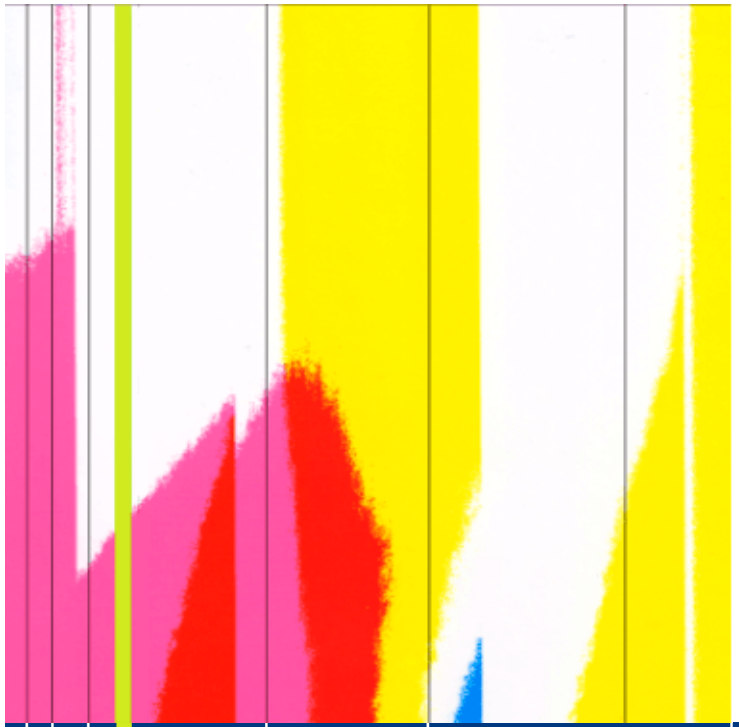


Models and algorithms for the genome

2011-01-05

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

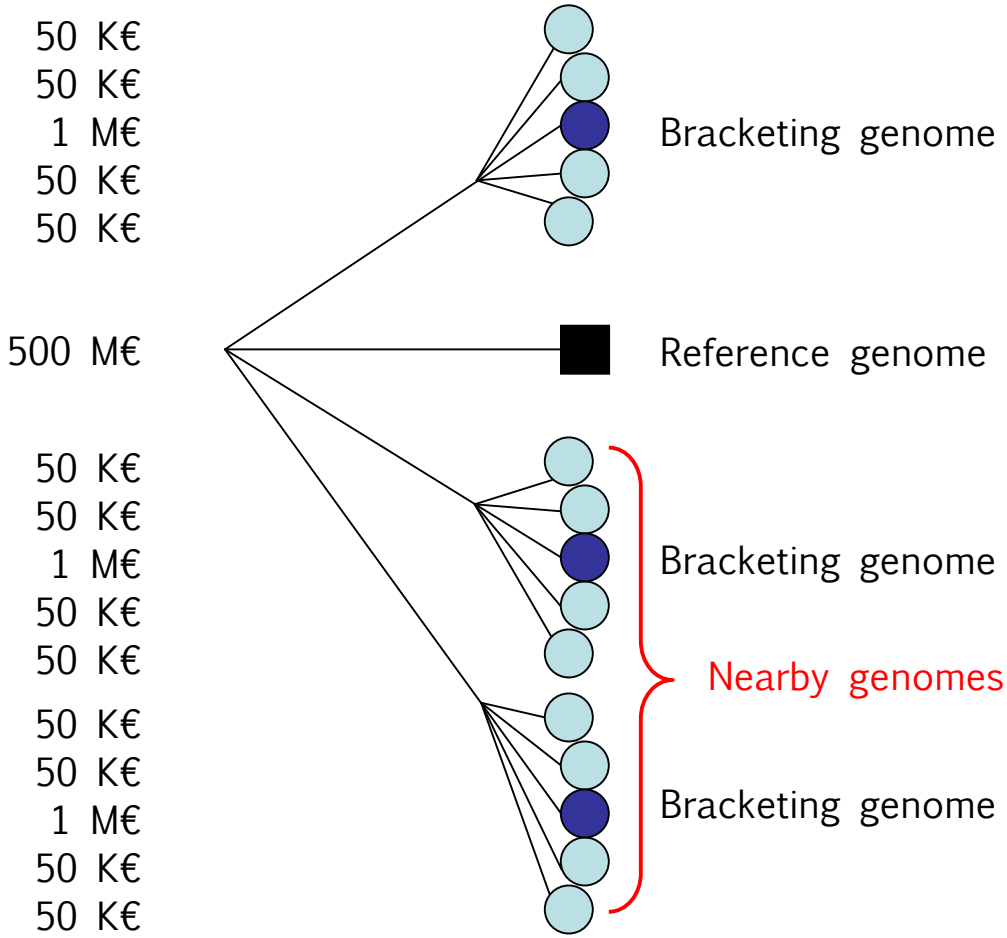
INRIA

centre de recherche
BORDEAUX - SUD-OUEST

David J. Sherman, DR INRIA

INRIA project-team « MAGNOME »
joint with CNRS & U. Bordeaux

Comparative genomics: a growing challenge

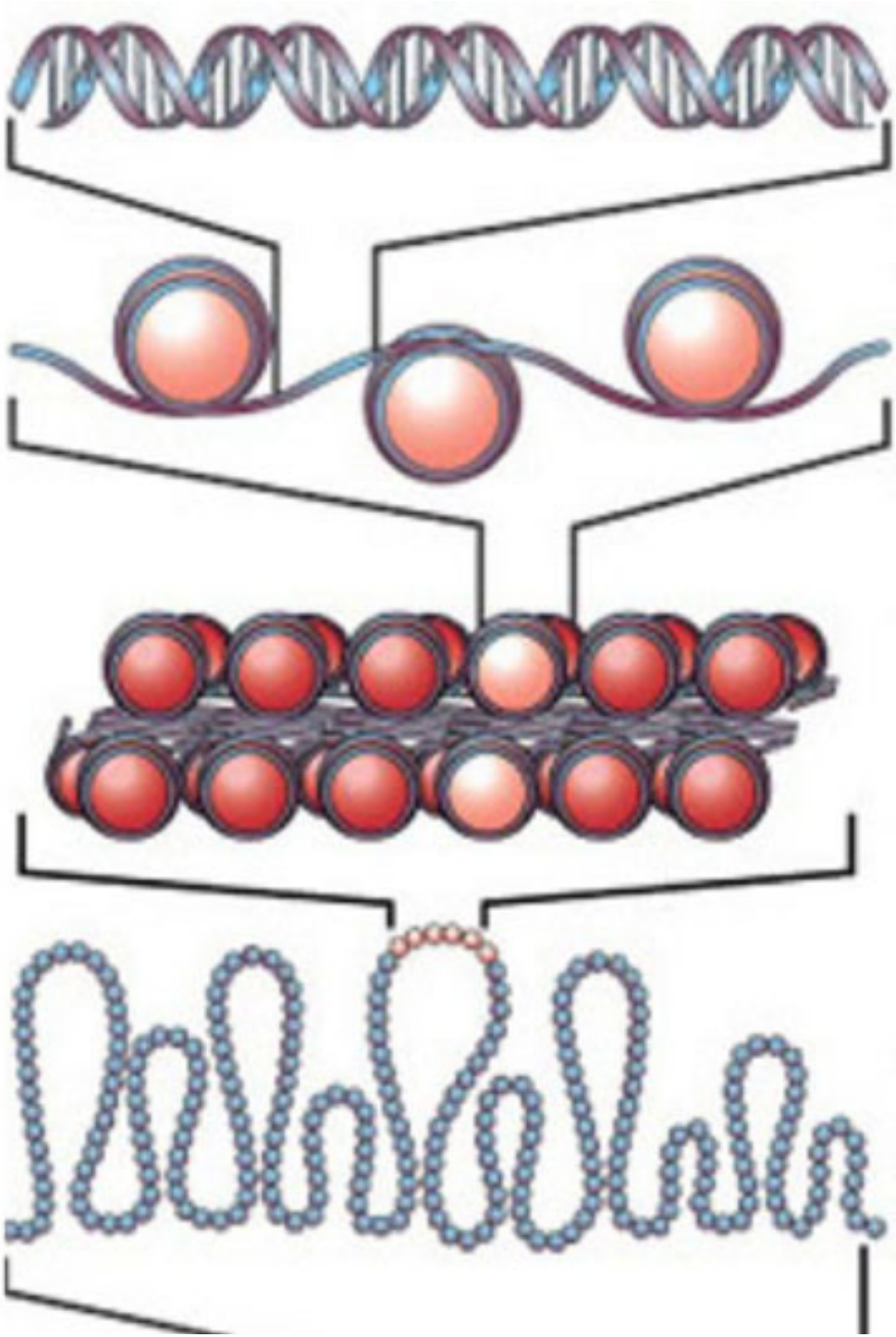


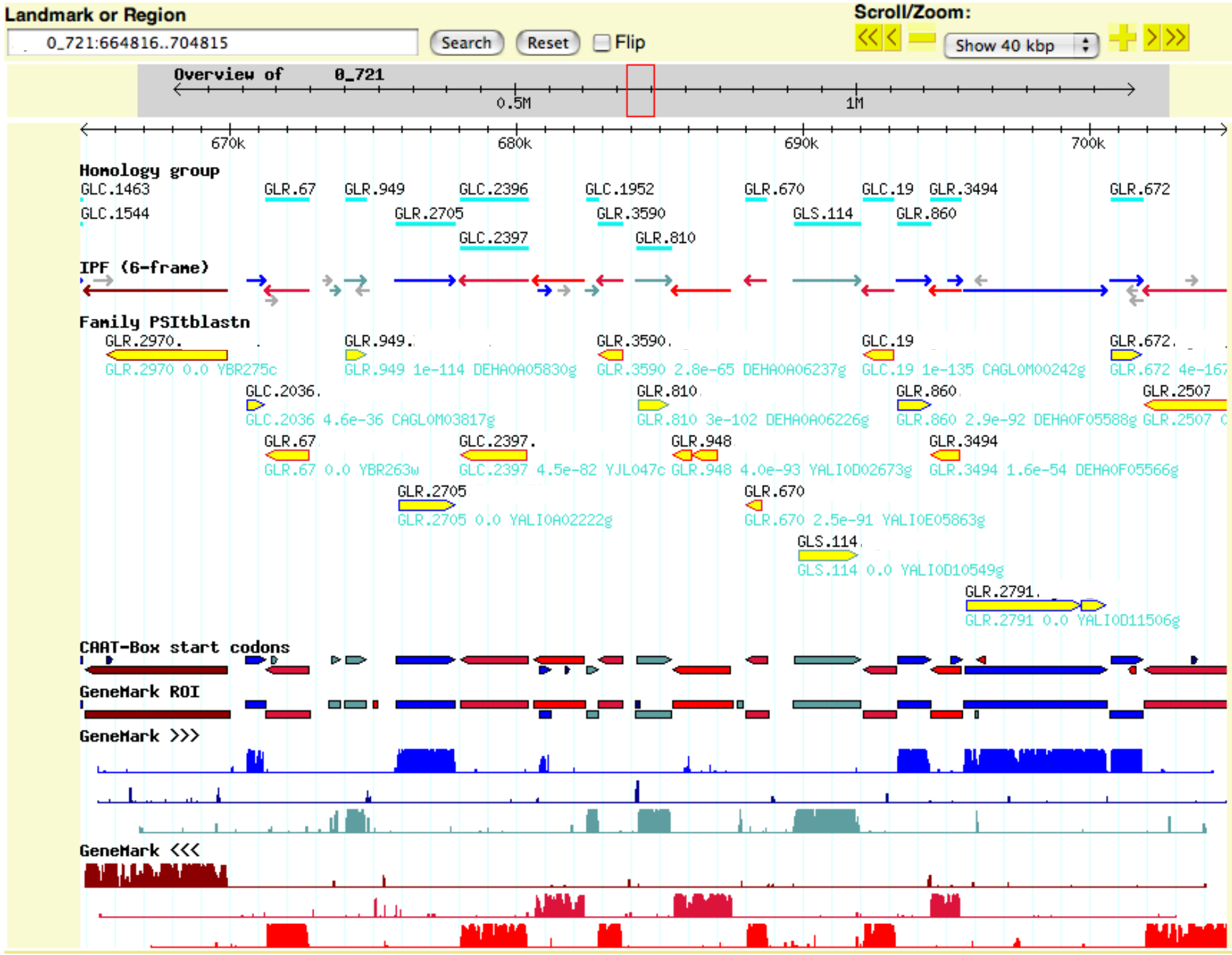
Comparative genomics makes most sense in this vicinity

Very many applications

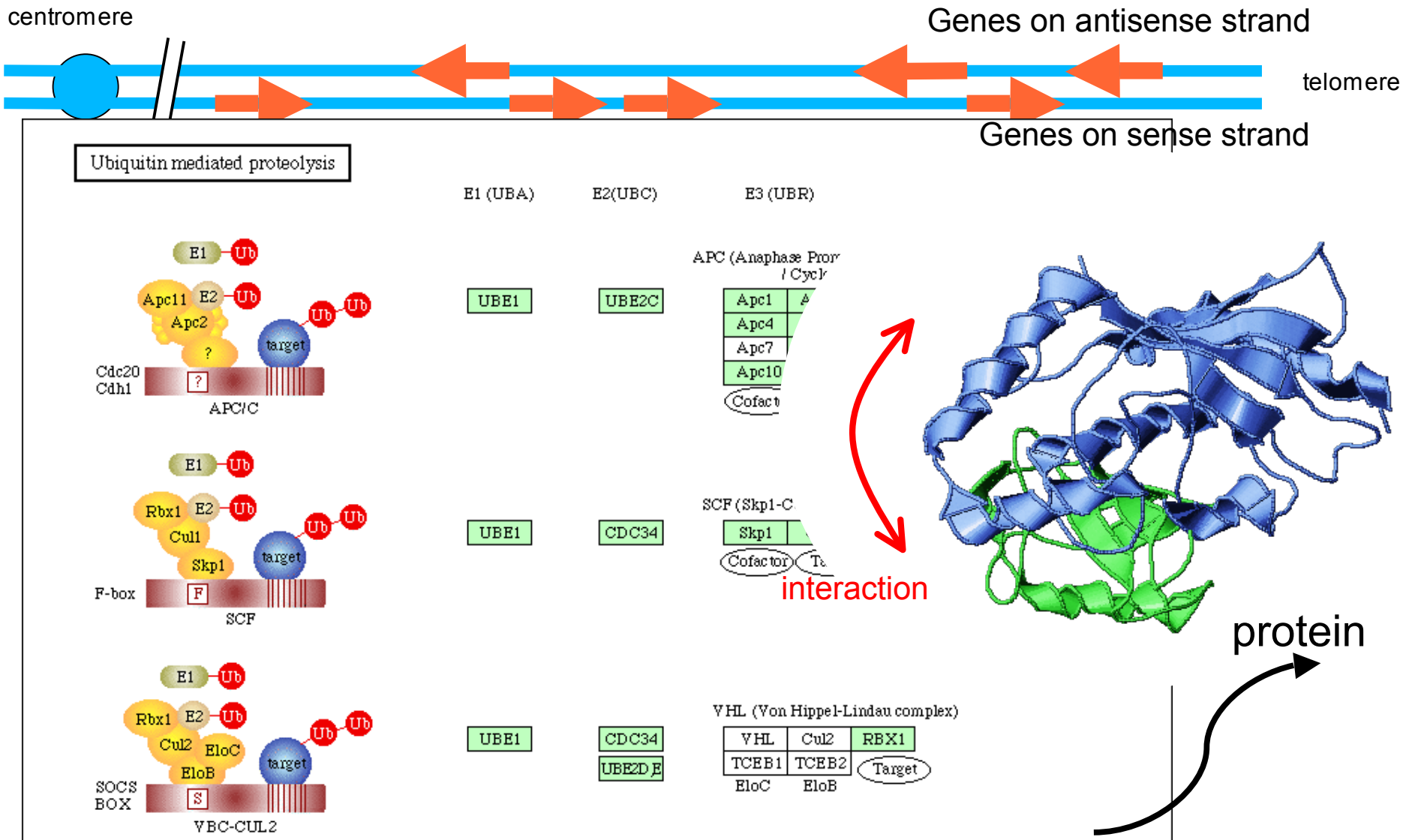
- Health
- Biotechnologies
- Biodiversity

Essentially scaling challenges for computer science


naked duplex DNA
"beads-on-a-string" created by formation of nucleosomes
30nm solenoid
extended form of chromosome

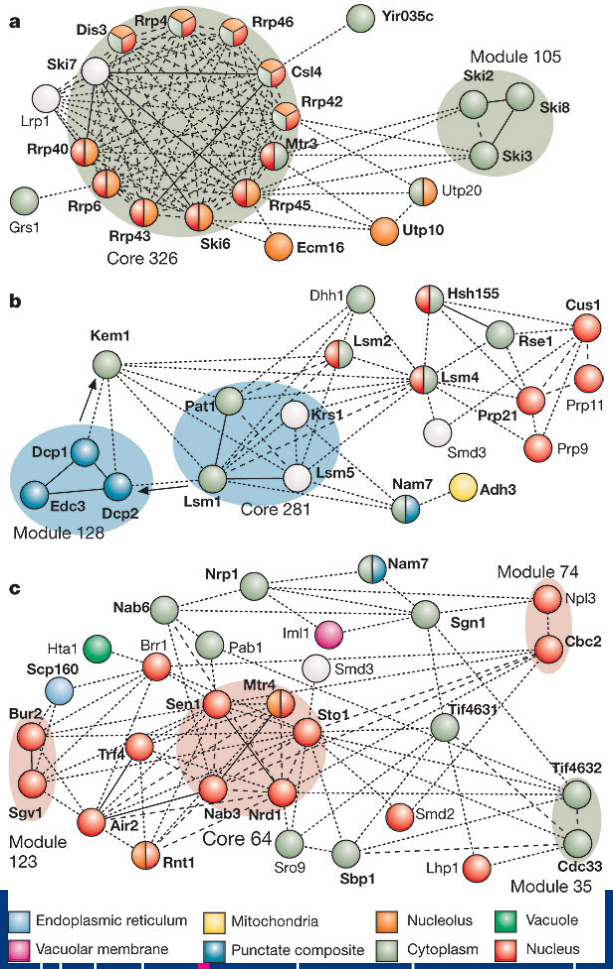

Landmark or Region
0_721:664816..704815
Search
Reset
Flip
Scroll/Zoom:
Show 40 kbp
Overview of 0_721
0.5M
1M
670k
680k
690k
700k
Homology group
GLC.1463
GLC.1544
GLR.67
GLR.949
GLC.2396
GLR.2705
GLC.2397
GLC.1952
GLR.3590
GLR.810
GLR.670
GLS.114
GLC.19
GLR.3494
GLR.860
GLR.672
IPF (6-frame)
Family PSItblastn
GLR.2970.
GLR.2970 0.0 YBR275c
GLR.949.
GLR.949 1e-114 DEHA0A05830g
GLR.3590.
GLR.3590 2.8e-65 DEHA0A06237g
GLC.19
GLC.19 1e-135 CAGL0M00242g
GLR.672.
GLR.672 4e-167
GLC.2036.
GLC.2036 4.6e-36 CAGL0M03817g
GLR.810
GLR.810 3e-102 DEHA0A06226g
GLR.860
GLR.860 2.9e-92 DEHA0F05588g
GLR.2507
GLR.2507 C
GLR.67
GLR.67 0.0 YBR263w
GLC.2397.
GLC.2397 4.5e-82 YJL047c
GLR.948
GLR.948 4.0e-93 YALI0D02673g
GLR.3494
GLR.3494 1.6e-54 DEHA0F05566g
GLR.2705
GLR.2705 0.0 YALI0A02222g
GLR.670
GLR.670 2.5e-91 YALI0E05863g
GLS.114
GLS.114 0.0 YALI0D10549g
GLR.2791.
GLR.2791 0.0 YALI0D11506g
CAAT-Box start codons
GeneMark ROI
GeneMark >>>
GeneMark <<<

From genome to cellular function



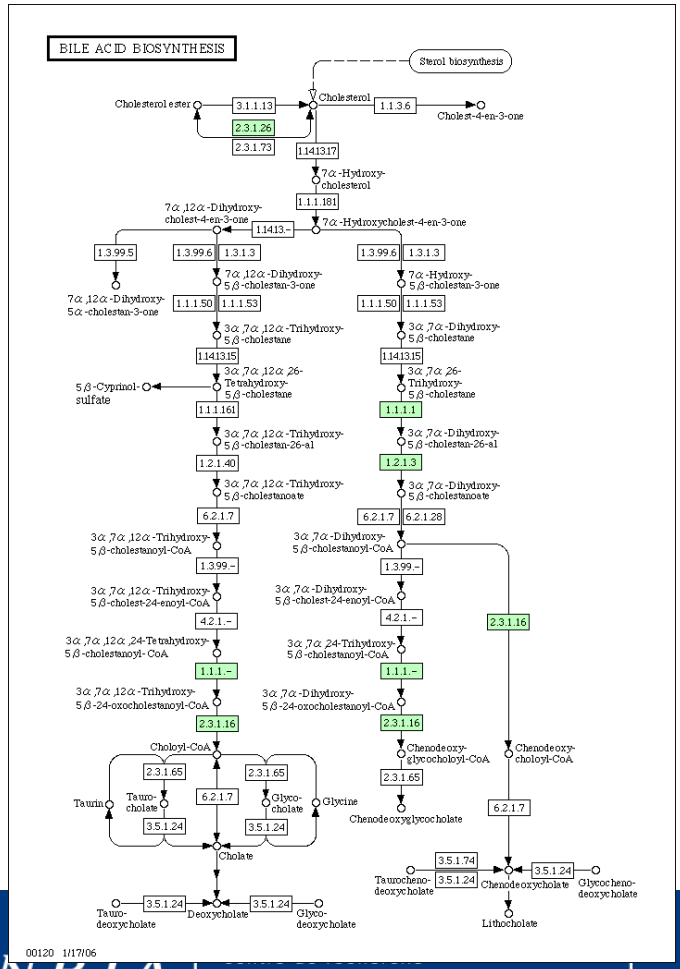
Interactions between components

interactions



A.-M. Gavin *et al.*

metabolism

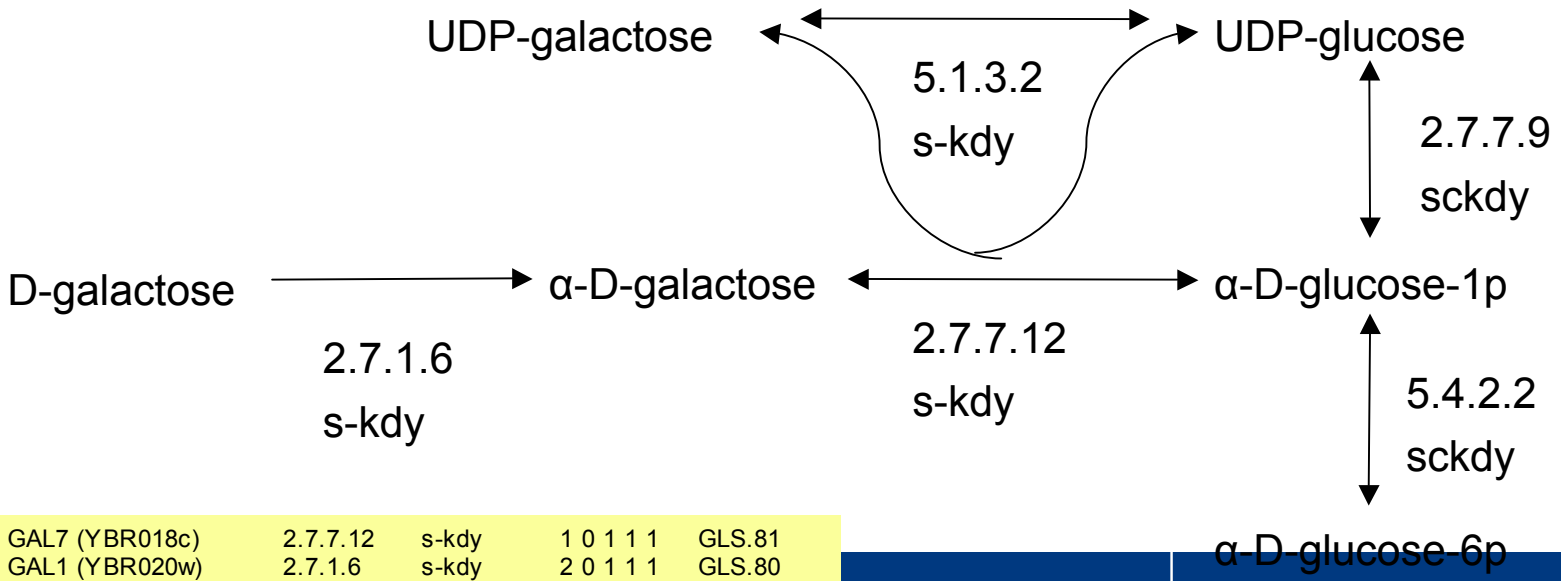


KEGG

Links from genome to function

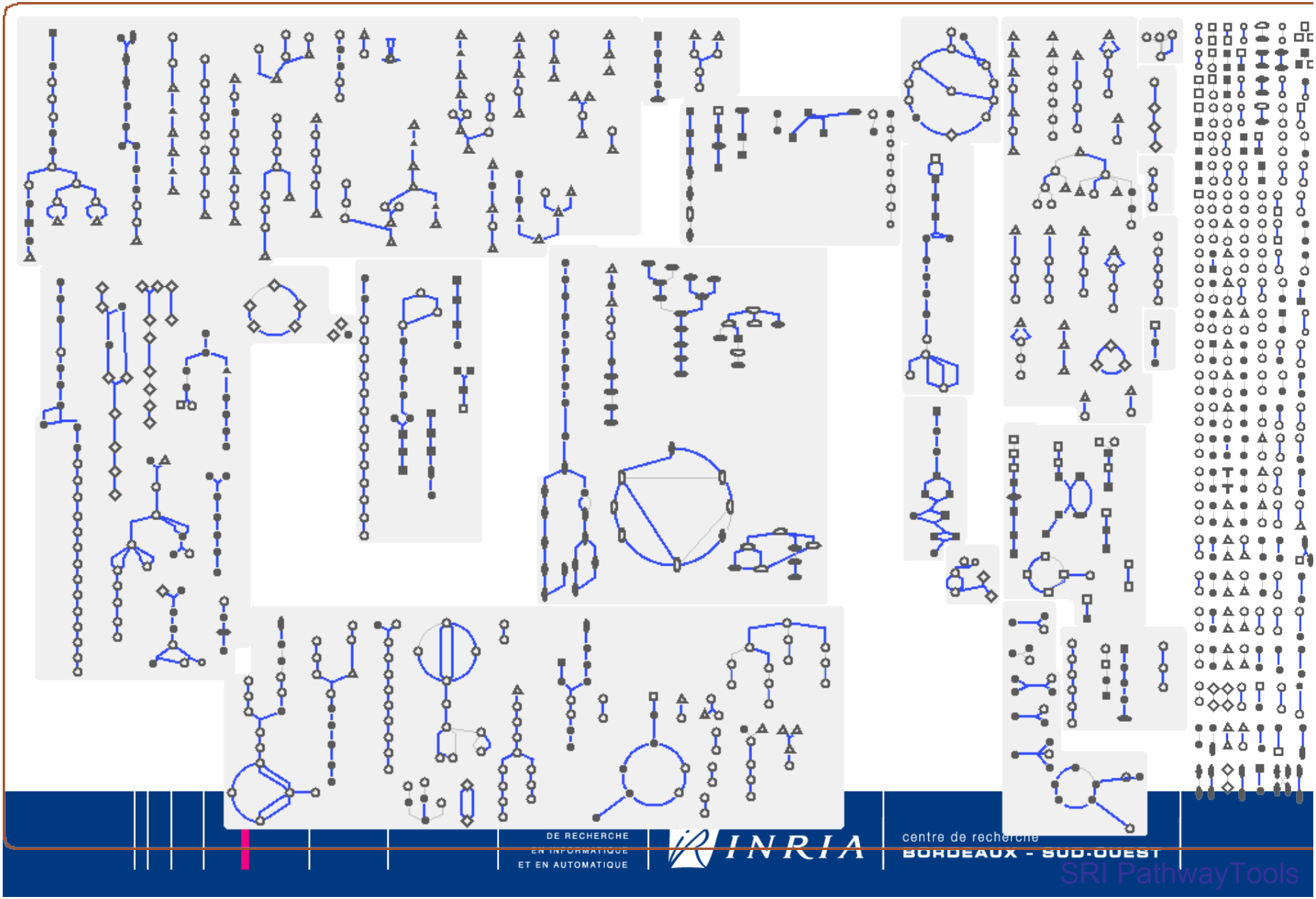
YHL012w (YHL012w)	2.7.7.9	sckdy	3 2 2 2 2	GLC.2368
UGP1 (YKL035w)	2.7.7.9	sckdy	3 2 2 2 2	GLC.2368
YHL012w (YHL012w)	2.7.7.9	sckdy	3 2 2 2 2	GLC.2368
UGP1 (YKL035w)	2.7.7.9	sckdy	3 2 2 2 2	GLC.2368

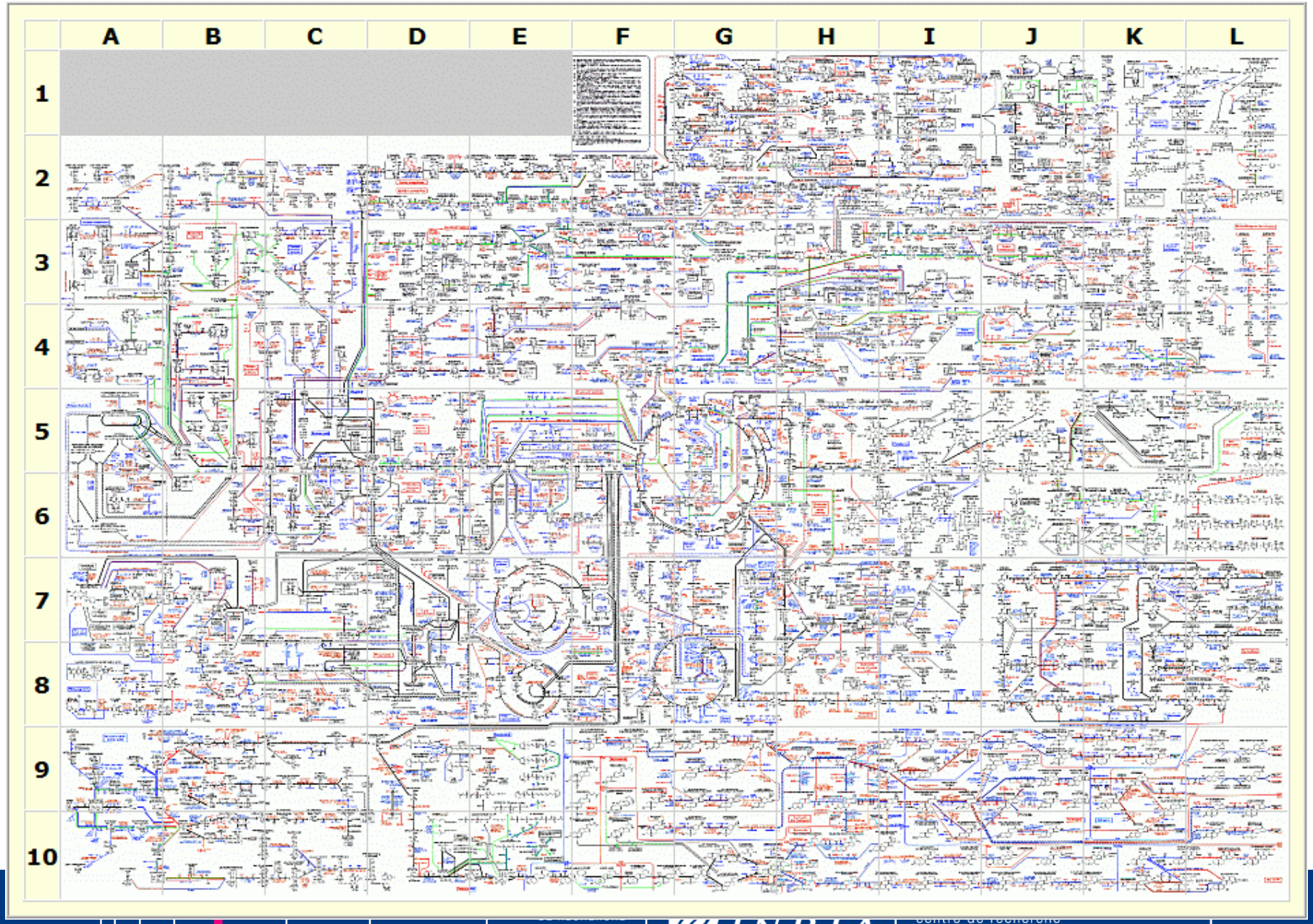
GAL10 (YBR019c)	5.1.3.2	s-kdy	3 0 1 1 1	GLS.83



GAL7 (YBR018c)	2.7.7.12	s-kdy	1 0 1 1 1	GLS.81
GAL1 (YBR020w)	2.7.1.6	s-kdy	2 0 1 1 1	GLS.80
GAL3 (YDR009w)	2.7.1.6	s-kdy	2 0 1 1 1	GLS.80

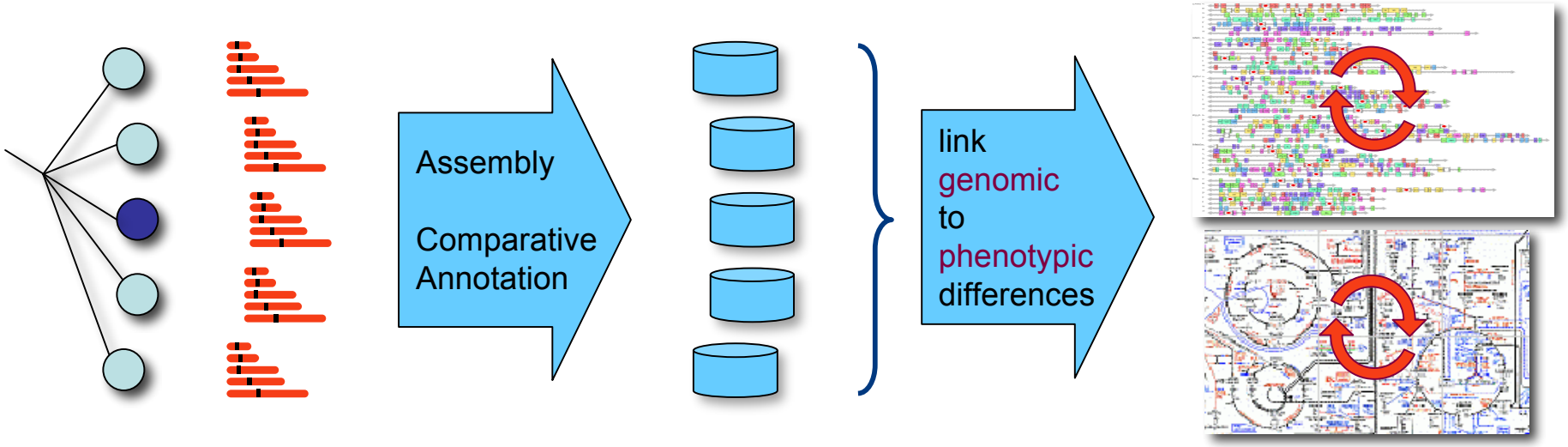
Metabolic paths in *S. cerevisiae*





MAGNOME

Algorithms and models for the genome



Related genomes compared to a reference species or strain
- cell factories
- biotechnologies
- pathogens

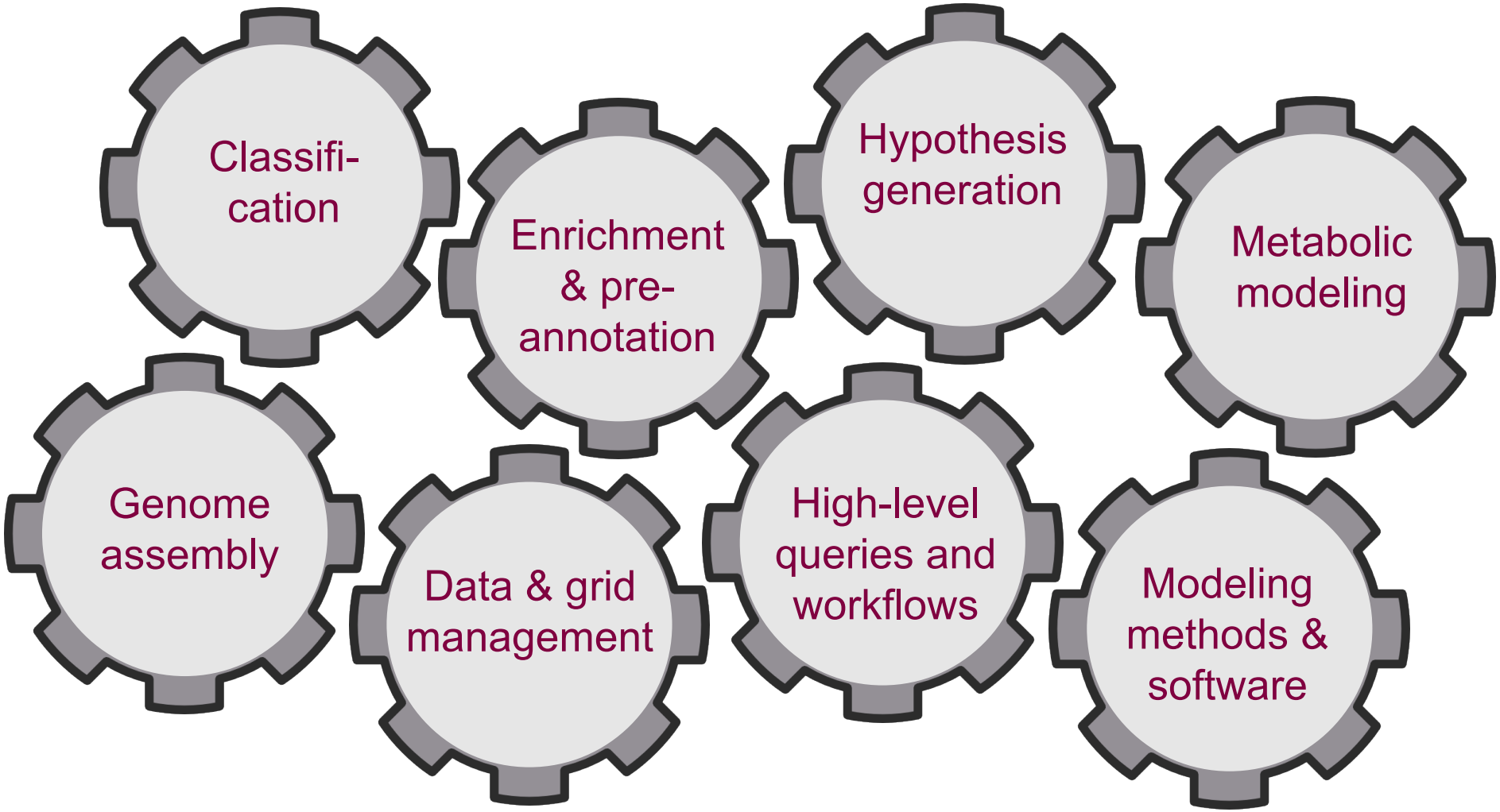
High-throughput sequencing

Genome-scale models

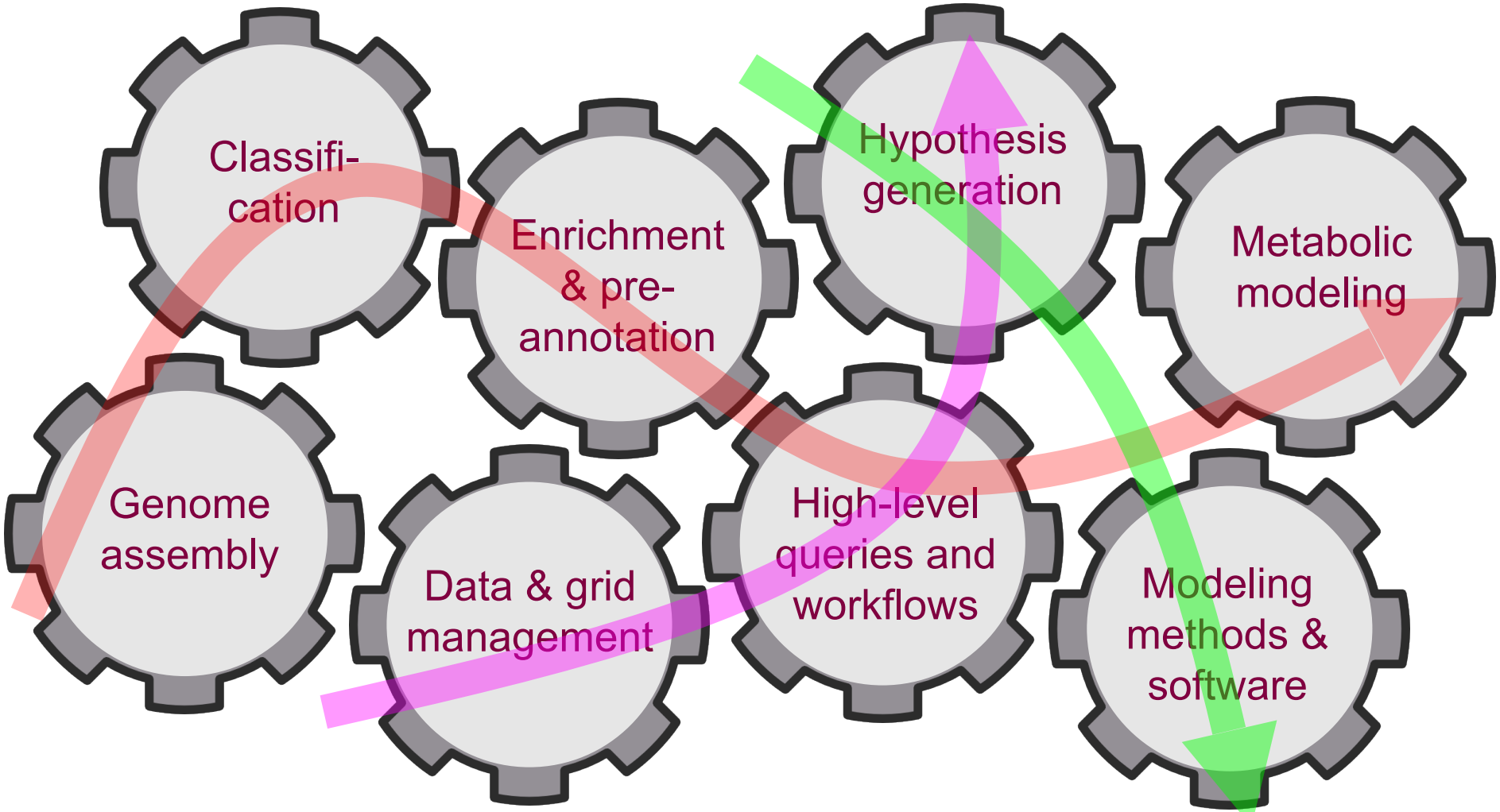
Comparative genomics
Hierarchical stochastic modeling
Classification & learning

➜ http://magnome.bordeaux.inria.fr

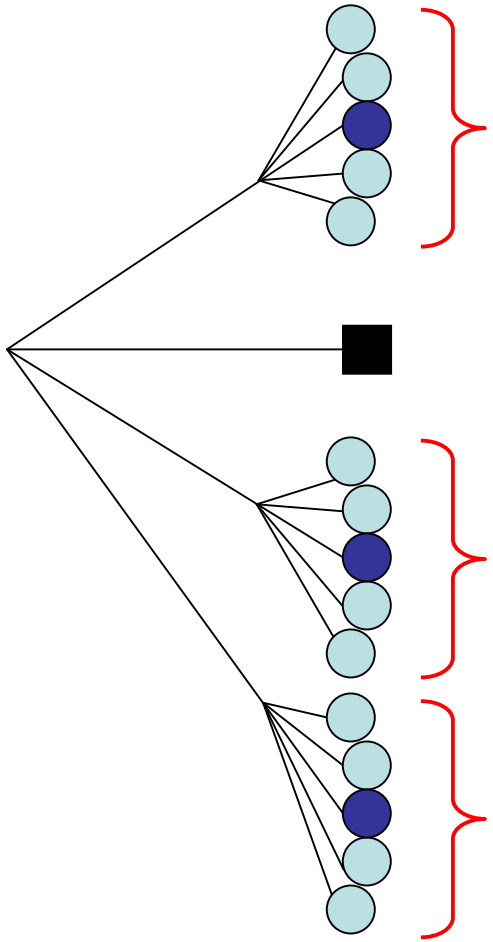
What techniques do we use?



What techniques do we use?



Current projects for closely related species

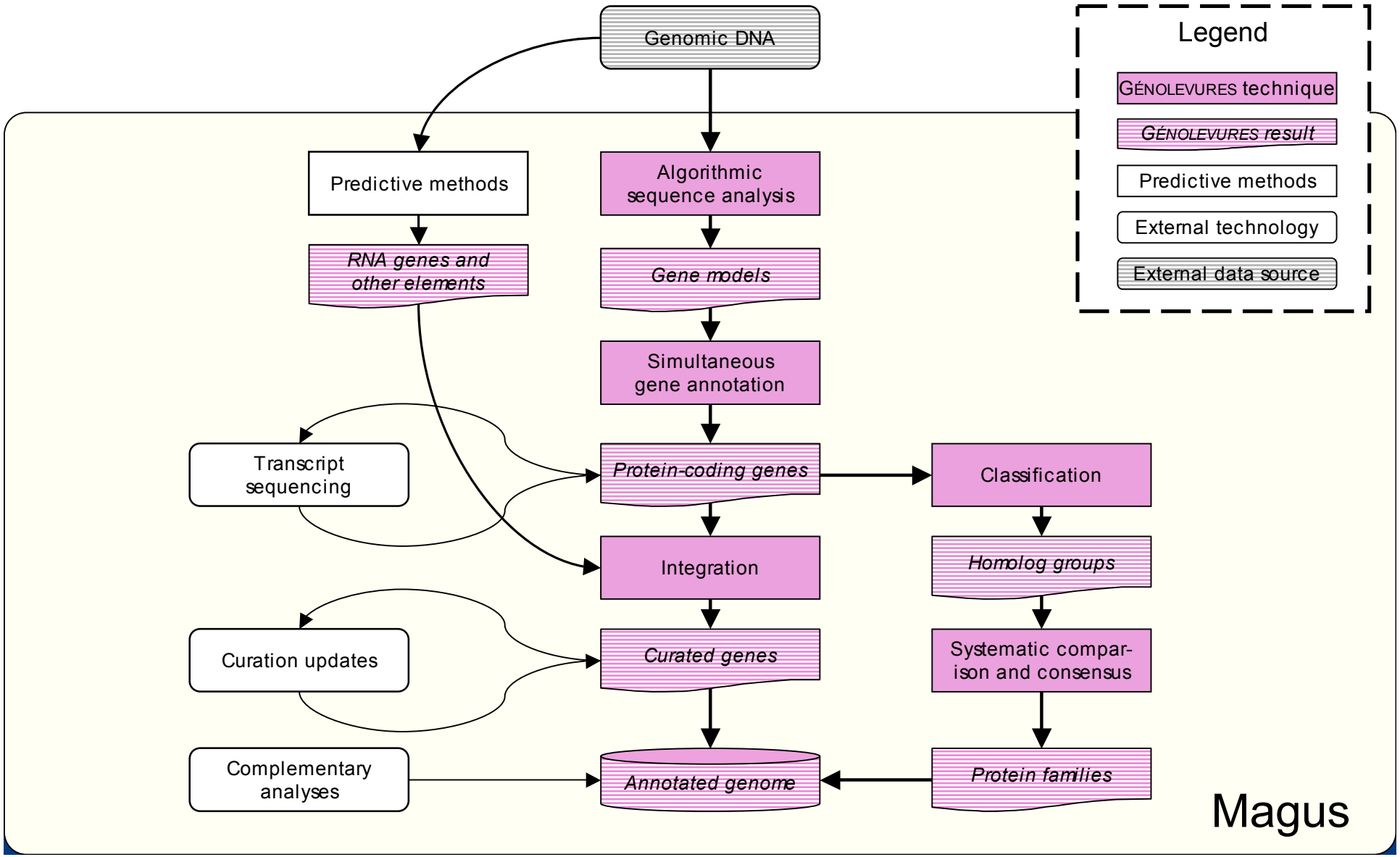


Pathogenic & nonpathogenic micro-organisms from the same clade

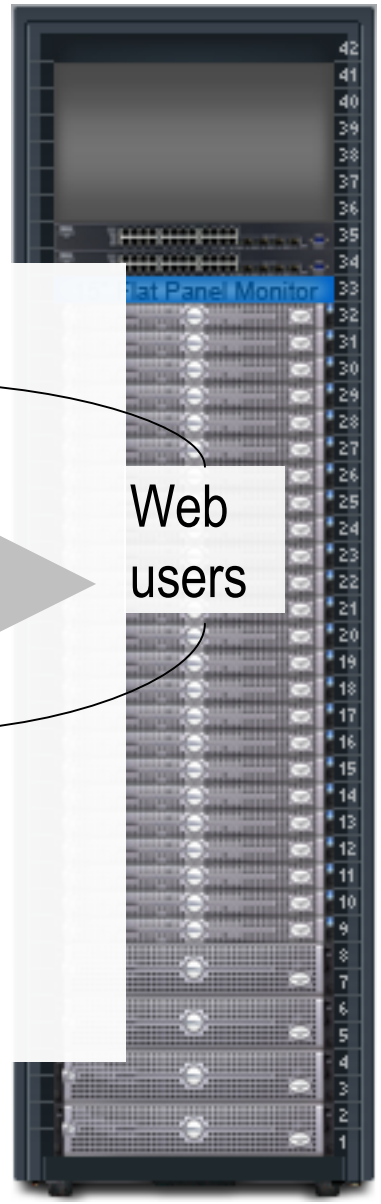
Industrial strains more or less adapted to a given wine must

Oleaginous yeasts with different efficiency or effect

The Annotation Process



The “big iron”



In silico predictions

Alignments & DB search

Rule checker

Rules

U.I. components

Web users

Web Service Bus

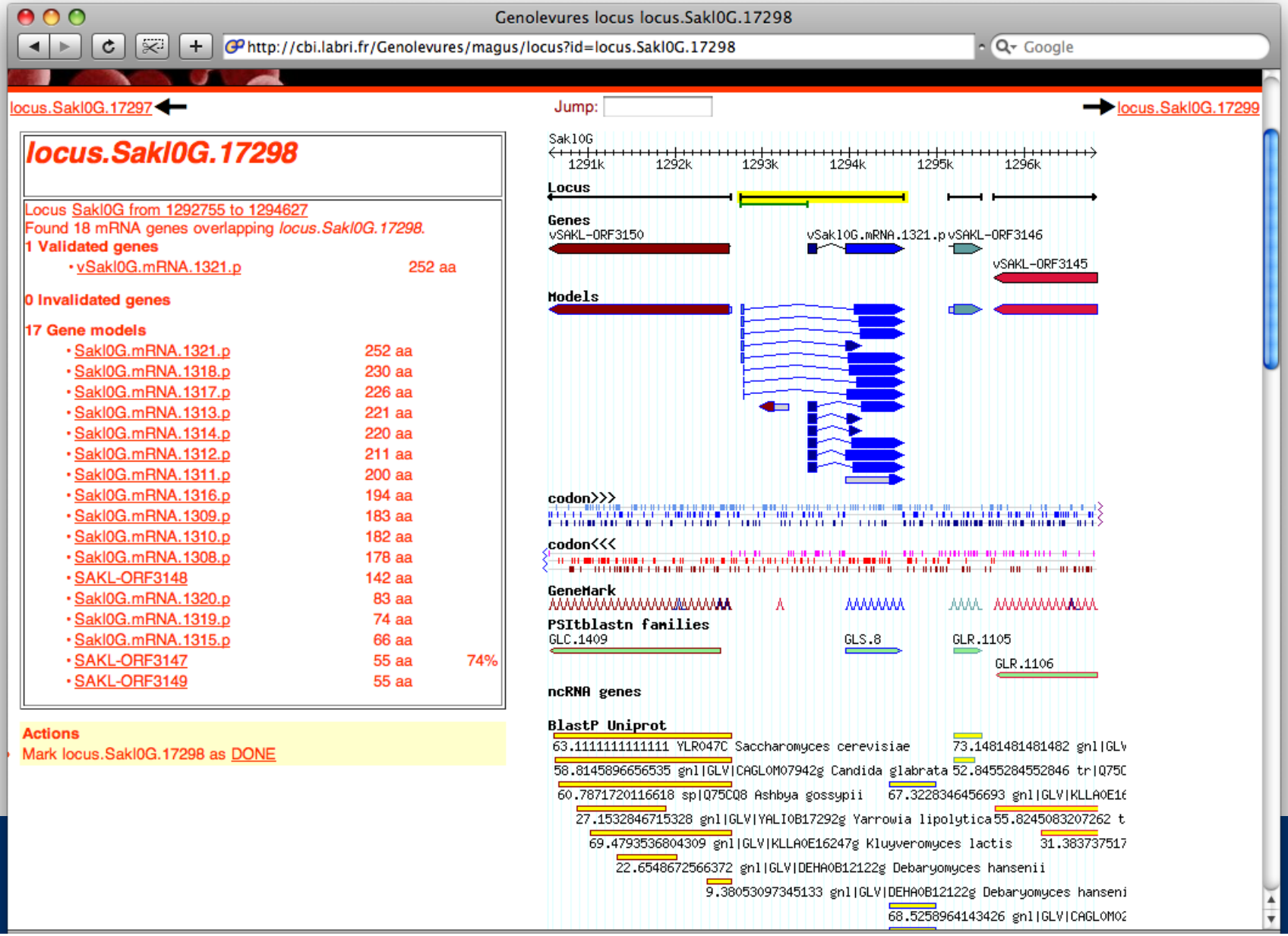
Genomes database

Computed results KB

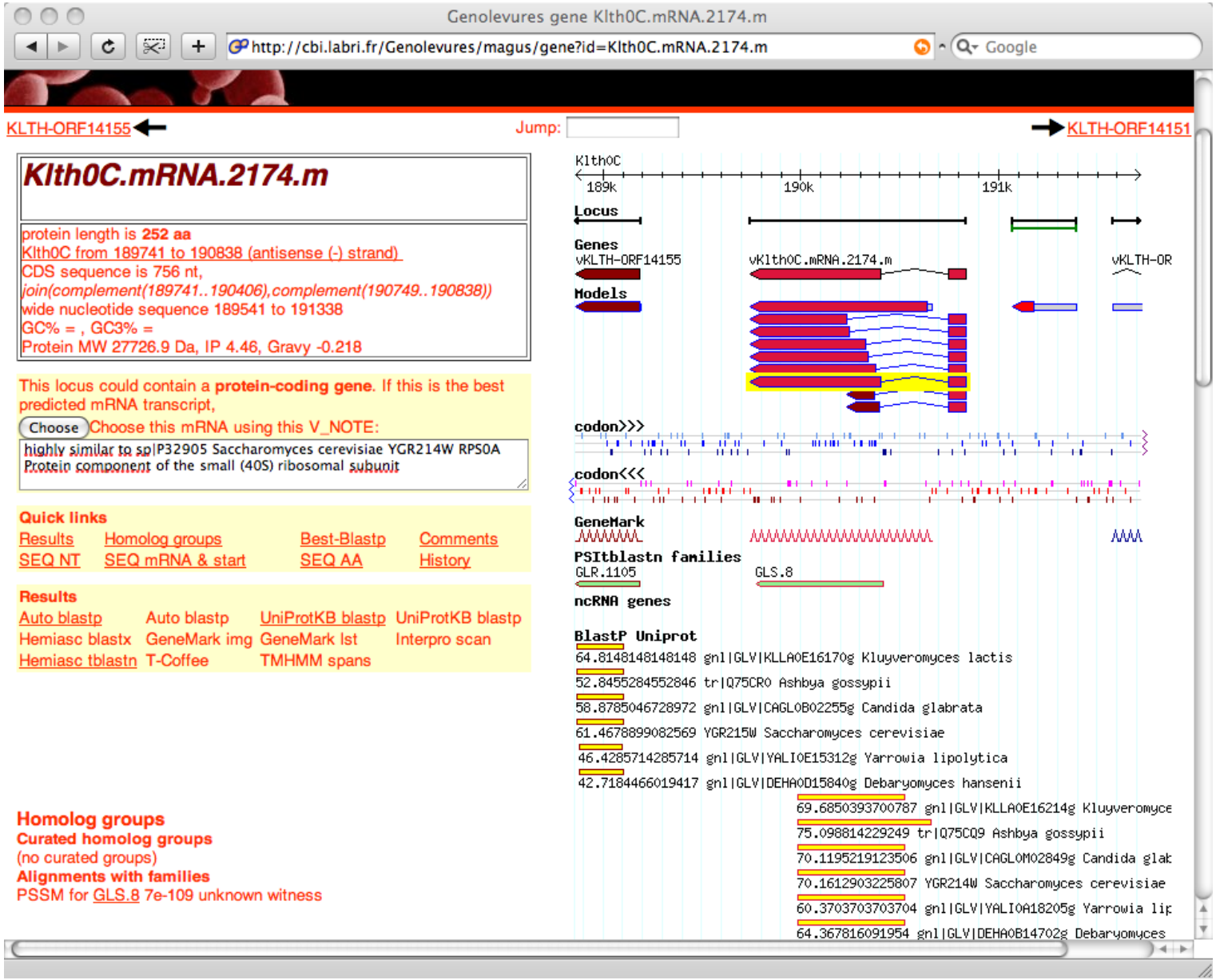
Genome Browser

Fast browser database

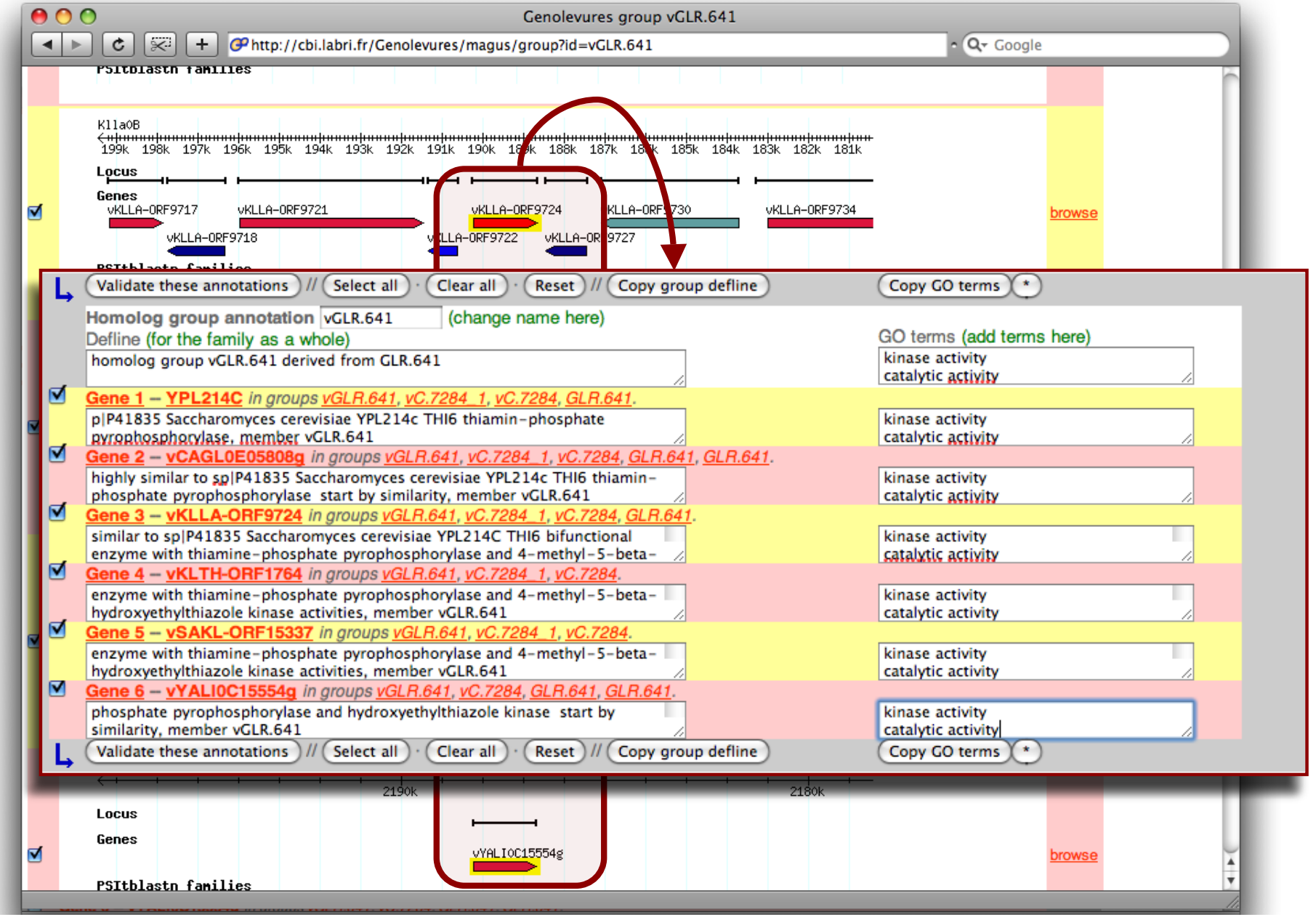
Viewing a Locus on a Genome



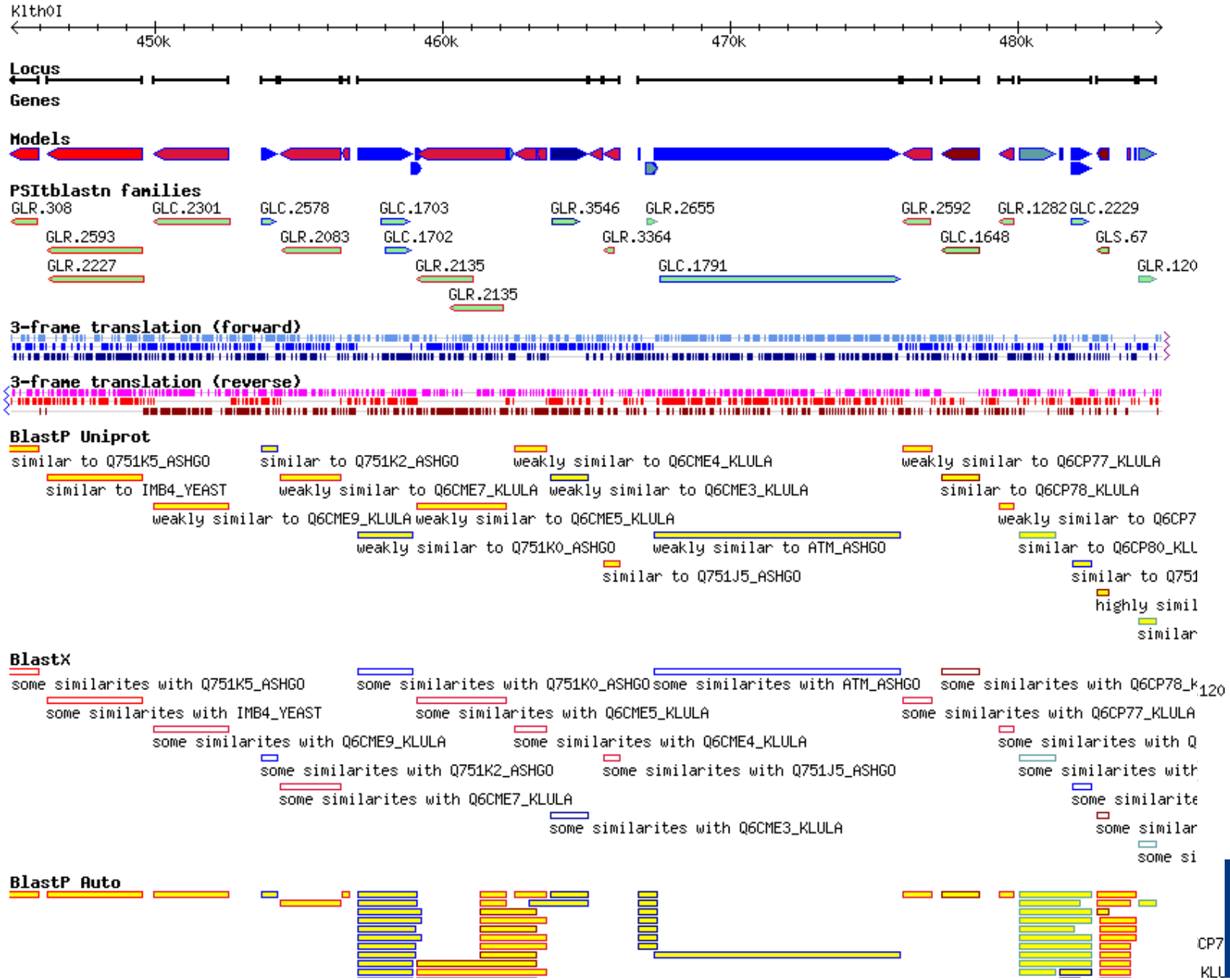
Validating a Gene Model



Annotating Homolog Groups



Browsing a genome region



Which one are you?

The hard part

Data

A solved problem

Push button

The hard part

Biological

–

otherwise

not interesting

Results

Algorithmic

–

otherwise

not interesting

MAGNOME ——— Algorithms and models for the genome

Permanent

- D. Sherman DR INRIA HDR
- P. Durrens* CR CNRS HDR
- E. Bon* MCF U.Bordeaux 2
- T. Martin* IR CNRS

**biologist or pluridisc. training*

Non permanent

- A. Garcia, IJD INRIA
- A. Goulielmakis, IE ANR
- N. Loira, PhD CONICYT (4/4)
- R. Assar, PhD INRIA (3/3)
- A. Sarkar, PhD EMMA (3/3)
- N. Golenetskaya, PhD INRIA (2/3)
- R. Issa, PhD PAB Syria (1/4)
- Alexander Makarov (intern DRI)
- Anna Zhukova (intern DRI)

Some open questions

Collaborative genome annotation

- Interaction design and human-computer interfaces

Mathematical modeling

- Hierarchical stochastic models of fermentation
- Metaheuristic optimization fo evolutionary scenarios

High-performance computing

- Deploying grid-aware web services
- Distributed NoSQL databases

Contact

- David Sherman david.sherman@inria.fr
- http://magnome.bordeaux.inria.fr

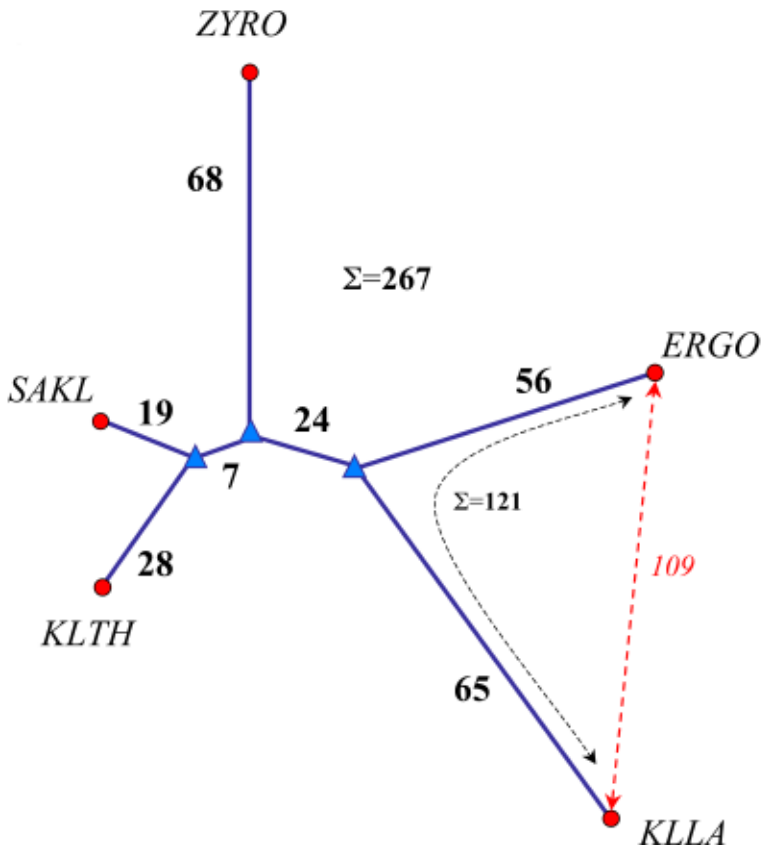
Two examples

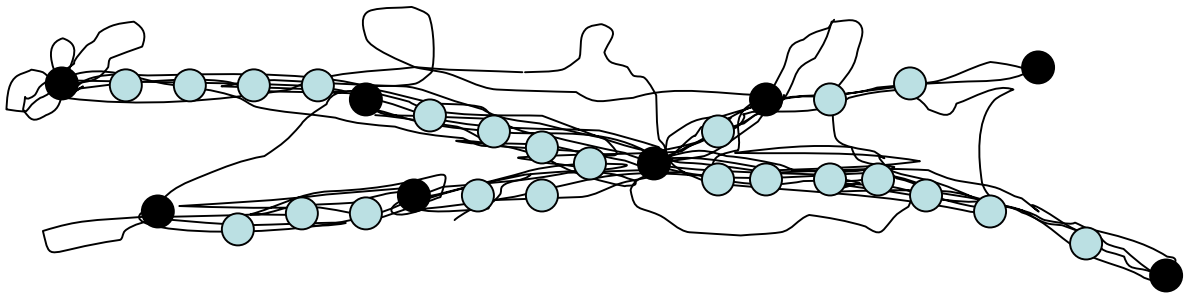
Finding median genomes

- Population
- Stochastic local search
- Experimental results

Finding parsimonious scenarios

- Ant colony optimization
- Hybridization with random walk
- Experimental results





Finding median genomes

[Goëffon *et al* GECCO 2008]

A *chromosome* $\pi = (\pi_1, \ldots, \pi_m)$ is represented by a sequence of signed gene markers whose sign indicates their relative direction on the chromosome. A size-n *multichromosomal genome* Π is defined as a set of chromosomes $\{\pi^1, \ldots, \pi^N\}$ such that $\sum_i |\pi^i| = n$. Markers take their value from the set of ordinals $1, \ldots, n$; no given marker appears in more than one chromosome.

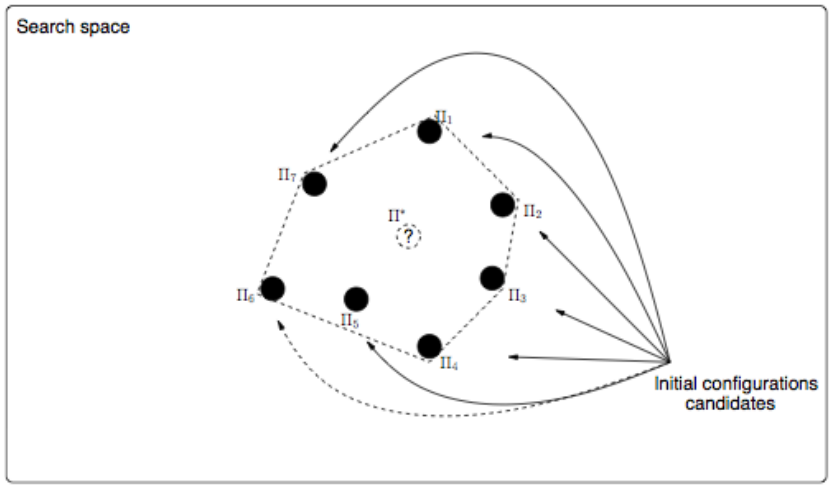
Given a set of size-n genomes $\{\Pi_i\}$ and a genome distance function d, an instance of the combinatorial minimization problem MGP is defined by two elements $\langle \tau_n, \phi \rangle$:

1. a search space, τ_n, composed of the set of all possible size-n genomes, and
2. an objective function $\phi : \tau_n \rightarrow \mathbb{N}$ (*score*) defined by $\phi(\Pi) = \sum_{\Pi_i} d(\Pi, \Pi_i)$.

A *median genome* for a given set of genomes $\{\Pi_i\}$ is a genome Π that minimizes $\phi(\Pi)$. Every optimal solution to MGP is a median genome.

Population-based local search

No crossing-over



1. the *initial configuration* is taken from $\{\Pi_i\}$,
2. the *evaluation function* is the same as the objective function ϕ: the rearrangement distance d,
3. the *neighborhood relation* we call $\mathcal{R}^1$ is a 1-step rearrangement: $\mathcal{R}^1(\Pi) = \{\Pi' \in \tau_n, d(\Pi, \Pi') = 1\}$,
4. the *move strategy* is a *first-improve selection* (FI) which accepts better and equivalent configurations (*side-walk* mechanism, SW [16]), given a specified number of iterations *nbit*.

Probabilistic neighbor selection

Need to select the most pertinent neighbors at each step

- Simultaneous descent
- Dependant replications

Probabilistic neighborhood

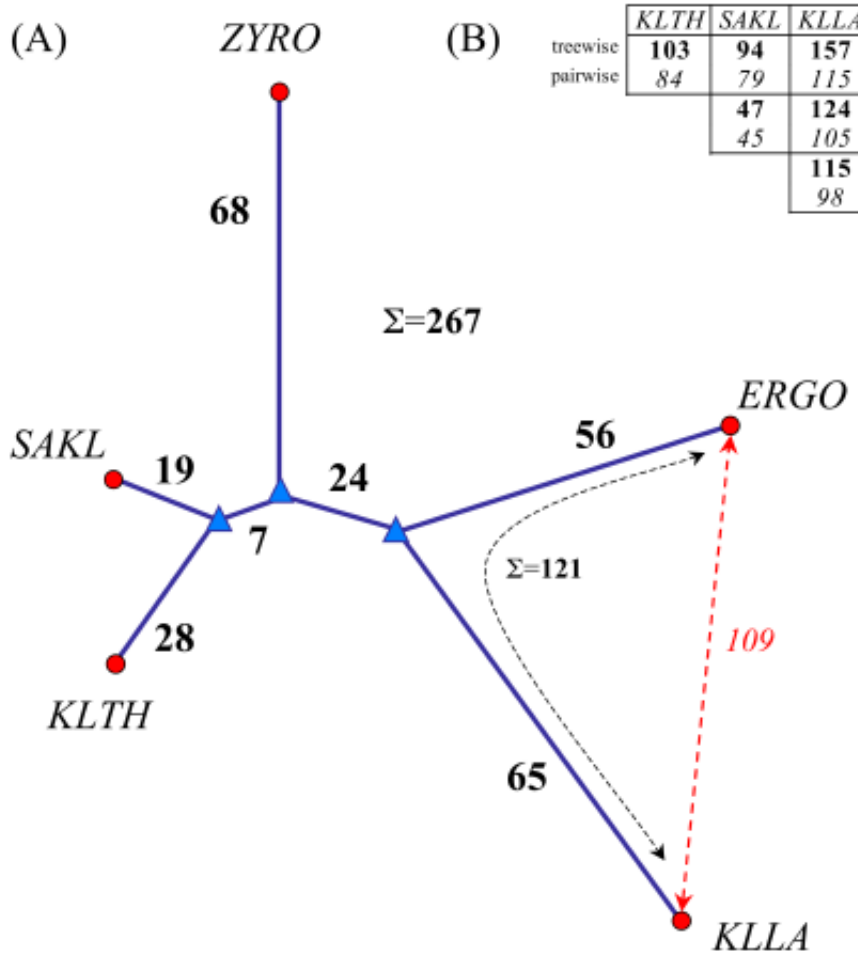
- Encourages adjacencies which are not, or are less represented in the population to be broken
- Replaces as a function of the proportional representation of the broken adjacencies

Algorithm 1

Require: $\{\Pi_1, \dots \Pi_k\}$: a set of k multichromosomal genomes of size n; $l.k$: the size of the population; $nbit$: the number of descent iterations.
Ensure: an approximate median genome set $\hat{\mathcal{P}}$
  let $\mathcal{P}$ be the multiset of the current genomes population, which initially contains l copies of each Π_i.
  let $numit \leftarrow 0$ be the number of performed local search iterations
  while $numit < nbit$ **do**
    for all $\Pi \in \mathcal{P}$ **do**
      loop
        randomly select $\Pi' \in \mathcal{R}^1(\Pi)$
        break with probability $p(\Pi, \Pi', \mathcal{P} \setminus \{\Pi\})$
      end loop
      if $\phi(\Pi') \leqslant \phi(\Pi)$ **then**
        $\Pi \leftarrow \Pi'$
      end if
    end for
    $numit \leftarrow numit + 1$
  end while
  return $\hat{\mathcal{P}} = \text{argmin}_{\Pi \in \mathcal{P}} \phi(\Pi)$

$$p(\Pi, \Pi', \mathcal{P} \setminus \{\Pi\}) = 1 - \frac{|\{\Pi'' \in \mathcal{P} \setminus \{\Pi\}, (\mathcal{A}(\Pi) \setminus \mathcal{A}(\Pi')) \cap \mathcal{A}(\Pi'') \neq \emptyset\}|}{|\mathcal{P}| - 1}$$

Protoploid Hemiascomycete yeasts



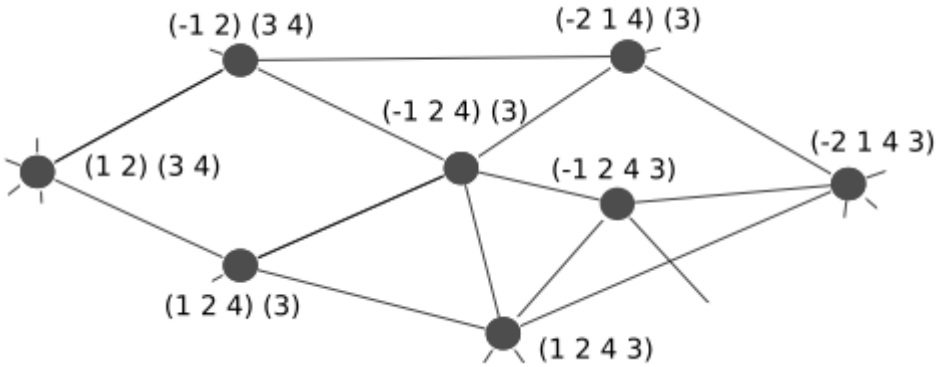
(B)

	KLTH	SAKL	KLLA	ERGO	
treewise	**103**	**94**	**157**	**148**	ZYRO
pairwise	*84*	*79*	*115*	*101*	
		47	**124**	**115**	KLTH
		45	*105*	*88*	
			115	**106**	SAKL
			98	*85*	
				121	KLLA
				109	

[Souciet *et al* Genome Res. 2009]

Multiple scenarios

[Vyahhi *et al* GECCO 2009]



Explore solution space

- No **one** scenario is best
- Suboptimal solutions may be more biologically plausible

ACO formulation

- $\Delta\tau_\varepsilon = 1/(l-d+1)$
- Evaporation rate ρ
- Edge detection threshold ε
- Select k random neighbors (except in the endgame)

DEFINITION 1. *A* ***scenario*** *of length m between two size-n multichromosomal genomes Π and Γ is a sequence $S = (\Pi_0, \ldots, \Pi_m)$ s.t.*

- $\forall\, i \in [0, m], \Pi_i$ *is a size-n multichromosomal genome,*
- $\Pi_0 = \Pi$ *and* $\Pi_m = \Gamma$, *and*
- $\forall\, i \in [0, m-1], d(\Pi_i, \Pi_{i+1}) = 1.$

If $m = d(\Pi, \Gamma)$, then S is a ***parsimonious scenario****. A sequence S s.t. $m = d(\Pi, \Gamma) + \alpha$ without cycles, that is, where $\Pi_i \neq \Pi_j \forall i, j \in \{0, \ldots, m\}(i \neq j)$, is called an* ***$\alpha$−parsimonious scenario***