Systems Biology: introduction

PhD Program, 2023

Michele Caselle – University of Torino and INFN
caselle@to.infn.it

Plan of the lecture

1. Introduction

2. The last twenty years: The “genomic revolution”

3. New tools and ideas: Computational Biology and Systems biology

4. Example 1: Evolutionary models

5. Example 2: Gene Regulation

6. Example 3: Identification of Cancer Driver Genes

References

1. U. Alon: An introduction to Systems Biology,
   (second edition) CRC press

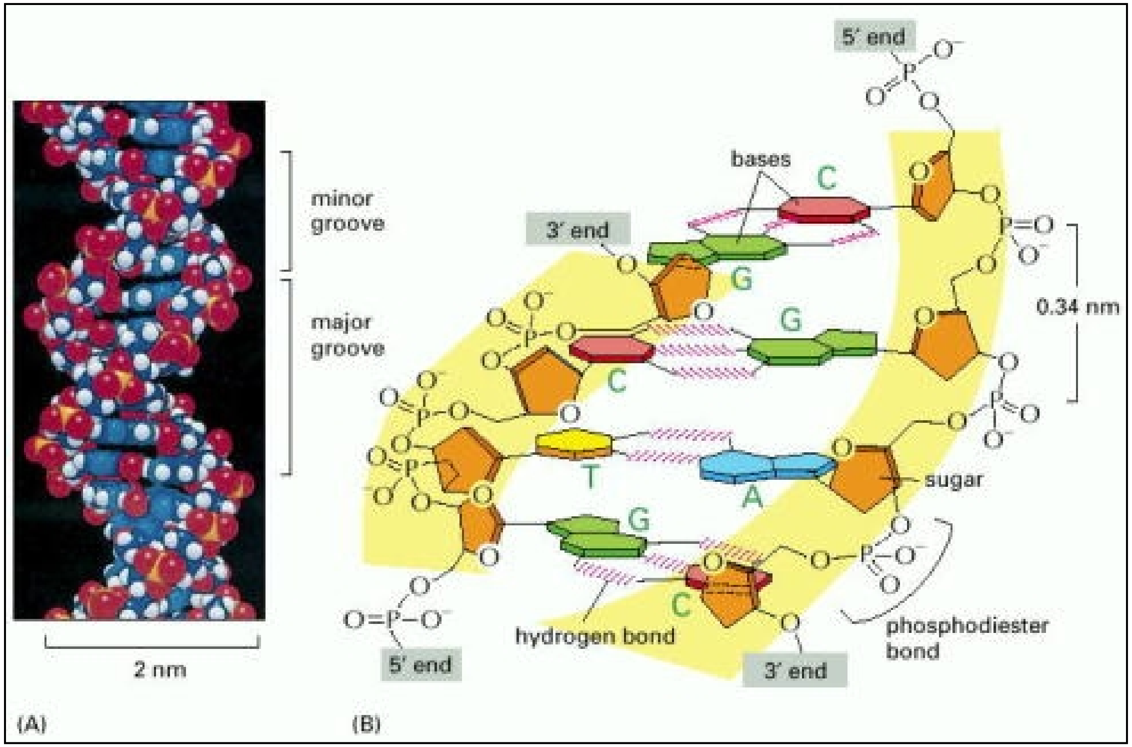
2. E.Klipp, W. Liebermeister, C. Wierling, A Kowald, R. Herwig
   Systems Biology: A Textbook Wiley ed.

“There is no precise technical definition of a “**complex system**”, but most researchers in the field would probably agree that it is a system composed of many interacting parts, such that the collective behavior of those parts together is more than the sum of their individual behaviors. The collective behaviors are sometimes also called “emergent” behaviors, and a complex system can thus be said to be a system of interacting parts that displays emergent behavior.”

M.E.Newman, Complex Systems: a Survey
http://arxiv.org/abs/1112.1440

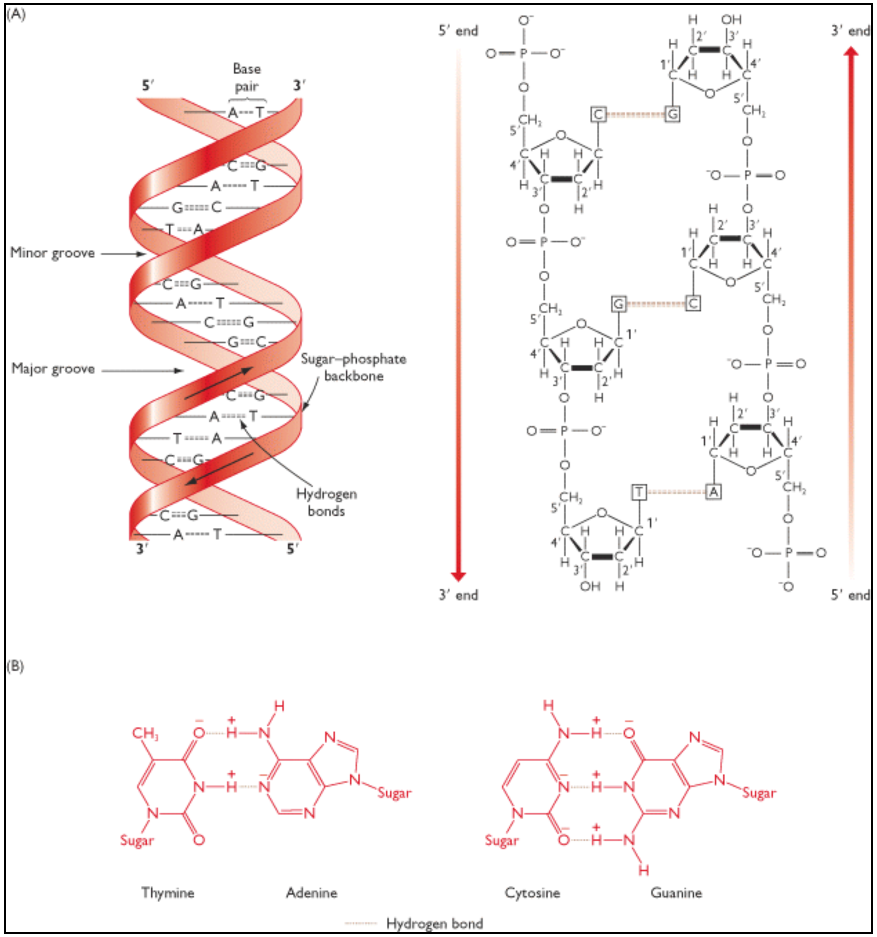
DNA

- Genomic information is encoded in the **DNA** chain.
- In the human case the genome is composed by 3x10^9 base pairs which may take four possible values: **A,C,G,T**



DNA

The main property of the **DNA** chain is base pairing: (A,T) and (C,G). This allows both DNA replication and the use of the chain as a template for protein production.



Genome size

Organism	Genome length (Mbp)	Number of coding genes
M. genitalium	0.58	470
E. coli	4.6	4288
S. Cerevisiae	12.2	6692
C. Elegans	103	20447
D. Melanogaster	144	13918
M. Musculus	3500	22606
H. Sapiens	3300	20300

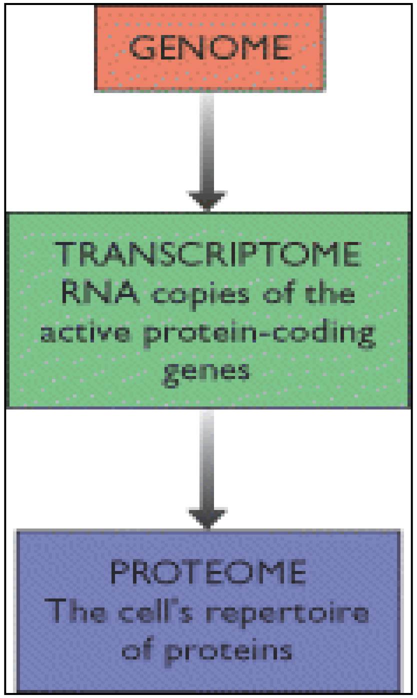
Genome Organization

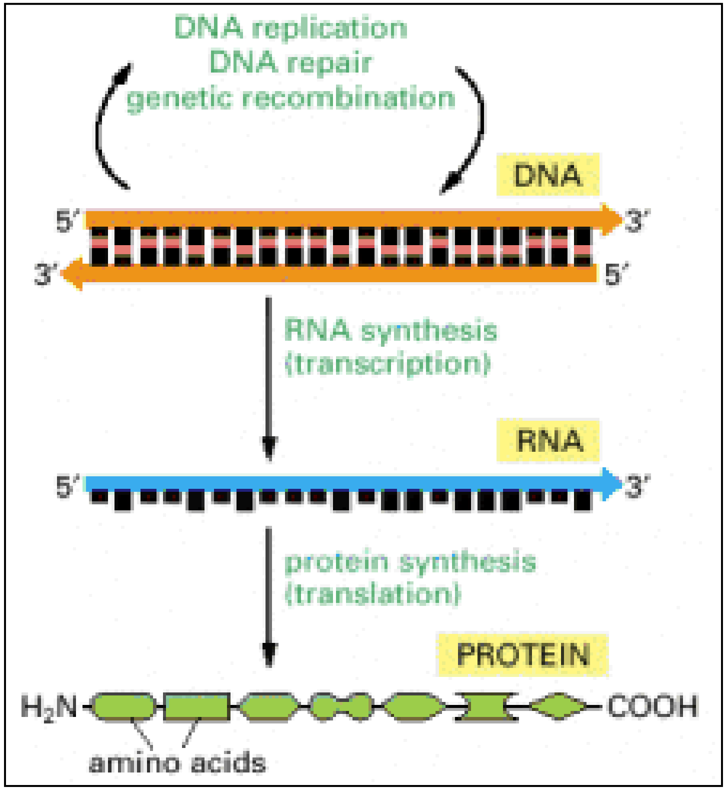
- While the genome size increases with the complexity of the organism. The **number of genes is almost costant!**
- The portion of the genome coding for proteins decreases as the complexity of the organism increases. It is very high in procaryotes and yeast but very low in mammalian.

  97% of the human genome is non-coding!!
- Most of this non-coding DNA is involved in the **regulation of gene expression**

"Old Paradigm": Information flow in the cell

"Central Dogma" of molecular biology

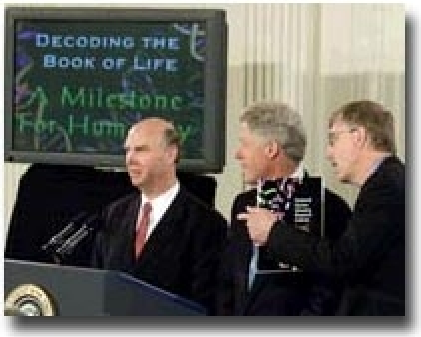

GENOME
TRANSCRIPTOME
RNA copies of the active protein-coding genes
PROTEOME
The cell's repertoire of proteins

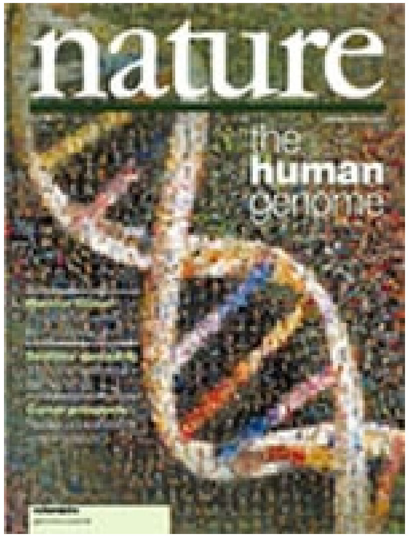


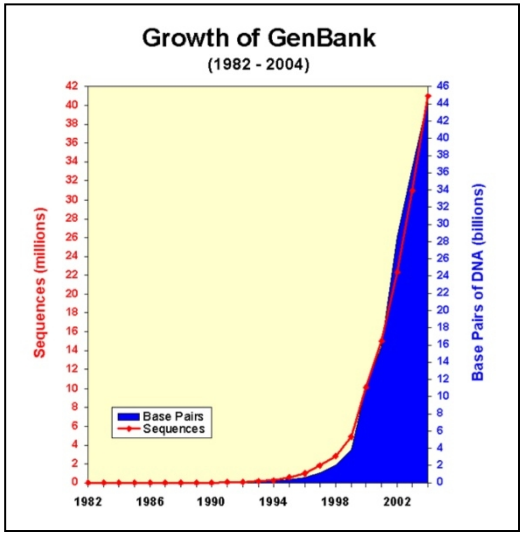
Genomic Revolution

The main driving force of the Genomic Revolution was the Human Genome Project (2000)

> *homo_sapiens*
ACTTTTTTACCCTCGTGTGTTGC
AGACTTTTTGCCACTTTTAAAAC
GCTGACAATTCGACCCTTTCCAA
GTGCAAAAAGTGCCAAGATTTA
CGATAAAATTCCCCCGAGAGAC
GTGTGCA………







New Paradigm:

- **Alternative Splicing:** one gene → many proteins

- **RNA regulatory genes:** miRNA, lncRNA → “RNA World”

- **Retrotransposition:** more than 50 % of the genome is composed by Transposons

- **Cell to cell variability:** several mRNAs are produced in few units: stochastic fluctuations are important

- **Network paradigm:** Complex functions are performed by a complex interplay of several genes

The Present Revolution:

- **Metagenomics:** Single bacterial strain → microbiome

- **Horizontal Transfer:** Genome → Pangenome

- **Single Cell Sequencing:** Each cell is unique! How can we classify cells? Thousands of different neurons in the brain!

- **Epigenomics:** Heritable, reversible genetic information without DNA mutations

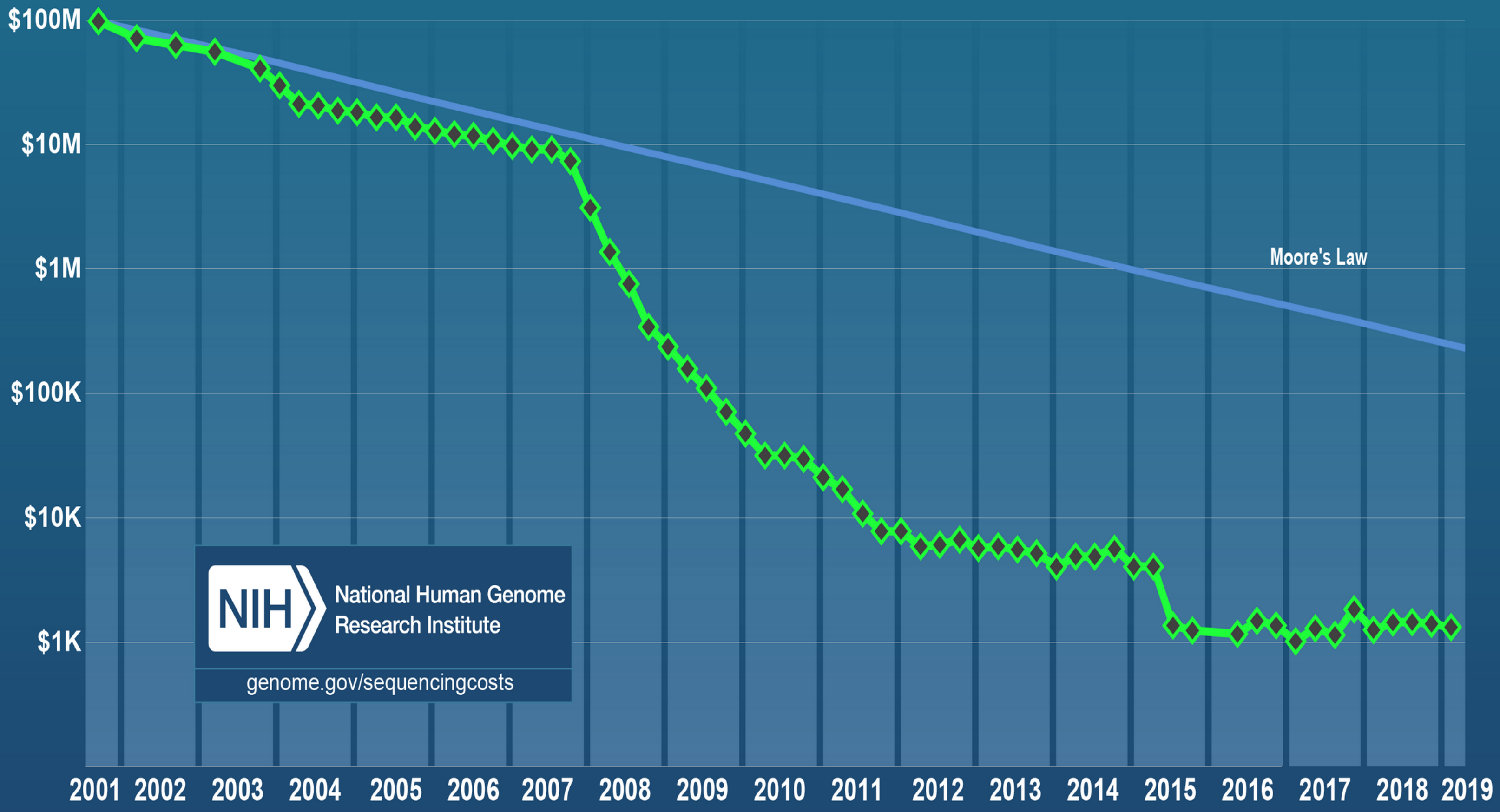
- **Hi-C techniques:** Tridimensional structure of Chromatine controls gene expression

A central role in this revolution was played by physics.

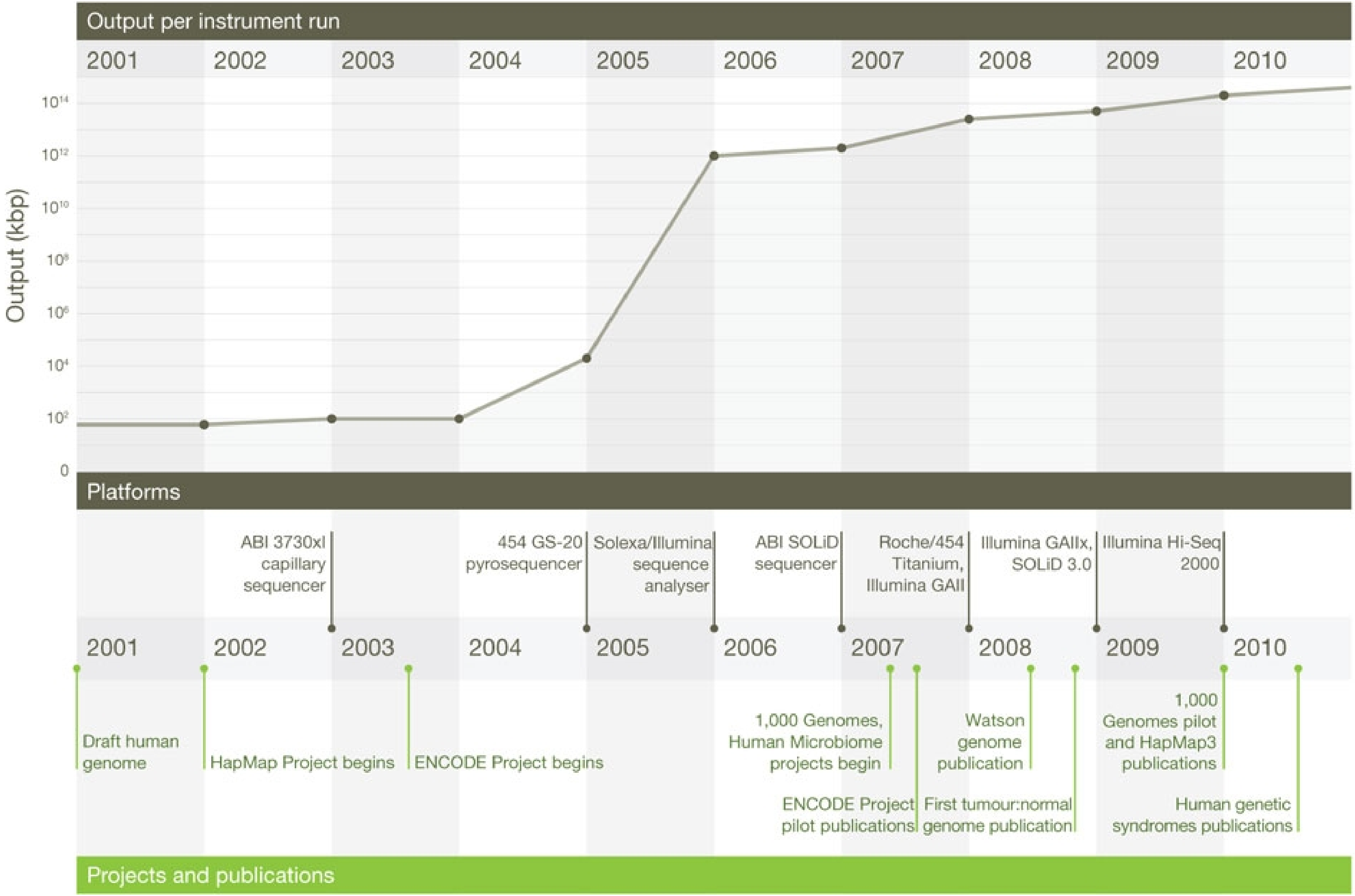
On the experimental side:

- nanotechnolgy
- microfluidics

Cost per Genome



Changes in instrument capacity



Timing of the major sequencing projects

A central role in this revolution was played by physics.

On the experimental side:

- nanotechnolgy
- microfluidics

A central role in this revolution was played by physics.

Both on the experimental side:

- nanotechnolgy
- microfluidics

And on the theoretical side:

- new inference methods
- modeling of complex systems
- network theory
- alignment tools

That is:

Computational Biology and Systems Biology

New Theoretical Tools:

Systems biology and Computational Biology

Computational Biology

With the terms "Computational Biology" or "Bioinformatics"
one usually refers to all the data mining tool based
on methods and ideas coming from
mathematics / physics / statistics / computer-science .

Genomic data   (both sequences and annotations)
Can be easily downloaded from huge "open access" data banks.

These data contain a lot of hidden information.
In general only a fraction of it has been recognized and published
by the authors of the experiments.

Relevant original results can be obtained with no need of new
costly experiments but simply using in a clever way existing data.

Systems Biology

Network theory: Complex functions, must be described at the network level and not at the level of single genes, proteins or neurons.

Modeling: These networks can be decomposed in elementary circuits. (“network motifs”) which may be modeled using differential or stochastic equations.

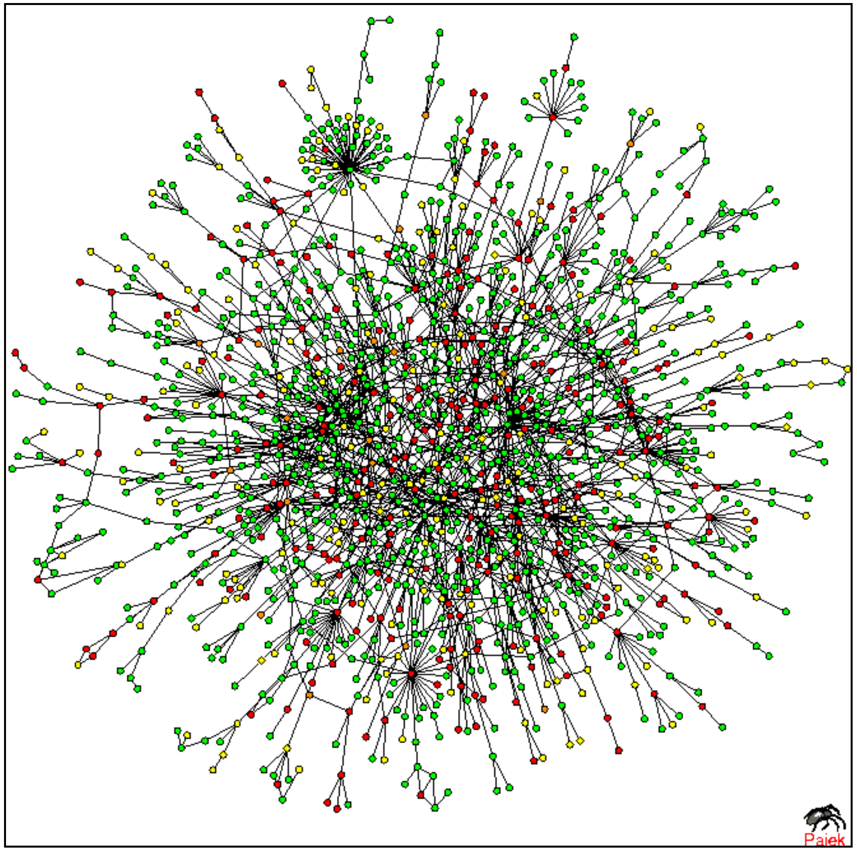
Ontologies: biological (and medical) information must be organized in a quantitative and standardized way

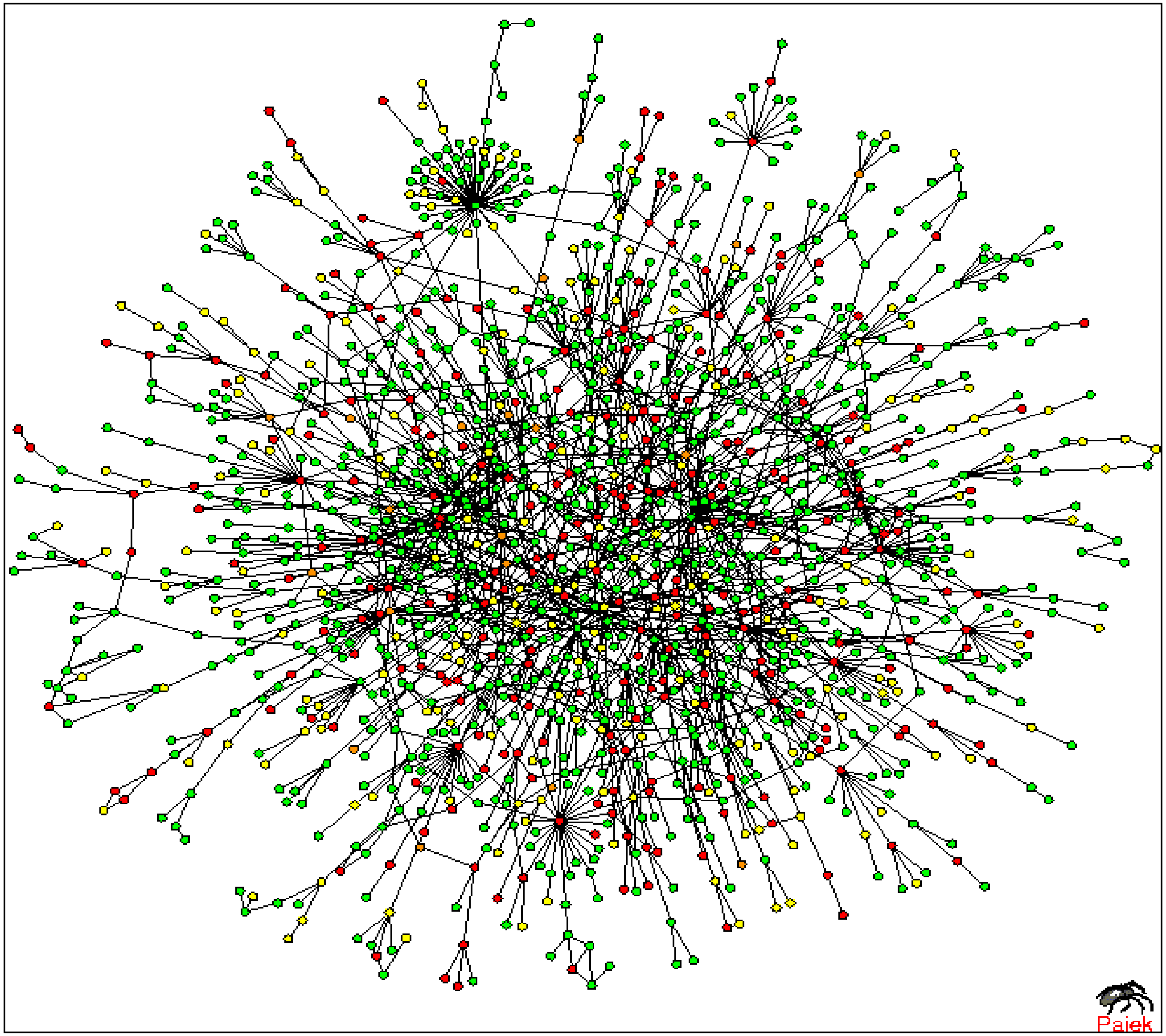
Modern Genomics: *networks*

- genes and proteins of a given organism are organized in networks .

- Cells react to external stimuli in a “global” way.

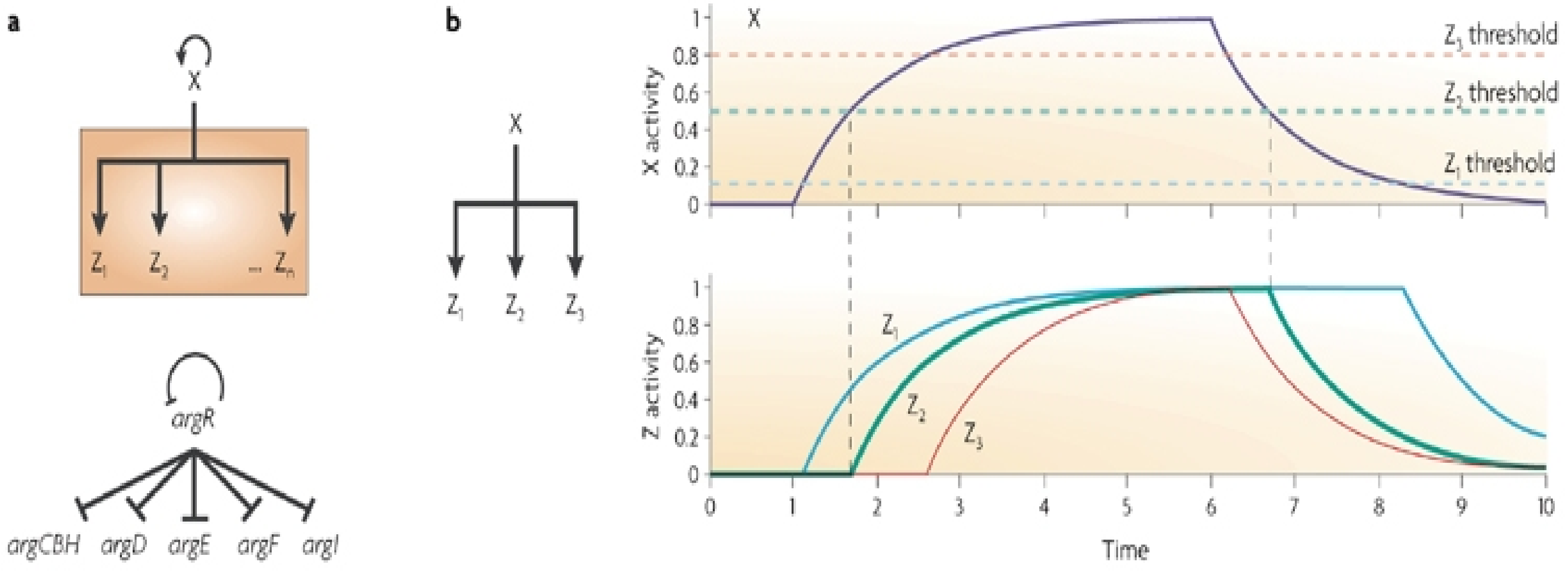
H.Jeong et al.
Nature, 411 (2001) 41




Pajek

Network motifs

Example: SIM (Single Input Module) (a) experimental realization: arginine biosynthesis b) Circuit behaviour: different genes are activated at different times as a function of their different activation threshold as the concentration of X (master regulator) changes in time R.Milo et al. Science 298 (2002) 824



Nature Reviews | Genetics

Modern Genomics:
Gene Ontology

- **Gene Ontology** is an example of standardization of biological data.

- The goal is the construction of a controlled vocabulary to describe:

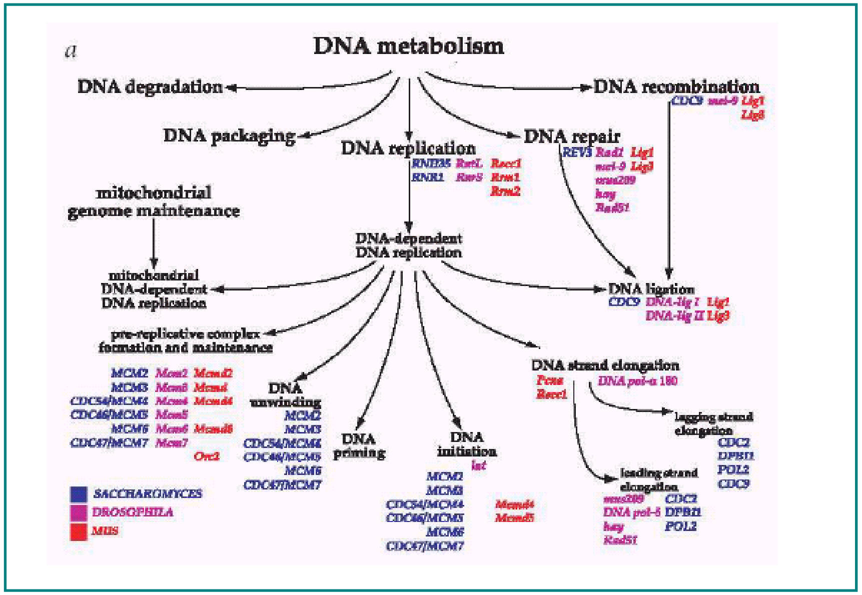
  - Molecular function

  - Biological process

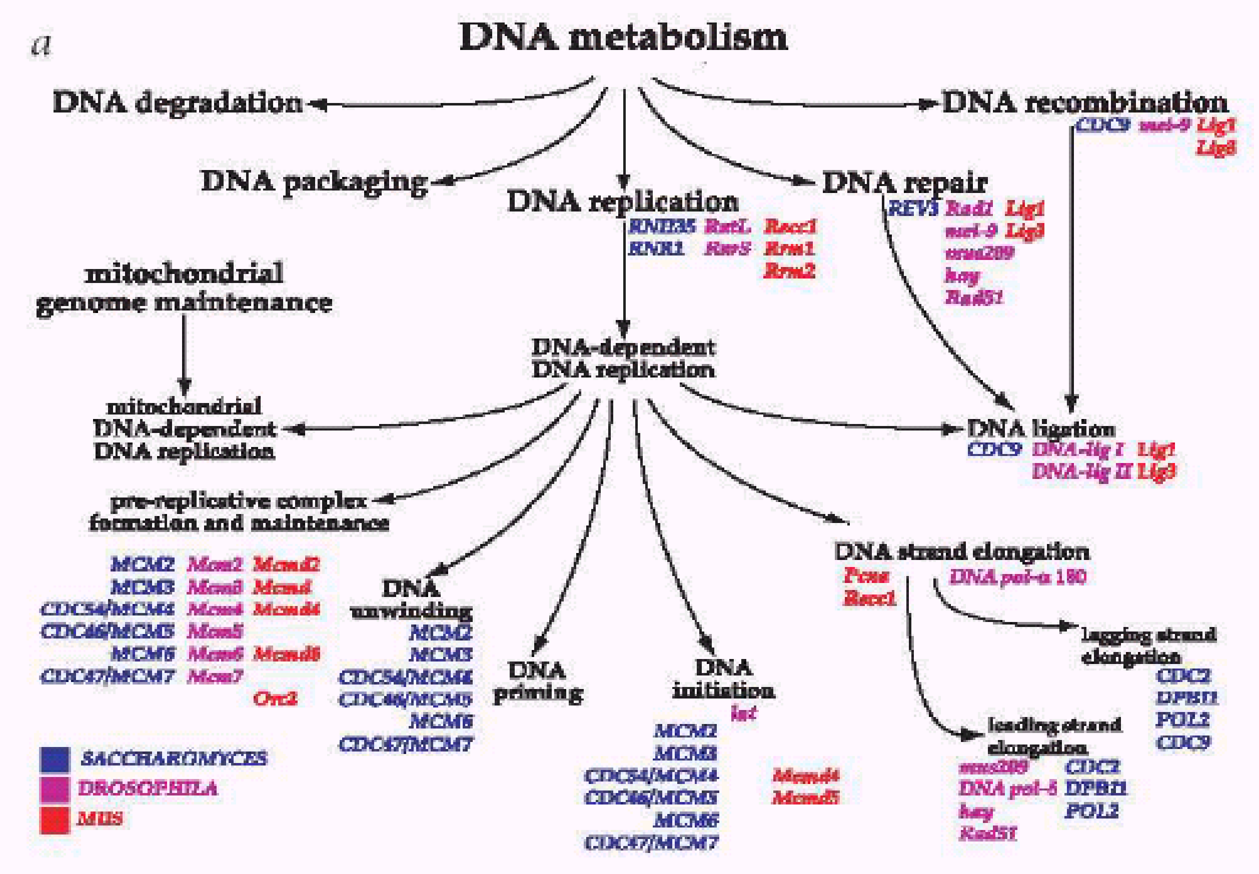
  - Cellular component

  of a given gene.

- The ontologies are organized as hierarchical networks (Directed acyclic graphs)



The G.O. Consortium
Nature Genet. 25 (2000) 25


a
DNA metabolism
DNA degradation
DNA packaging
DNA replication
RNR35 RnrL Recc1
RNR1 RnrS Rrm1
Rrm2
DNA repair
REV3 Rad1 Lig1
mei-9 Lig3
mus209
hay
Rad51
DNA recombination
CDC9 mei-9 Lig1
Lig3
mitochondrial
genome maintenance
mitochondrial
DNA-dependent
DNA replication
DNA-dependent
DNA replication
DNA ligation
CDC9 DNA-lig I Lig1
DNA-lig II Lig3
pre-replicative complex
formation and maintenance
MCM2 Mcm2 Mcmd2
MCM3 Mcm3 Mcmd
CDC54/MCM4 Mcm4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7
Orc2
DNA
unwinding
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
DNA
priming
DNA
initiation
lat
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
Mcmd4
Mcmd5
DNA strand elongation
Pcna
Recc1
DNA pol-α 180
lagging strand
elongation
CDC2
DPB11
POL2
CDC9
leading strand
elongation
mus209
DNA pol-δ
hay
Rad51
CDC2
DPB11
POL2
SACCHAROMYCES
DROSOPHILA
MUS

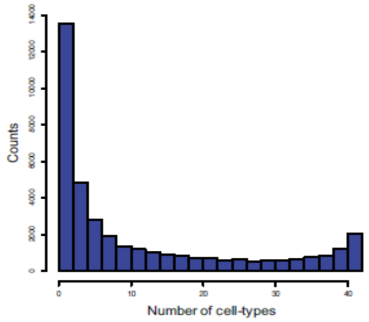
Systems Biology: Regulatory Networks

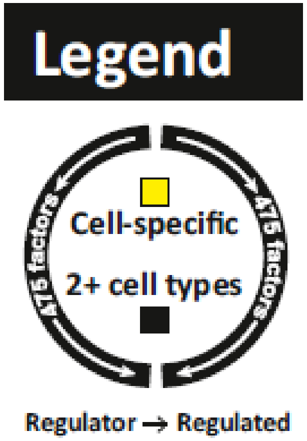
Example : **“Circuitry and Dynamics of Human Transcription Factor Regulatory Network**” Neph et al. CELL (2012) 150, 1274 (ENCODE collaboration).

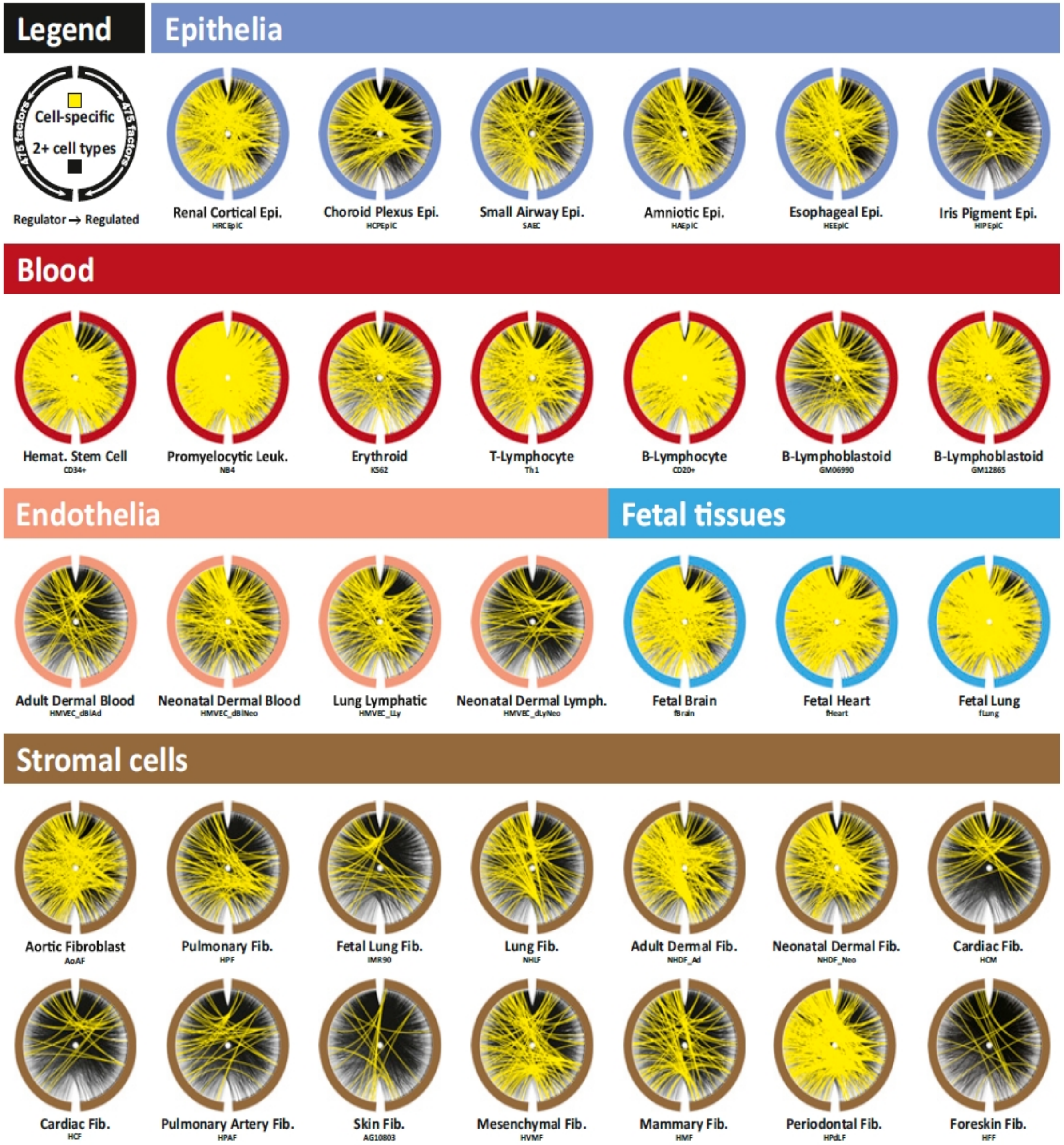
Regulatory network in 41 different human cell lines among 475 TFs using DNAse footprinting

Cell-Specific versus Shared Regulatory Interactions in TF Networks of 41 Diverse Cell Types

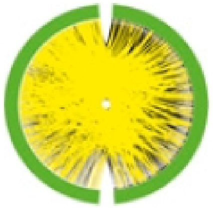
Number of cell-types that a transcriptional regulatory interaction was observed in



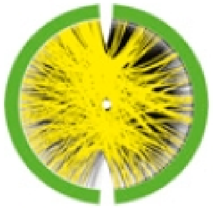




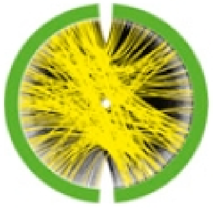
Visceral cells



Hippocampal Astrocyte
HA-h



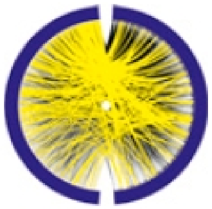
Skeletal Myoblast
HSMM



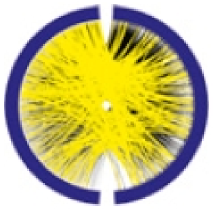
Skeletal Muscle
SKMC

Astrocyte
NH-A

Cancer

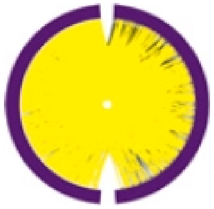


Neuroblastoma
SK-N-SH_RA



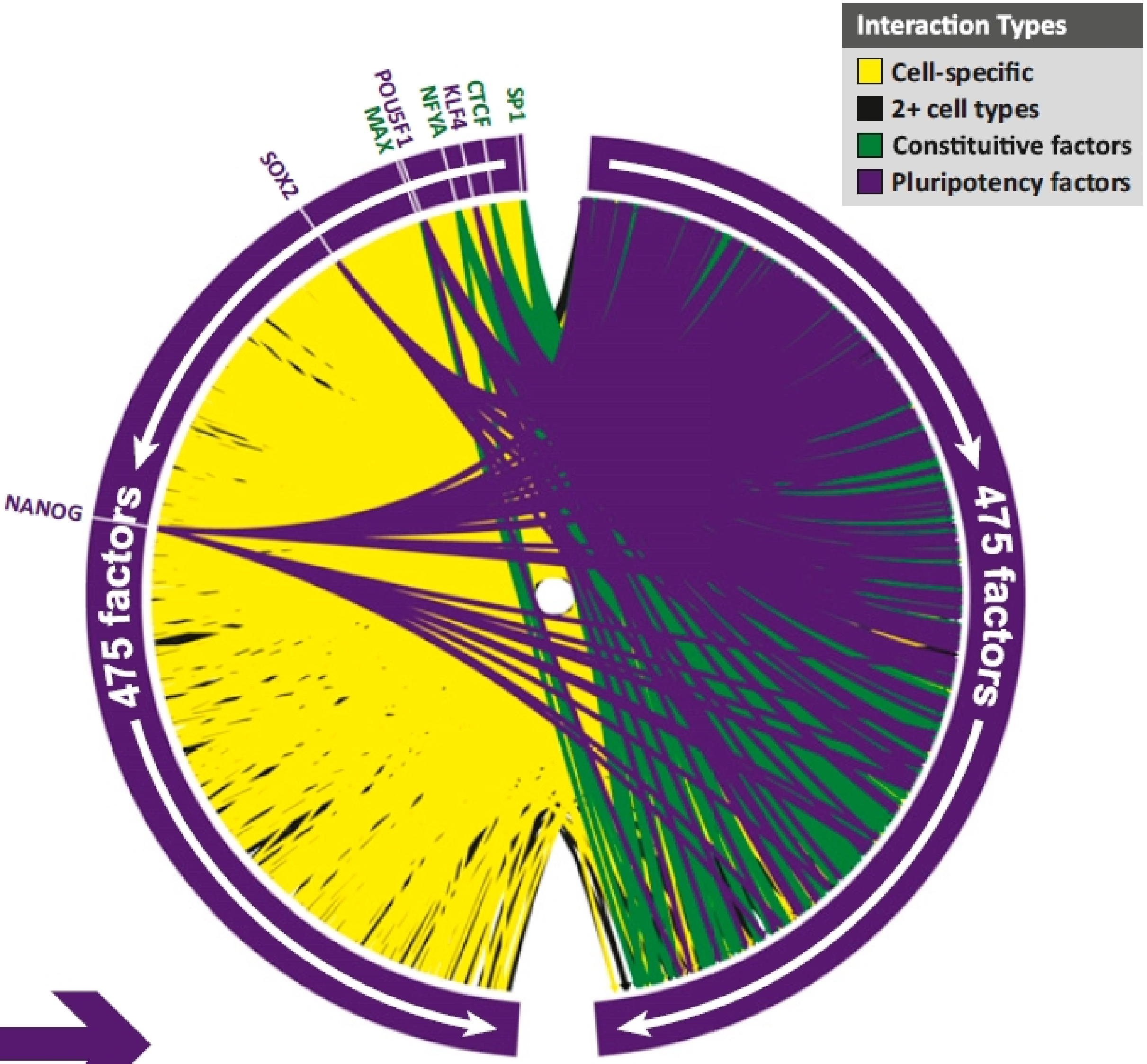
Hepatoblastoma
HepG2

Embryonic Stem Cells



Embryonic Stem Cells
H7-hESC

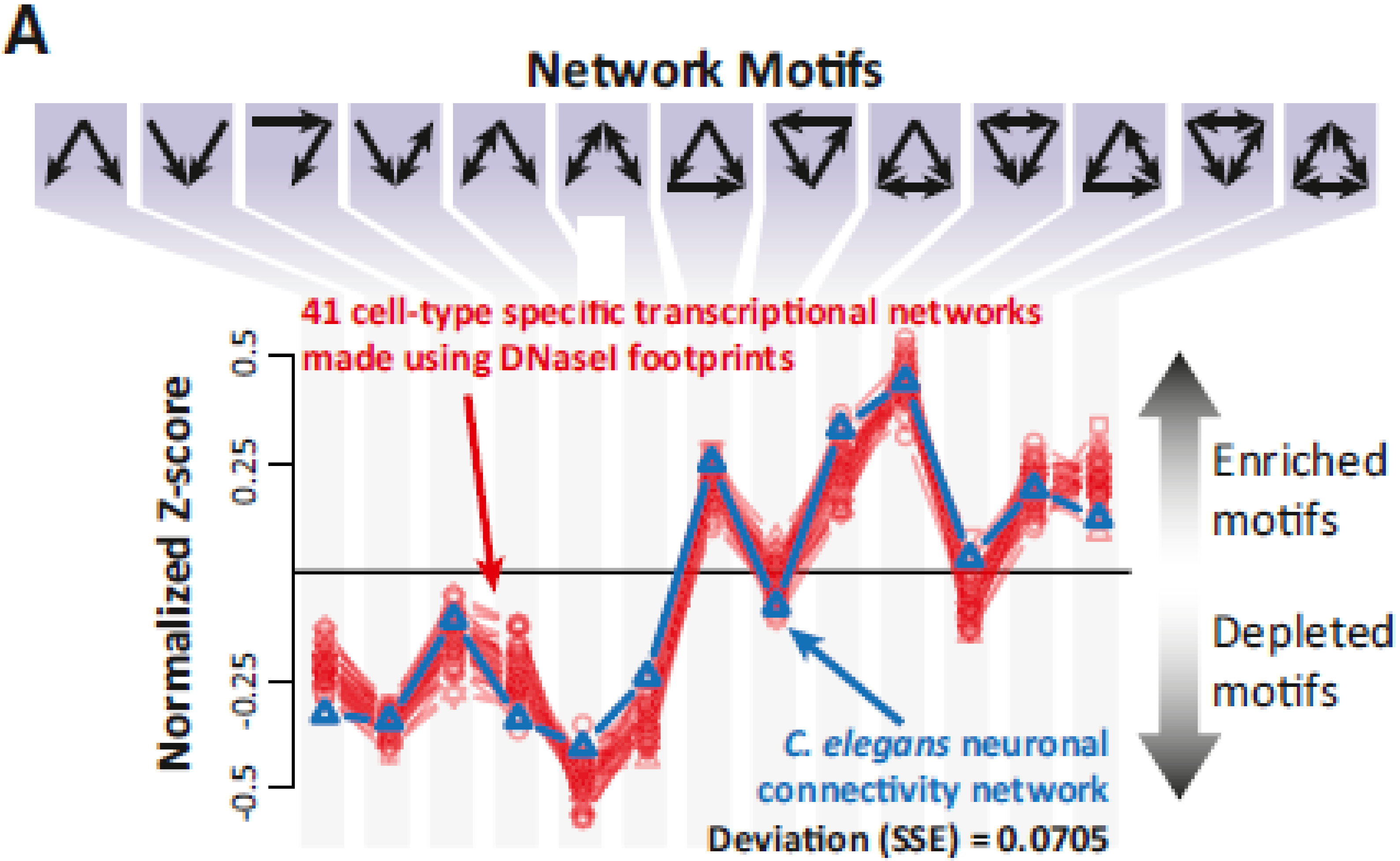
Regulatory interactions in human ES cells (detail)



Regulator → Regulated

475 transcription factors, ordered from high (SP1) to low degree (ZNF354C) in H7-hES Cells

Conserved Architecture of Human TF Regulatory Networks



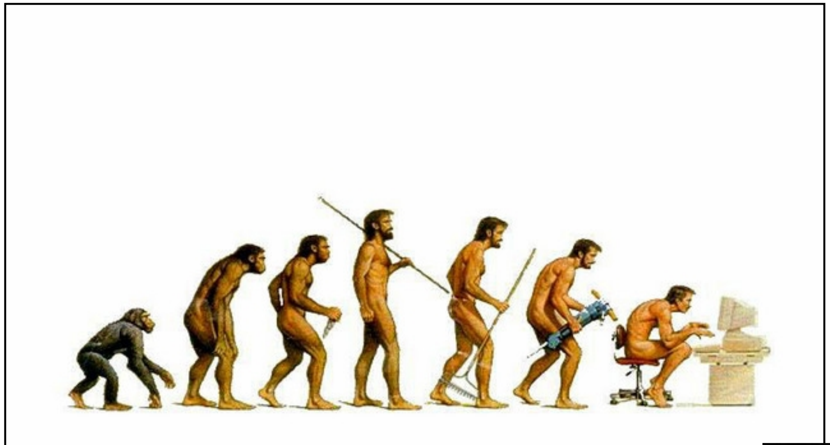
Three paradigmatic examples

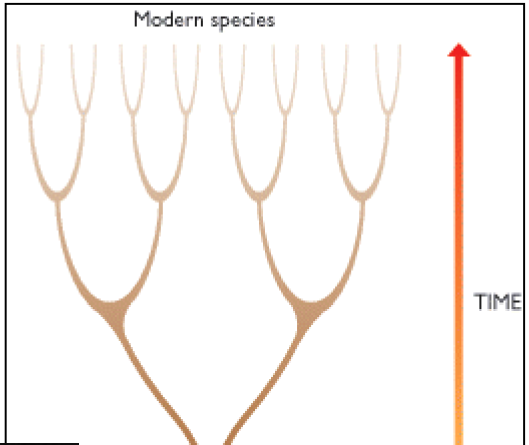
§ Evolutionary models

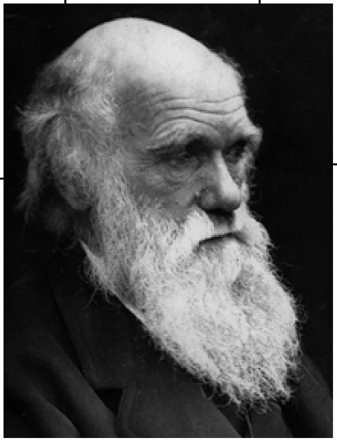
§ Gene Regulation

§ Identification of Cancer Driver Genes

Evolutionary models







Evolution at the Genomic level

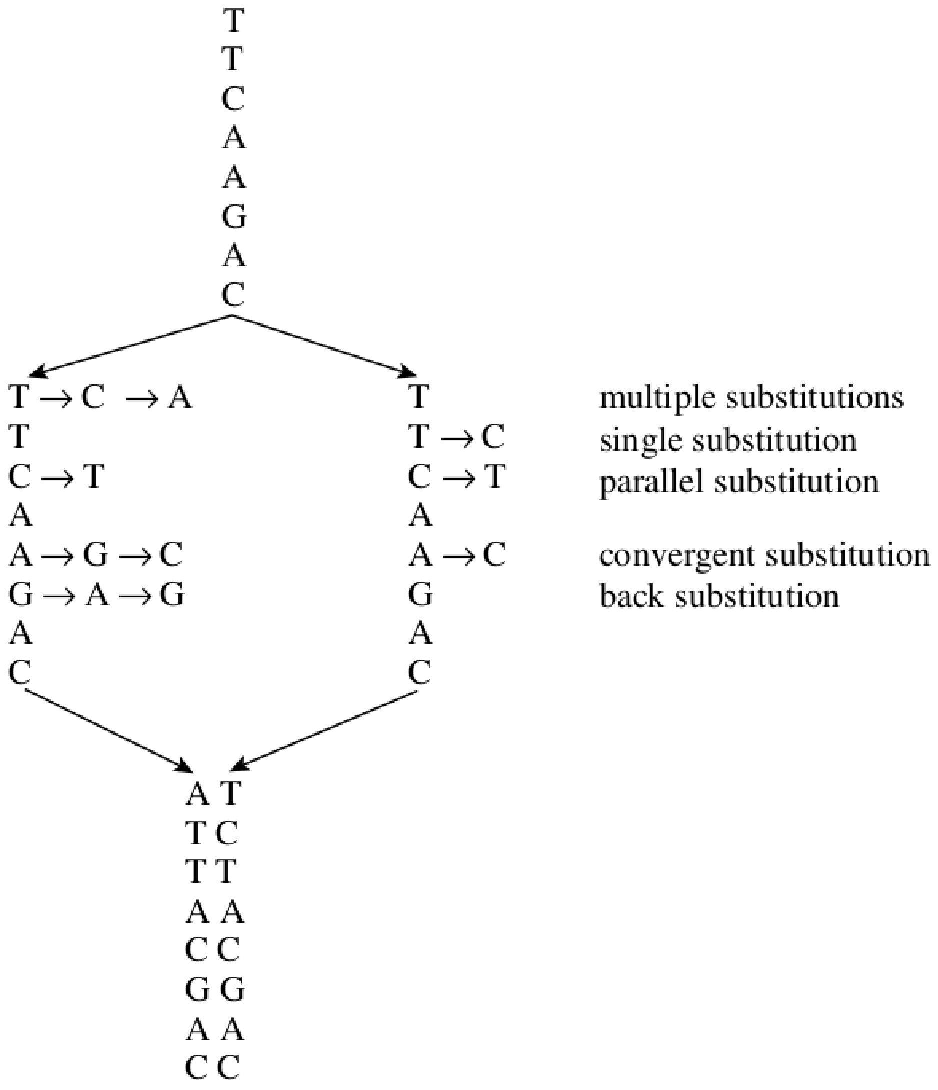
There are three different processes which drive sequence evolution and accordingly there are three different **scales** at which the DNA sequence can evolve.

- Single Nucleotide Mutations (SNP)

- Gene duplications

- Whole Genome Duplication (WGD)

Single Point Mutation



T
T
C
A
A
G
A
C

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

T
T → C
C → T
A
A → C
G
A
C

multiple substitutions
single substitution
parallel substitution
convergent substitution
back substitution

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

Single Point Mutation

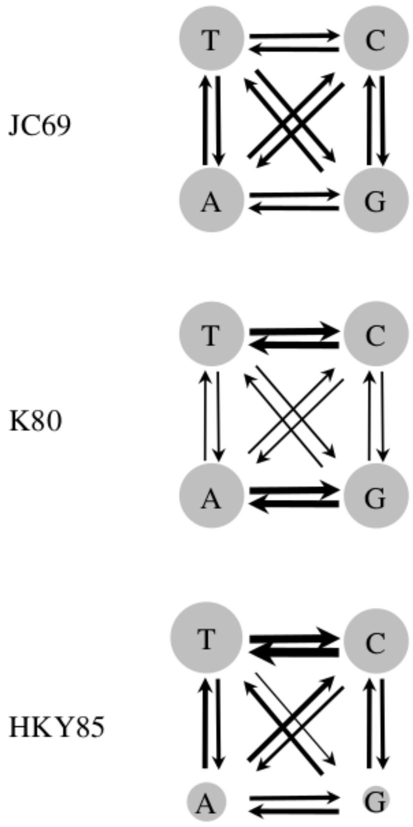


Fig. 1.2 Relative substitution rates between nucleotides under three Markov-chain models of nucleotide substitution: JC69 (Jukes and Cantor 1969), K80 (Kimura 1980), and HKY85 (Hasegawa *et al.* 1985). The thickness of the lines represents the substitution rates while the sizes of the circles represent the steady-state distribution.

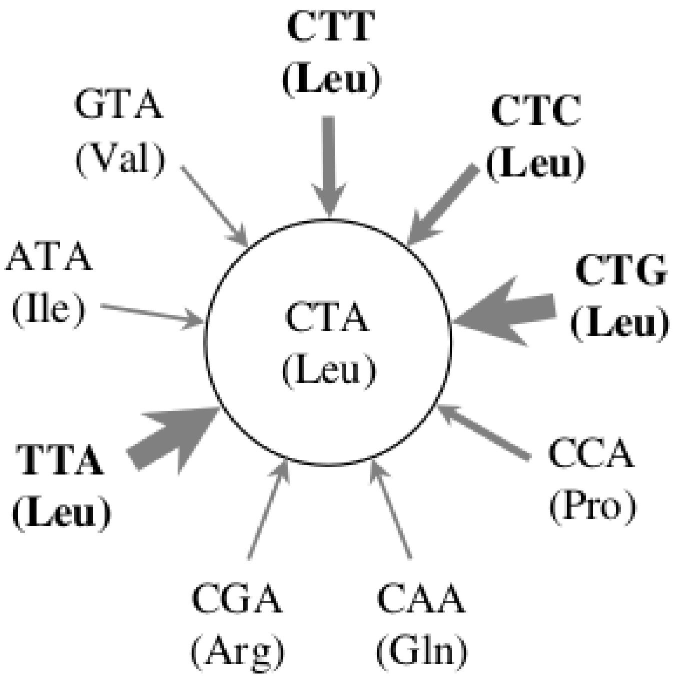
Table 1.1 Substitution-rate matrices for commonly used Markov models of nucleotide substitution

	From	To			
		T	C	A	G
JC69 (Jukes and Cantor 1969)	T	·	λ	λ	λ
	C	λ	·	λ	λ
	A	λ	λ	·	λ
	G	λ	λ	λ	·
K80 (Kimura 1980)	T	·	α	β	β
	C	α	·	β	β
	A	β	β	·	α
	G	β	β	α	·
F81 (Felsenstein 1981)	T	·	π_C	π_A	π_G
	C	π_T	·	π_A	π_G
	A	π_T	π_C	·	π_G
	G	π_T	π_C	π_A	·
HKY85 (Hasegawa *et al.* 1984, 1985)	T	·	$\alpha\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$\alpha\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	·	$\alpha\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_A$	·
F84 (Felsenstein, DNAML program since 1984)	T	·	$(1+\kappa/\pi_Y)\beta\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$(1+\kappa/\pi_Y)\beta\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	·	$(1+\kappa/\pi_R)\beta\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$(1+\kappa/\pi_R)\beta\pi_A$	·
TN93 (Tamura and Nei 1993)	T	·	$\alpha_1\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$\alpha_1\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	·	$\alpha_2\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$\alpha_2\pi_A$	·
GTR (REV) (Tavaré 1986; Yang 1994*b*; Zharkikh 1994)	T	·	$a\pi_C$	$b\pi_A$	$c\pi_G$
	C	$a\pi_T$	·	$d\pi_A$	$e\pi_G$
	A	$b\pi_T$	$d\pi_C$	·	$f\pi_G$
	G	$c\pi_T$	$e\pi_C$	$f\pi_A$	·
UNREST (Yang 1994*b*)	T	·	q_{TC}	q_{TA}	q_{TG}
	C	q_{CT}	·	q_{CA}	q_{CG}
	A	q_{AT}	q_{AC}	·	q_{AG}
	G	q_{GT}	q_{GC}	q_{GA}	·

The diagonals of the matrix are determined by the requirement that each row sums to 0. The equilibrium distribution is $\pi = (1/4, 1/4, 1/4, 1/4)$ under JC69 and K80, and $\pi = (\pi_T, \pi_C, \pi_A, \pi_G)$ under F81, F84, HKY85, TN93, and GTR. Under the general unrestricted (UNREST) model, it is given by the equations $\pi Q = 0$ under the constraint $\sum_i \pi_i = 1$.

Mutations in the coding regions

Synonymous versus nonsynonymous mutations



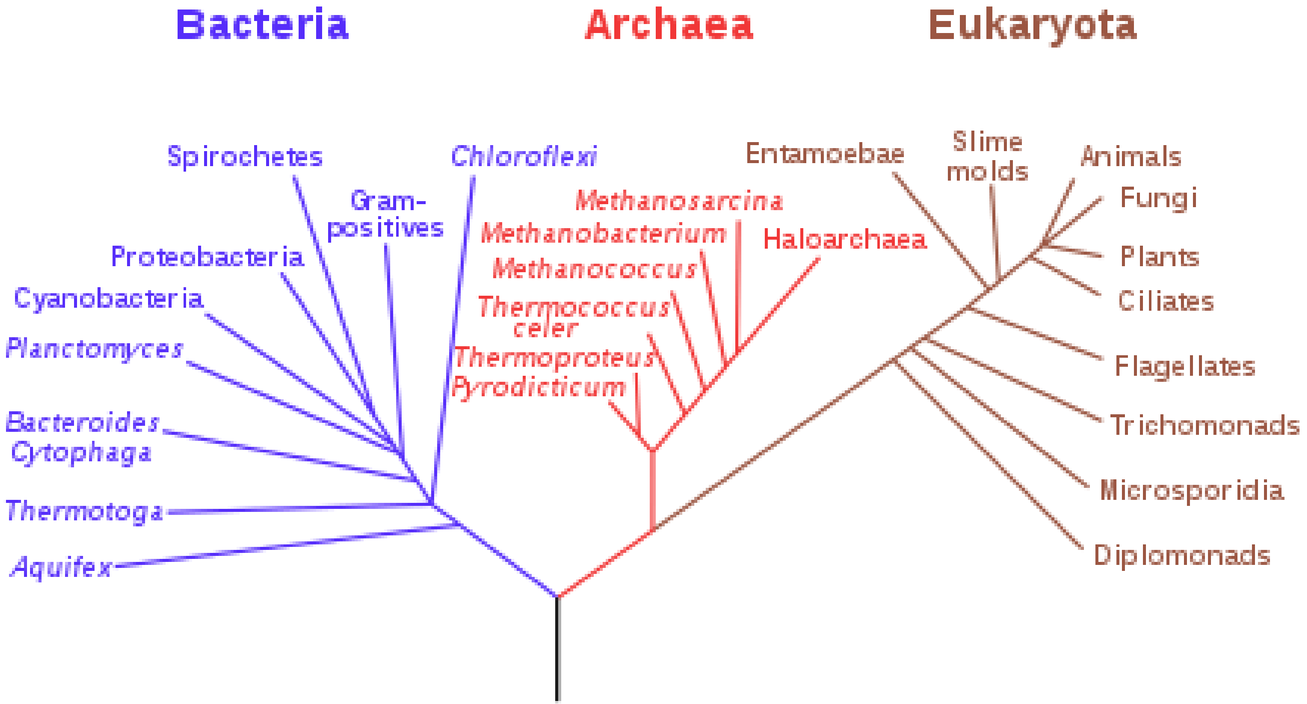
Important observations

- Most of the SNPs are neutral and are not selected by evolution. Since they occur at a fixed rate, by counting their number we can infer the time at which the process started. This the **“Molecular clock!”**

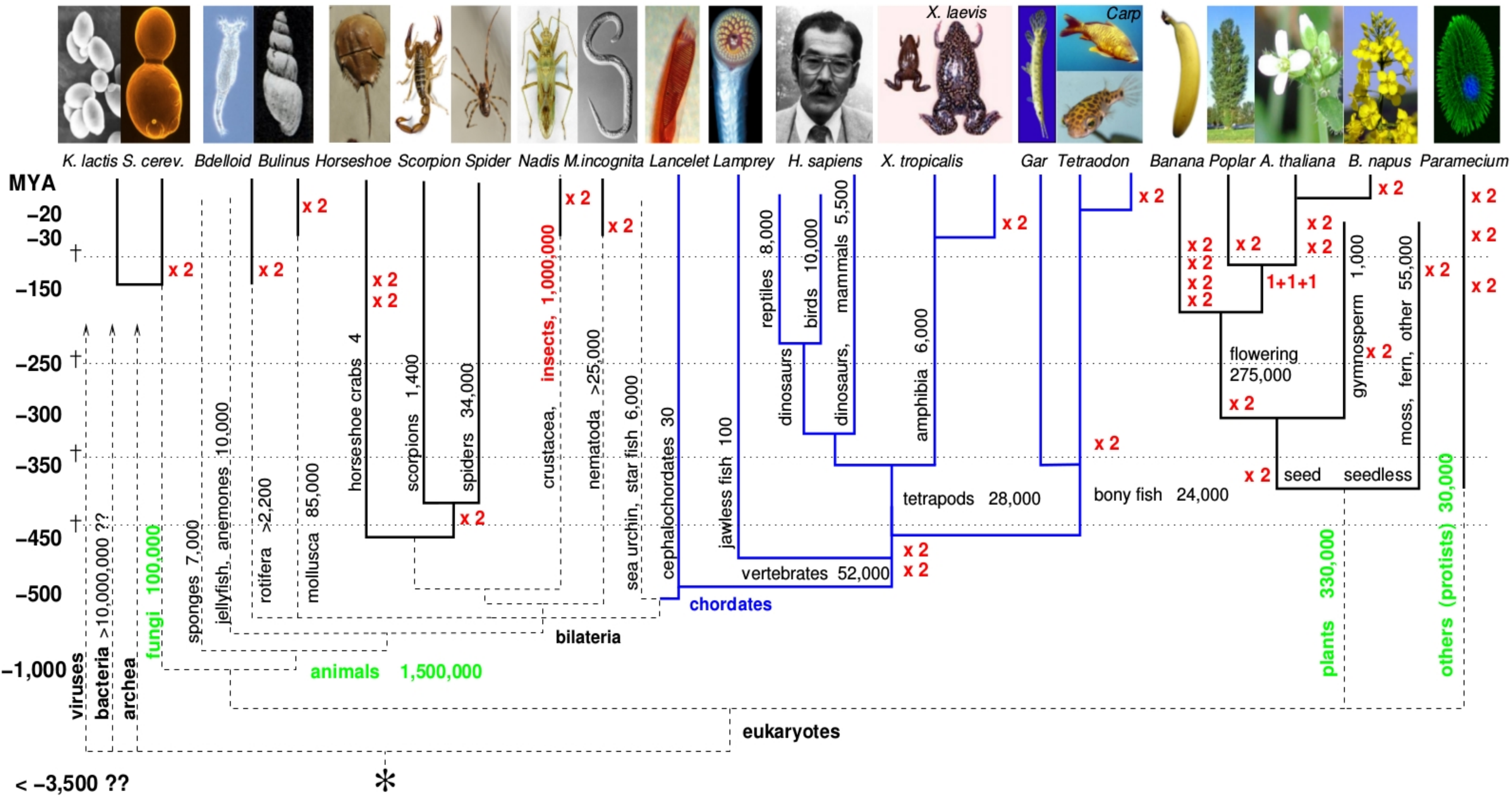
- By comparing the sequences of two species with **“alignment algorthms”** we can infer the number of mutations and thus the time of speciation. This is the **Genomic Tree of Life!**

- By comparing synonymous versus non-synonymous mutations we can measure the effect of selection: “**important regions of DNA are kept conserved under evolution**” . The same holds also for non coding regulatory regions. **Sequence conservation is a hallmark of Biological relevance!**

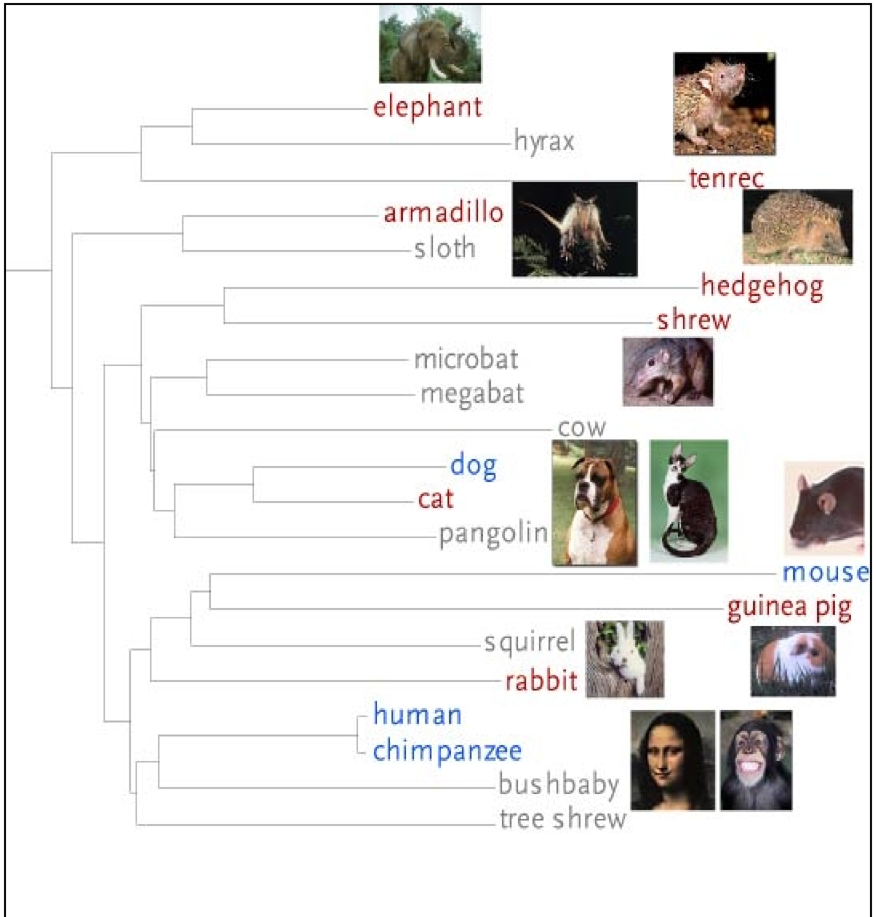
The tree of life



The tree of life: Zoom



Taxonomic versus Genomic trees

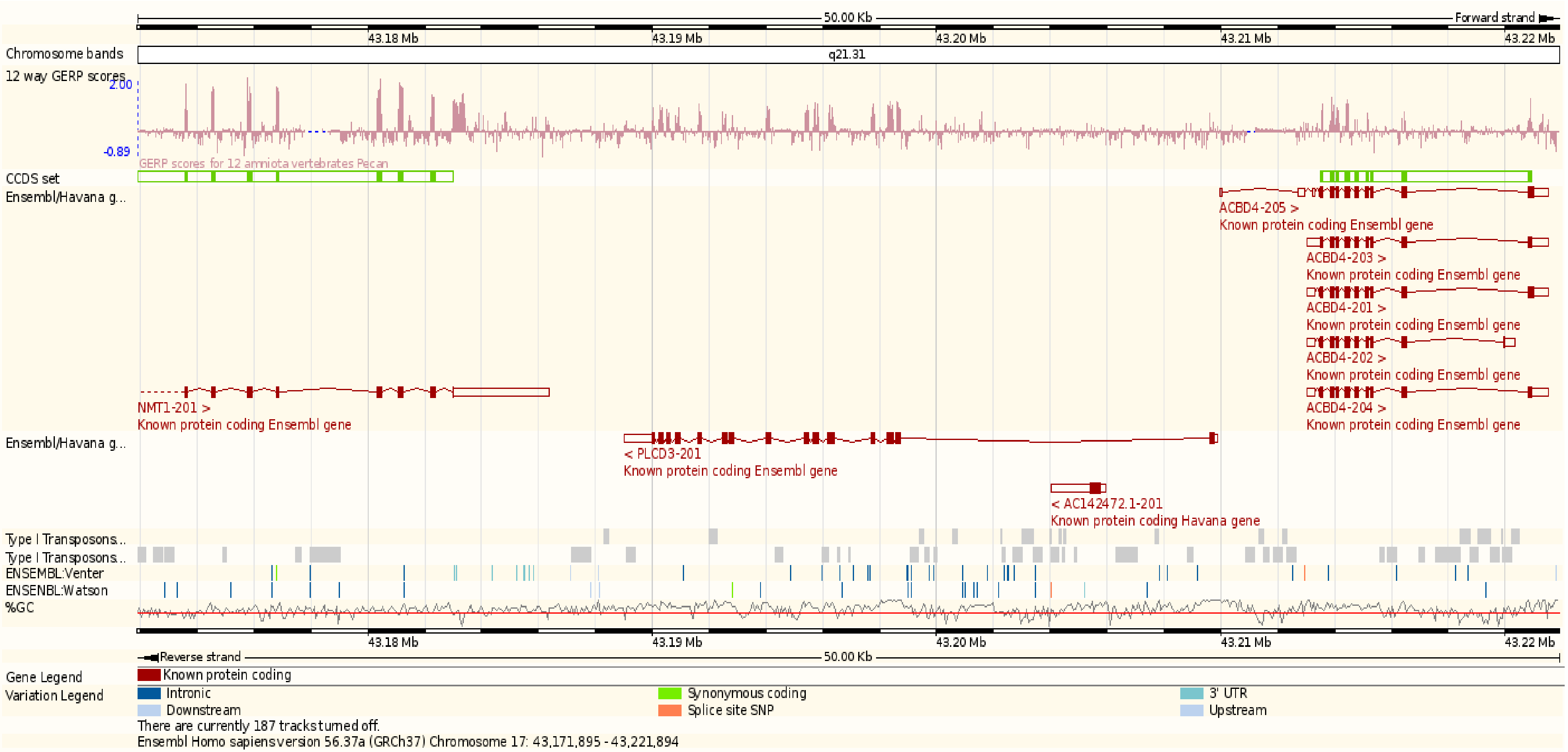


Genomic trees may be obtained using alignment algorithms. They are impressively similar to the taxonomic trees!!

This is a highly non trivial test of Evolution theory.

```
                                    Err-α
Human   GCCTGGCCGAAAATCTCTCCCGCGCGCCTGACCTTGGGTTGCCCCAGCCA
Mouse   ----------AAGCCTGTGGCGCGC-CGTGACCTTGGGCTGCCCCAGGCG
Rat     ----------AAGTTTCT---CTGC-CCTGACCTTGGGTTGCCCCAGGCG
Dog     GGCTGC----AGACCTGCCCTGAGGGAATGACCTTGGGCGGCCGCAGCGG
                  *    *       *    **********  *** ***
```

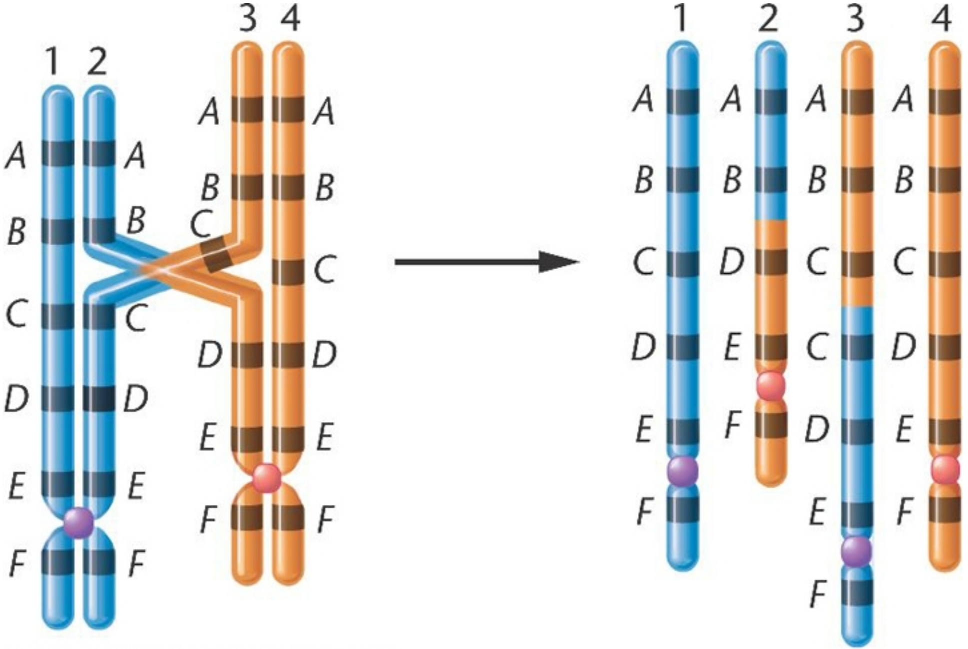
Coding and regulatory regions are conserved!



There are currently 187 tracks turned off.

Ensembl Homo sapiens version 56.37a (GRCh37) Chromosome 17: 43,171,895 - 43,221,894

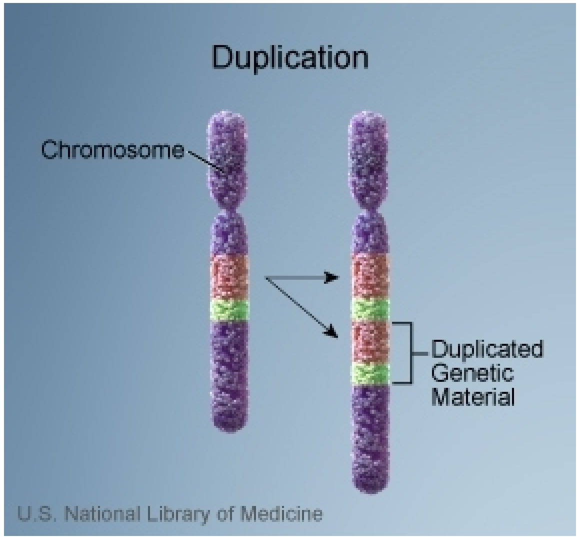
Gene duplication via unequal crossingover

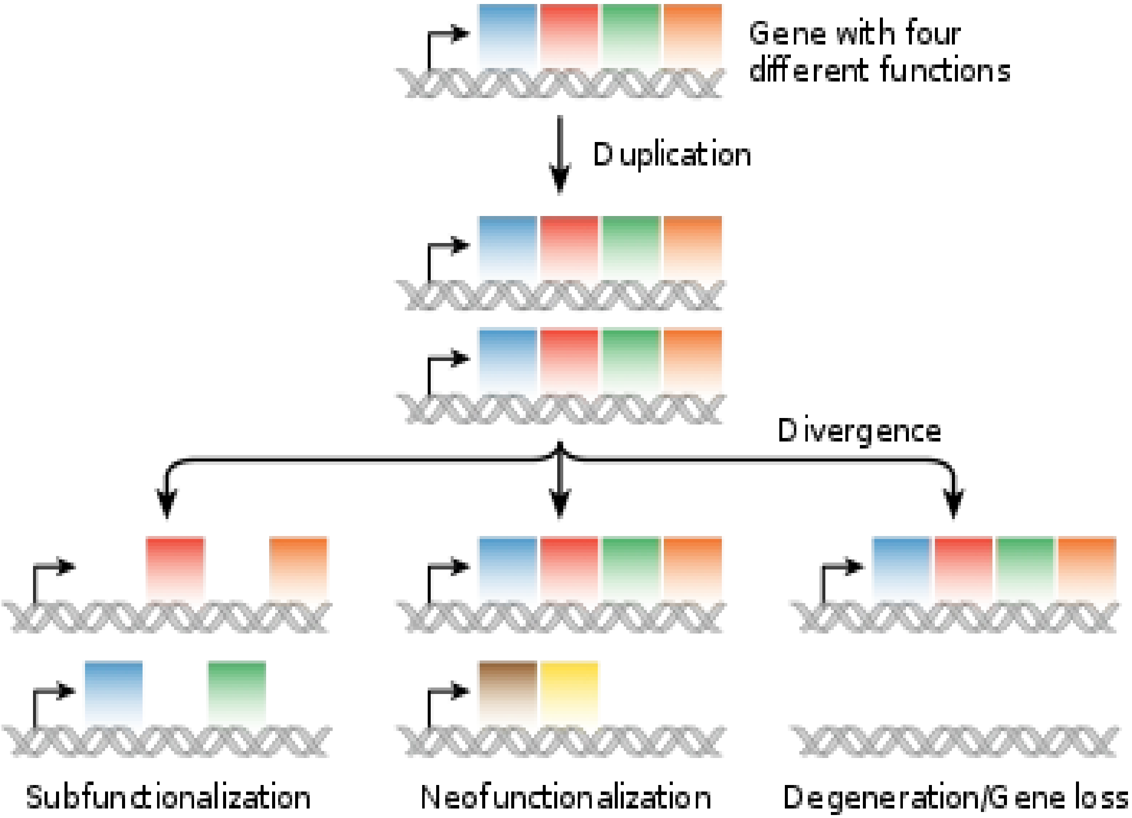


This tetrad is mispaired at meiotic synapsis.

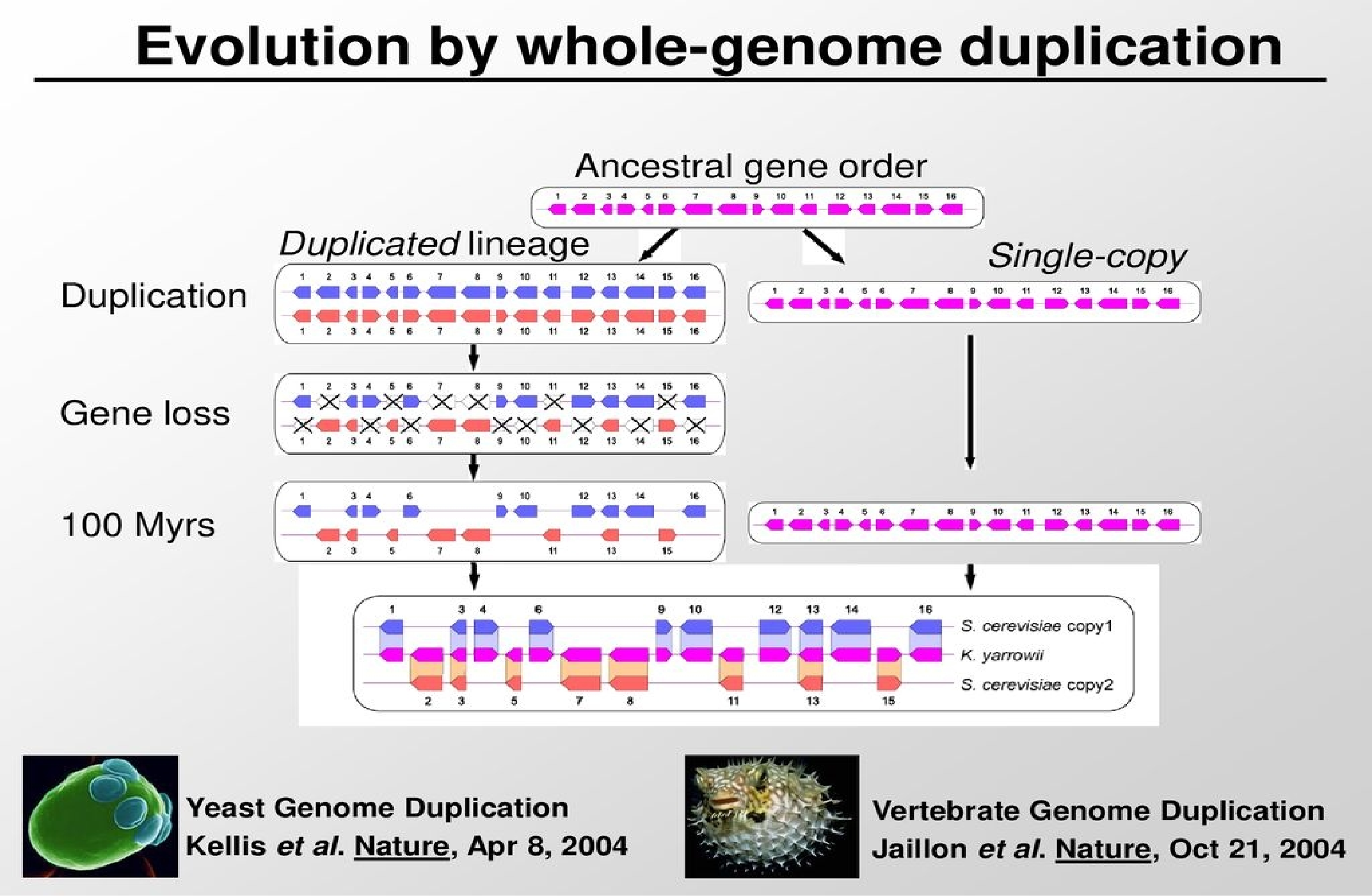
The result, after crossing over, is two unequal chromosomes: one with a duplication (3) and one with a deletion (2).

Gene duplication: moving in the phenotype space


Duplication
Chromosome
Duplicated Genetic Material
U.S. National Library of Medicine



Whole Genome duplication: jumping in the phenotype space!

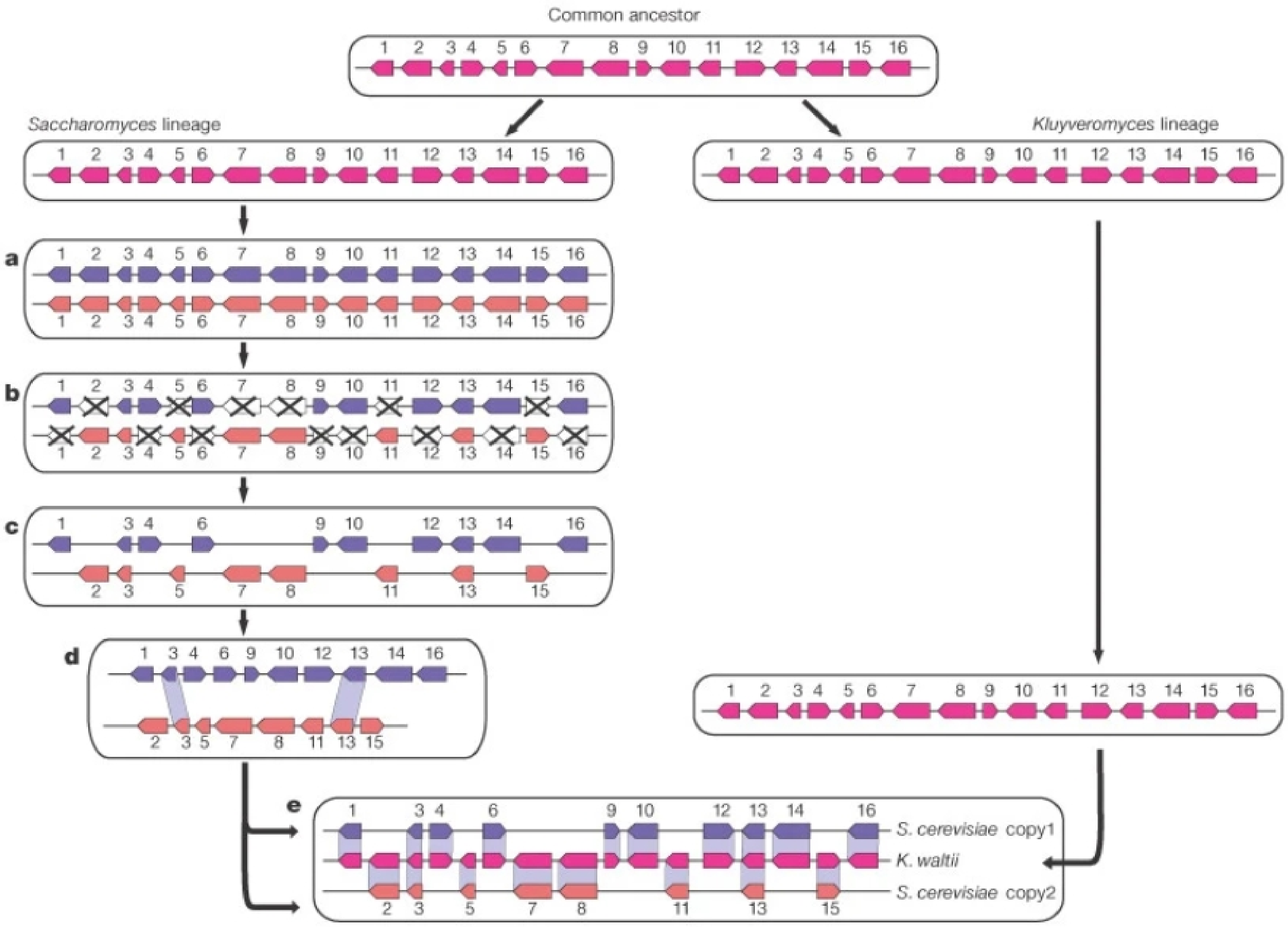


Evolution by whole-genome duplication

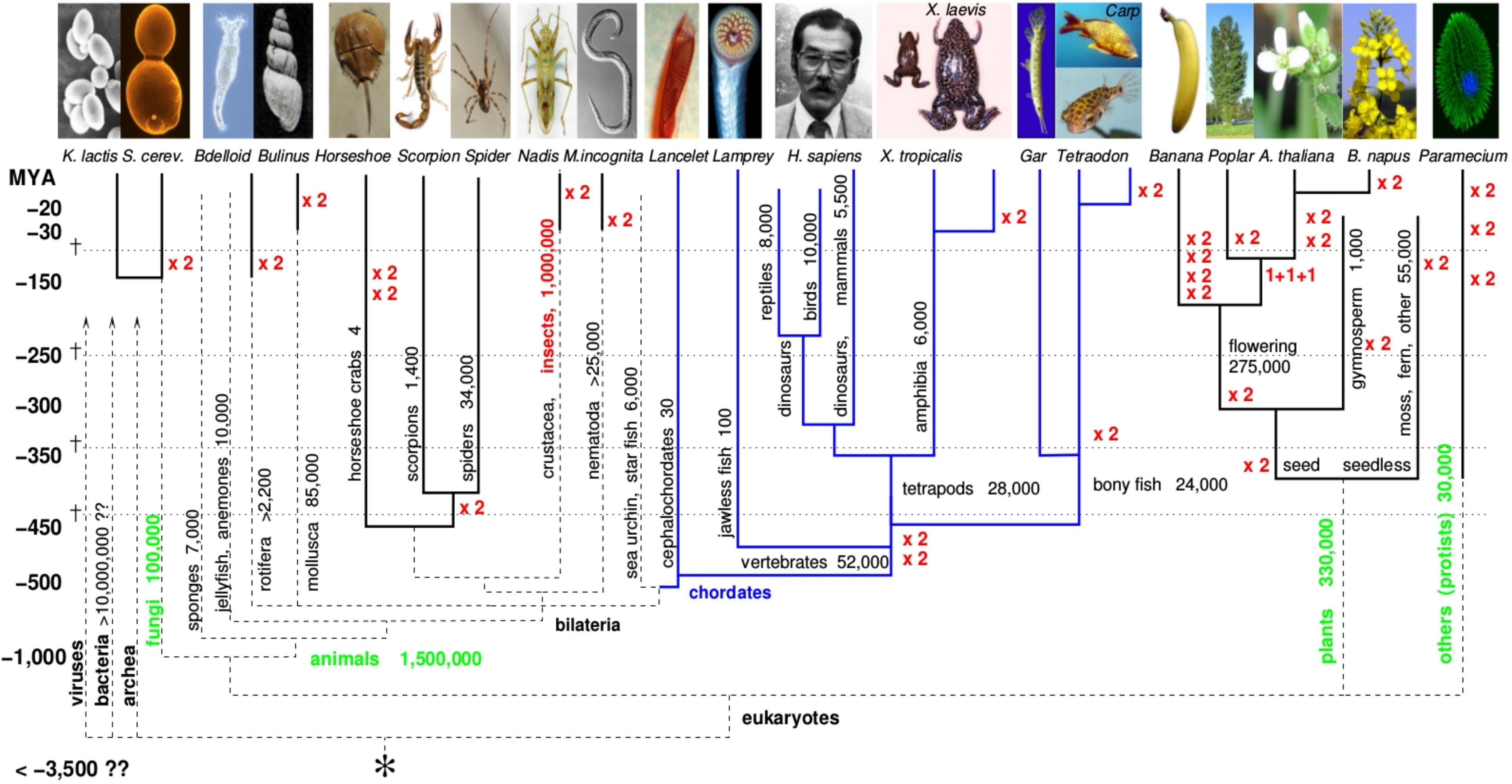
Yeast Genome Duplication
Kellis *et al.* Nature, Apr 8, 2004

Vertebrate Genome Duplication
Jaillon *et al.* Nature, Oct 21, 2004

Whole Genome duplication: the yeast case



Two rounds of WGD at the beginning of the vertebrate lineage!



Two examples of applications

§ Human/Chimp Comparison

§ Evolution of the protein-protein interaction network: the copying model

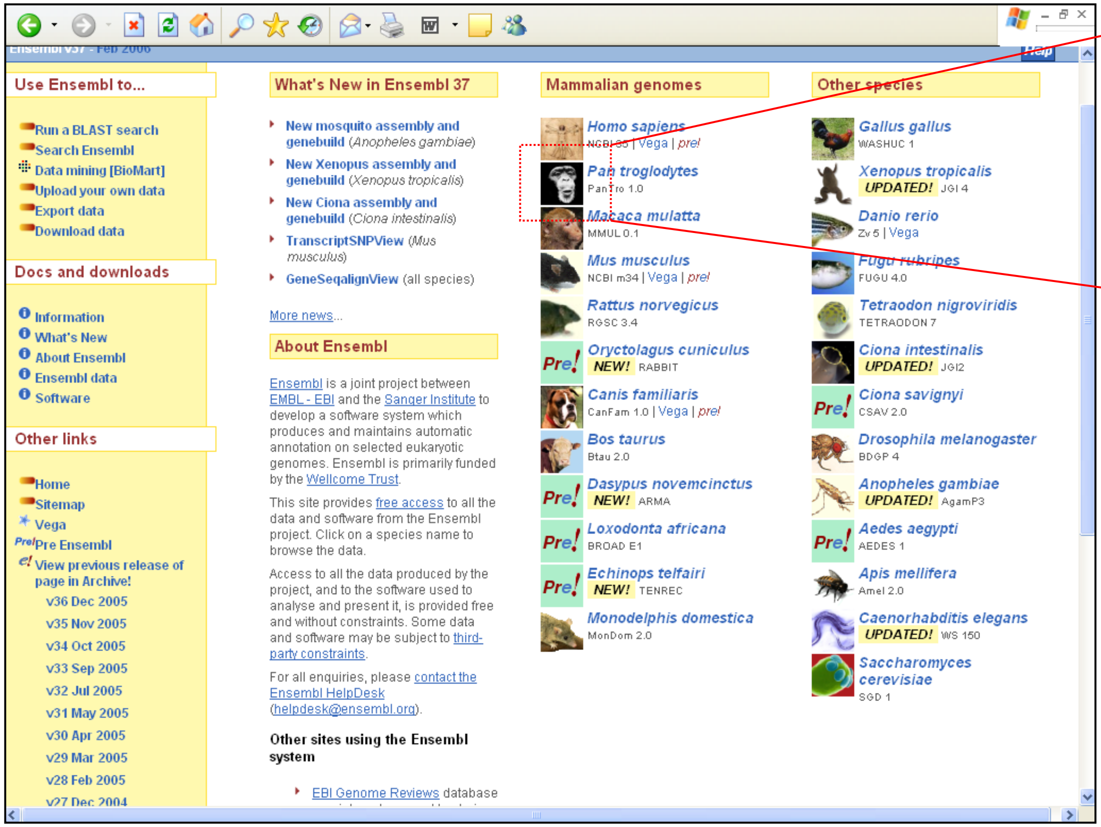
Evolution and gene regulation

- **Genomic conservation can be used as an indicator of the functional importance of a given sequence !**

The so called “Ultraconserved regions” have been protecetd by mutations for hundreds of millions of years and they are almost the same in all vertebrates.
Most likely they have a crucial functional role !

Most of them are involved in **gene regulation.** Their identification is the first step toward the construction of a regulatory voculabory in higher eukaryotes.

Humans and Chimps



Ensembl v37 - Feb 2006

Help

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
  - v36 Dec 2005
  - v35 Nov 2005
  - v34 Oct 2005
  - v33 Sep 2005
  - v32 Jul 2005
  - v31 May 2005
  - v30 Apr 2005
  - v29 Mar 2005
  - v28 Feb 2005
  - v27 Dec 2004

What's New in Ensembl 37

- New mosquito assembly and genebuild (*Anopheles gambiae*)
- New Xenopus assembly and genebuild (*Xenopus tropicalis*)
- New Ciona assembly and genebuild (*Ciona intestinalis*)
- TranscriptSNPView (*Mus musculus*)
- GeneSeqalignView (all species)

More news...

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to third-party constraints.

For all enquiries, please contact the Ensembl HelpDesk (helpdesk@ensembl.org).

Other sites using the Ensembl system

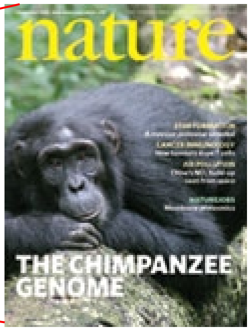
- EBI Genome Reviews database

Mammalian genomes

	Homo sapiens	NCBI 35 | Vega | *pre!*
	Pan troglodytes	PanTro 1.0
	Macaca mulatta	MMUL 0.1
	Mus musculus	NCBI m34 | Vega | *pre!*
	Rattus norvegicus	RGSC 3.4
Pre!	*Oryctolagus cuniculus*	*NEW!* RABBIT
	Canis familiaris	CanFam 1.0 | Vega | *pre!*
	Bos taurus	Btau 2.0
Pre!	*Dasypus novemcinctus*	*NEW!* ARMA
Pre!	*Loxodonta africana*	BROAD E1
Pre!	*Echinops telfairi*	*NEW!* TENREC
	Monodelphis domestica	MonDom 2.0

Other species

	Gallus gallus	WASHUC 1
	Xenopus tropicalis	*UPDATED!* JGI 4
	Danio rerio	Zv 5 | Vega
	Fugu rubripes	FUGU 4.0
	Tetraodon nigroviridis	TETRAODON 7
	Ciona intestinalis	*UPDATED!* JGI2
Pre!	*Ciona savignyi*	CSAV 2.0
	Drosophila melanogaster	BDGP 4
	Anopheles gambiae	*UPDATED!* AgamP3
Pre!	*Aedes aegypti*	AEDES 1
	Apis mellifera	Amel 2.0
	Caenorhabditis elegans	*UPDATED!* WS 150
	Saccharomyces cerevisiae	SGD 1



96% of the human genome coincides with the chimp’s one! Most of the differences are random non-coding SNPs!

Humans and Chimps

Big Problem: Human and Chimps are too similar from a genomic point of view!!

- **Idea: use evolutionary conservation in the other way around: Look for a region ultraconserved in all vertebrates but mutated in human!**

One of these regions is found within the coding sequence of the FOXP2 gene

FOXP2 !!

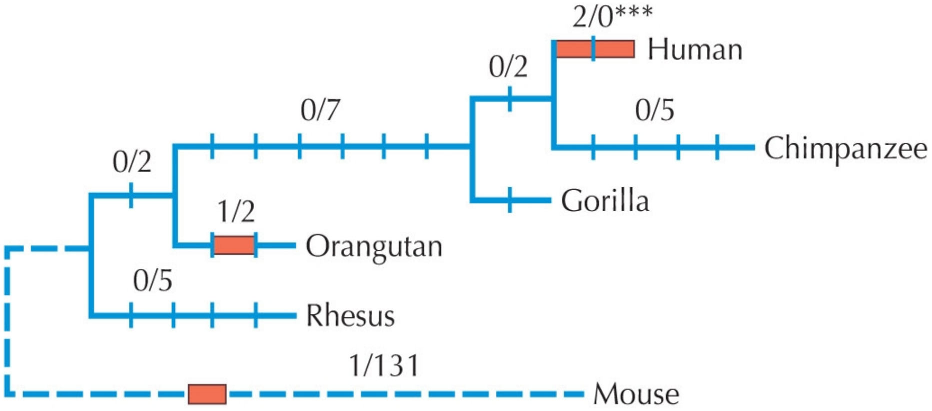


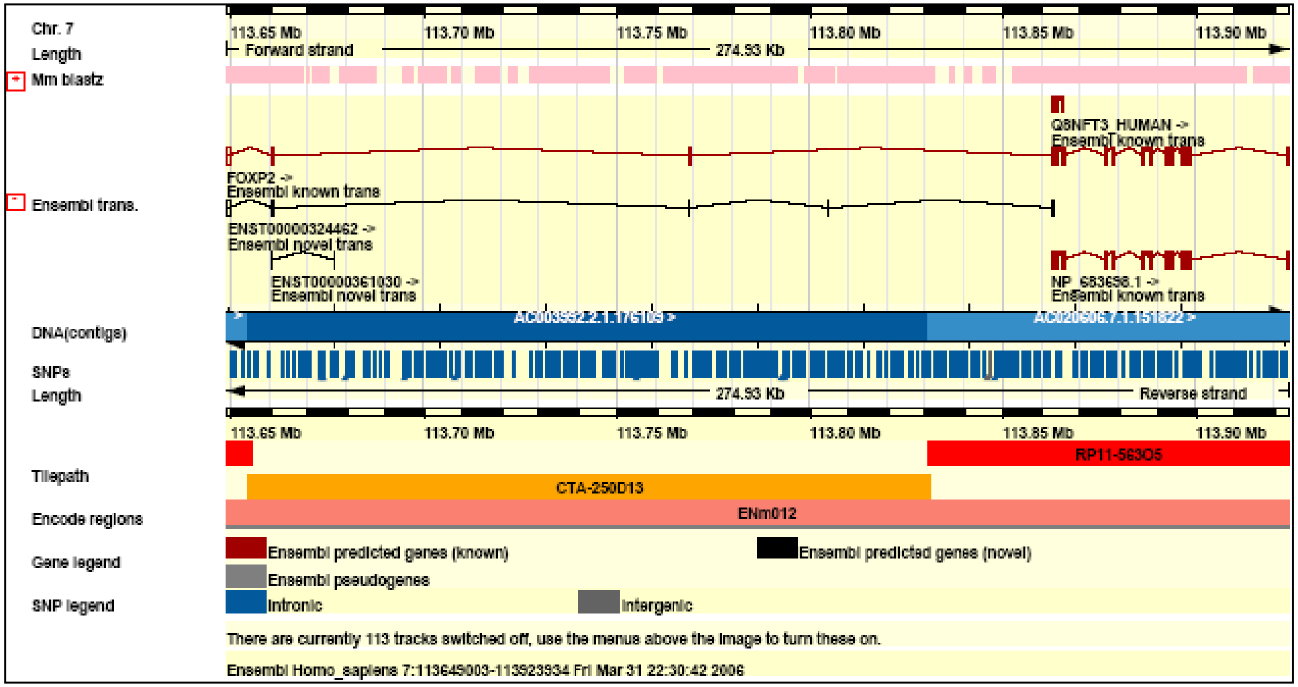
FIGURE 25.27. The *FOXP2* gene is the first gene identified that carries a mutation that causes a specific language deficit in humans. The silent and replacement nucleotide substitutions in this gene as mapped on a primate phylogeny are shown. (*Red bars*) Amino acid changes; (*blue tick marks*) nucleotide changes. Data suggest that the *FOXP2* gene has been the target of selection during recent human evolution after the separation of the human lineage from the common ancestor with the chimpanzee. Numbers show how many nonsynonymous/synonymous changes have occurred along each branch.

25.27, adapted from Enard W. et al., *Nature* **418:** 869–872, © 2002 Macmillan, www.nature.com

Evolution © 2007 Cold Spring Harbor Laboratory Press

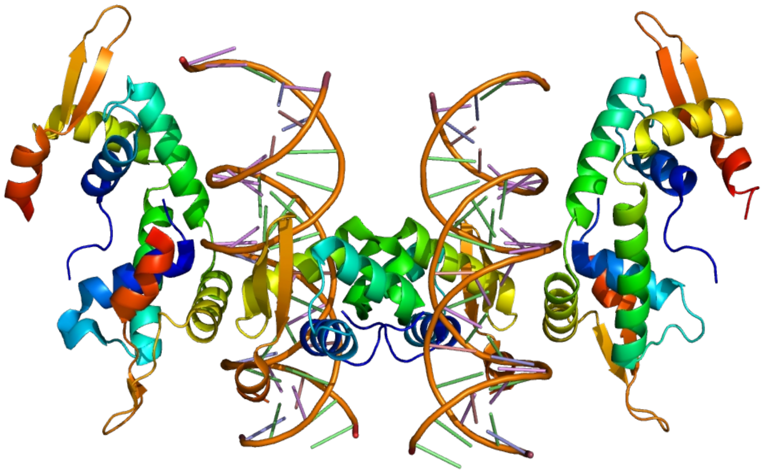
FOXP2 !!

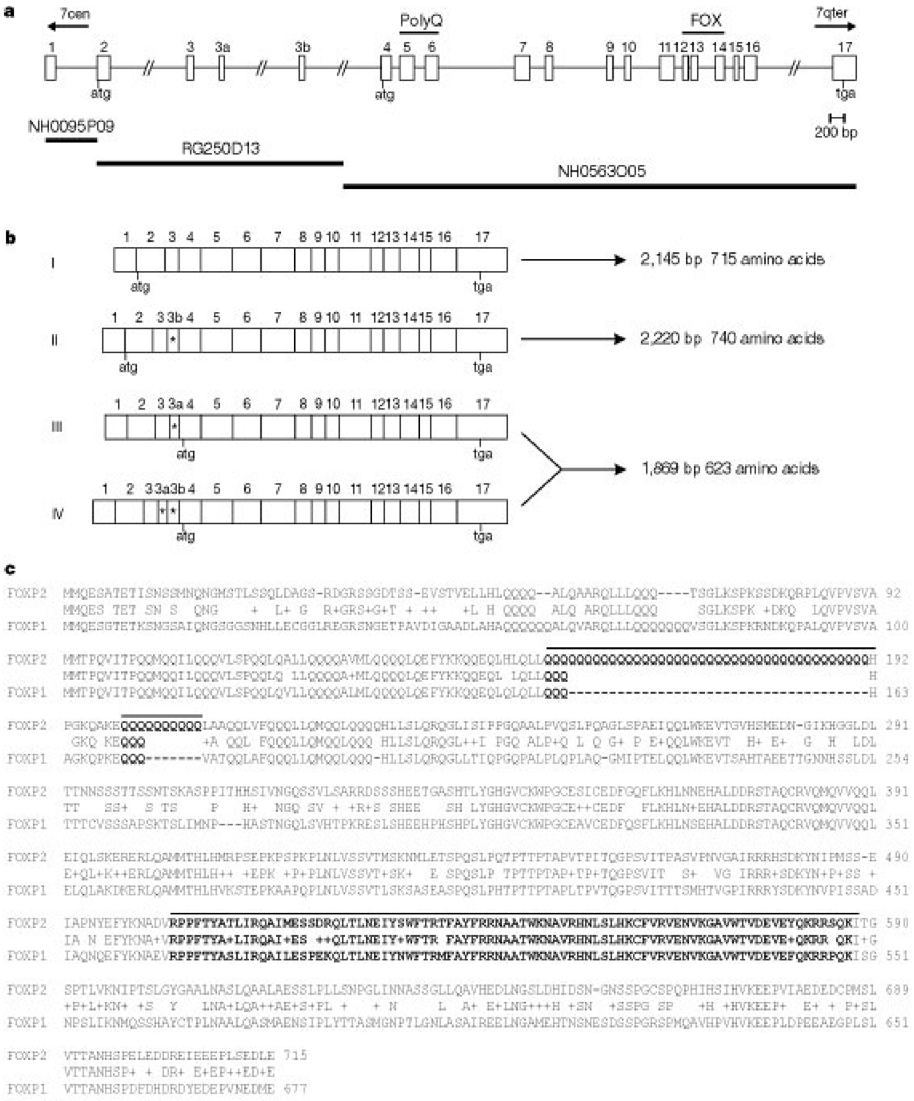
Mutations (SNPs) in the FOXP2 gene are associated to deep alterations in speaking ability.



There are currently 113 tracks switched off, use the menus above the image to turn these on.

Ensembl Homo_sapiens 7:113649003-113923934 Fri Mar 31 22:30:42 2006





```
FOXP2  MMQESATETISNSSMNQNGMSTLSSQLDAGS-RDGRSSGDTSS-EVSTVELLHLQQQQ--ALQAARQLLLQQQ----TSGLKSPKSSDKQRPLQVPVSVA 92
       MMQES TET SN S  QNG     +  L+ G  R+GRS+G+T + ++   +L H QQQQ  ALQ ARQLLLQQQ     SGLKSPK +DKQ  LQVPVSVA
FOXP1  MMQESGTETKSNGSAIQNGSGGSNHLLECGGLREGRSNGETPAVDIGAADLAHAQQQQQALQVARQLLLQQQQQQVSGLKSPKRNDKQPALQVPVSVA 100

FOXP2  MMTPQVITPQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQH 192
       MMTPQVITPQQMQQILQQQVLSPQQLQ LLQQQQA+MLQQQQLQEFYKKQQEQL LQLLQQQ                             H
FOXP1  MMTPQVITPQQMQQILQQQVLSPQQLQVLLQQQQALMLQQQQLQEFYKKQQEQLQLQLLQQQ-----------------------------H 163

FOXP2  PGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQGLISIPPGQAALPVQSLPQAGLSPAEIQQLWKEVTGVHSMEDN-GIKHGGLDL 291
        GKQ KEQQQ       +A QQL FQQQLLQMQQLQQQ HLLSLQRQGL++I PGQ ALP+Q L Q G+ P E+QQLWKEVT  H+ E+  G  H  LDL
FOXP1  AGKQPKEQQQ-------VATQQLAFQQQLLQMQQLQQQ-HLLSLQRQGLLTIQPGQPALPLQPLAQ-GMIPTELQQLWKEVTSAHTAEETTGNNHSSLDL 254

FOXP2  TTNNSSSTTSSNTSKASPPITHHSIVNGQSSVLSARRDSSSHEETGASHTLYGHGVCKWPGCESICEDFGQFLKHLNNEHALDDRSTAQCRVQMQVVQQL 391
       TT    SS+  S TS     P    H+   NGQ SV + +R+S SHEE     SH LYGHGVCKWPGCE++CEDF   FLKHLN+EHALDDRSTAQCRVQMQVVQQL
FOXP1  TTTCVSSSAPSKTSLIMNP---HASTNGQLSVHTPKRESLSHEEHPHSHPLYGHGVCKWPGCEAVCEDFQSFLKHLNSEHALDDRSTAQCRVQMQVVQQL 351

FOXP2  EIQLSKERERLQAMMTHLHMRPSEPKPSPKPLNLVSSVTMSKNMLETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVGAIRRRHSDKYNIPMSS-E 490
       E+QL+K++ERLQAMMTHLH++  +EPK  +P+PLNLVSSVT+SK+   E SPQSLP TPTTPTAP+TP+TQGPSVIT   S+   VG IRRR+SDKYN+P+SS +
FOXP1  ELQLAKDKERLQAMMTHLHVKSTEPKAAPQPLNLVSSVTLSKSASEASPQSLPHTPTTPTAPLTPVTQGPSVITTTSMHTVGPIRRRYSDKYNVPISSAD 451

FOXP2  IAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFAYFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITG 590
       IA N EFYKNA+VRPPFTYA+LIRQAI+ES ++QLTLNEIY+WFTR FAYFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVE+QKRR QKI+G
FOXP1  IAQNQEFYKNAEVRPPFTYASLIRQAILESPEKQLTLNEIYNWFTRMFAYFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEFQKRRPQKISG 551

FOXP2  SPTLVKNIPTSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSN-GNSSPGCSPQPHIHSIHVKEEPVIAEDEDCPMSL 689
       +P+L+KN+ +S  Y   LNA+LQA++AE+S+PL +    + N      L  A+ E+LNG+++H +SN   +SSPG SP    +H +HVKEEP+  E+ + P+SL
FOXP1  NPSLIKNMQSSHAYCTPLNAALQASMAENSIPLYTTASMGNPTLGNLASAIREELNGAMEHTNSNESDGSPGRSPMQAVHPVHVKEEPLDPEEAEGPLSL 651

FOXP2  VTTANHSPELEDDREIEEEPLSEDLE 715
       VTTANHSP+ + DR+ E+EP++ED+E
FOXP1  VTTANHSPDFDHDRDYEDEPVNEDME 677
```

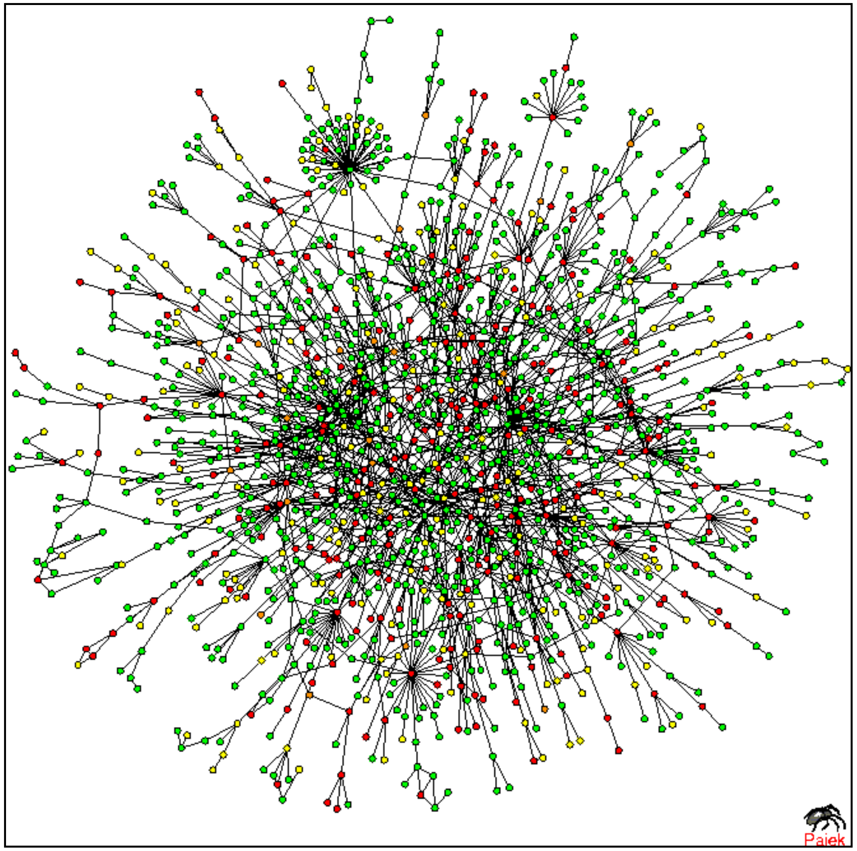
The human version of FOXP2 could be at the origin of the ability of H. Sapiens to articulate words and thus organize complex chasing strategies...

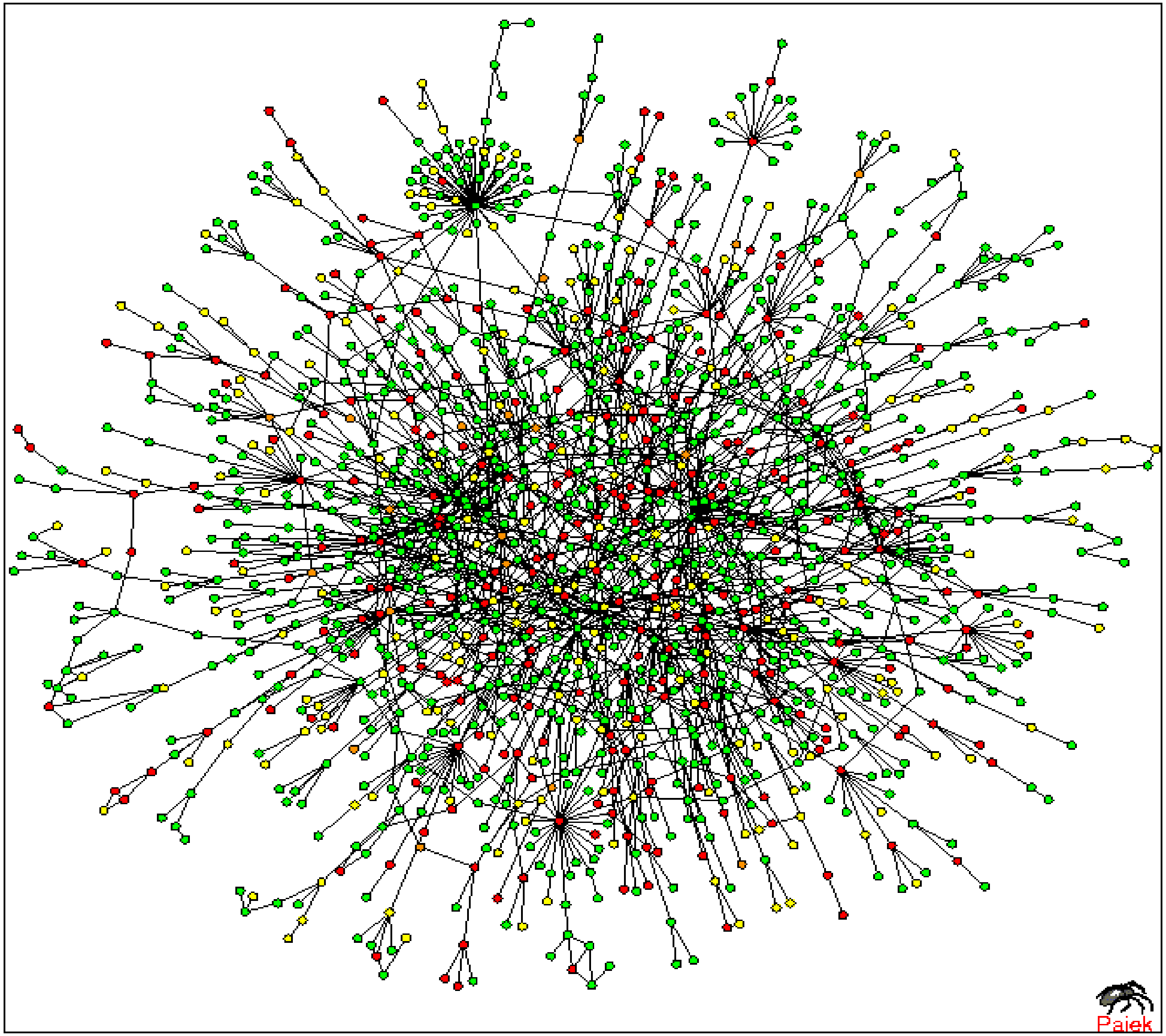
Modern Genomics: *networks*

- genes and proteins of a given organism are organized in networks .

- Cells react to external stimuli in a “global” way.

H.Jeong et al.
Nature, 411 (2001) 41




Pajek

Interaction Networks

- All the interaction networks in biological systems are "**heterogeneous**", with a "fat-tailed" connettivity distribution: few hubs and several peripheral links.

- Standard explanations such as "**preferential attachment**" models used to describe WWW topology cannot work for biological systems.

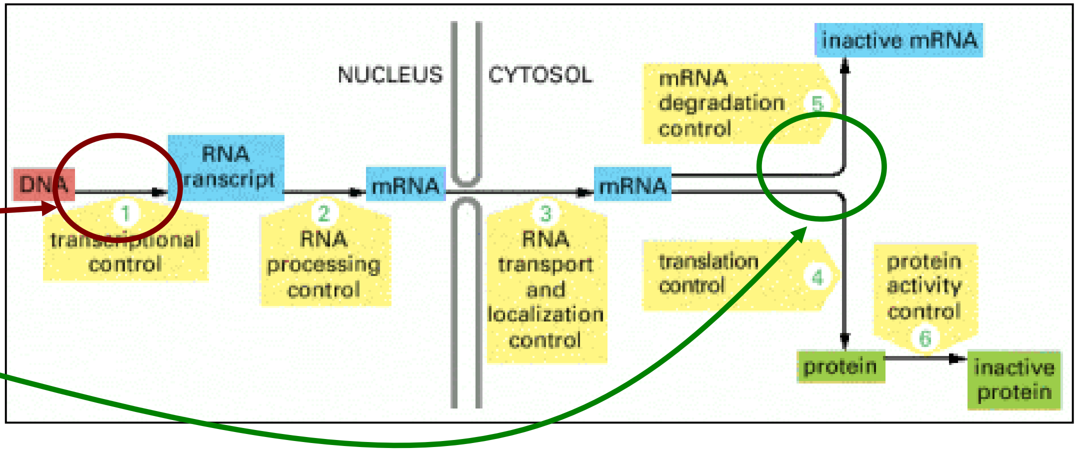
- Evolutionary models based on gene duplication, the so called "**copying models**" lead to a biologically rooted explanation which fits the available data very well

Gene Regulation

Gene expression is tightly regulated. All cells in the body carry the full set of genes, but only express about 20% of them at any particular time. Different proteins are expressed in different cells (neurons, muscle cells....) according to the different functions of the cell.

Among the various regulatory steps the most important ones are:

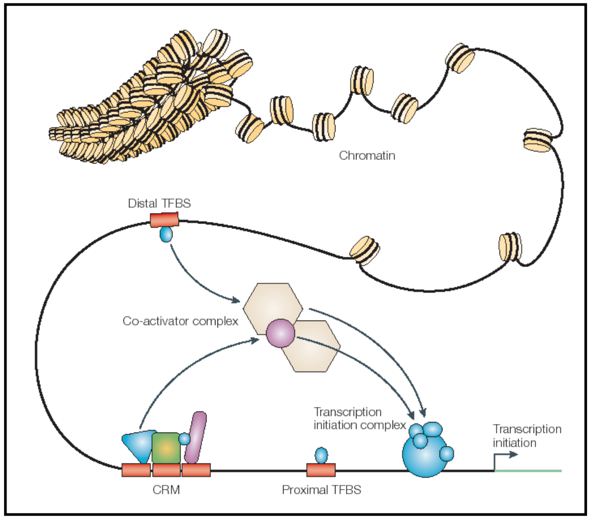
- transcriptional control, by **Transcription Factors.**
- post-transcriptional control, by **microRNAs.**



Alberts, *Molecular Biology of the Cell*

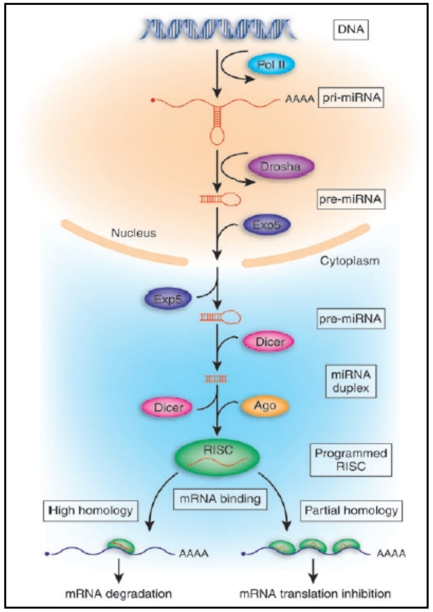
Transcription Factors and miRNAs

Transcription Factors (TFs): proteins binding to specific recognition **motifs (TFBSs)** usually short (5-10 bp) and located **upstream** of the coding region of the regulated gene.

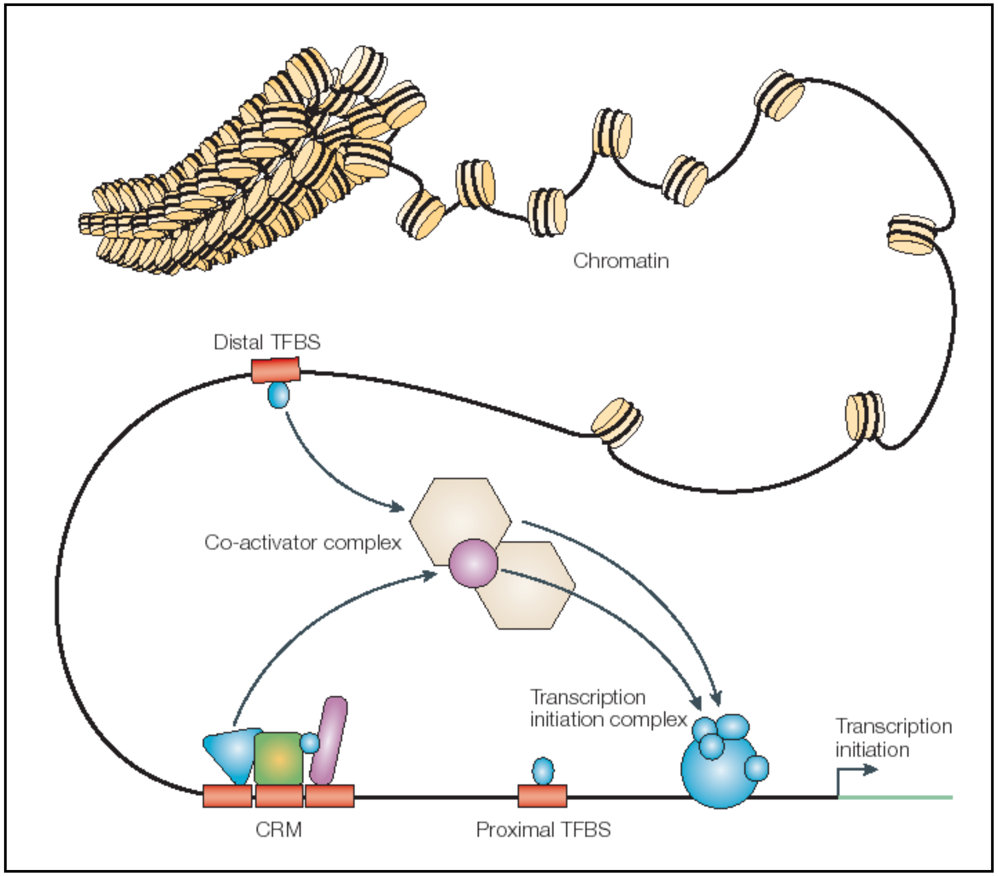


Wassermann, Nat. Rev. Genetics

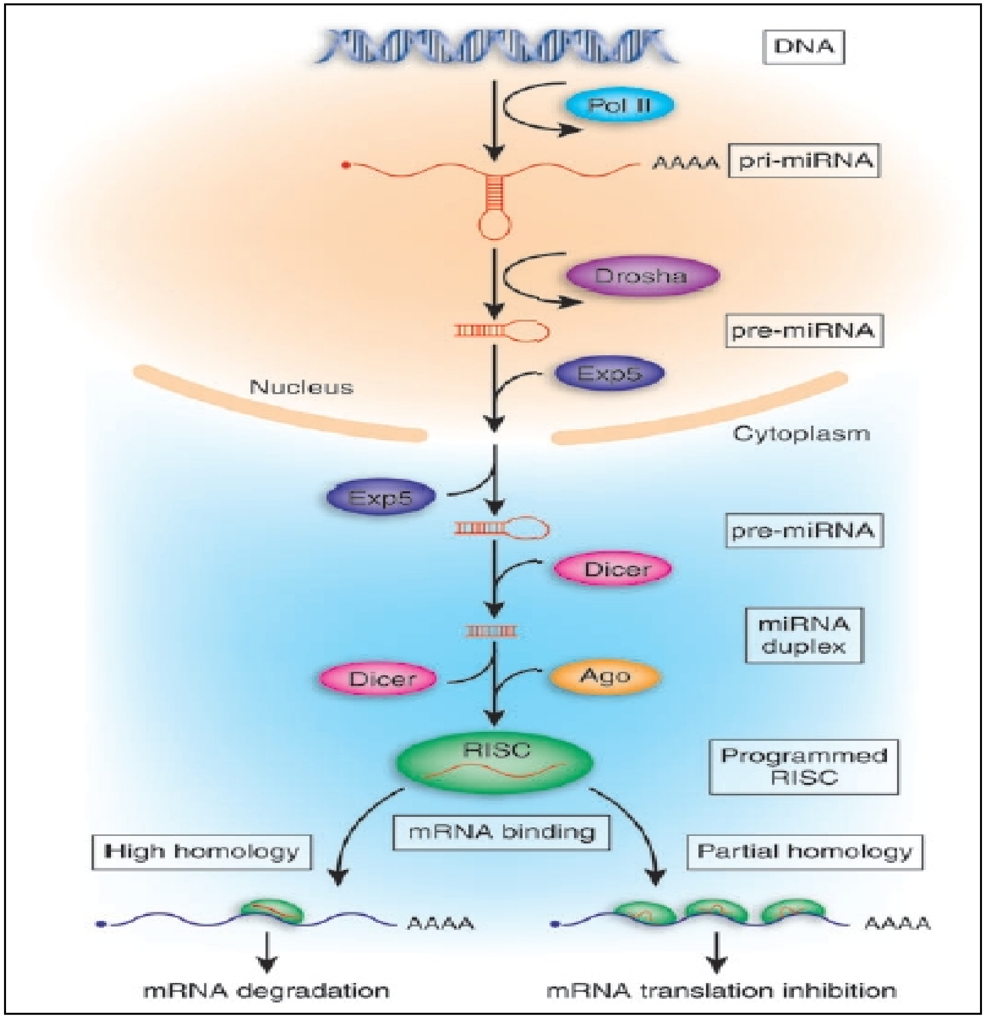
MicroRNAs (miRNAs) are a family of small RNAs (typically **21 - 25** nucleotide long) that **negatively regulate gene expression at the posttranscriptional level**, (usually) thanks to the “seed” region in 3’-UTR regions.



Transcription Factors



MicroRNAs

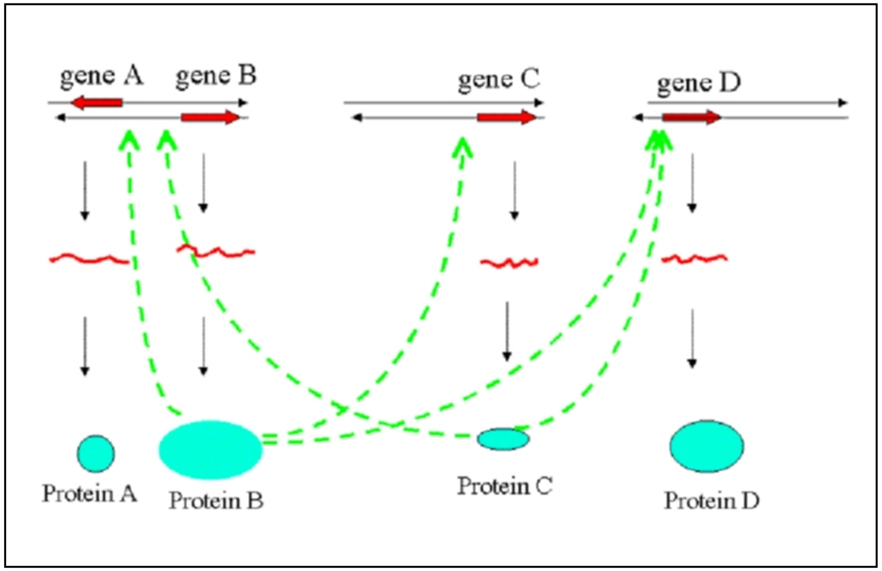


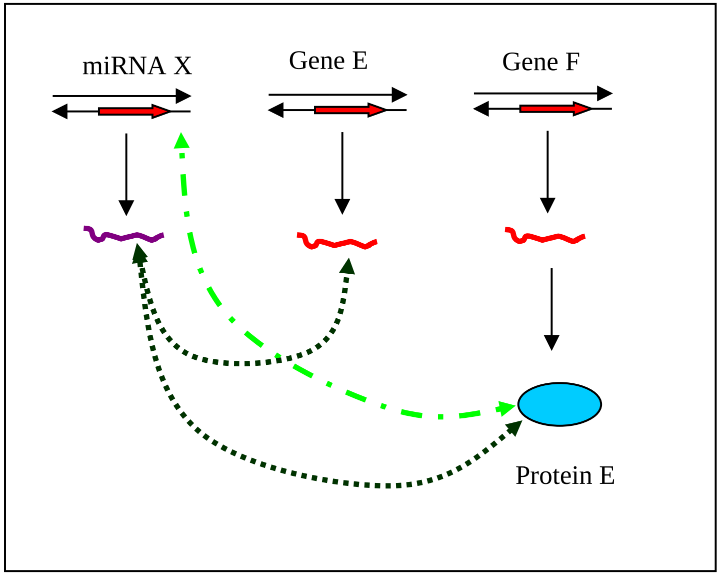
He L. , Hannon GJ. Nature Review Genetics 5, 522 - 531 (2004)

Regulatory Networks 1

Key 1 --> **TFs** are themselves proteins produced by other genes, and they act in a combinatorial way, resulting in a complex network of interactions between genes and their products.
--> Transcriptional Network

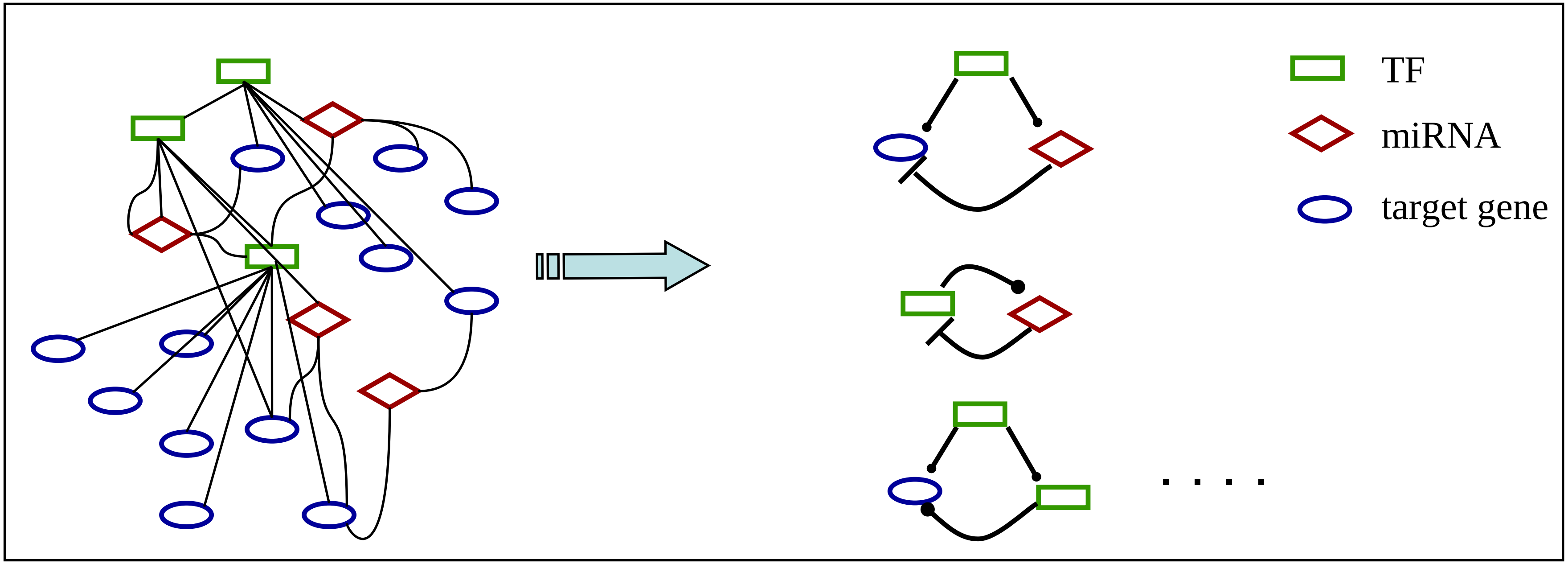
miRNAs also act in a combinatorial and one-to-many way, and, moreover, <u>are transcribed from same POL-II promotes of TFs</u>.
--> Post-Transcriptional Network





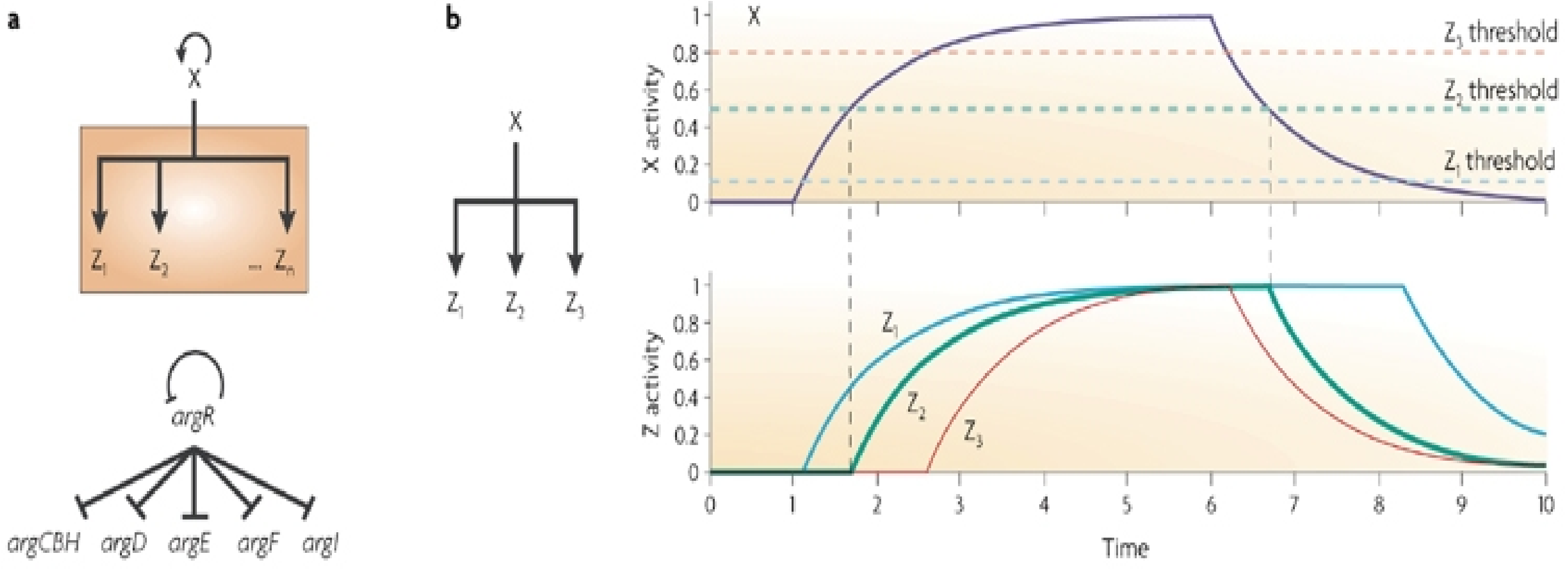
Regulatory Networks 2

Key 2 --> Biological functions are performed by groups of genes which act in an interdependent and synergic way. A complex network can be divided into simpler, distinct regulatory patterns called **network motifs**, typically composed by 3 or 4 interacting components which are able to perform elementary signal processing functions.



Network motifs

Example: SIM (Single Input Module) (a) experimental realization: arginine biosynthesis b) Circuit behaviour: different genes are activated at different times as a function of their different activation threshold as the concentration of X (master regulator) changes in time R.Milo et al. Science 298 (2002) 824



Nature Reviews | Genetics

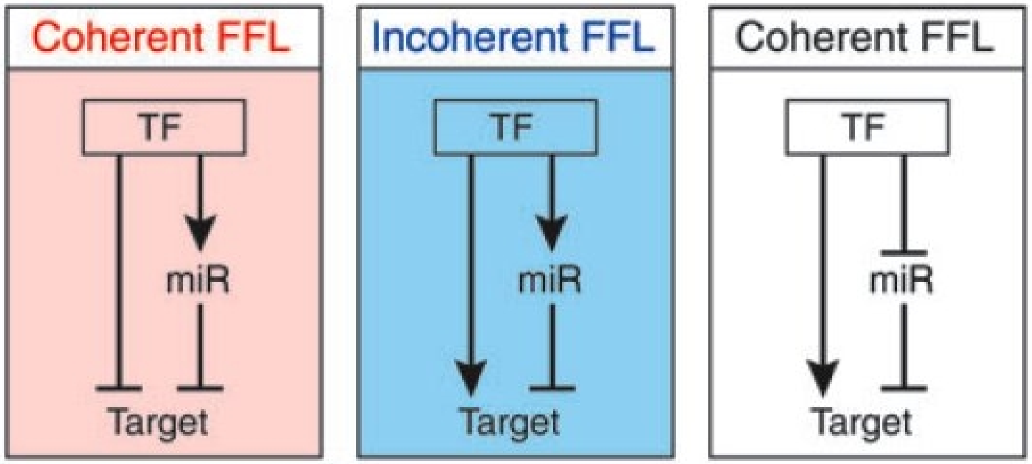
Network Motifs II

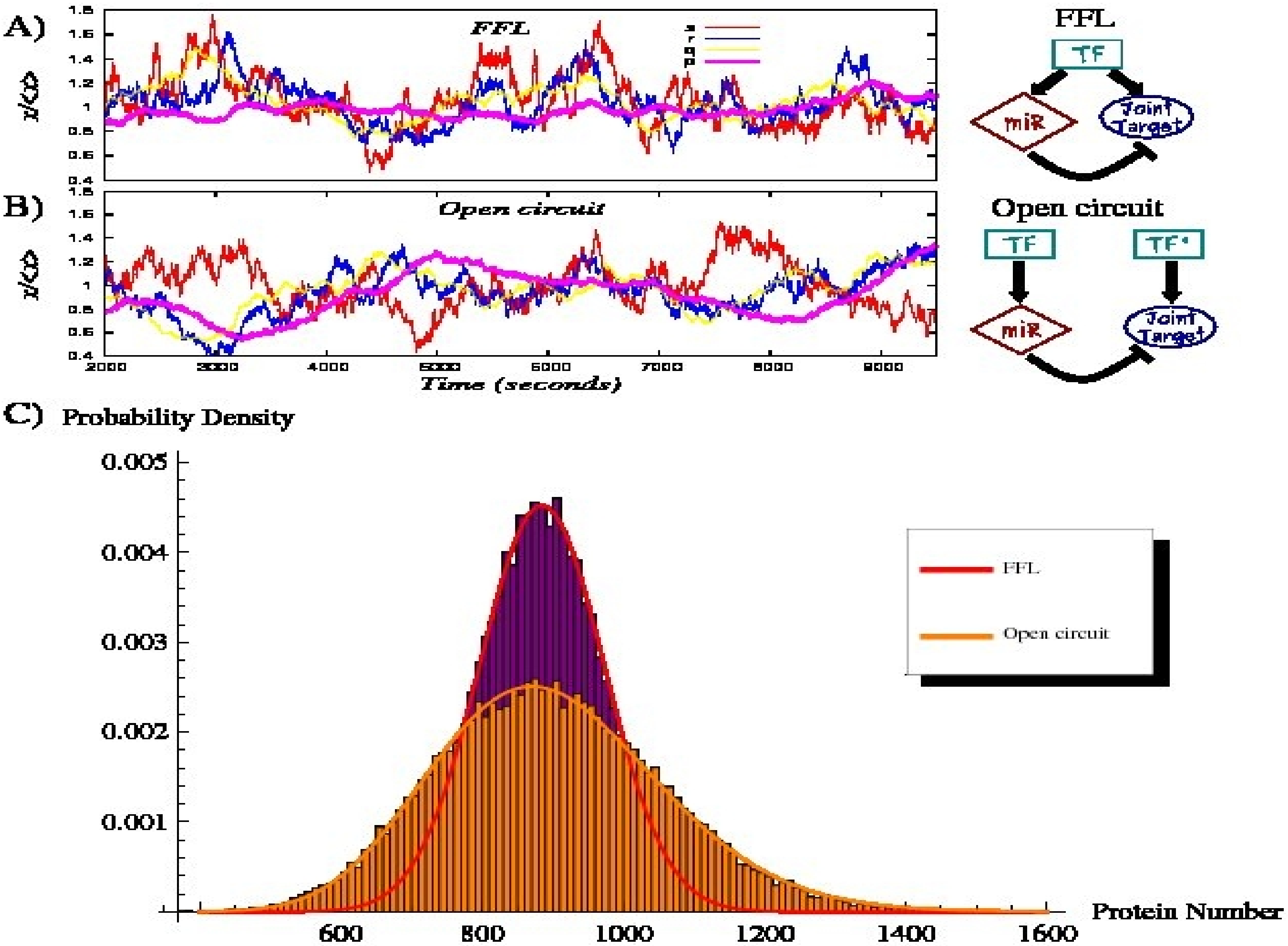
Network motifs can be studied using standard tools of theoretical physics:

- Ordinary differential equations
- Stochastic equations
- Montecarlo (Gillespie) simulations.

- Goal: understand the functional role of the motif and why it was selected by evolution

- Example 1: incoherent feedforward loops can reduce the noise in the amount of produced proteins.




A)
FFL
s
r
g
p
x/<x>
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
FFL
TF
miR
Joint Target
B)
Open circuit
x/<x>
2000
3000
4000
5000
6000
7000
8000
9000
Time (seconds)
Open circuit
TF
TF'
miR
Joint Target
C)
Probability Density
0.005
0.004
0.003
0.002
0.001
600
800
1000
1200
1400
1600
Protein Number
FFL
Open circuit

Cancer Driver Genes

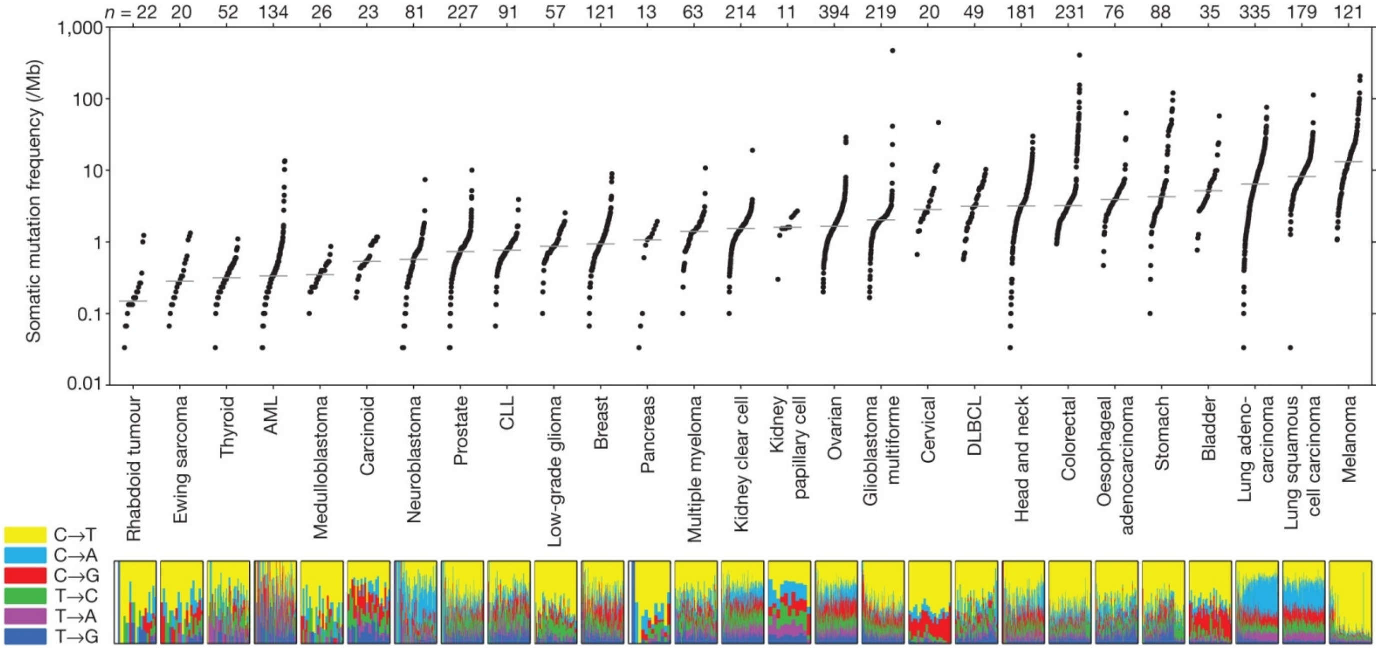
Cancer is the result of a pathological alteration of the regulatory network induced by somatic mutations (to be distinguished from the germinal mutations) and chromosomal alterations (Copy Number Variations)

Main outcome of the recent genomic studies: each tumour is unique!
→ new therapeutic approach: “Personalized Medicine / Precision Medicine”

Can we identify on purely computational basis the drivers of this disregulation process?

Problem: in a typical cancer cell we find thousands of altered genes. Can we identify the real drivers?

Somatic mutation frequencies observed in exomes from 3,083 tumour–normal pairs.



MS Lawrence *et al. Nature* **000**, 1-5 (2013) doi:10.1038/nature12213

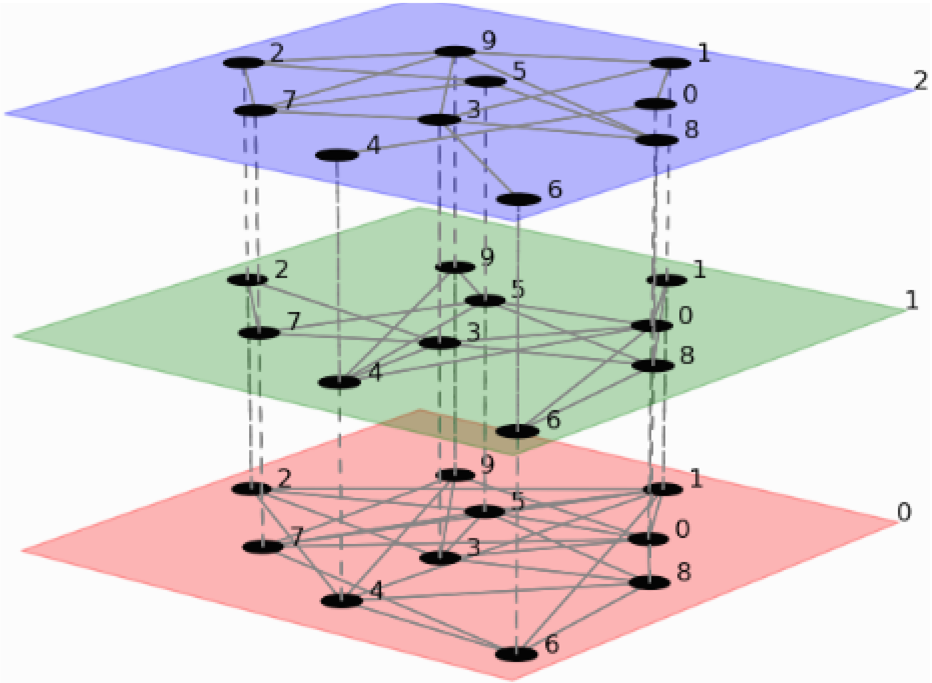
Cancer driver genes

Goal: integrate different sources of regulatory information using Network Theory to identify driver genes in cancer

TOOLS:

- Multiplex Theory
- Mutual Information (ARACNE)
- Filtering Methods (Disparity Filter)
- Community detection algorithms (Infomap, OSLOM, Label propagation, Louvain and Modularity optimization via simulated annealing)
- Consensus Clustering

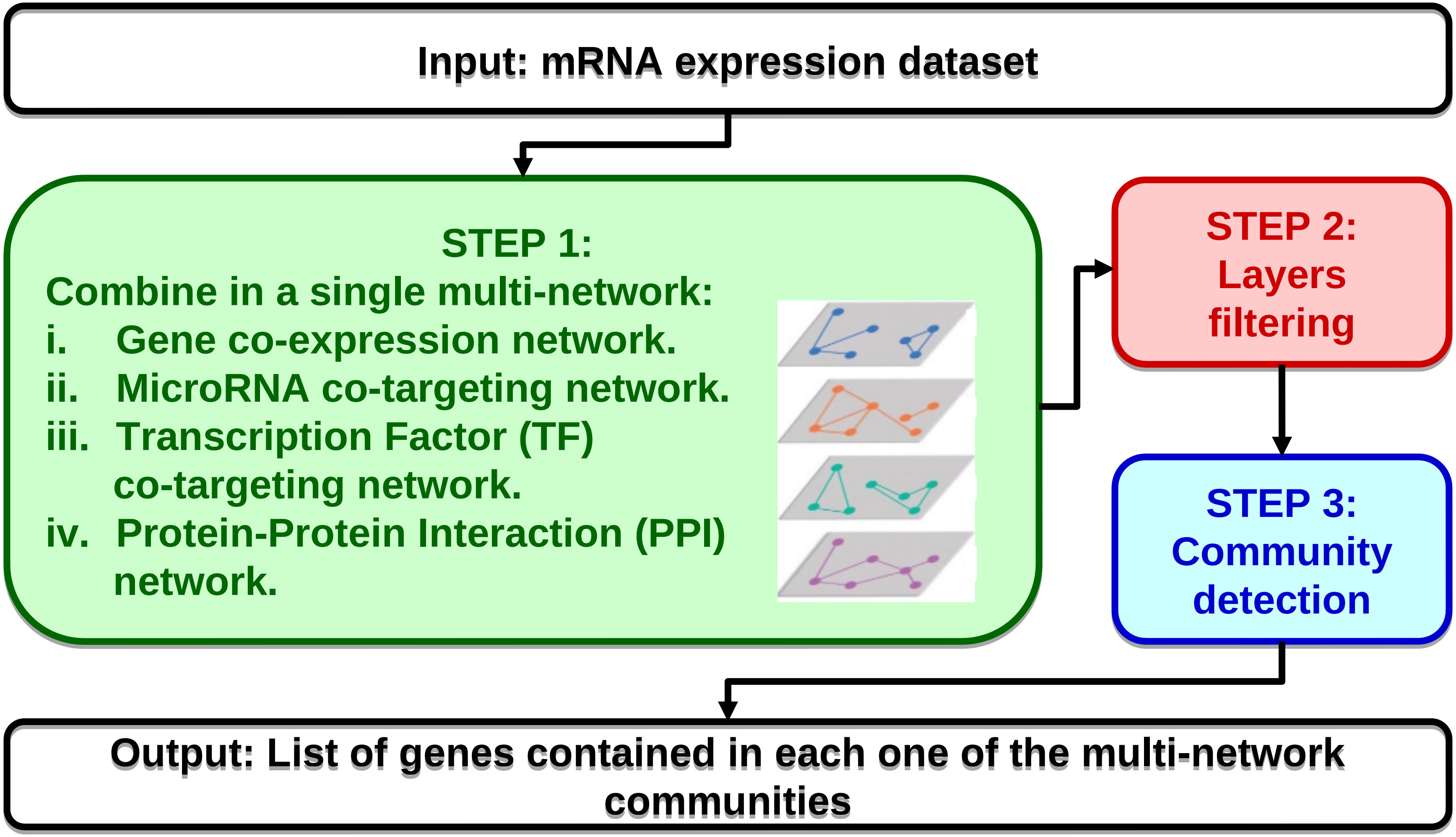
Proposal: Multi-Network Integration of Regulatory Information



Cantini L., Medico E., Fortunato S. and Caselle M.,

"Detection of gene communities in multi-networks reveals cancer drivers."

Scientific Reports (2015) **5**: 17386



Input: mRNA expression dataset

STEP 1:

Combine in a single multi-network:

i. Gene co-expression network.

ii. MicroRNA co-targeting network.

iii. Transcription Factor (TF) co-targeting network.

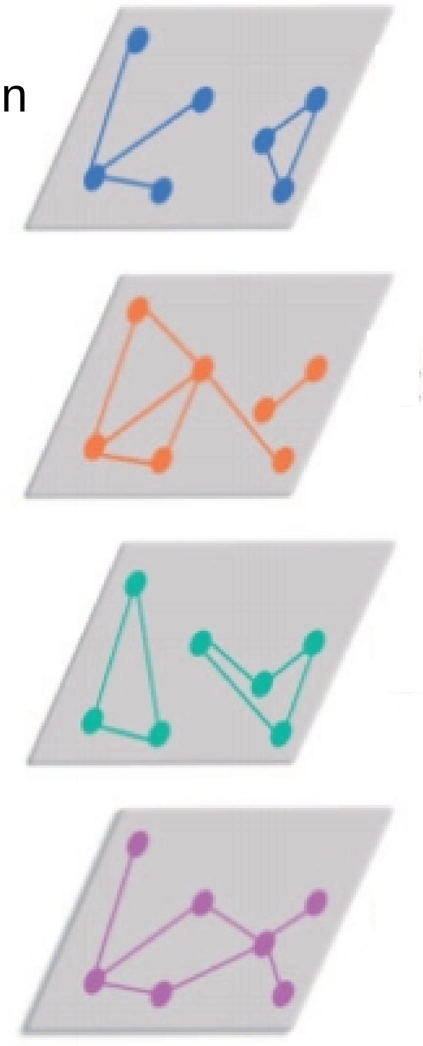
iv. Protein-Protein Interaction (PPI) network.

STEP 2: Layers filtering

STEP 3: Community detection

Output: List of genes contained in each one of the multi-network communities

Proposal: Multi-Network Integration of Regulatory Information



L1: expression

L2: miRNA

L3: TF

L4: PPI

Multi-network communities reconstruction:

1. Community detection within each network layer
2. Consensus clustering across the four layers.

Open source Community detection algorithms:

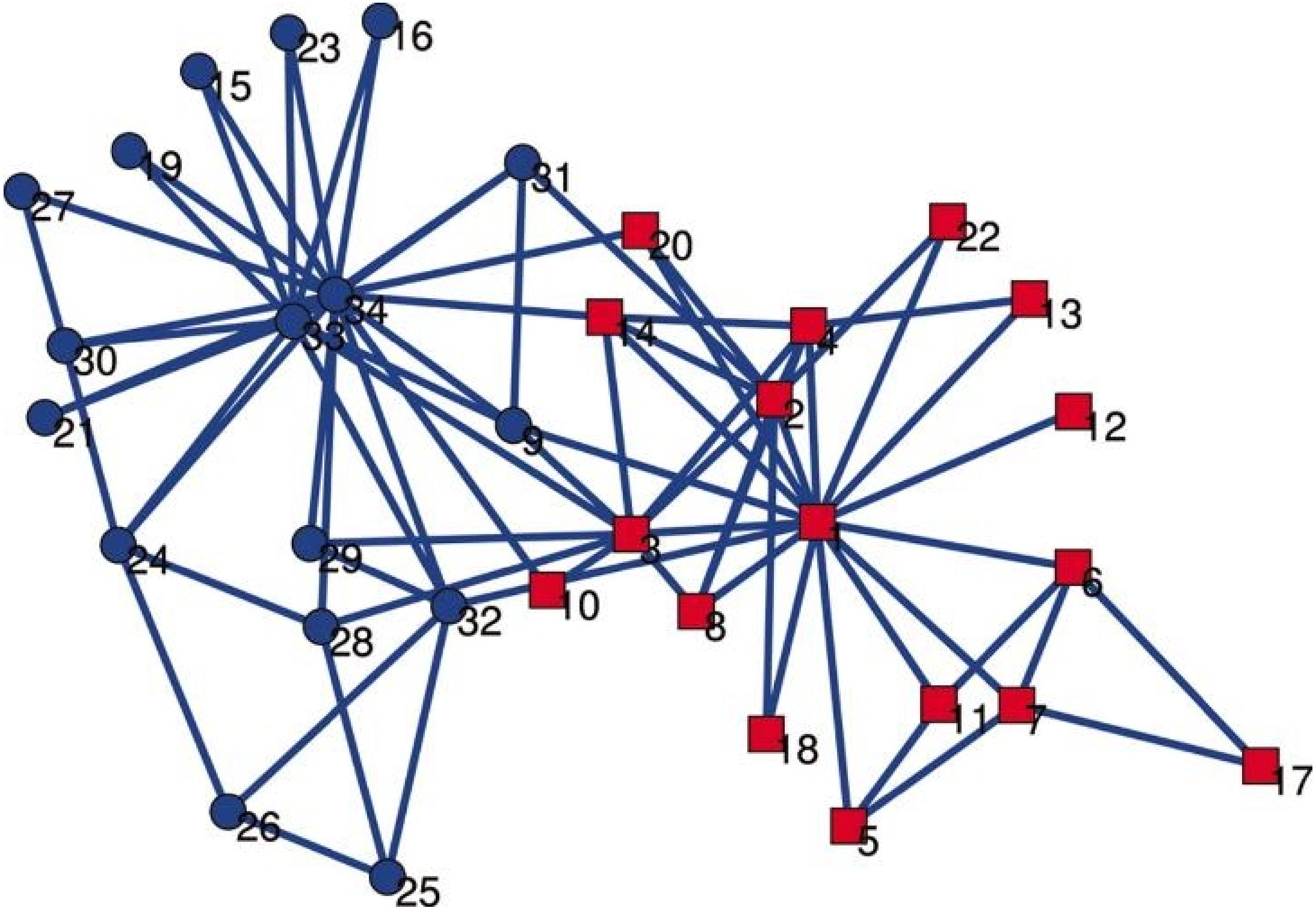
- Infomap,
- OSLOM,
- Label propagation,
- Louvain
- Modularity optimization via simulated annealing.

→ The rationale behind this choice is that gene coexpression and protein-protein interactions require a tight coregulation of the partners and that such a fine tuned regulation can be obtained only combining both the transcriptional and post-transcriptional layers of regulation. .

→ Our procedure is valid in principle for any pathology but is particularly suited for cancer, due to role that disregulation plays in cancer. We studied in particular gastric, lung, pancreas and colorectal cancer

→ To extract the relevant biological information we constructed for each tumor two multiplex networks: one using expression data for the normal tissue and one for the tumor and then compared their partition into ***communities.***

A **community** is a group of nodes that are densely connected to each other, but sparsely connected to the other nodes of the network.



Results: Chromosomal Locations

Three major outcomes:

1) Out of the hundreds of genes contained in each enriched chromosomal location, with the Multiplex recosntruction *only the few genes which are involved in a common co-regulatory scheme are selected.* and thus are likely to be the real drivers of the cancer.

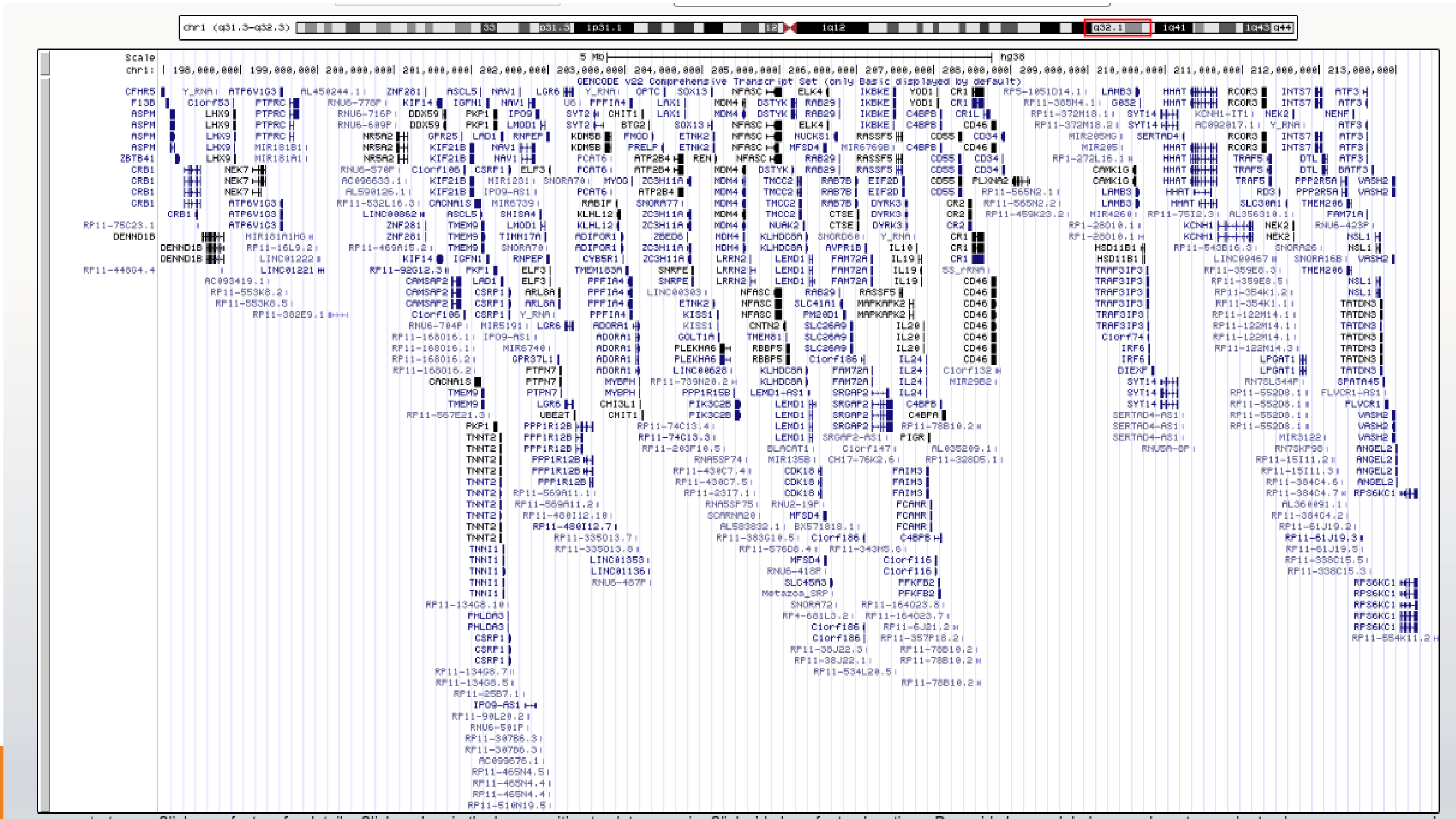
2) In the communities one find also *genes outside the enriched chromosomal locus, related to them not only by a coexpression link but also by regulatory relations* and this suggests that they could be part of a common biological pathway which is disregulated in the tumour.

3) *In some cases the community is also characterized by a GO or KEGG enriched category* which may give some hint to identify the above pathway.

Enriched Chromosomal locations

Locus 1q32 enriched in 43rd community of Pancreatic Cancer Data

- Genes in the intersection: ATF3, BTG2, CD46, IRF6 and PPP1R15B.
- 43rd community also enriched in DREAM pathway



Conclusions:
a set of open questions

- Which is the role of **Non coding DNA**?
- Which is the genomic origin of the difference between **Human and Chimps**?
- Which is the genomic origin of the difference between **different human beings**?
- Can we identify the genomic source of the impressive **complexity of multicellular organisms**?
- Can we identify the disregulated pathways leading to **complex deseases, in particular cancer**?