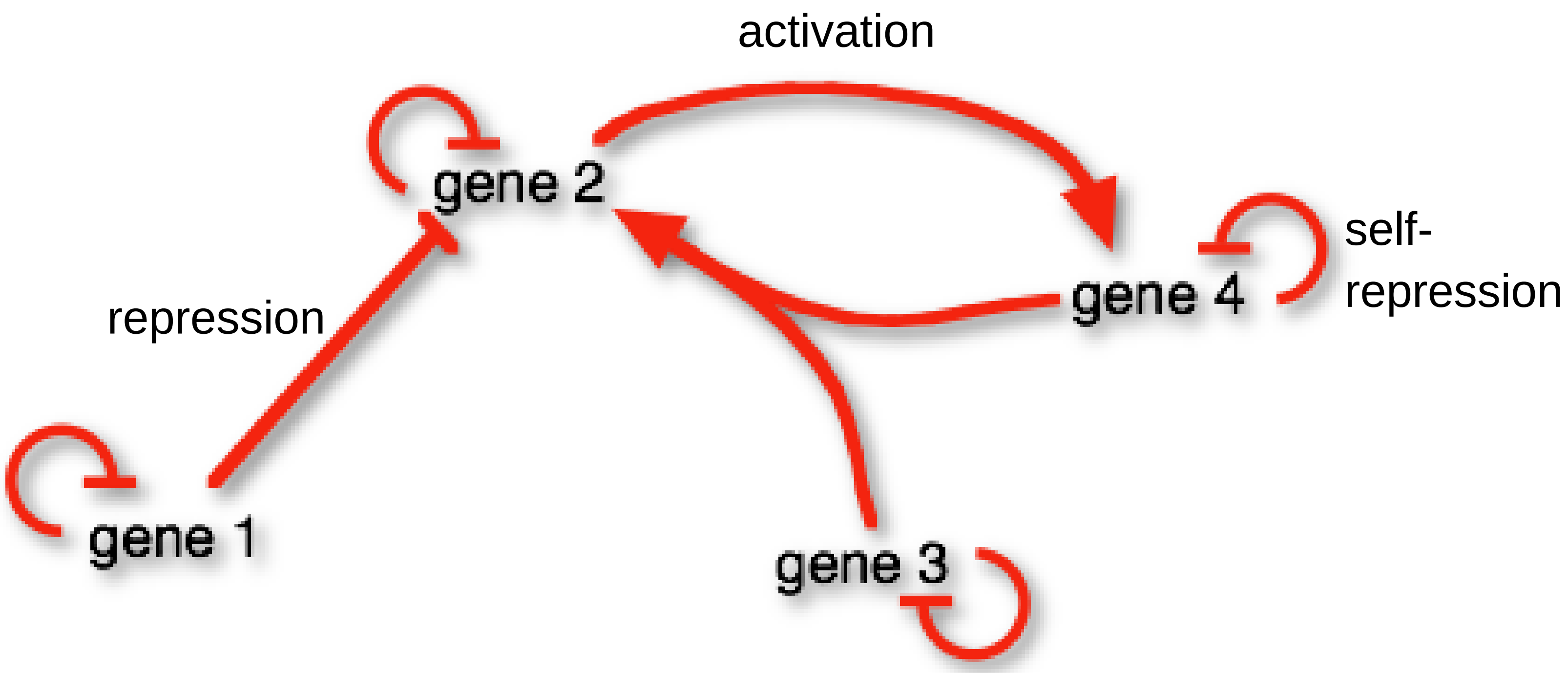
Note that genes do not interact directly

Conventions for GRN Graphs

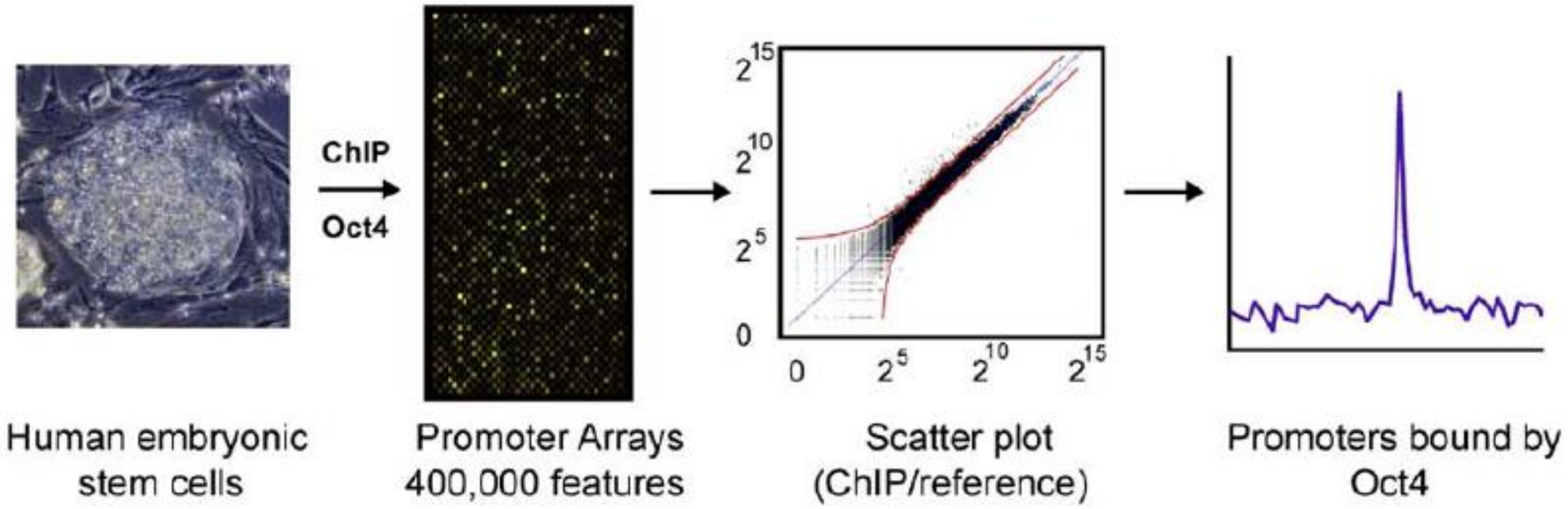
Nodes: genes that code for proteins which catalyze products …
→ everything is projected onto respective gene

Gene regulation networks have "cause and action"
→ **directed** networks

A gene can enhance or suppress the expression of another gene
→ **two types** of arrows



Which TF binds where?


ChIP
Oct4
2^{15}
2^{10}
2^{5}
0
0
2^{5}
2^{10}
2^{15}
Human embryonic stem cells
Promoter Arrays 400,000 features
Scatter plot (ChIP/reference)
Promoters bound by Oct4

Chromatin immuno precipitation: use e.g. antibody against Oct4

➔ ”fish“ all DNA fragments that bind Oct4

➔ sequence DNA fragments bound to Oct4

➔ align them + extract characteristic sequence features

➔ Oct4 binding motif

Boyer et al. Cell 122, 947 (2005)

Sequence logos represent binding motifs

A **logo** represents each column of the alignment by a stack of letters.

The height of each letter is proportional to the **observed frequency** of the corresponding amino acid or nucleotide.

The overall height of each stack is proportional to the **sequence conservation** at that position.

Sequence conservation is defined as difference between the maximum possible entropy and the entropy of the observed symbol distribution:

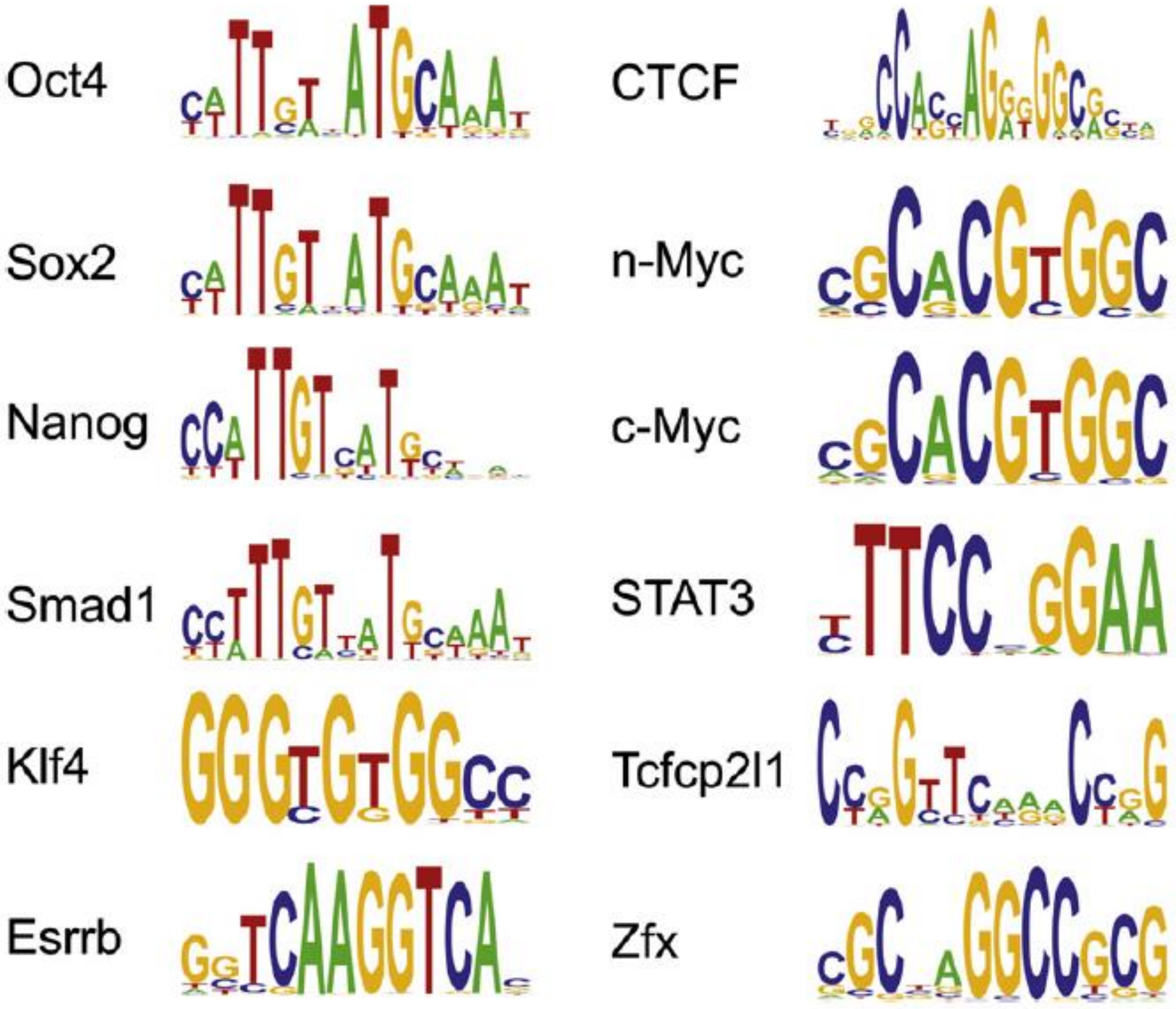
$$R_{seq} = S_{max} - S_{obs} = \log_2 N - \left(- \sum_{n=1}^{N} p_n \log_2 p_n \right)$$

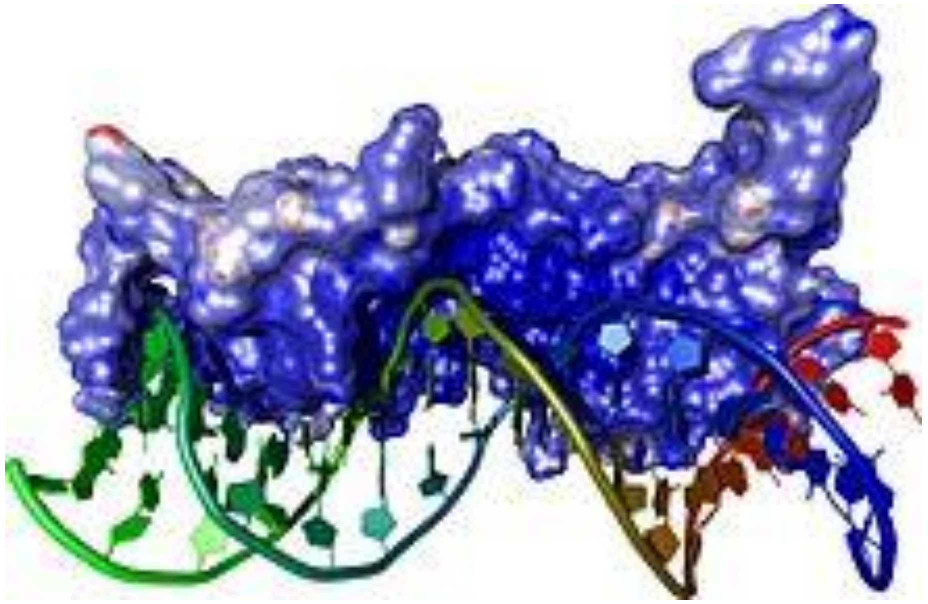
p_n : observed frequency of symbol n at a particular sequence position

N : number of distinct symbols for the given sequence type, either 4 for DNA/RNA or 20 for protein.

Crooks et al., Genome Research
14:1188–1190 (2004)

Construct preferred binding motifs





DNA-binding domain of a glucocorticoid - receptor from *Rattus norvegicus* with the matching DNA fragment ; www.wikipedia.de

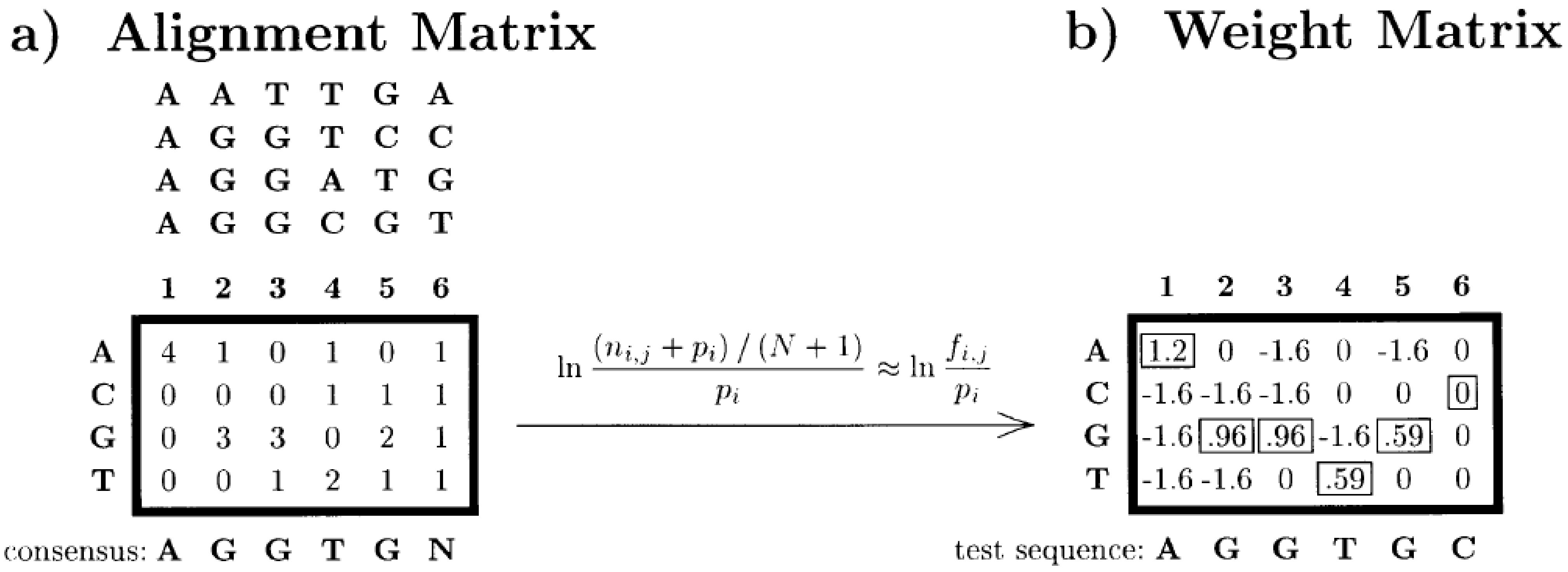
Chen et al., Cell 133, 1106-1117 (2008)

Position specific weight matrix

Build list of genes that share a TF binding motif.

Generate multiple sequence alignment of their sequences.

Alignment matrix: how often does each letter occur at each position in the alignment?



a) Alignment Matrix

A A T T G A
A G G T C C
A G G A T G
A G G C G T

	1	2	3	4	5	6
A	4	1	0	1	0	1
C	0	0	0	1	1	1
G	0	3	3	0	2	1
T	0	0	1	2	1	1

consensus: A G G T G N

$$\xrightarrow{\ln \frac{(n_{i,j}+p_i)\,/\,(N+1)}{p_i} \approx \ln \frac{f_{i,j}}{p_i}}$$

b) Weight Matrix

	1	2	3	4	5	6
A	$\boxed{1.2}$	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	$\boxed{0}$
G	-1.6	$\boxed{.96}$	$\boxed{.96}$	-1.6	$\boxed{.59}$	0
T	-1.6	-1.6	0	$\boxed{.59}$	0	0

test sequence: A G G T G C

Fig. 1. Examples of the simple matrix model for summarizing a DNA alignment. (**a**) An alignment matrix describing the alignment of the four 6-mers on top. The matrix contains the number of times, $n_{i,j}$, that letter i is observed at position j of this alignment. Below the matrix is the consensus sequence corresponding to the alignment (N indicates that there is no nucleotide preference). (**b**) A weight matrix derived from the alignment in (a). The formula used for transforming the alignment matrix to a weight matrix is shown above the arrow. In this formula, N is the total number of sequences (four in this example), p_i is the *a priori* probability of letter i (0.25 for all the bases in this example) and $f_{i,j} = n_{i,j}/N$ is the frequency of letter i at position j. The numbers enclosed in blocks are summed to give the overall score of the test sequence. The overall score is 4.3, which is also the maximum possible score with this weight matrix.

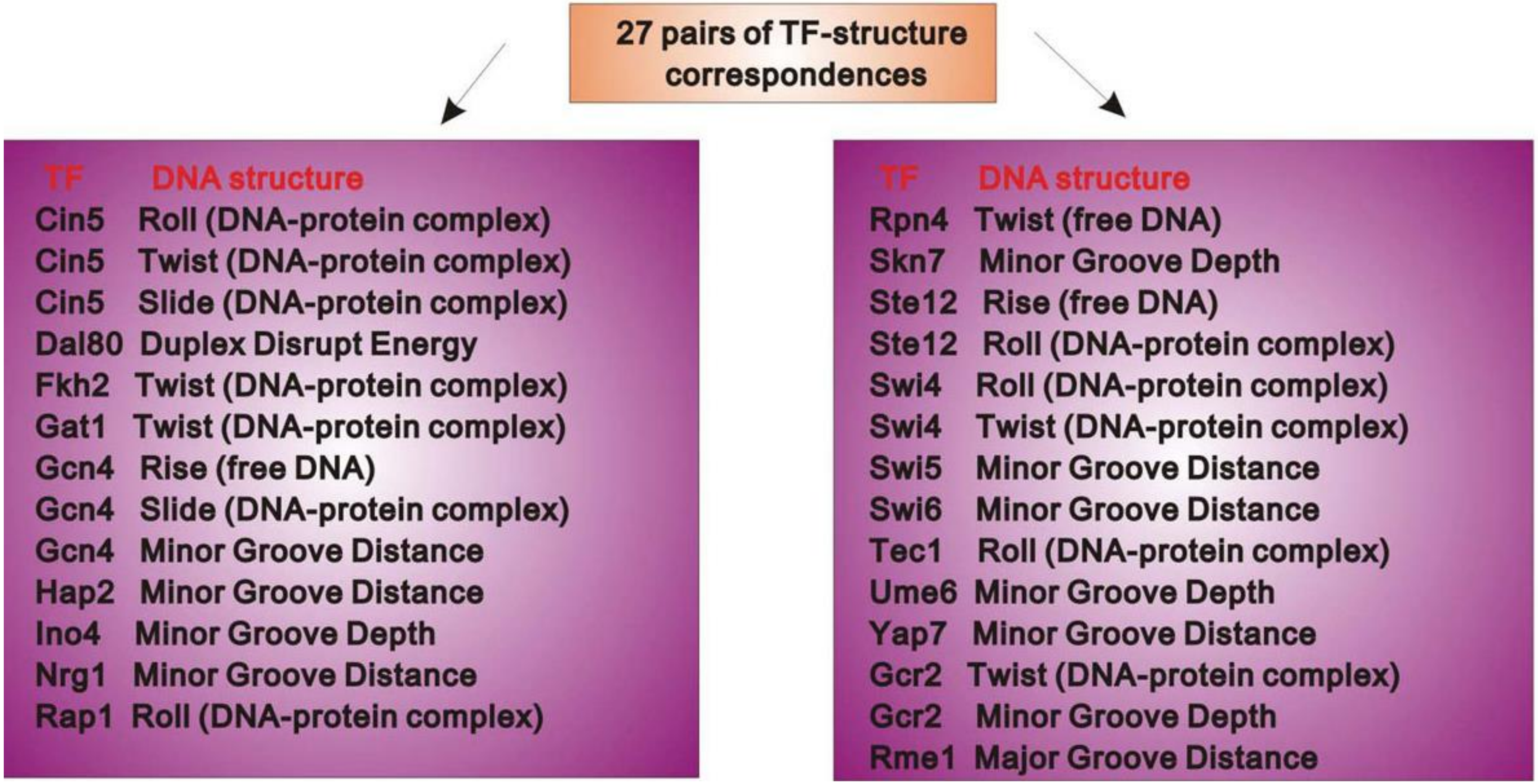
Hertz, Stormo (1999) Bioinformatics 15, 563

What do TFs recognize?

(1) Amino acids of TFs make specific contacts (e.g. hydrogen bonds) with DNA base pairs

(2) DNA conformation depends on its sequence

→ Some TFs „measure“ different aspects of the DNA conformation



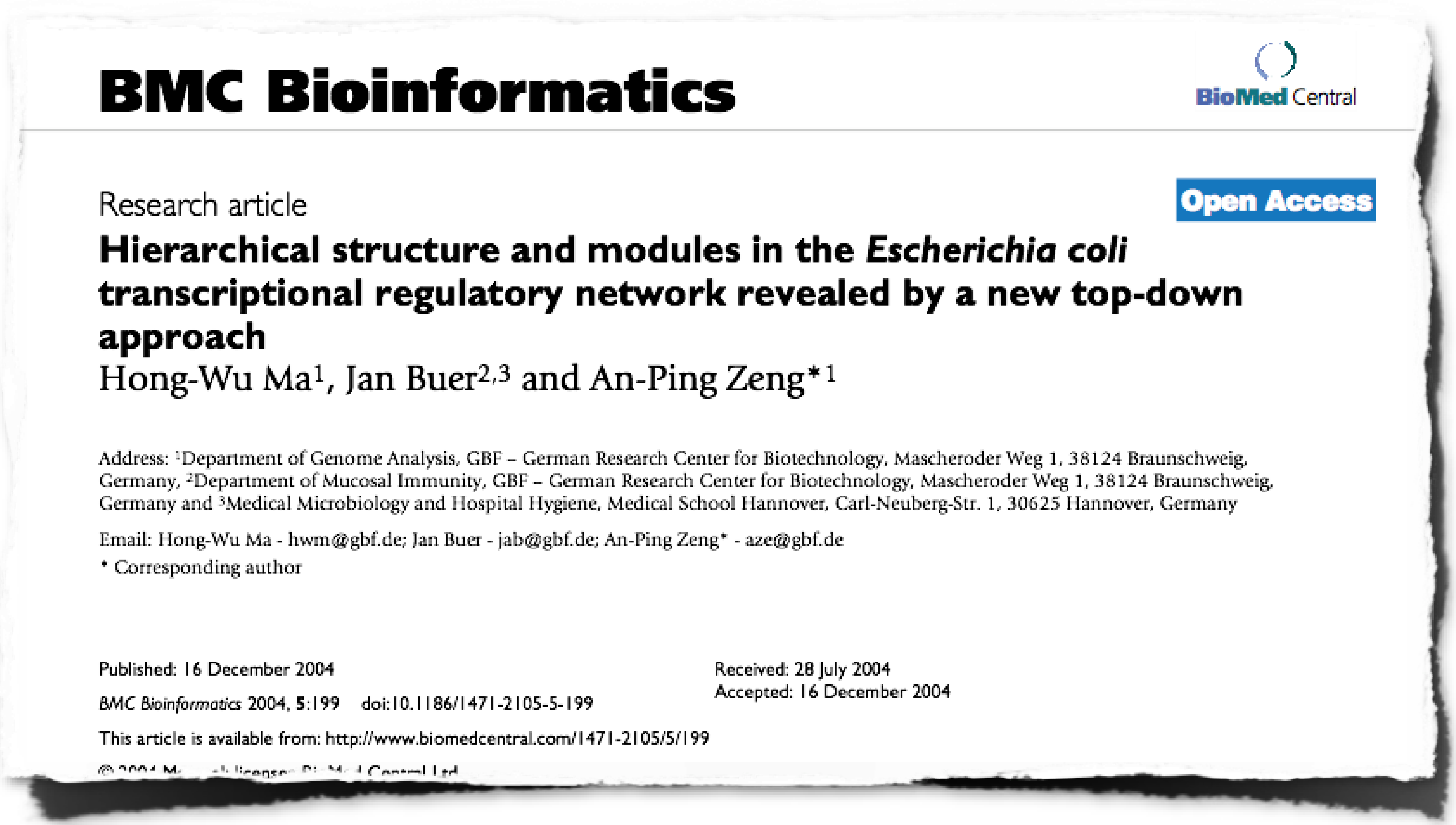
27 pairs of TF-structure correspondences

TF	DNA structure
Cin5	Roll (DNA-protein complex)
Cin5	Twist (DNA-protein complex)
Cin5	Slide (DNA-protein complex)
Dal80	Duplex Disrupt Energy
Fkh2	Twist (DNA-protein complex)
Gat1	Twist (DNA-protein complex)
Gcn4	Rise (free DNA)
Gcn4	Slide (DNA-protein complex)
Gcn4	Minor Groove Distance
Hap2	Minor Groove Distance
Ino4	Minor Groove Depth
Nrg1	Minor Groove Distance
Rap1	Roll (DNA-protein complex)

TF	DNA structure
Rpn4	Twist (free DNA)
Skn7	Minor Groove Depth
Ste12	Rise (free DNA)
Ste12	Roll (DNA-protein complex)
Swi4	Roll (DNA-protein complex)
Swi4	Twist (DNA-protein complex)
Swi5	Minor Groove Distance
Swi6	Minor Groove Distance
Tec1	Roll (DNA-protein complex)
Ume6	Minor Groove Depth
Yap7	Minor Groove Distance
Gcr2	Twist (DNA-protein complex)
Gcr2	Minor Groove Depth
Rme1	Major Groove Distance

Dai et al. *BMC Genomics* 2015, **16**(Suppl 3):S8

E. coli Regulatory Network



BMC Bioinformatics

BioMed Central

Research article

Open Access

Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach

Hong-Wu Ma[1], Jan Buer[2,3] and An-Ping Zeng*[1]

Address: [1]Department of Genome Analysis, GBF – German Research Center for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany, [2]Department of Mucosal Immunity, GBF – German Research Center for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany and [3]Medical Microbiology and Hospital Hygiene, Medical School Hannover, Carl-Neuberg-Str. 1, 30625 Hannover, Germany

Email: Hong-Wu Ma - hwm@gbf.de; Jan Buer - jab@gbf.de; An-Ping Zeng* - aze@gbf.de

* Corresponding author

Published: 16 December 2004

BMC Bioinformatics 2004, **5**:199 doi:10.1186/1471-2105-5-199

Received: 28 July 2004
Accepted: 16 December 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/199

BMC Bioinformatics **5** (2004) 199

Global Regulators in *E. coli*

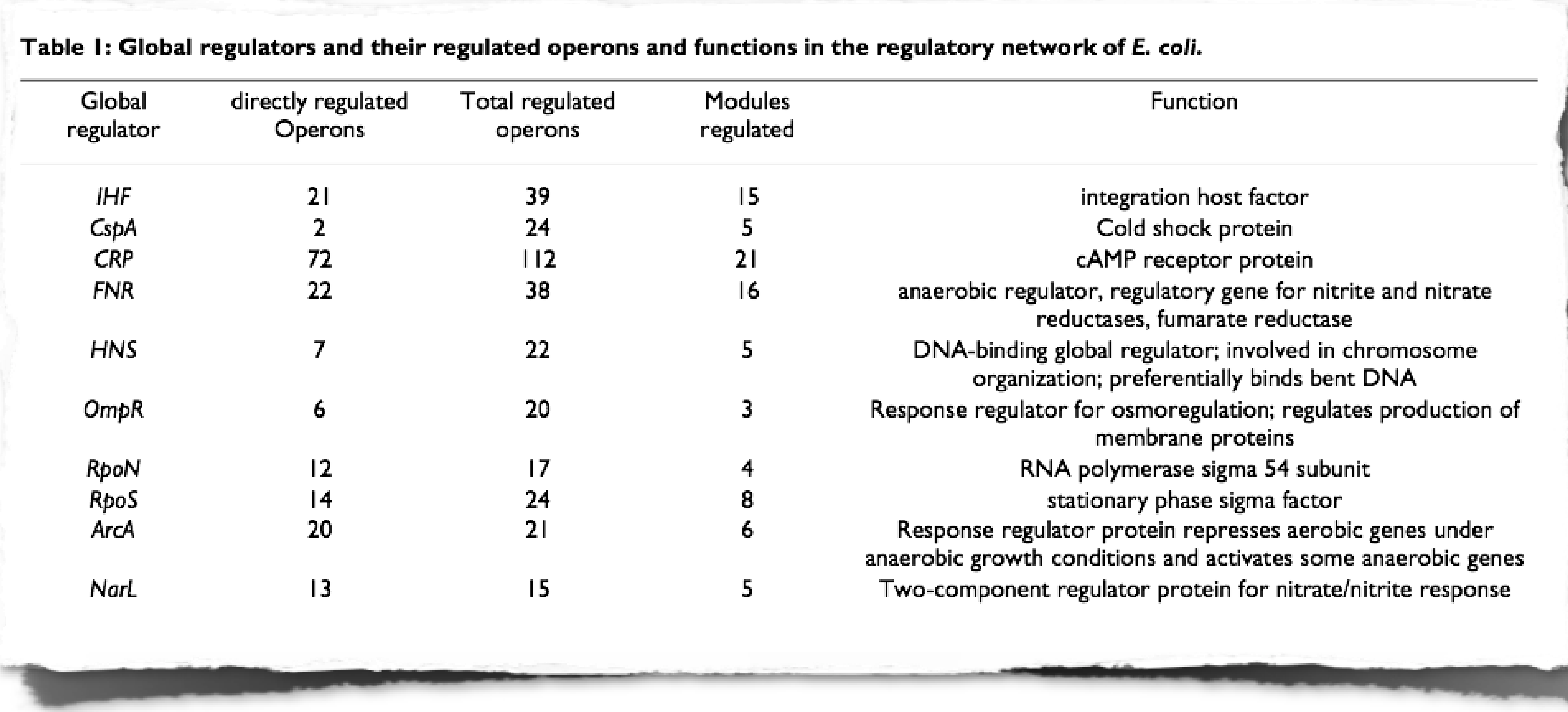


Table 1: Global regulators and their regulated operons and functions in the regulatory network of *E. coli*.

Global regulator	directly regulated Operons	Total regulated operons	Modules regulated	Function
IHF	21	39	15	integration host factor
CspA	2	24	5	Cold shock protein
CRP	72	112	21	cAMP receptor protein
FNR	22	38	16	anaerobic regulator, regulatory gene for nitrite and nitrate reductases, fumarate reductase
HNS	7	22	5	DNA-binding global regulator; involved in chromosome organization; preferentially binds bent DNA
OmpR	6	20	3	Response regulator for osmoregulation; regulates production of membrane proteins
RpoN	12	17	4	RNA polymerase sigma 54 subunit
RpoS	14	24	8	stationary phase sigma factor
ArcA	20	21	6	Response regulator protein represses aerobic genes under anaerobic growth conditions and activates some anaerobic genes
NarL	13	15	5	Two-component regulator protein for nitrate/nitrite response

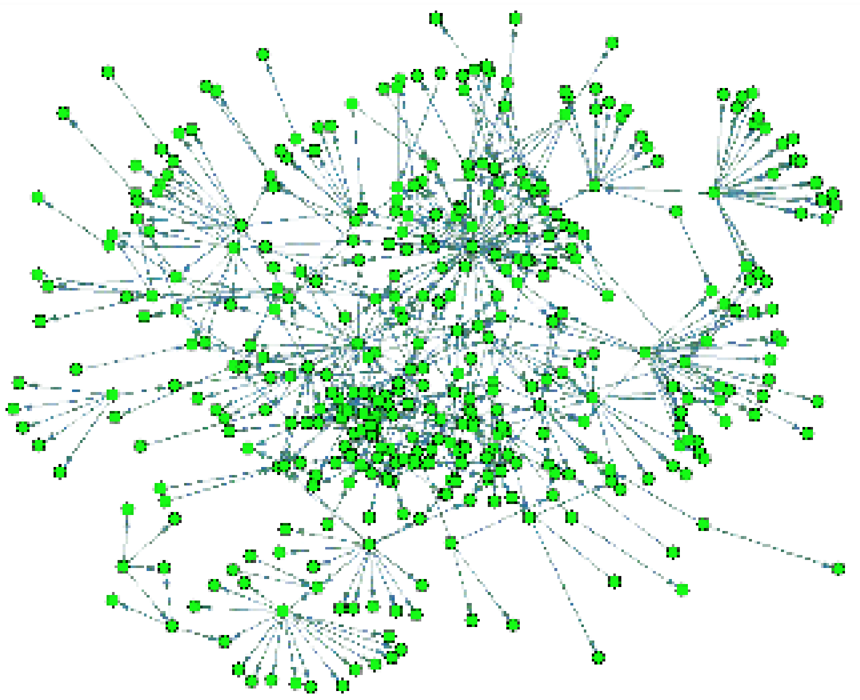
Ma et al., *BMC Bioinformatics* **5** (2004) 199

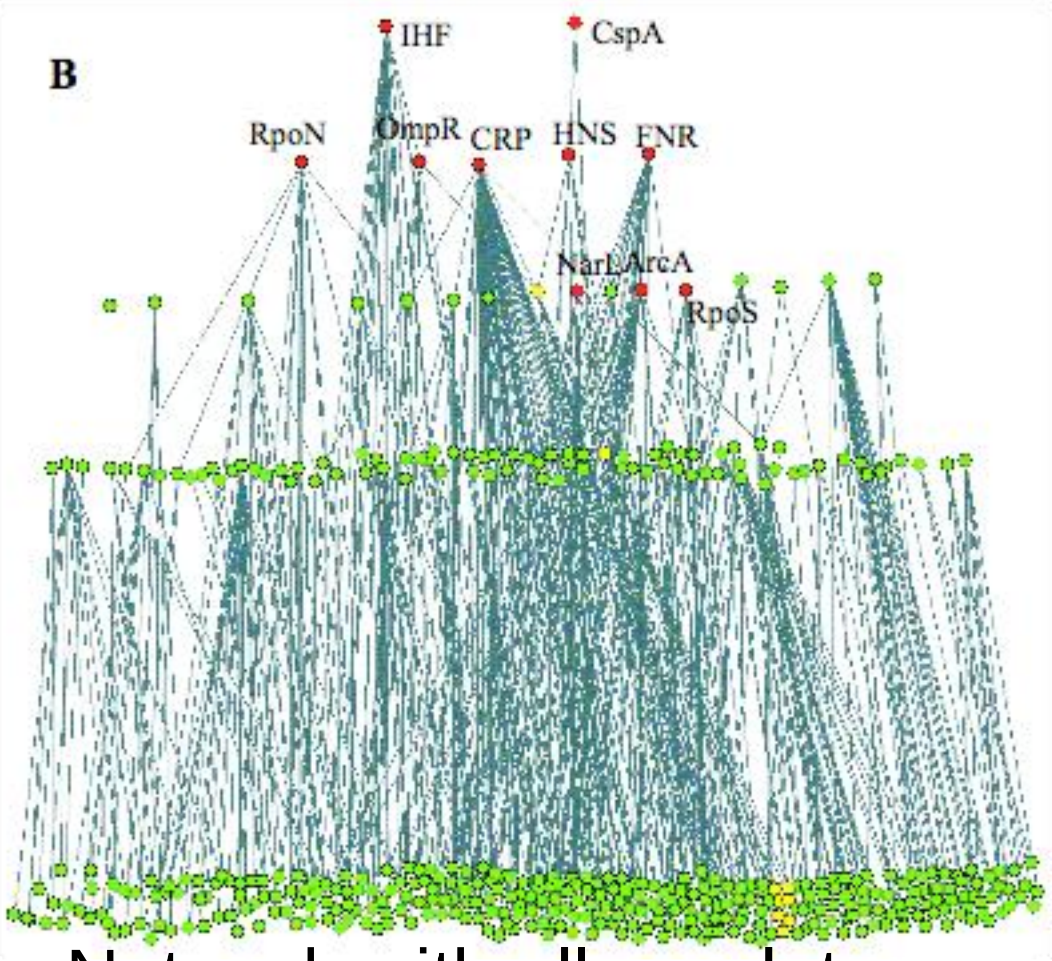
Simple organisms have hierarchical GRNs

Largest weakly connected component (WCC)
(ignore directions of regulation):
325 operons
(3/4 of the complete network)

Lowest level: operons that code for TFs with only auto-regulation, or no TFs

Next layer: delete nodes of lower layer, identify TFs that do not regulate other operons in this layer (only lower layers)
Continue …





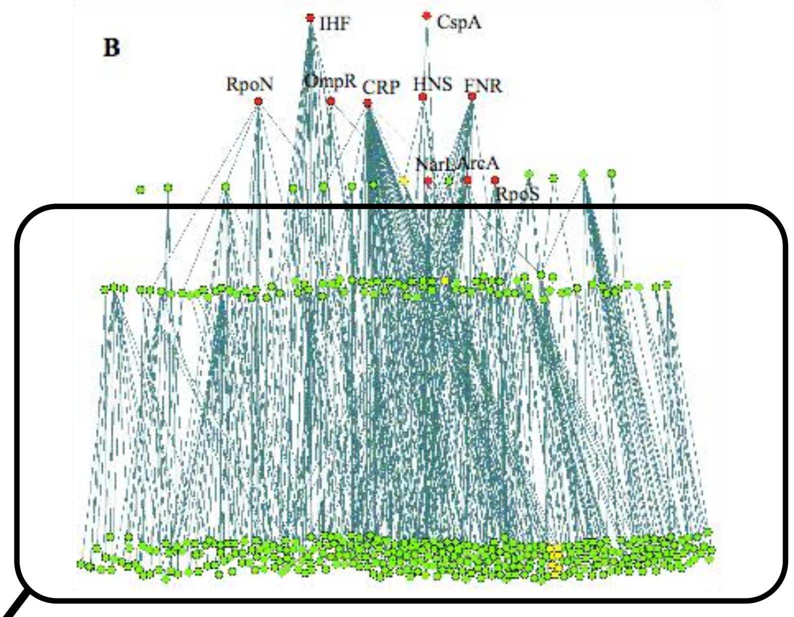
Network from standard layout algorithm → Network with all regulatory edges pointing downwards

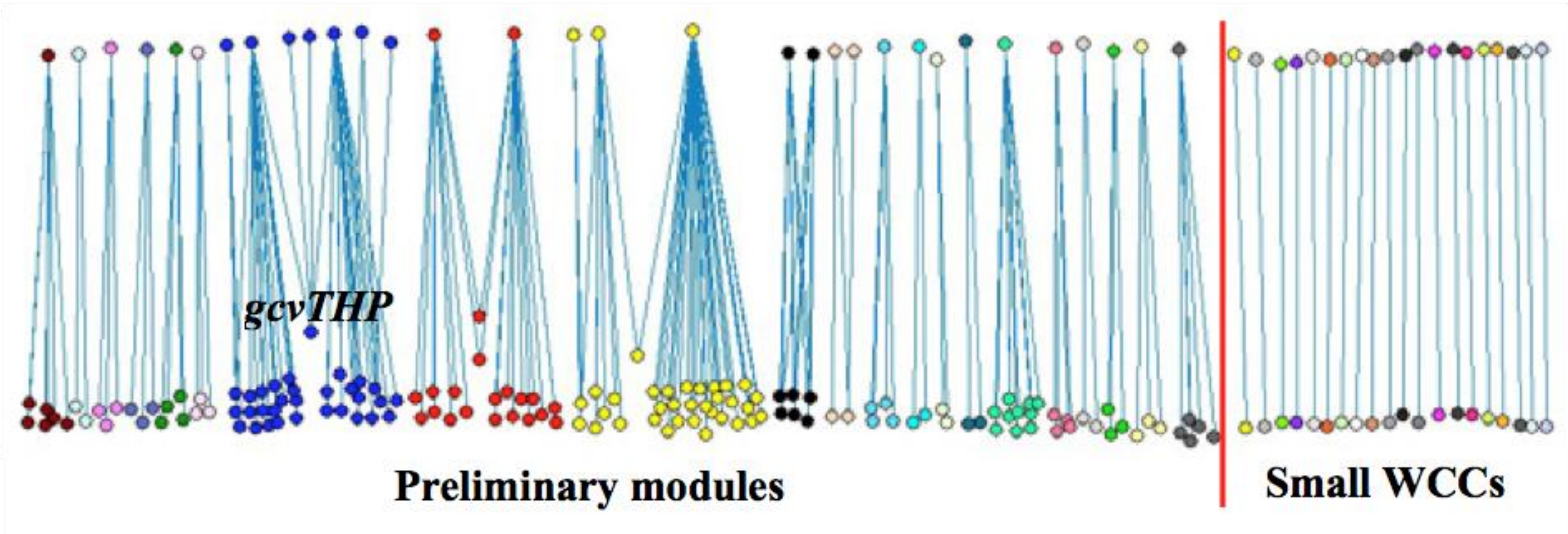
→ a few global regulators (•) control all the details

Ma et al., *BMC Bioinformatics* **5** (2004) 199

E.coli GRN modules

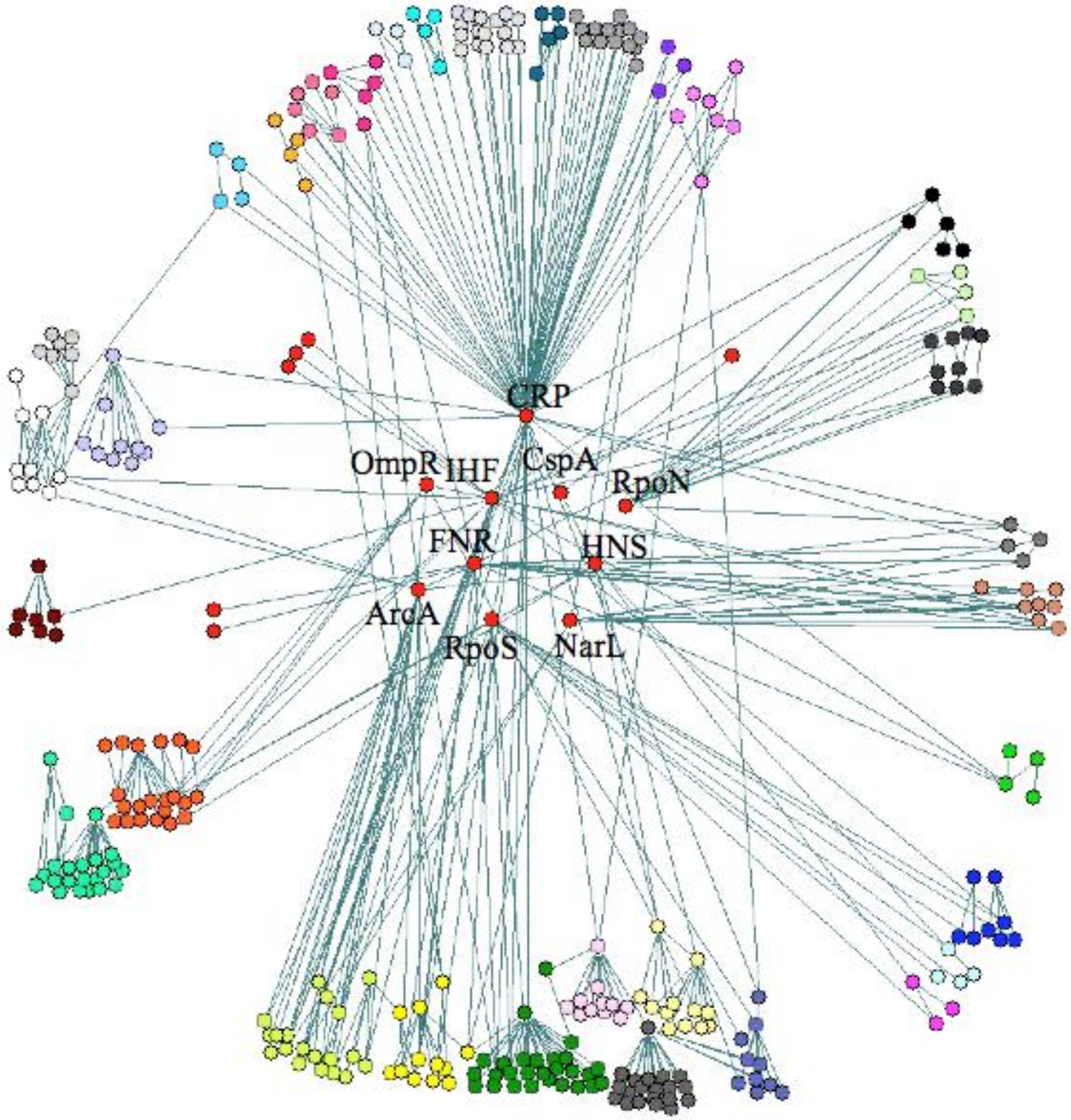
Remove top 3 layers and determine WCCs
→ just a few modules





Ma et al., *BMC Bioinformatics* **5** (2004) 199

Putting it back together



The 10 global regulators are at the core of the network,

some hierarchies exist between the modules

Ma et al., *BMC Bioinformatics* **5** (2004) 199

Modules have specific functions

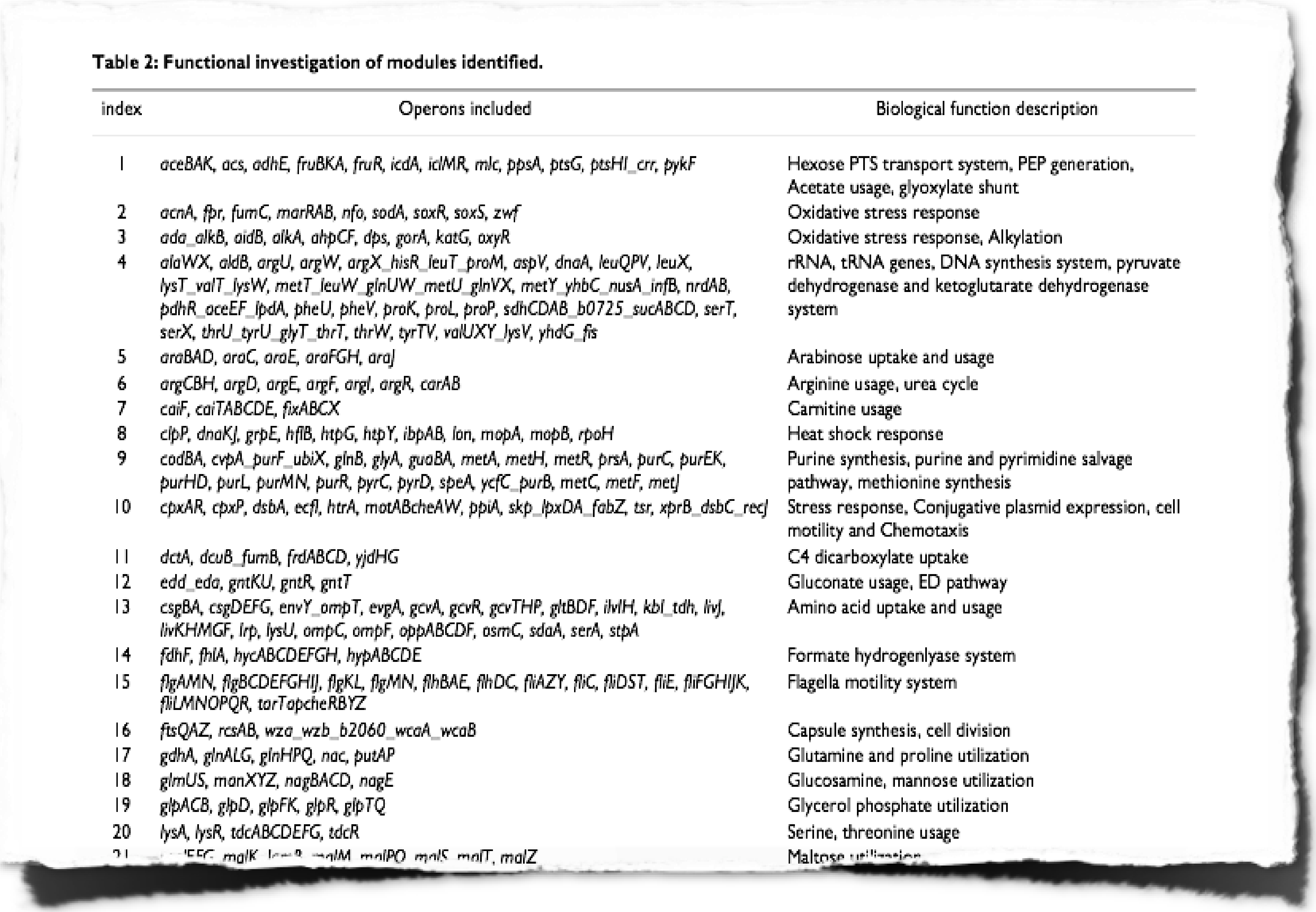


Table 2: Functional investigation of modules identified.

index	Operons included	Biological function description
1	aceBAK, acs, adhE, fruBKA, fruR, icdA, iclMR, mlc, ppsA, ptsG, ptsHI_crr, pykF	Hexose PTS transport system, PEP generation, Acetate usage, glyoxylate shunt
2	acnA, fpr, fumC, marRAB, nfo, sodA, soxR, soxS, zwf	Oxidative stress response
3	ada_alkB, aidB, alkA, ahpCF, dps, gorA, katG, oxyR	Oxidative stress response, Alkylation
4	alaWX, aldB, argU, argW, argX_hisR_leuT_proM, aspV, dnaA, leuQPV, leuX, lysT_valT_lysW, metT_leuW_glnUW_metU_glnVX, metY_yhbC_nusA_infB, nrdAB, pdhR_aceEF_lpdA, pheU, pheV, proK, proL, proP, sdhCDAB_b0725_sucABCD, serT, serX, thrU_tyrU_glyT_thrT, thrW, tyrTV, valUXY_lysV, yhdG_fis	rRNA, tRNA genes, DNA synthesis system, pyruvate dehydrogenase and ketoglutarate dehydrogenase system
5	araBAD, araC, araE, araFGH, araJ	Arabinose uptake and usage
6	argCBH, argD, argE, argF, argI, argR, carAB	Arginine usage, urea cycle
7	caiF, caiTABCDE, fixABCX	Carnitine usage
8	clpP, dnaKJ, grpE, hflB, htpG, htpY, ibpAB, lon, mopA, mopB, rpoH	Heat shock response
9	codBA, cvpA_purF_ubiX, glnB, glyA, guaBA, metA, metH, metR, prsA, purC, purEK, purHD, purL, purMN, purR, pyrC, pyrD, speA, ycfC_purB, metC, metF, metJ	Purine synthesis, purine and pyrimidine salvage pathway, methionine synthesis
10	cpxAR, cpxP, dsbA, ecfI, htrA, motABcheAW, ppiA, skp_lpxDA_fabZ, tsr, xprB_dsbC_recJ	Stress response, Conjugative plasmid expression, cell motility and Chemotaxis
11	dctA, dcuB_fumB, frdABCD, yjdHG	C4 dicarboxylate uptake
12	edd_eda, gntKU, gntR, gntT	Gluconate usage, ED pathway
13	csgBA, csgDEFG, envY_ompT, evgA, gcvA, gcvR, gcvTHP, gltBDF, ilvIH, kbl_tdh, livJ, livKHMGF, lrp, lysU, ompC, ompF, oppABCDF, osmC, sdaA, serA, stpA	Amino acid uptake and usage
14	fdhF, fhlA, hycABCDEFGH, hypABCDE	Formate hydrogenlyase system
15	flgAMN, flgBCDEFGHIJ, flgKL, flgMN, flhBAE, flhDC, fliAZY, fliC, fliDST, fliE, fliFGHIJK, fliLMNOPQR, tarTapcheRBYZ	Flagella motility system
16	ftsQAZ, rcsAB, wza_wzb_b2060_wcaA_wcaB	Capsule synthesis, cell division
17	gdhA, glnALG, glnHPQ, nac, putAP	Glutamine and proline utilization
18	glmUS, manXYZ, nagBACD, nagE	Glucosamine, mannose utilization
19	glpACB, glpD, glpFK, glpR, glpTQ	Glycerol phosphate utilization
20	lysA, lysR, tdcABCDEFG, tdcR	Serine, threonine usage
21	malEFG, malK_lamB_malM, malPQ, malS, malT, malZ	Maltose utilization

Ma et al., *BMC Bioinformatics* **5** (2004) 199

Frequency of co-regulation

Martinez-Antonio,
Collado-Vides,
Curr Opin Microbiol
6, 482 (2003)

Half of all target genes are regulated by multiple TFs.

In most cases, a „gobal“ regulator (with > 10 interactions) works together with a more specific local regulator.

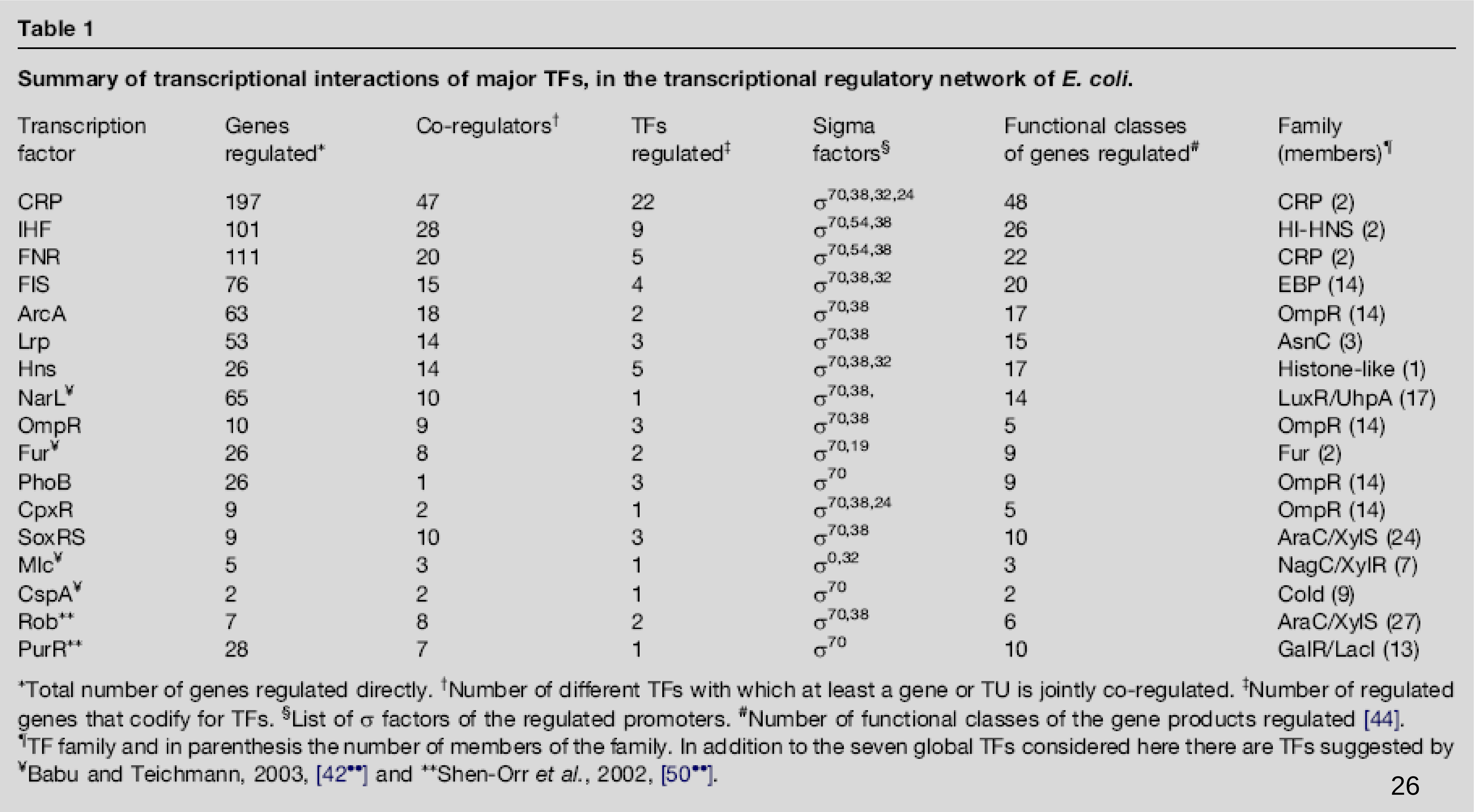


Table 1

Summary of transcriptional interactions of major TFs, in the transcriptional regulatory network of *E. coli*.

Transcription factor	Genes regulated*	Co-regulators†	TFs regulated‡	Sigma factors§	Functional classes of genes regulated#	Family (members)¶
CRP	197	47	22	$\sigma^{70,38,32,24}$	48	CRP (2)
IHF	101	28	9	$\sigma^{70,54,38}$	26	HI-HNS (2)
FNR	111	20	5	$\sigma^{70,54,38}$	22	CRP (2)
FIS	76	15	4	$\sigma^{70,38,32}$	20	EBP (14)
ArcA	63	18	2	$\sigma^{70,38}$	17	OmpR (14)
Lrp	53	14	3	$\sigma^{70,38}$	15	AsnC (3)
Hns	26	14	5	$\sigma^{70,38,32}$	17	Histone-like (1)
NarL¥	65	10	1	$\sigma^{70,38,}$	14	LuxR/UhpA (17)
OmpR	10	9	3	$\sigma^{70,38}$	5	OmpR (14)
Fur¥	26	8	2	$\sigma^{70,19}$	9	Fur (2)
PhoB	26	1	3	σ^{70}	9	OmpR (14)
CpxR	9	2	1	$\sigma^{70,38,24}$	5	OmpR (14)
SoxRS	9	10	3	$\sigma^{70,38}$	10	AraC/XylS (24)
Mlc¥	5	3	1	$\sigma^{0,32}$	3	NagC/XylR (7)
CspA¥	2	2	1	σ^{70}	2	Cold (9)
Rob**	7	8	2	$\sigma^{70,38}$	6	AraC/XylS (27)
PurR**	28	7	1	σ^{70}	10	GalR/LacI (13)

*Total number of genes regulated directly. †Number of different TFs with which at least a gene or TU is jointly co-regulated. ‡Number of regulated genes that codify for TFs. §List of σ factors of the regulated promoters. #Number of functional classes of the gene products regulated [44]. ¶TF family and in parenthesis the number of members of the family. In addition to the seven global TFs considered here there are TFs suggested by ¥Babu and Teichmann, 2003, [42••] and **Shen-Orr *et al.*, 2002, [50••].

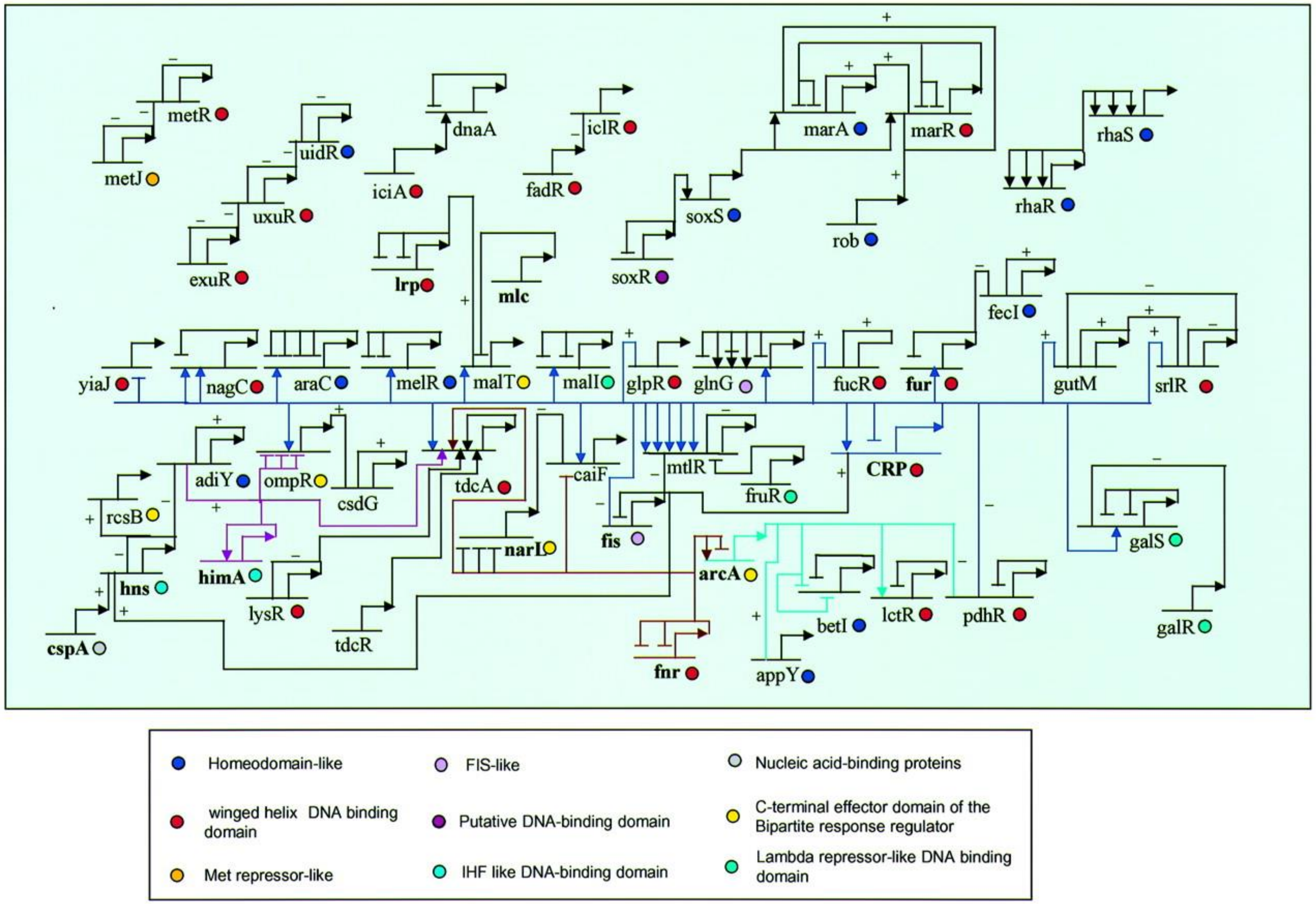
TF regulatory network in *E.coli*

When more than one TF regulates a gene, the order of their binding sites is as given in the figure.

Arrowheads and **horizontal bars** indicate positive / negative regulation when the position of the binding site is known.

In cases where only the nature of regulation is known, without binding site information, + and – are used to indicate positive and negative regulation.

Regulation of transcription factors in E. coli

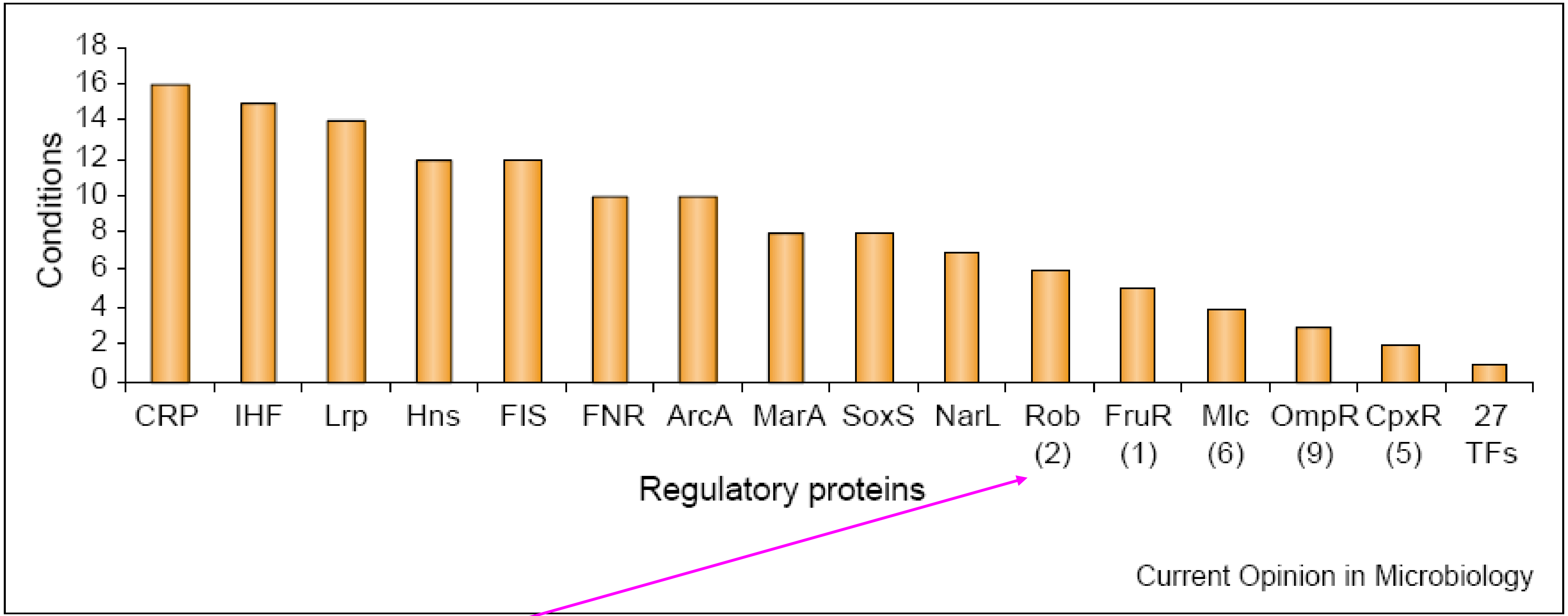


The names of **global regulators** are in **bold**.

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

Response to changes in environmental conditions

TFs also sense changes in environmental conditions or other internal signals encoding changes.



Current Opinion in Microbiology

Global environment growth conditions in which TFs are regulating.
# in brackets indicates how many additional TFs participate in the same number of conditions.

Martinez-Antonio, Collado-Vides, Curr Opin Microbiol 6, 482 (2003)

Structural view at *E. coli* TFs

Determine homology between the domains and protein families of TFs and regulated genes and proteins of known 3D structure.

→ Determine uncharacterized *E.coli* proteins with DNA-binding domains (DBD)

→ identify large majority of *E.coli* TFs.

Sarah Teichmann
EBI

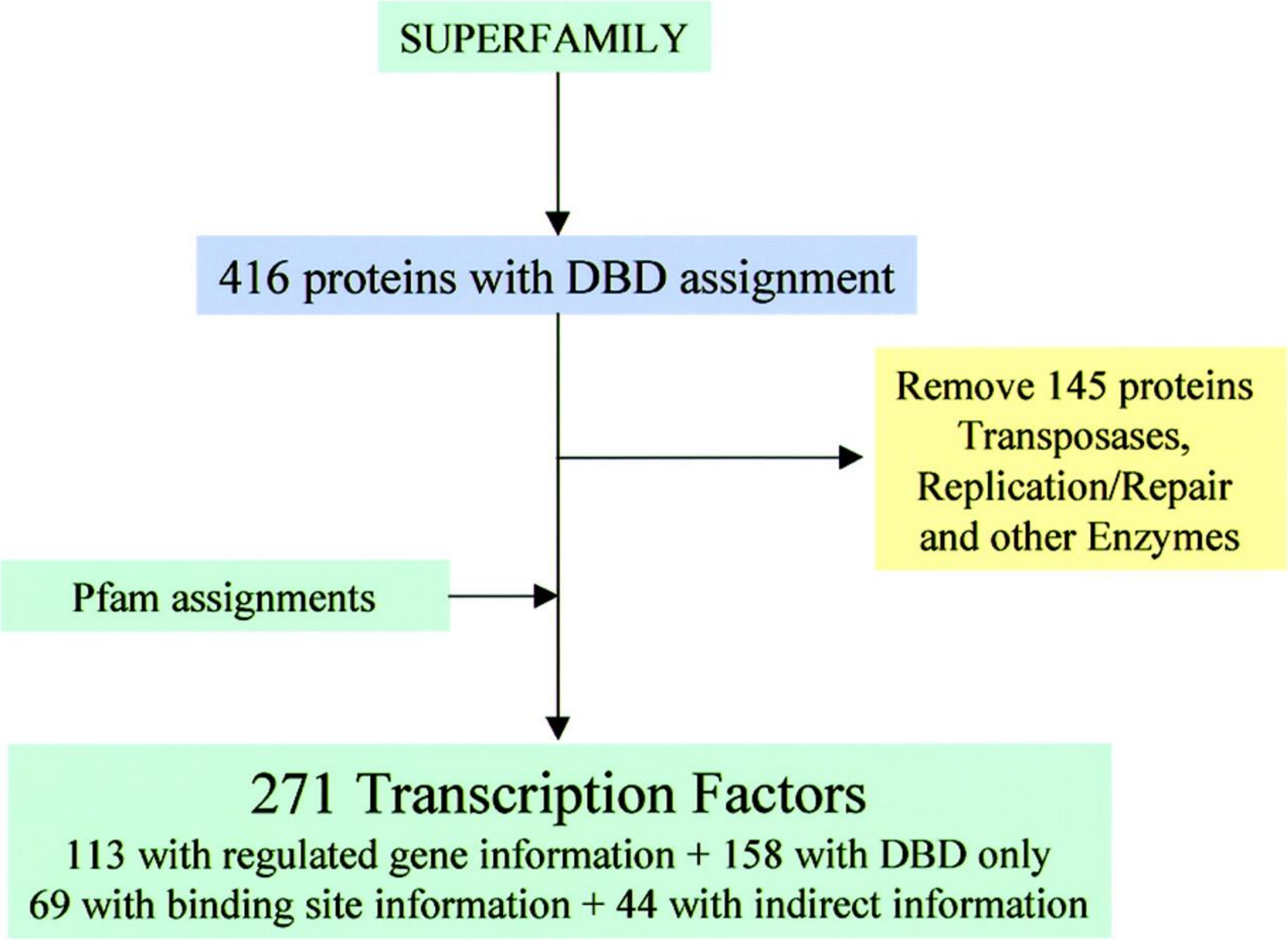
Madan Babu,
MRC

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

Flow chart of method to identify TFs in E.coli

SUPERFAMILY database (C. Chothia) contains a library of HMM models based on the sequences of proteins in SCOP for predicted proteins of completely sequenced genomes.

Remove all DNA-binding proteins involved in replication/repair etc.


SUPERFAMILY
416 proteins with DBD assignment
Remove 145 proteins Transposases, Replication/Repair and other Enzymes
Pfam assignments
271 Transcription Factors
113 with regulated gene information + 158 with DBD only
69 with binding site information + 44 with indirect information

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

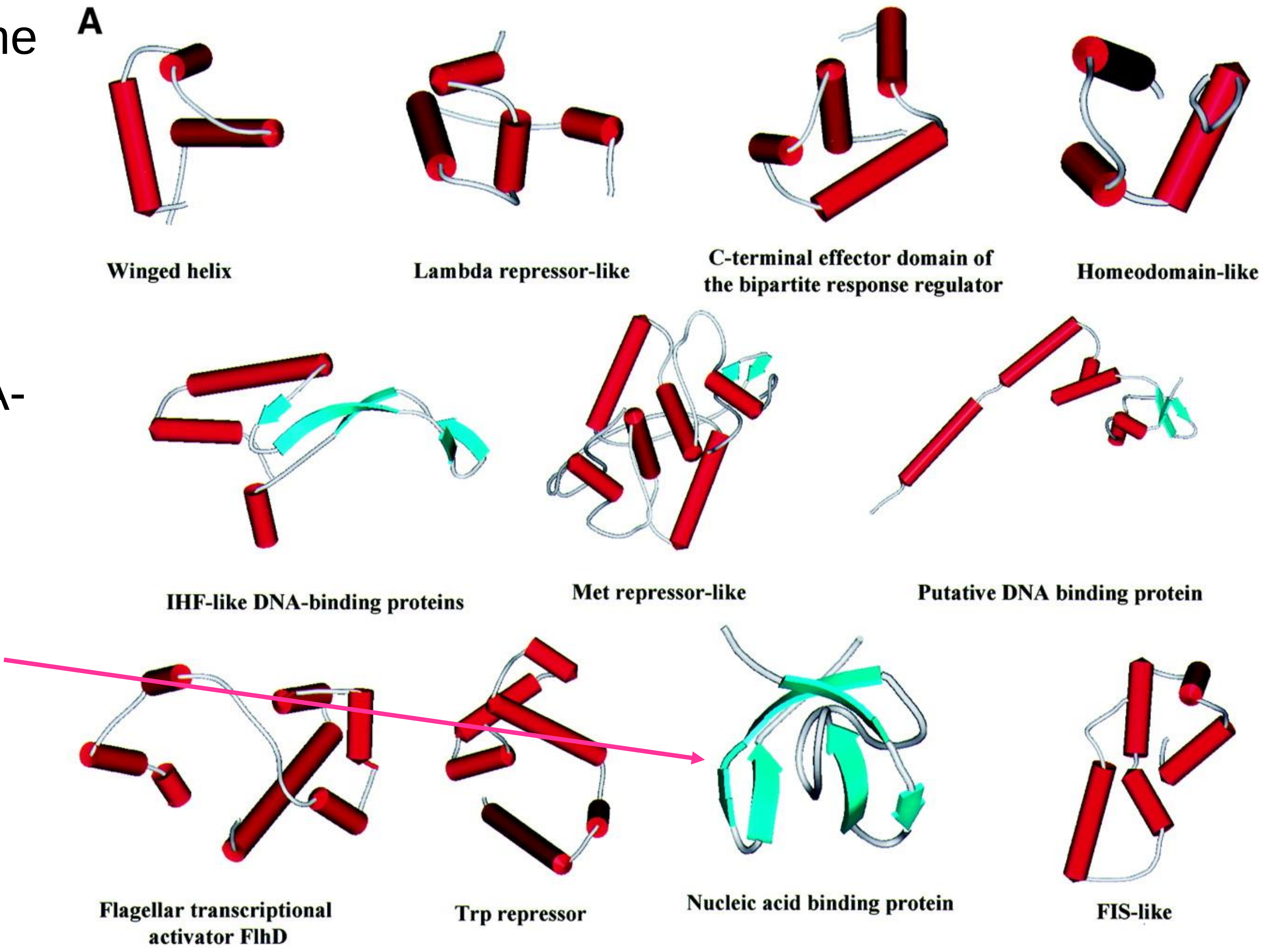
3D structures of putative (and real) TFs in *E.coli*

3D structures of the 11 DBD families seen in the 271 identified TFs in *E.coli*.

The **helix–turn–helix motif** is typical for DNA-binding proteins.

It occurs in all families except the nucleic acid binding family.

Still the scaffolds in which the motif occurs are very different.


A
Winged helix
Lambda repressor-like
C-terminal effector domain of the bipartite response regulator
Homeodomain-like
IHF-like DNA-binding proteins
Met repressor-like
Putative DNA binding protein
Flagellar transcriptional activator FlhD
Trp repressor
Nucleic acid binding protein
FIS-like

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

Domain architectures of TFs

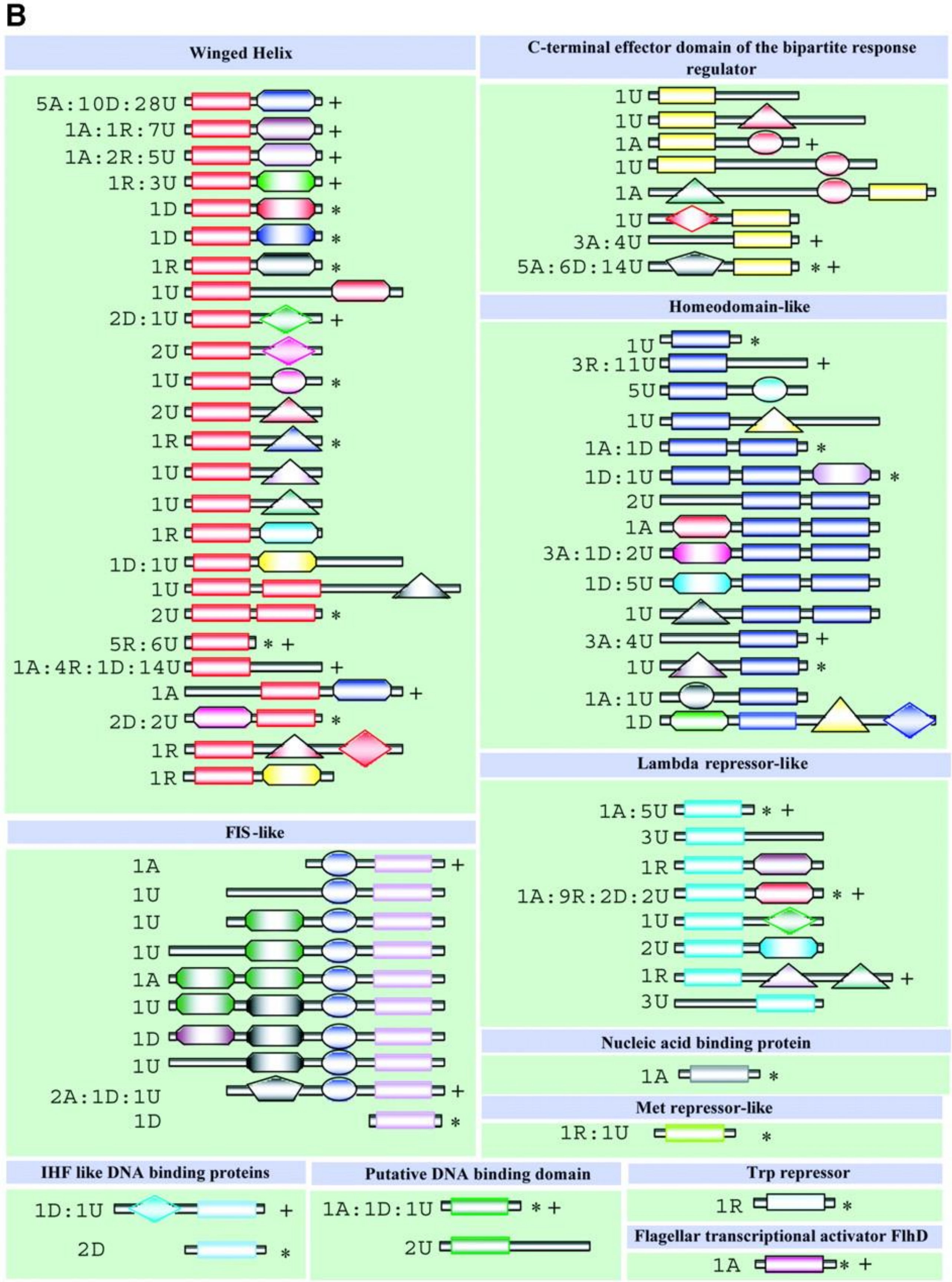
The 74 unique domain architectures of the 271 TFs.

The **DBDs** are represented as rectangles. ▭
The partner domains are represented as hexagons (**small molecule-binding domain**), ⬡
triangles (**enzyme** domains), △
circles (protein interaction domain), ●
diamonds (domains of unknown function). ◇
The receiver domain has a pentagonal shape.

A, R, D and U stand for activators, repressors, dual regulators and TFs of unknown function.

The number of TFs of each type is given next to each domain architecture.

Architectures of known 3D structure are denoted by asterisks.
'+' are cases where the regulatory function of a TF has been inferred by indirect methods, so that the DNA-binding site is not known.



Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

Evolution of TFs

10%	1-domain proteins
75%	2-domain proteins
12%	3-domain proteins
3%	4-domain proteins

TFs have evolved by apparently extensive recombination of domains.

Proteins with the same sequential arrangement of domains
are likely to be direct duplicates of each other.

74 distinct domain architectures have duplicated to give rise to 271 TFs.

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

Evolution of the gene regulatory network

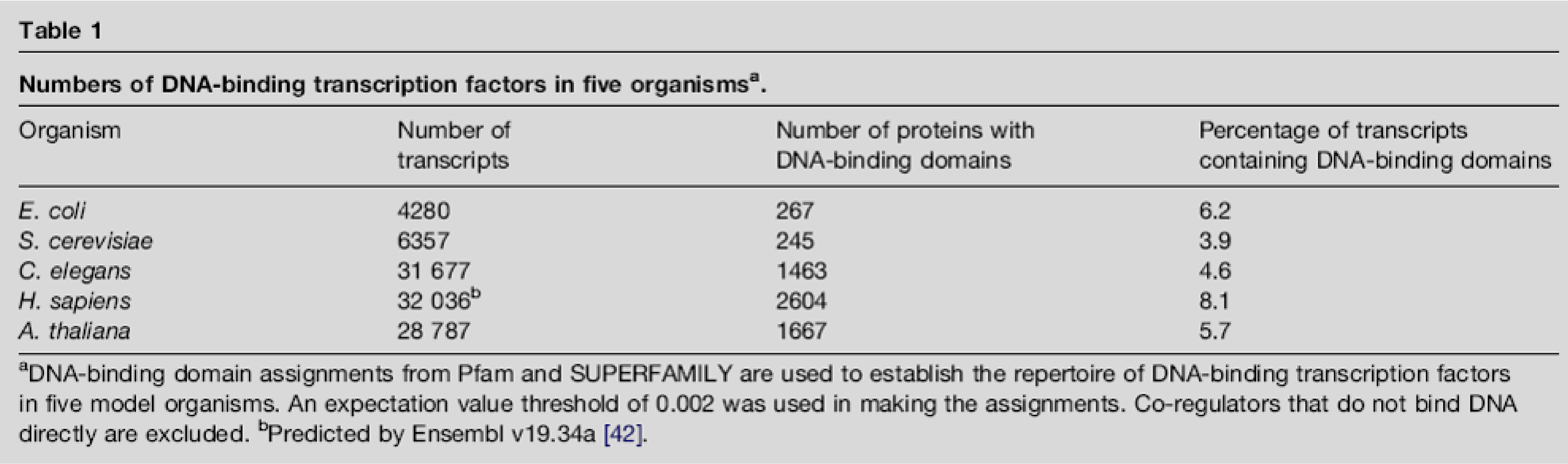


Table 1

Numbers of DNA-binding transcription factors in five organisms[a].

Organism	Number of transcripts	Number of proteins with DNA-binding domains	Percentage of transcripts containing DNA-binding domains
E. coli	4280	267	6.2
S. cerevisiae	6357	245	3.9
C. elegans	31 677	1463	4.6
H. sapiens	32 036[b]	2604	8.1
A. thaliana	28 787	1667	5.7

[a] DNA-binding domain assignments from Pfam and SUPERFAMILY are used to establish the repertoire of DNA-binding transcription factors in five model organisms. An expectation value threshold of 0.002 was used in making the assignments. Co-regulators that do not bind DNA directly are excluded. [b] Predicted by Ensembl v19.34a [42].

Larger genomes tend to have more TFs per gene.

Babu et al. Curr Opin Struct Biol. 14, 283 (2004)

Transcription factors in yeast *S. cereviseae*

Q: How can one define transcription factors?

Hughes & de Boer consider as TFs proteins that
(a) bind DNA directly and in a sequence-specific manner and
(b) function to regulate transcription nearby sequences they bind

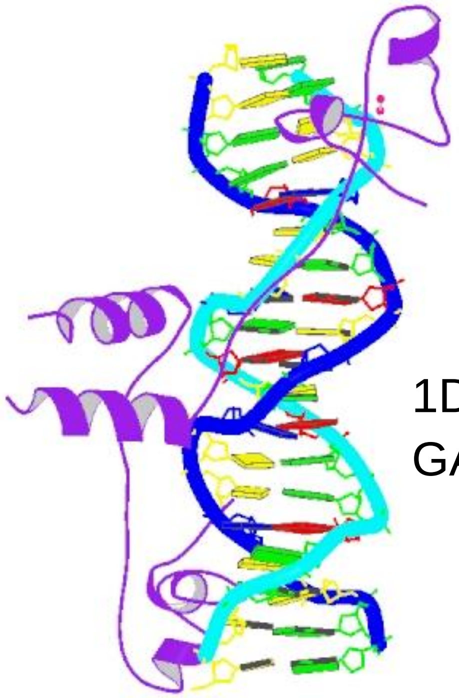
Q: Is this a good definition?

Yes. Only 8 of 545 human proteins that bind specific DNA sequences and regulate transcription lack a known DNA-binding domain (DBD).

Hughes, de Boer (2013) Genetics 195, 9-36

Transcription factors in yeast

Hughes and de Boer list 209 known and putative yeast TFs.
The vast majority of them contains a canonical DNA-binding domain.



1D66.pdb
GAL4 family

Most abundant:

- GAL4/zinc cluster domain (57 proteins),
  largely specific to fungi (e.g. yeast)
- zinc finger C2H2 domain (41 proteins),
  most common among all eukaryotes.

Other classes :

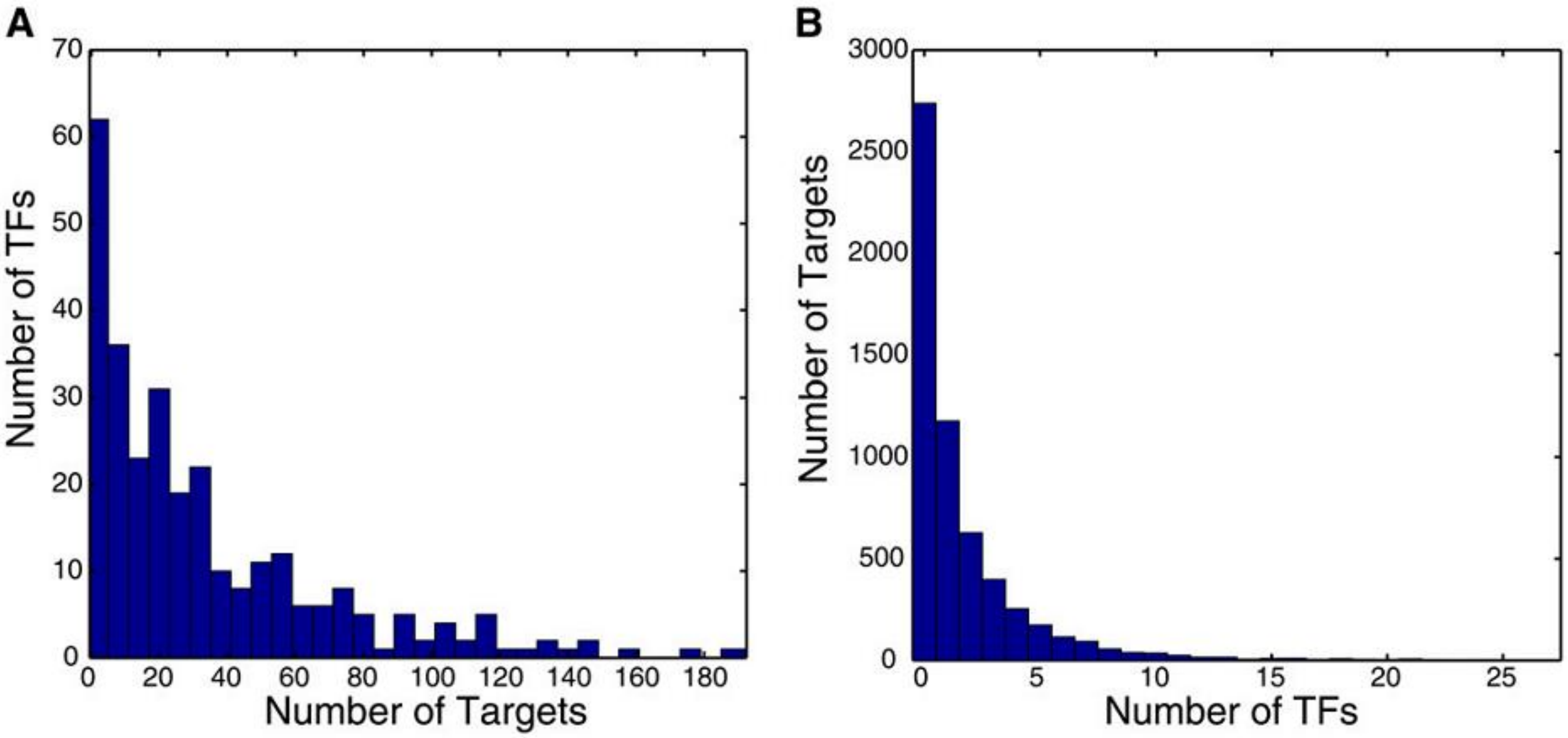
- bZIP (15),
- Homeodomain (12),
- GATA (10), and
- basic helix-loop-helix (bHLH) (8).

Hughes, de Boer (2013) Genetics 195, 9-36

TFs of *S. cereviseae*

(A) Most TFs tend to bind relatively few targets.
57 out of 155 unique proteins bind to ≤ 5 promoters in at least one condition.
17 did not significantly bind to any promoters under any condition tested.

In contrast, several TFs have hundreds of promoter targets.
These TFs include the general regulatory factors (GRFs), which play a global role in transcription under diverse conditions.



(B) # of TFs that bind to one promoter.

Hughes, de Boer (2013) Genetics 195, 9-36

Co-expression of TFs and target genes?

Overexpression of a TF often leads to induction or repression of target genes.

This suggests that many TFs can be regulated simply by the abundance (expression levels) of the TF.

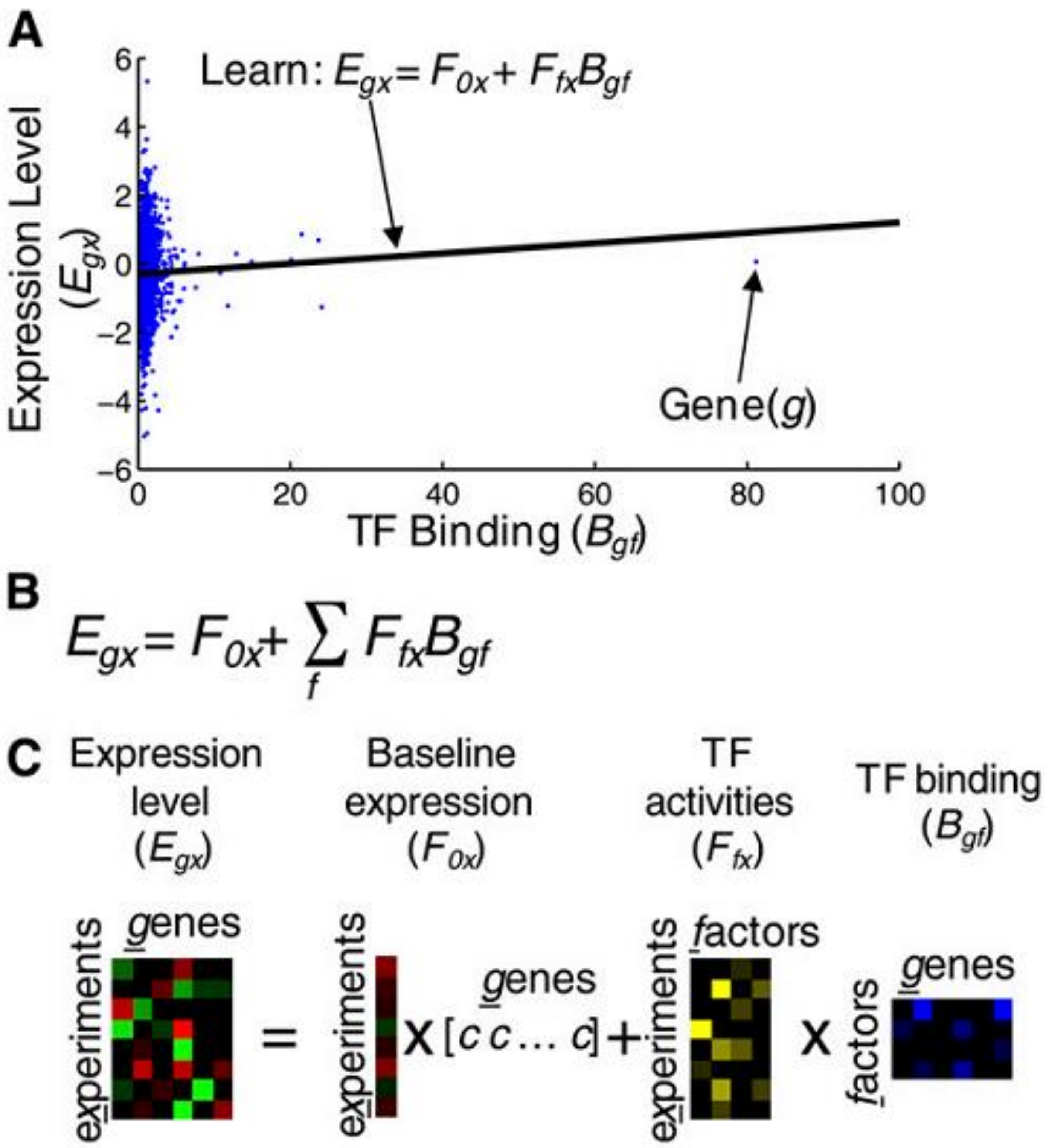
However, across 1000 microarray expression experiments for yeast, the **correlation** between a TF's expression and that of its ChIP-based targets was typically **very low** (only between 0 and 0.25)!

At least some of this (small) correlation can be accounted for by the fact that a subset of TFs autoregulate.

→ TF expression accounts for only a minority of the regulation of TF activity in yeast.

Hughes, de Boer (2013) Genetics 195, 9-36

Using regression to predict gene expression



(A) Example where the relationship between expression level (E_{gx}) and TF binding to promoters (B_{gf}) is found for a single experiment (x) and a single TF (f). Here, the model learns 2 parameters: the background expression level for all genes in the experiment (F_{0x}) and the activity of the transcription factor in the given experiment (F_{fx}).

(B) The generalized equation for multiple factors and multiple experiments.

(C) Matrix representation of the generalized equation.

Baseline expression is the same for all genes and so is represented as a single vector multiplied by a row vector of constants where c = 1/(no. genes).

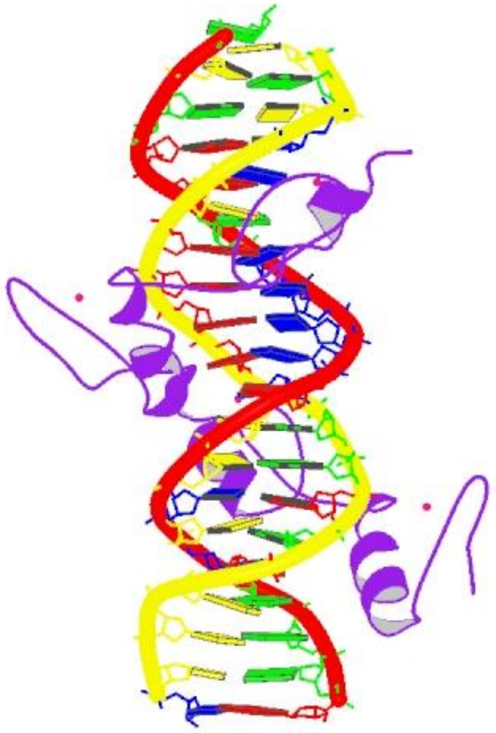
Hughes, de Boer (2013) Genetics 195, 9-36

Transcription factors in human: ENCODE

Some TFs can either activate or repress target genes.

The TF YY1 shows largest mixed group of target genes.

TF	Ubiquitously activated	Ubiquitously repressed
YY1	COQ5[cd]	AC091153.1
	CPNE1	ATP5O
	CPSF2 [cd]	BIRC6[d]
	CR613718	CAPZA2
	IP6K2[a]	CXorf26
	NARS[ac]	DKFZp434H247
	PAK4[d]	EFHA1
	PSMB4[ac]	MRPS10[c]
	UBR5	MRPS18B[acd]
		NUP160
		OXCT1
		PSMD8[ac]
		SNX27
		SNX3[ad]
		SRP68[ad]
		TNKS



1UBD.pdb
human YY1

Whitfield et al. Genome Biology 2012, 13:R50

YY1 binding motifs

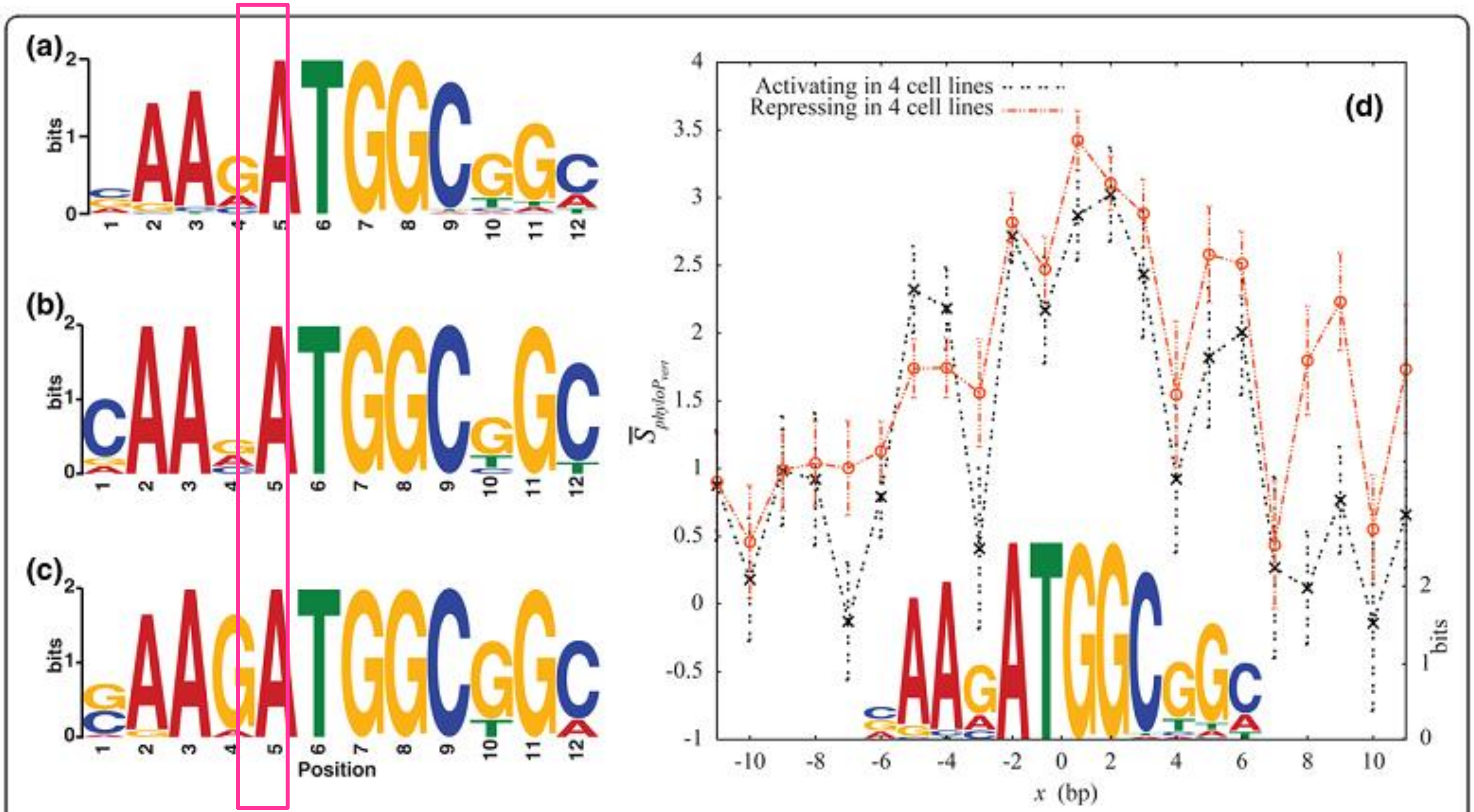
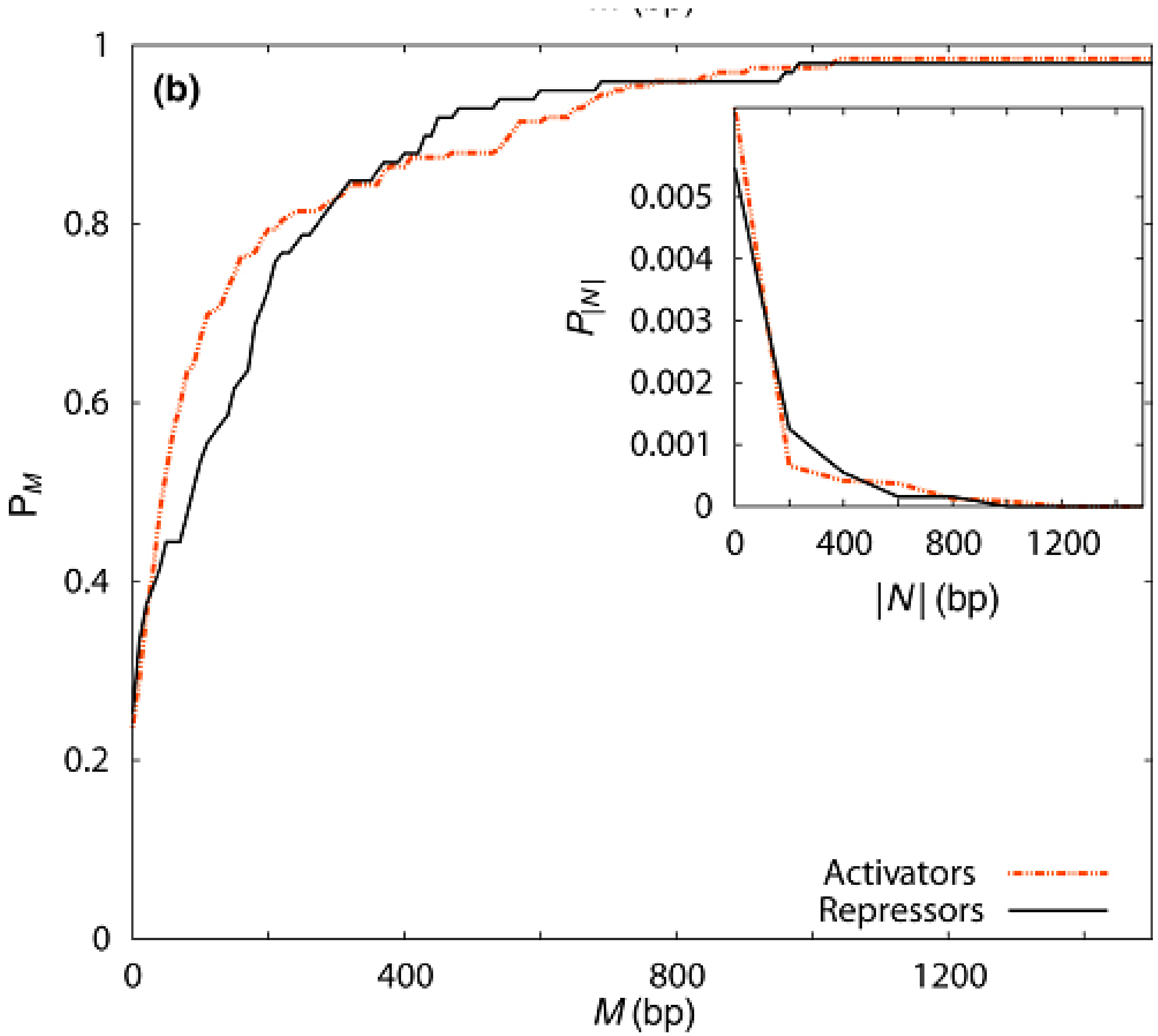


Figure 2 Characterization of functional YY1 binding sites. Sequence logo [102] for YY1 binding sites from **(a)** PWM and sites that are functionally **(b)** ubiquitously activating (9 BS) or **(c)** ubiquitously repressive (16 BS) in four human cell lines. In **(d)**, we plot the mean vertebrate phyloP conservation score [90] around functional YY1 binding sites. The mean score, $\bar{S}_{phyloP_{vert}}$, was computed at each base for sites where the binding event ubiquitously activated (black line) or repressed (red line) transcription in all four cell lines. The position weight matrix that was used to predict YY1 binding sites is shown (scale on the right axis).

No noticeable difference in binding motifs of activated or repressed target genes.

Whitfield et al. Genome Biology 2012, 13:R50

Where are TF binding sites wrt TSS?



Inset: probability to find binding site at position N from transcriptional start site (TSS)

Main plot: cumulative distribution.

activating TF binding sites are closer to the TSS than repressing TF binding sites ($p = 4.7 \times 10^{-2}$).

Whitfield et al. Genome Biology 2012, 13:R50

Summary transcription

- Gene transcription (mRNA levels) is controlled by transcription factors (activating / repressing) and by microRNAs (degrading)

- Binding regions of TFs are ca. 5 – 10 bp stretches of DNA

- Global TFs regulate hundreds of target genes

- Global TFs often act together with more specific TFs

- TF expression only weakly correlated with expression of target genes (yeast)

- Some TFs can activate or repress target genes. Use similar binding motifs for this.