Genomic Exploration of the Hemiascomycetous Yeasts

3rd Workshop on Algorithms in Bioinformatics

Moscow 2008-10-08

David J. Sherman

U. Bordeaux, France

LaBRI CNRS & INRIA team “MAGNOME”

Comparative genomics

Which ones do we sequence?

And what do we do after that?

Is certainly about comparison

But is also about the genomes

A caricature

The hard part

Data

A solved problem

Push button

The hard part

Biological

–

otherwise

not interesting

Results

Algorithmic

–

otherwise

not interesting

Hemiascomycetous yeasts

Eukaryotic genomes

Small and compact

Experimental model

Biotechnological interest
- beer, wine, bread
- assimilate hydrocarbons, tannin extracts
- horomones and vaccines

Medical interest

Biodiversity

Systems

Understand mechanisms of molecular evolution

Genome redundancy

Ortho-/para- log divergence

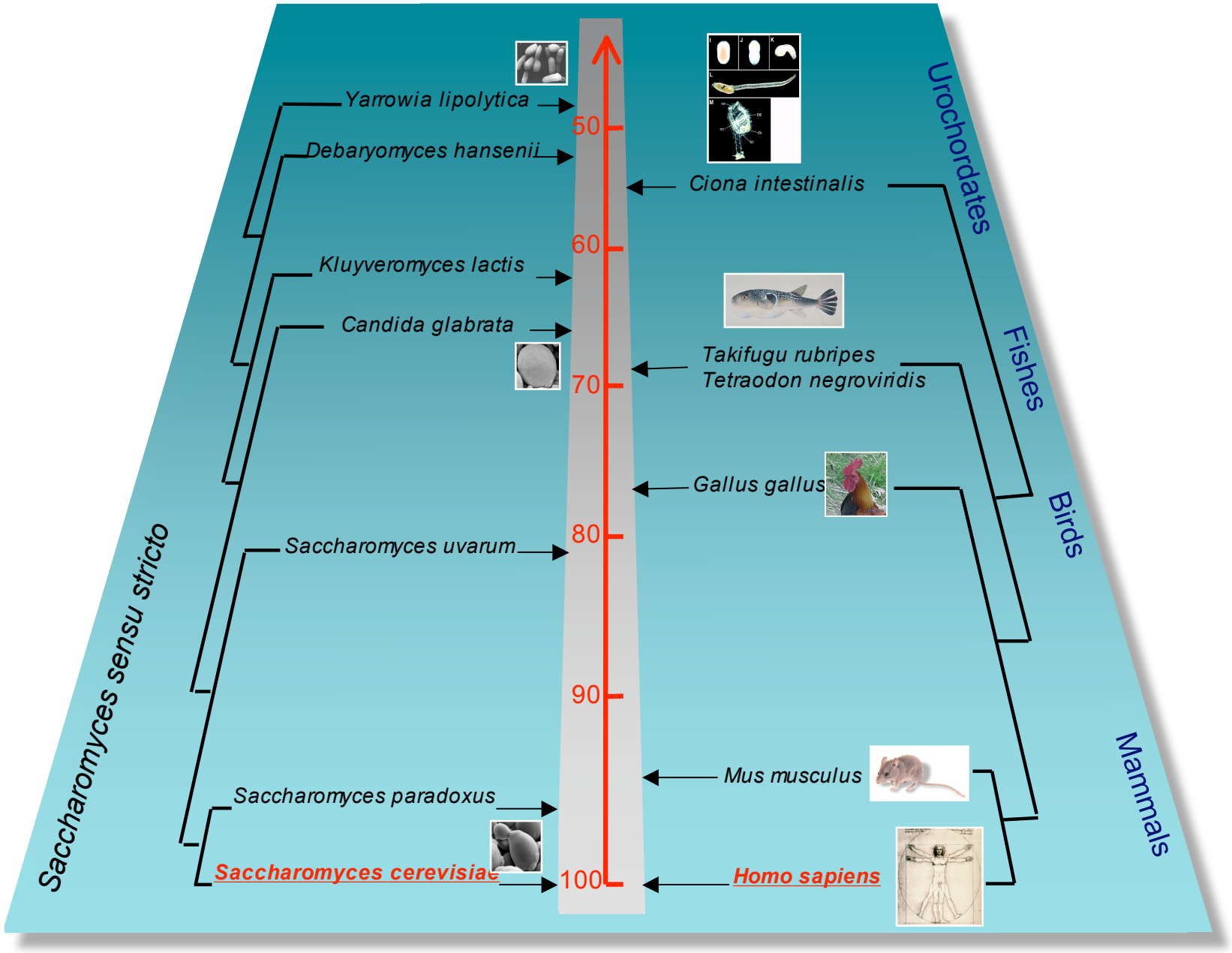
Expansion and contraction of universal families

Tandem duplications

Block duplication and rearrangement

Conservation of synteny

Comparison of evolutionary range of Hemiascomycetes and Chordates



Scale: average % of amino-acid identity between complete set of orthologous proteins

Dujon (2006) *Trends in Genetics* **22**: 375-387

Génolevures Sequencing Projects

Génolevures 1

- 13 species, partial 0.2-0.4X
- Souciet *et al* 2000 [21 papers] FEBS Letters 487

Génolevures 2

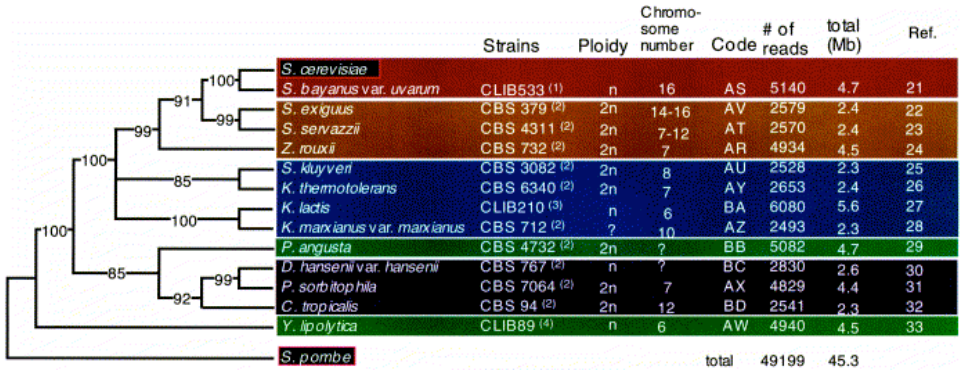
- 4 species complete 12X
- Dujon, Sherman *et al* 2004 Nature 430
- Sherman *et al* 2006 NAR 34

Génolevures 3

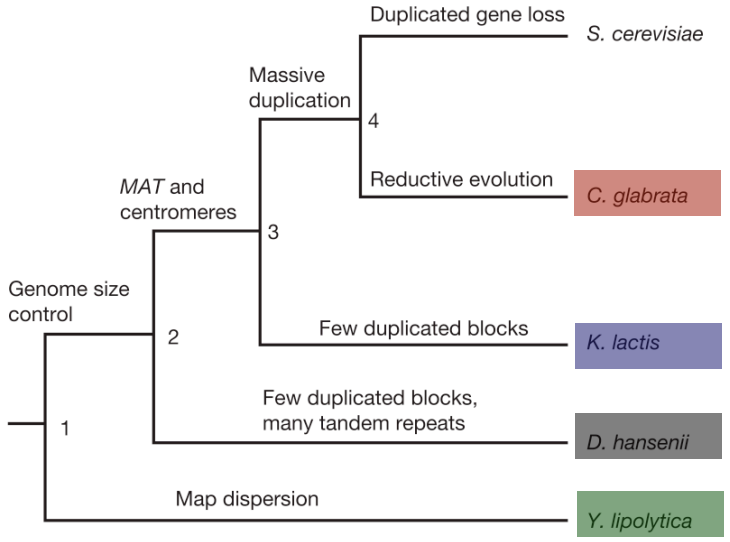
- 3 species complete 12X
- 2 species complete 7-12X

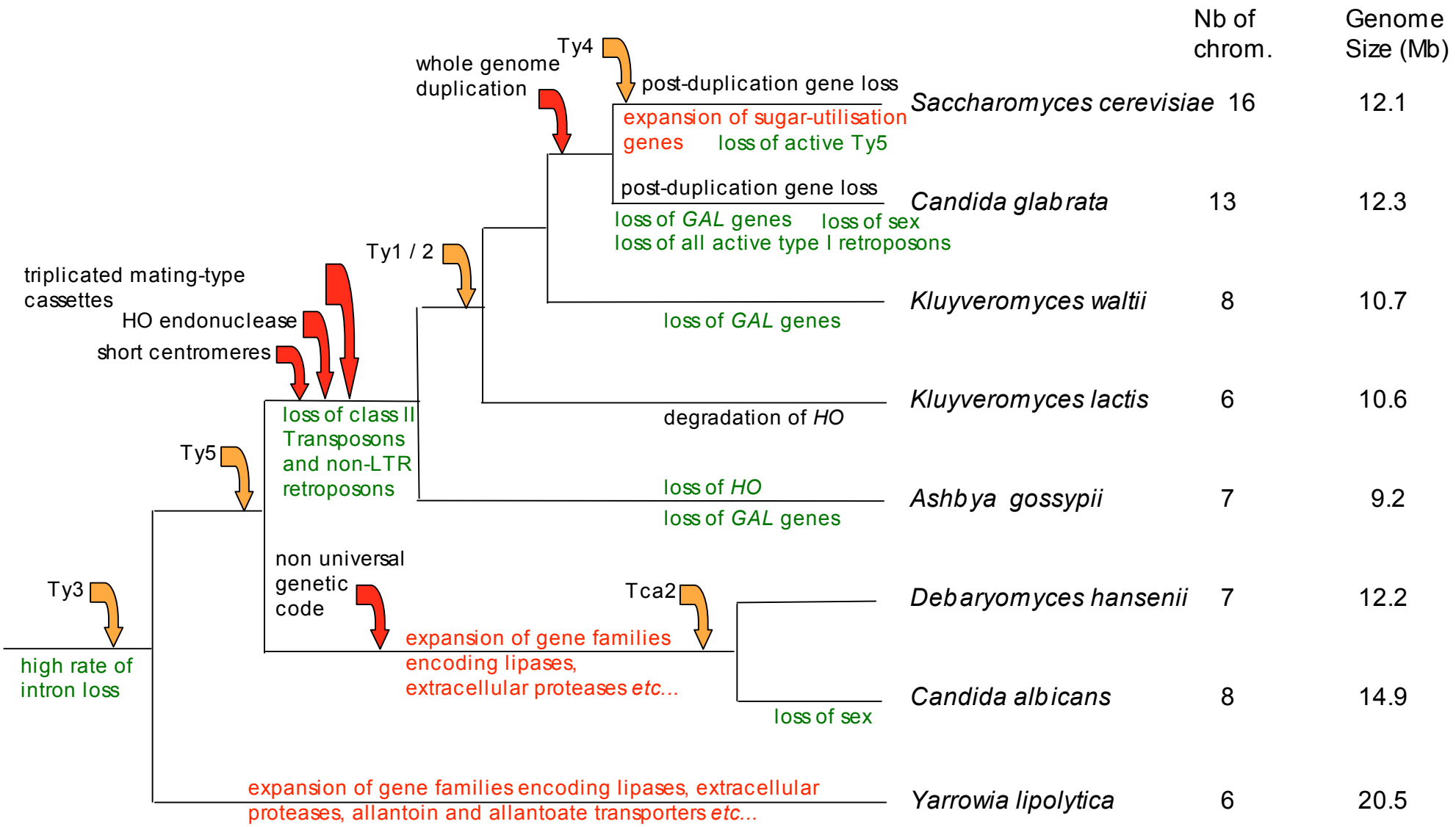
Génolevures 4

- 4 + 5 + 5 close species, NGS



Strains	Ploidy	Chromosome number	Code	# of reads	total (Mb)	Ref.
S. cerevisiae						
S. bayanus var. *uvarum* CLIB533 (1)	n	16	AS	5140	4.7	21
S. exiguus CBS 379 (2)	2n	14-16	AV	2579	2.4	22
S. servazzii CBS 4311 (2)	2n	7-12	AT	2570	2.4	23
Z. rouxii CBS 732 (2)	2n	7	AR	4934	4.5	24
S. kluyveri CBS 3082 (2)	2n	8	AU	2528	2.3	25
K. thermotolerans CBS 6340 (2)	2n	7	AY	2653	2.4	26
K. lactis CLIB210 (3)	n	6	BA	6080	5.6	27
K. marxianus var. *marxianus* CBS 712 (2)	?	10	AZ	2493	2.3	28
P. angusta CBS 4732 (2)	2n	?	BB	5082	4.7	29
D. hansenii var. *hansenii* CBS 767 (2)	n	?	BC	2830	2.6	30
P. sorbitophila CBS 7064 (2)	2n	7	AX	4829	4.4	31
C. tropicalis CBS 94 (2)	2n	12	BD	2541	2.3	32
Y. lipolytica CLIB89 (4)	n	6	AW	4940	4.5	33
S. pombe				total 49199	45.3	





Dujon (2006) *Trends in Genetics* **22**: 375-387

Genomic data for complete genomes

Complete genomes sequenced by the Génoscope

What is complete?

- Sequence subtelomere to subtelomere
- Fully assembled chromosomes
- Careful manual annotation

What can you do with a complete sequence?

- Track chromosomal rearrangements
- Analyze species- or clade-specific gain or loss
- Measure expansion and contraction of protein families
- Look for long-range correlations

What's next?

Genome Annotation

- Magus annotation system
- Simultaneous annotation of putative homologs

Classification into protein families

- Consensus ensemble clustering

Comparative maps

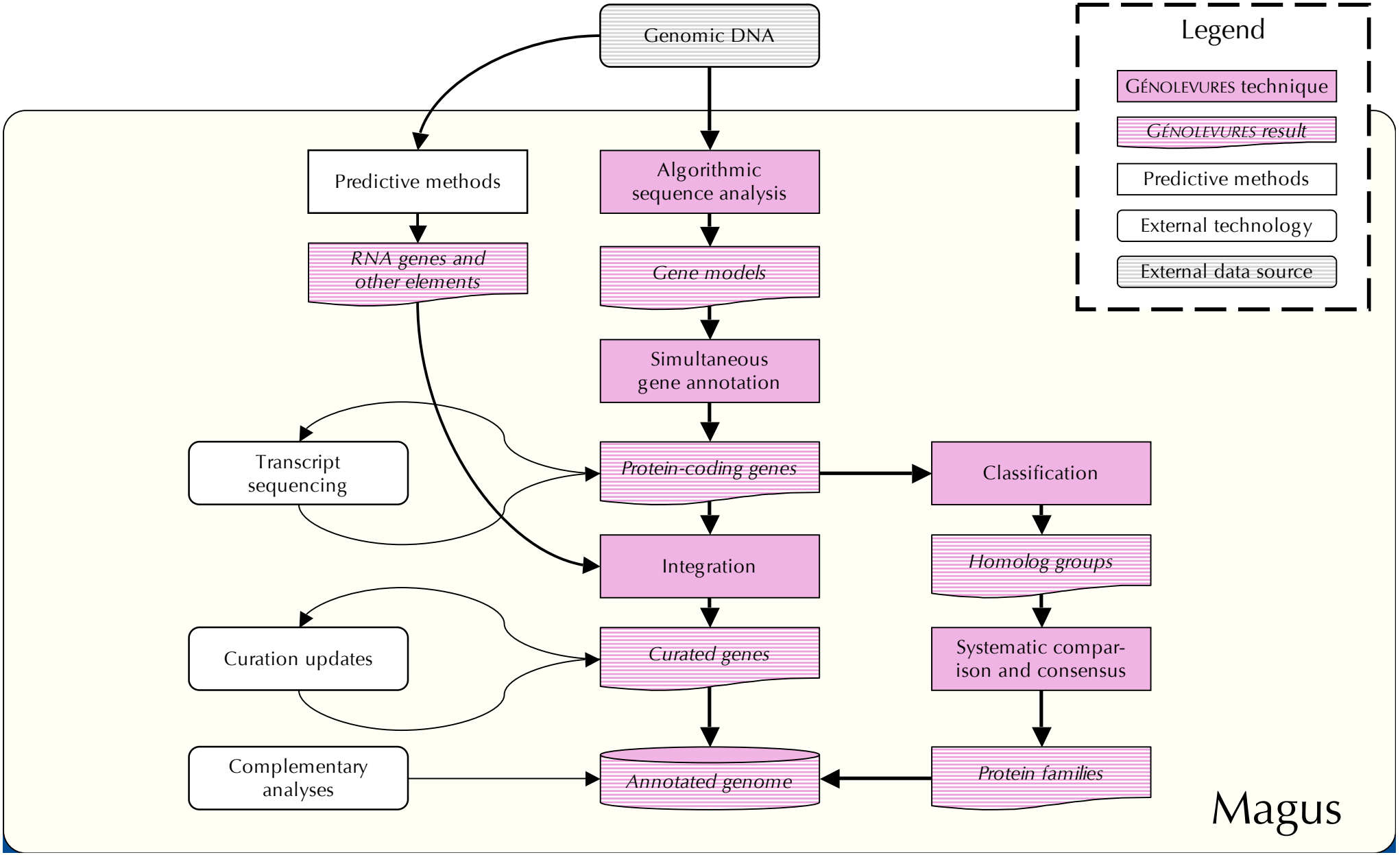
- Discovering synteny
- Identifying orthologs

Let's avoid teleology

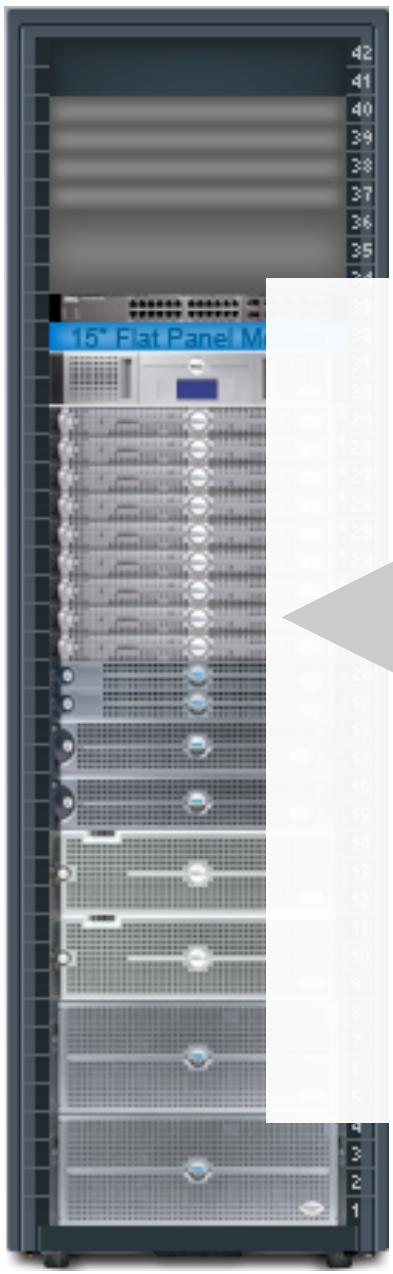
R. Greaves

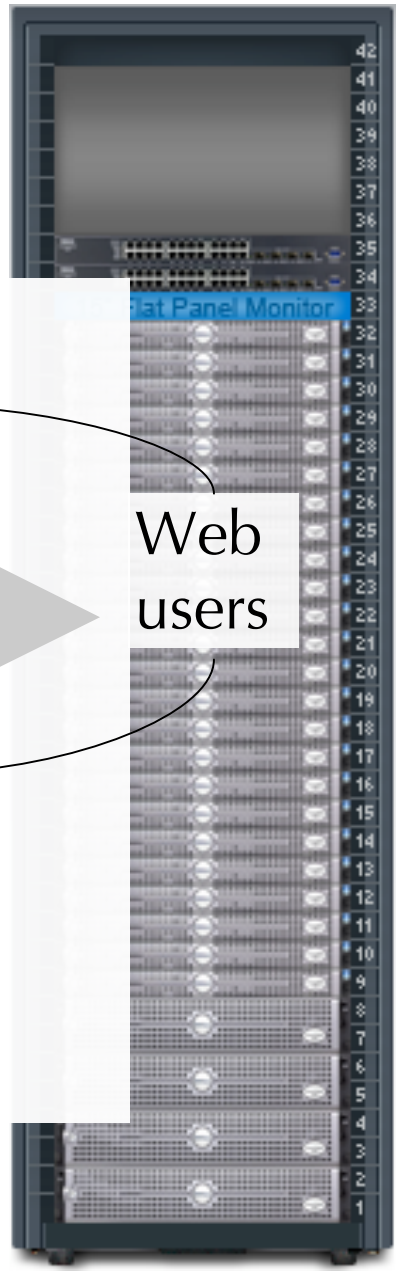
- Genomes are thrown together from bits and pieces of things that worked, once
- Genome annotation has to reflect reality and not expectations
- It is hard, painstaking work
- It is not fully automatic
- Good tools help

The Annotation Process



The “big iron”





In silico predictions

Alignments & DB search

Rule checker

U.I. components

Rules

Web Service Bus

Web users

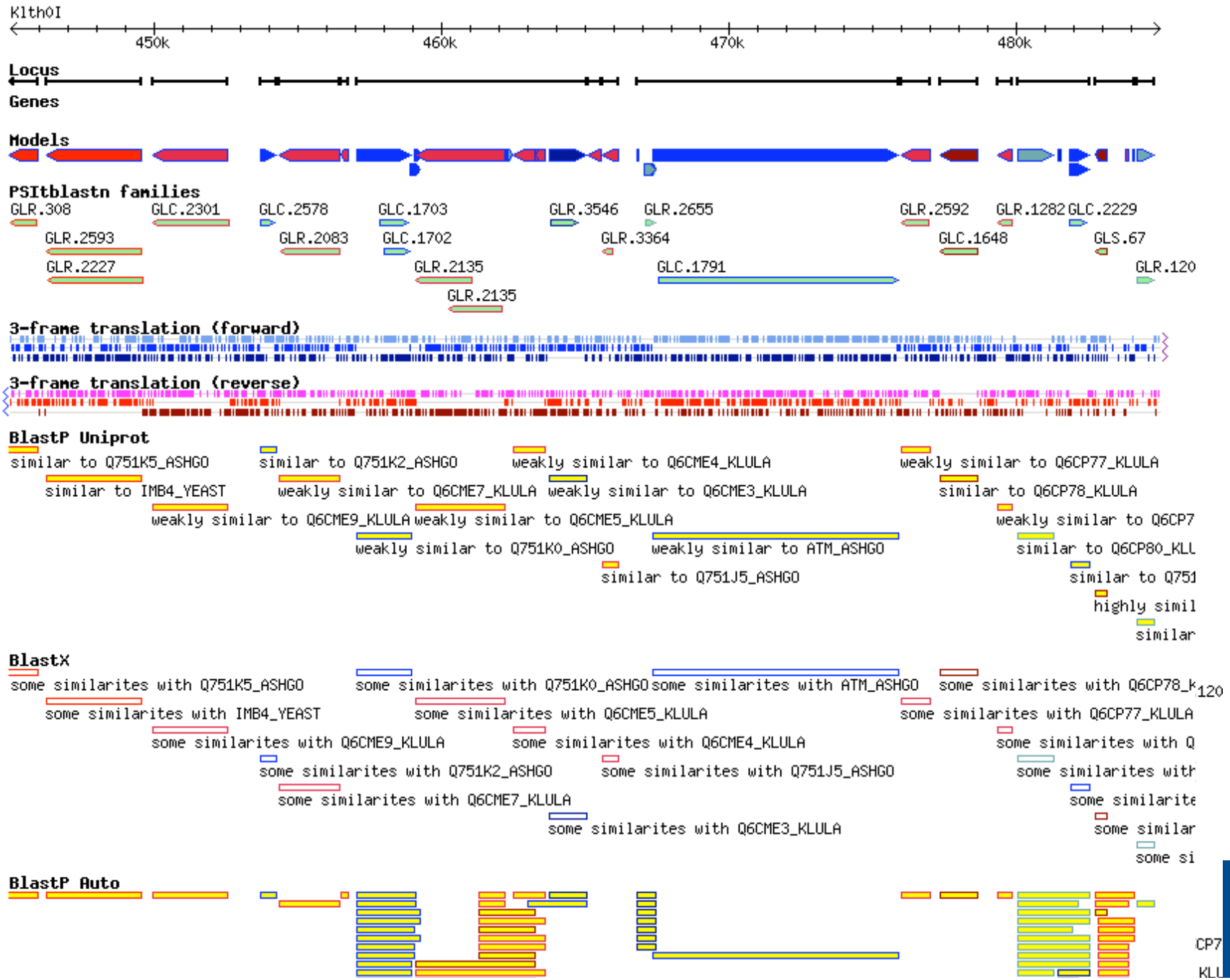
Genomes database

Compute d results KB

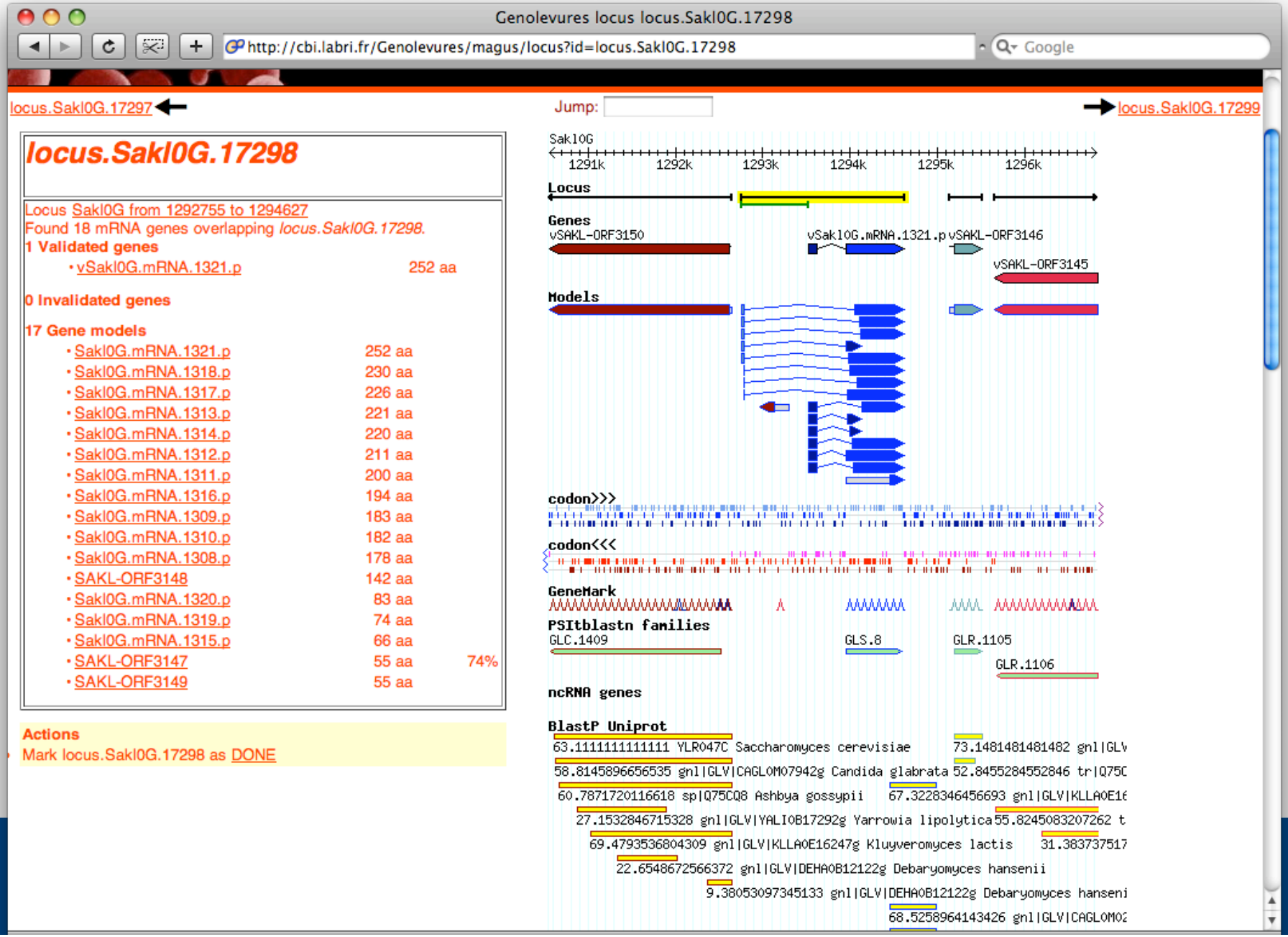
Genome Browser

Fast browser database

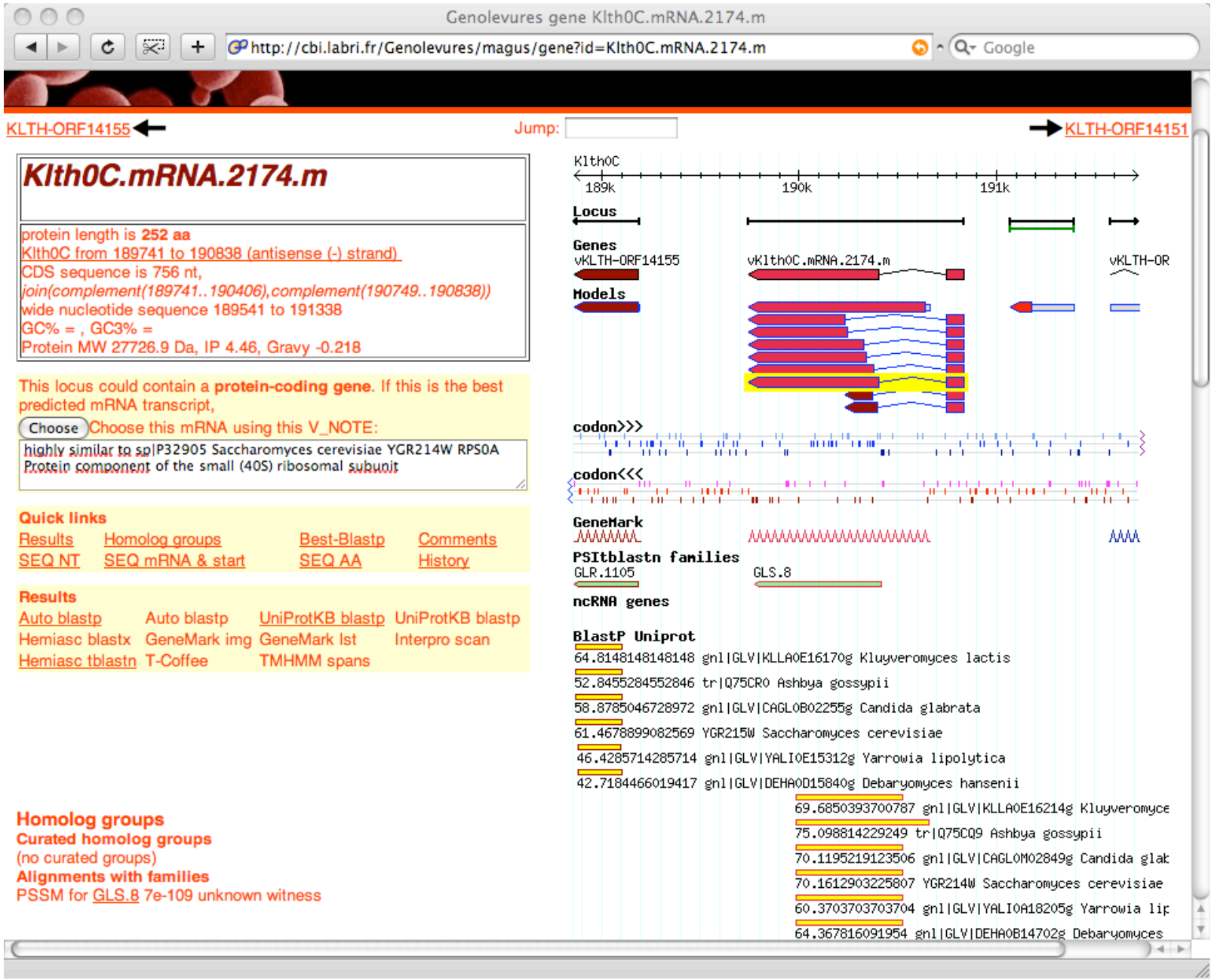
Browsing a genome region



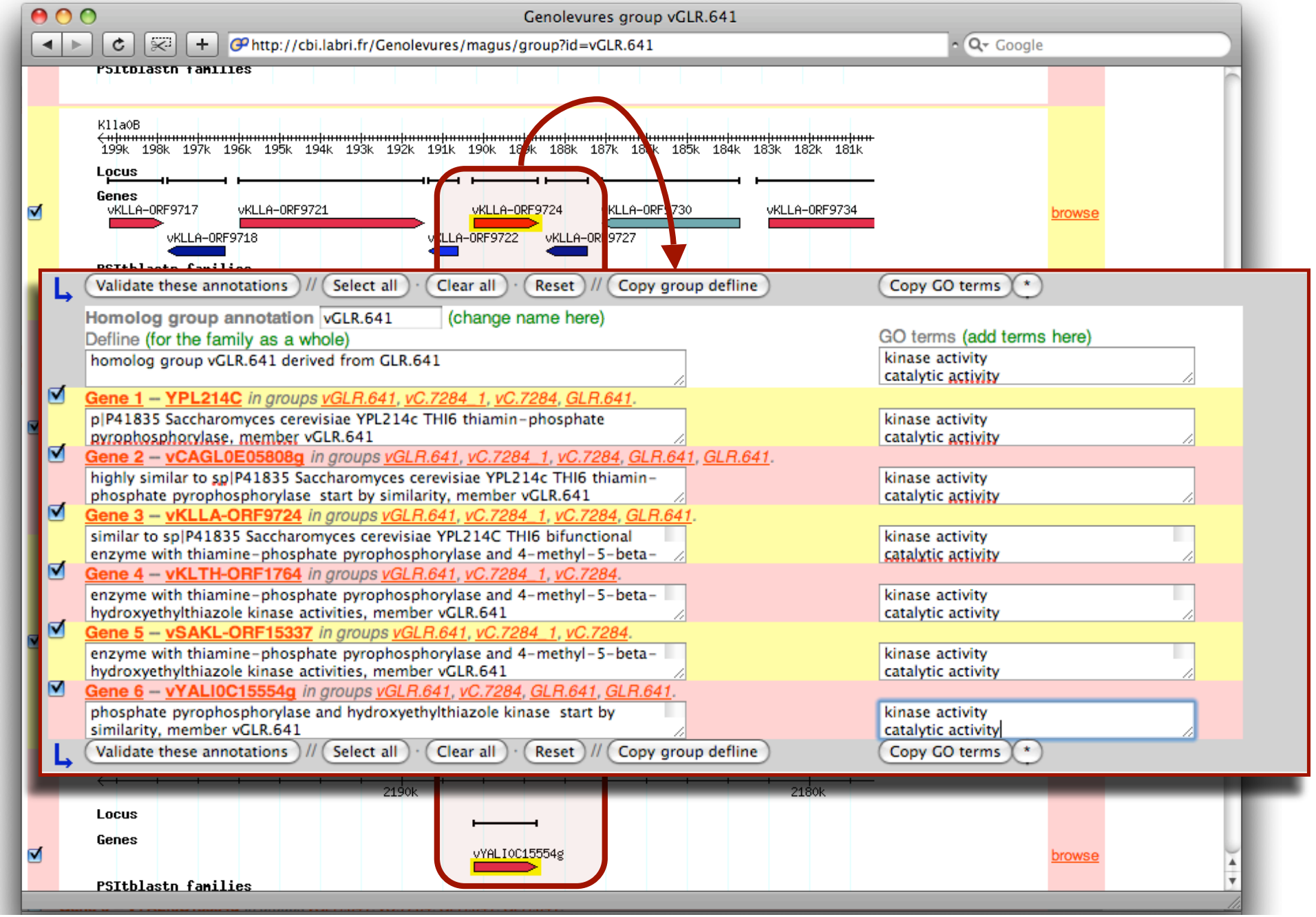
Viewing a Locus on a Genome



Validating a Gene Model



Annotating Homolog Groups



Protein families

Multi-species groups of related proteins

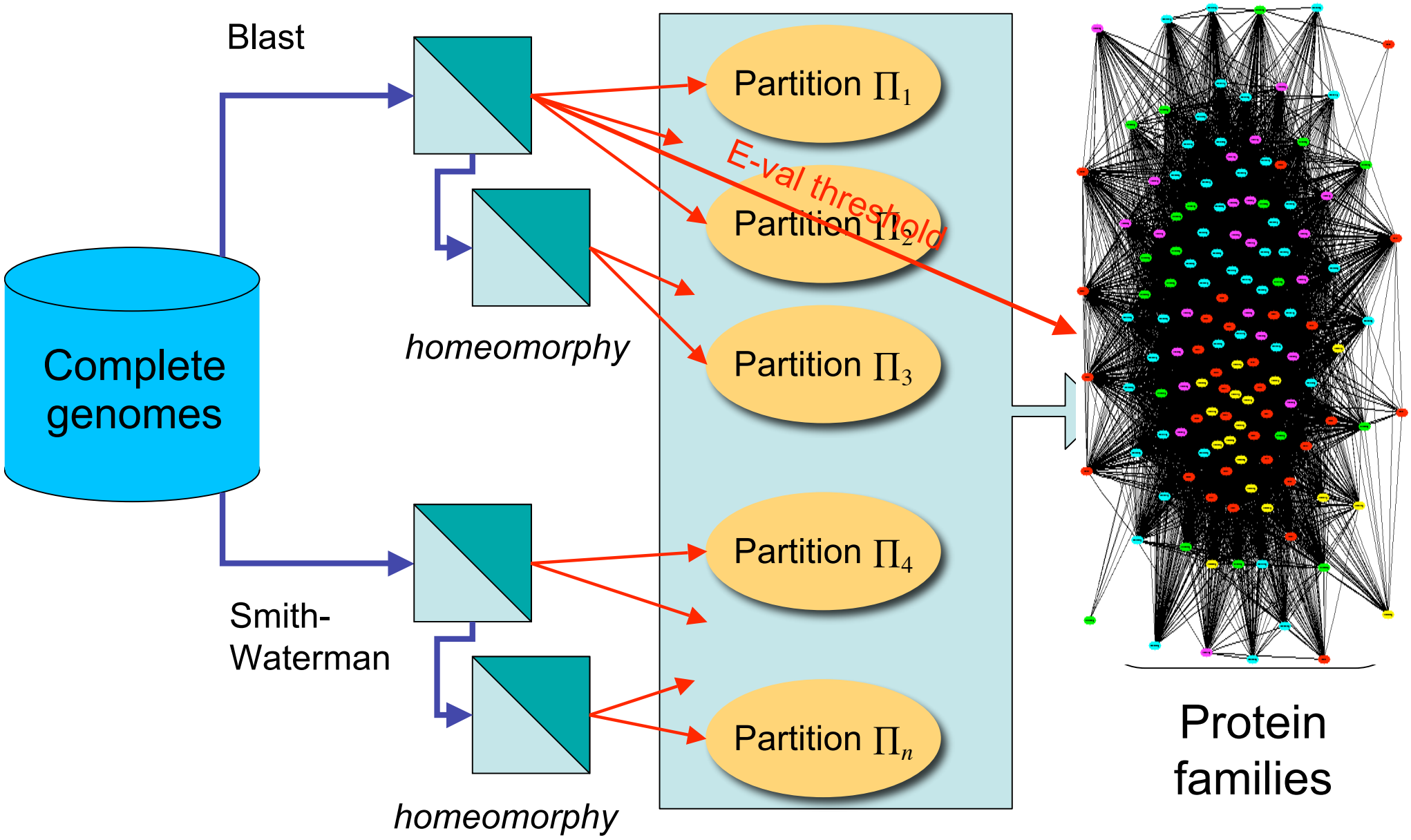
Phylogenetic relationship → functional similarity

Diversity of *in silico* results

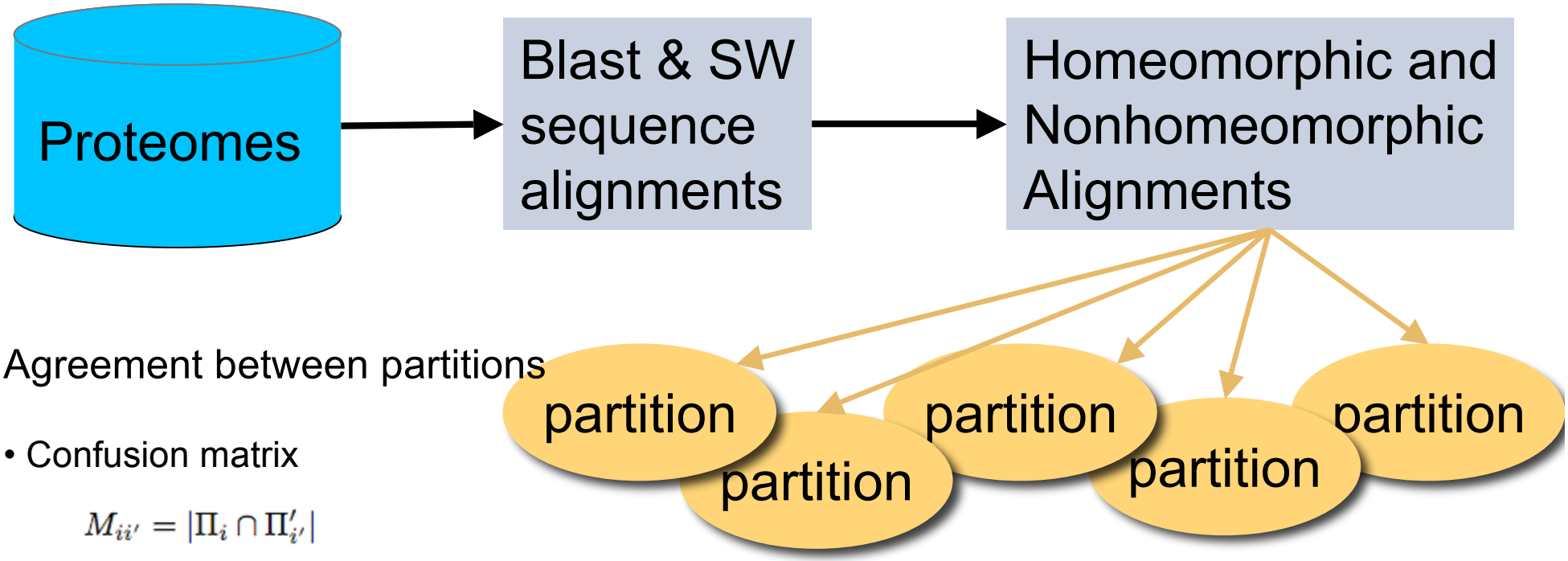
Need to calibrate or train methods for different phylogenetic groups

New algorithm for consensus clustering that is efficient in practice

What's the goal?



Reconciling different *in silico* predictions



Agreement between partitions

- Confusion matrix

$$M_{ii'} = |\Pi_i \cap \Pi'_{i'}|$$

- Distance between partitions

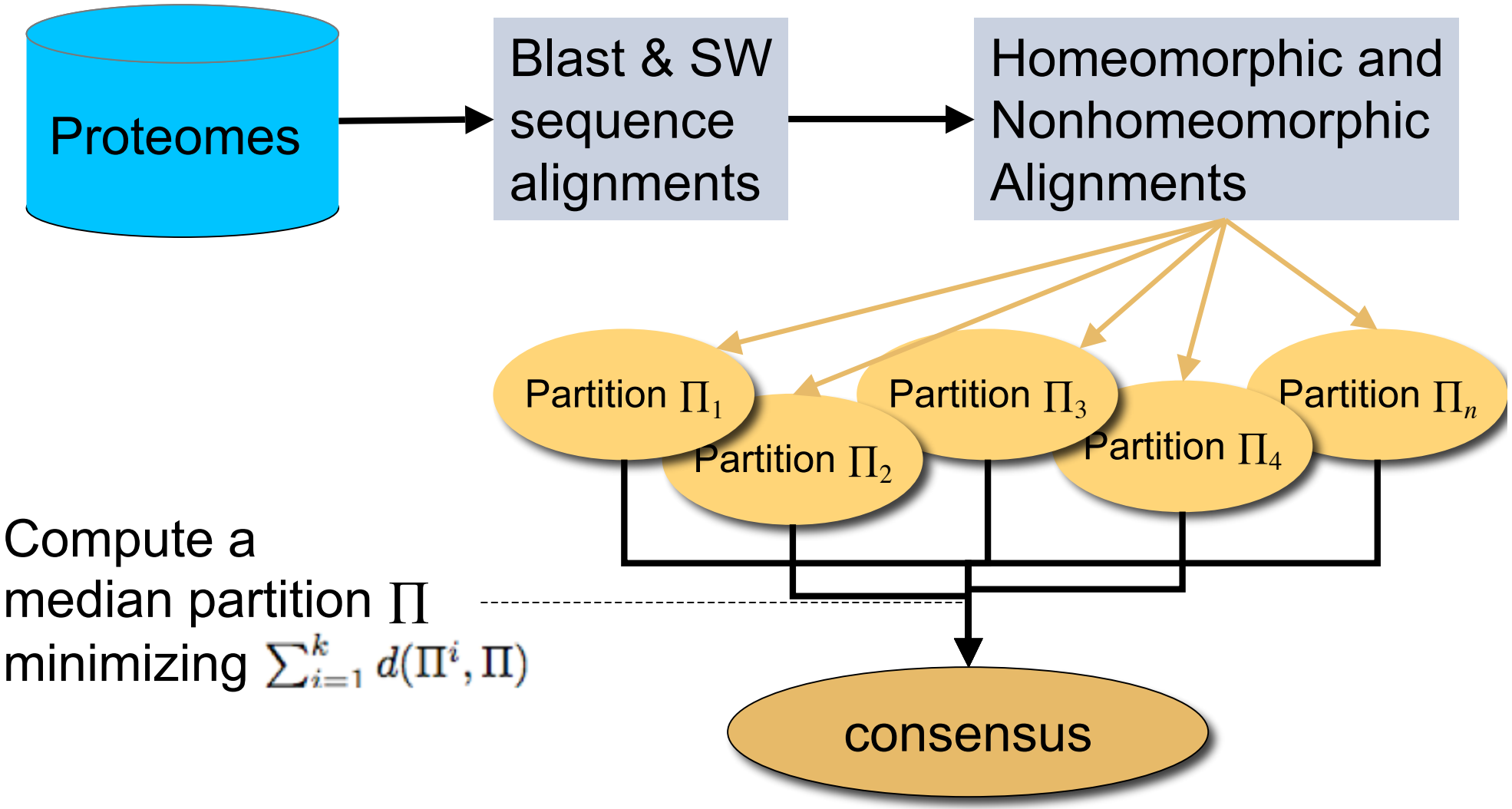
$$d(\Pi, \Pi') = 2n - \pi_{\Pi}(\Pi') - \pi_{\Pi'}(\Pi)$$

that is, a shortest path in
a graph of fusions/fissions

Consensus clustering (CC): Given k partitions, $\Pi^1, ..., \Pi^k$, find a consensus partition Π that minimizes $S = \sum_{i=1}^{k} d(\Pi^i, \Pi)$ where d is a distance.

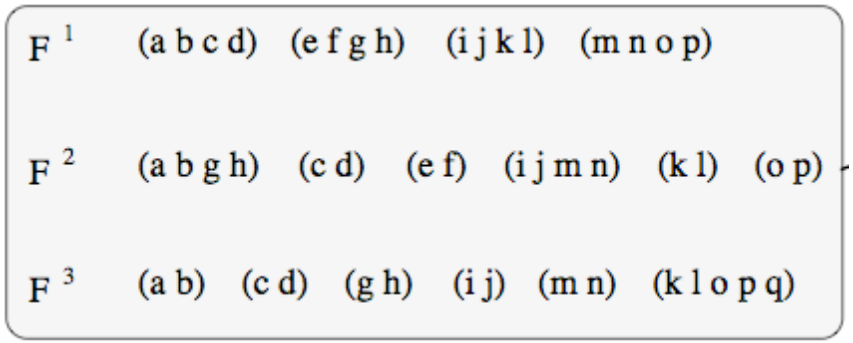
NP-complete

Median partitions by consensus clustering

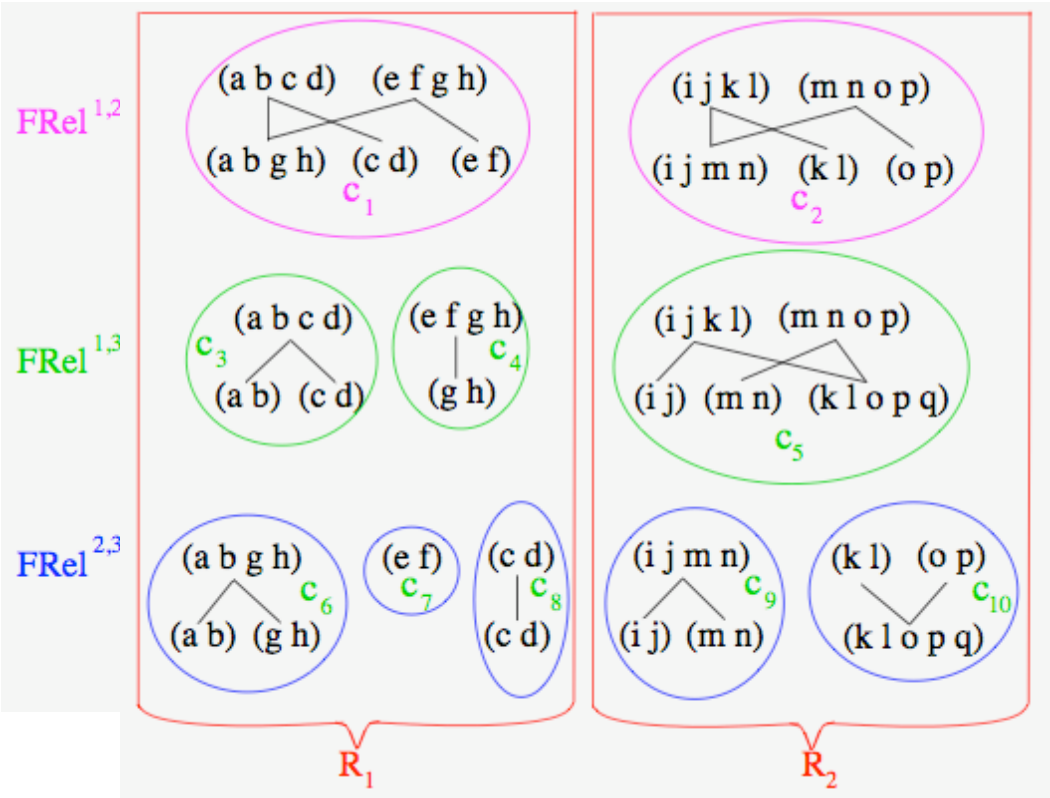


Compute a median partition Π minimizing $\sum_{i=1}^{k} d(\Pi^i, \Pi)$

Construction and algorithm



$\mathrm{FRel}^{i,j}$: encodes confusion matrix



Define a similarity measure based on the composants c_i

R_k maximal conflict regions

Select c_i in each R_k by MDC (*min. disjoint cover*)

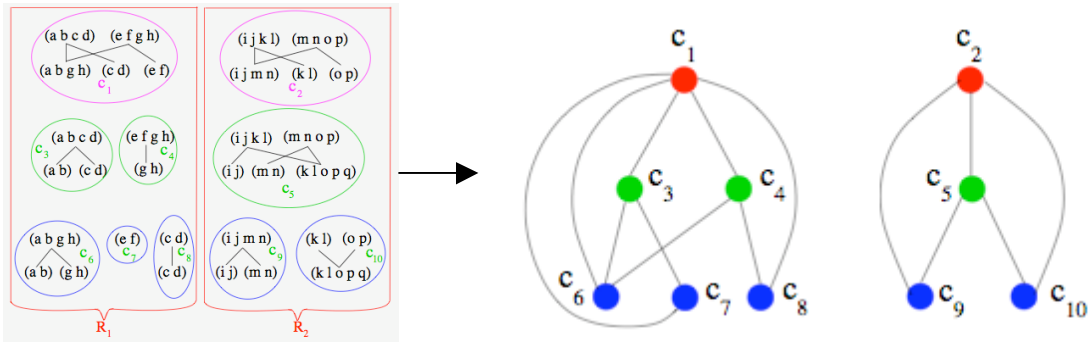
NP-complete

Efficient heuristic

Relaxation: admit *inexact* cover

(Not all proteins are in families)

Resolve conflicts by *election* + *policy*



Conflict regions ⟶ Conflict graph

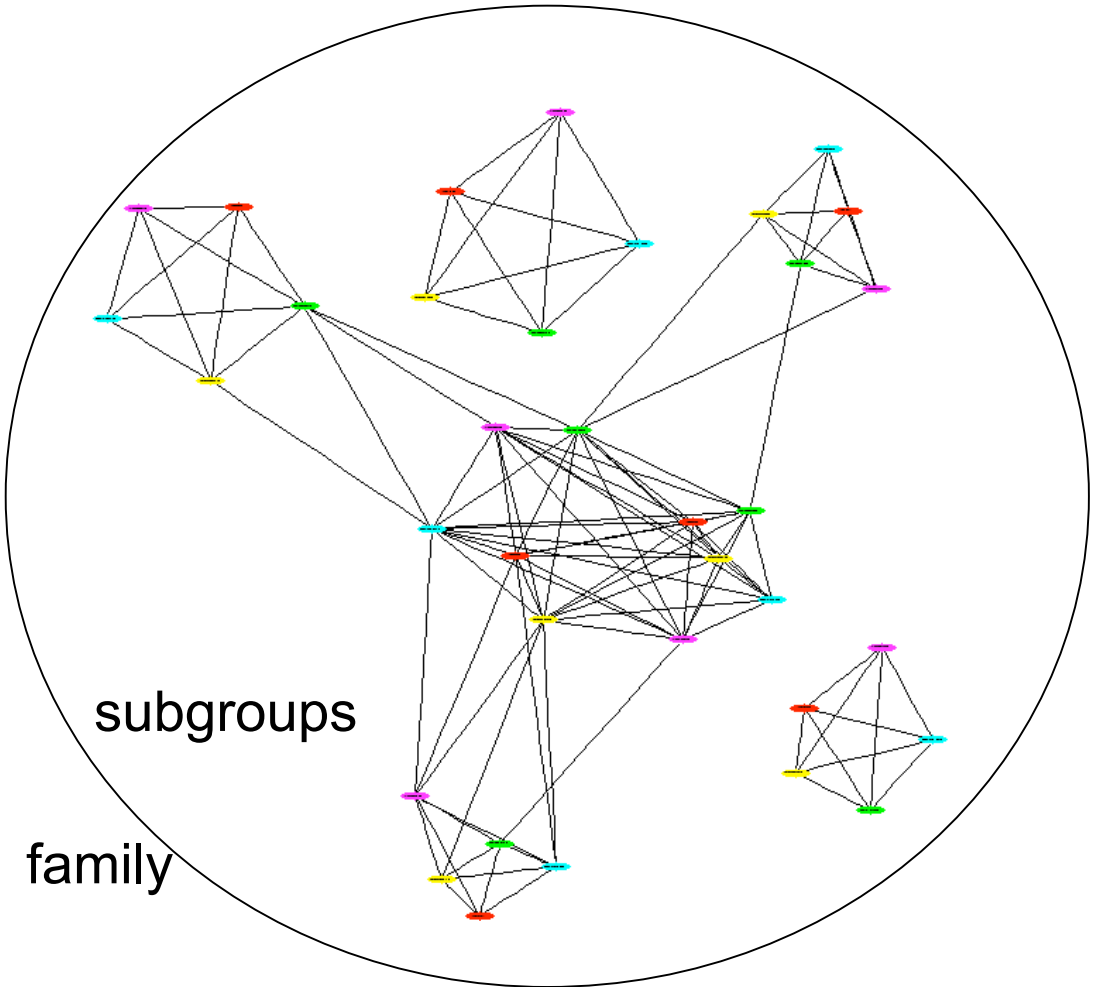
For each comp. C
  for each $c_i \in C$
    compute S_i et D_i
  each p votes for c_i
    in ordre $D_i \uparrow$ and $S_i \downarrow$
  take the winning c_i
    in order so as to
    cover the most
    proteins p

$$n^2 O(m^2) + O(|\mathbb{P}|^2)$$

GLC.1720 consensus family

index -- GLC.1720 -- cartoon -- deflines -- GO terms -- comparisons

Family	GLC.1720
Phyletic pattern	sckdy
Phylogen. profile	7 7 7 7 7
Proteins	YBL076c YGR094w YGR171c YGR264c YLR382c YPL040c YPL160w CAGL0F08459g CAGL0G03091g CAGL0G03311g CAGL0H05049g CAGL0I08723g CAGL0J03652g CAGL0L12606g KLLA0A09647g KLLA0D06105g KLLA0D11858g KLLA0D14971g KLLA0E05269g KLLA0E20625g KLLA0F14971g DEHA0A05060g DEHA0A06413g DEHA0C08525g DEHA0F08228g DEHA0F21648g DEHA0G21395g DEHA0G21549g YALI0A00264g YALI0C10780g YALI0D03619g YALI0E24607g YALI0F02299g YALI0F20218g YALI0F29843g (fasta)



Deflines

DEHA0F08228g	highly similar to trlQ9HGT2 Candida albicans Cytosolic leucyl-tRNA synthetase, start by similarity [Debaryomyces hansenii] Complete CDS. DEHA0F08228g|DEHA-IPF8483|DEHA-CDS0330.1
YGR264c	splP00958 Saccharomyces cerevisiae YGR264C MES1 Methionyl-tRNA synthetase, forms a complex with glutamyl-tRNA synthetase (Gus1p) and Arc1p, which increases the catalytic efficiency of both tRNA synthetases; also has a role in nuclear export of tRNAs
KLLA0A09647g	similar to splP00958 Saccharomyces cerevisiae YGR264c MES1 methionyl-tRNA synthetase singleton, start by similarity [Kluyveromyces lactis] Complete CDS. KLLA0A09647g|KLLA-IPF8320|KLLA-CDS0813.1
YLR382c	splP11325 Saccharomyces cerevisiae YLR382C NAM2 Mitochondrial leucyl-tRNA synthetase, also has a direct role in splicing of several mitochondrial group I introns; indirectly required for mitochondrial genome maintenance

Related GO terms (from *S. cerevisiae* genes)

GO:0000372 Group I intron splicing
GO:0003729 mRNA binding
GO:0004822 isoleucine-tRNA ligase activity
GO:0004823 leucine-tRNA ligase activity
GO:0004825 methionine-tRNA ligase activity

MBE Advance Access originally published online on January 12, 2005
Molecular Biology and Evolution 2005 22(4):1011-1023; doi:10.1093/molbev/msi083

Research Article

Comparative Genomics of Hemiascomycete Yeasts: Genes Involved in DNA Replication, Repair, and Recombination

Guy-Franck Richard, Alix Kerrest, Ingrid Lafontaine and Bernard Dujon

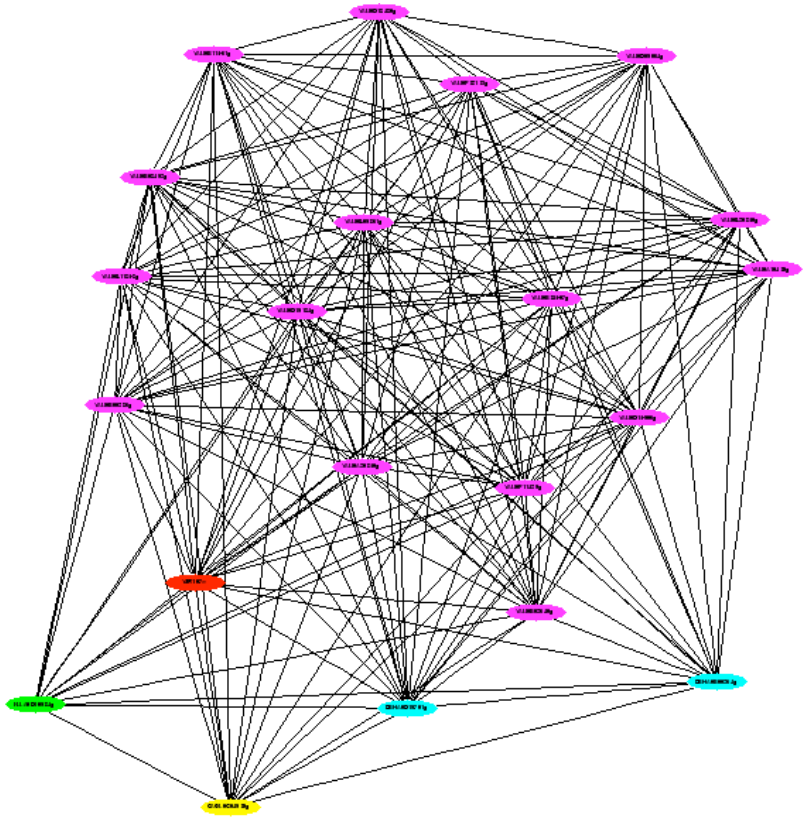
Unité de Génétique Moléculaire des Levures (URA 2171 CNRS, UFR 927 Université Pierre et Marie Curie), Institut Pasteur, Paris cedex, France

Correspondence: E-mail: gfrichar@pasteur.fr.

Correlated gain and loss and in networks and metabolic pathways

GLS.94 standard family

Family	GLS.94	*index* -- *GLS.94* -- *cartoon* -- *deflines* -- *GO terms* -- *comparisons*
Phyletic pattern	sckdy	
Phylogen. profile	1 1 1 2 16	
Proteins	YJR107w CAGL0C04939g KLLA0D10934g DEHA0D19701g DEHA0E00264g YALI0A10439g YALI0A20350g YALI0B09361g YALI0B11858g YALI0B20350g YALI0D09064g YALI0D15906g YALI0D18480g YALI0D19184g YALI0E00286g YALI0E02640g YALI0E08492g YALI0E11561g YALI0E34507g YALI0F11429g YALI0F32131g (fasta)	

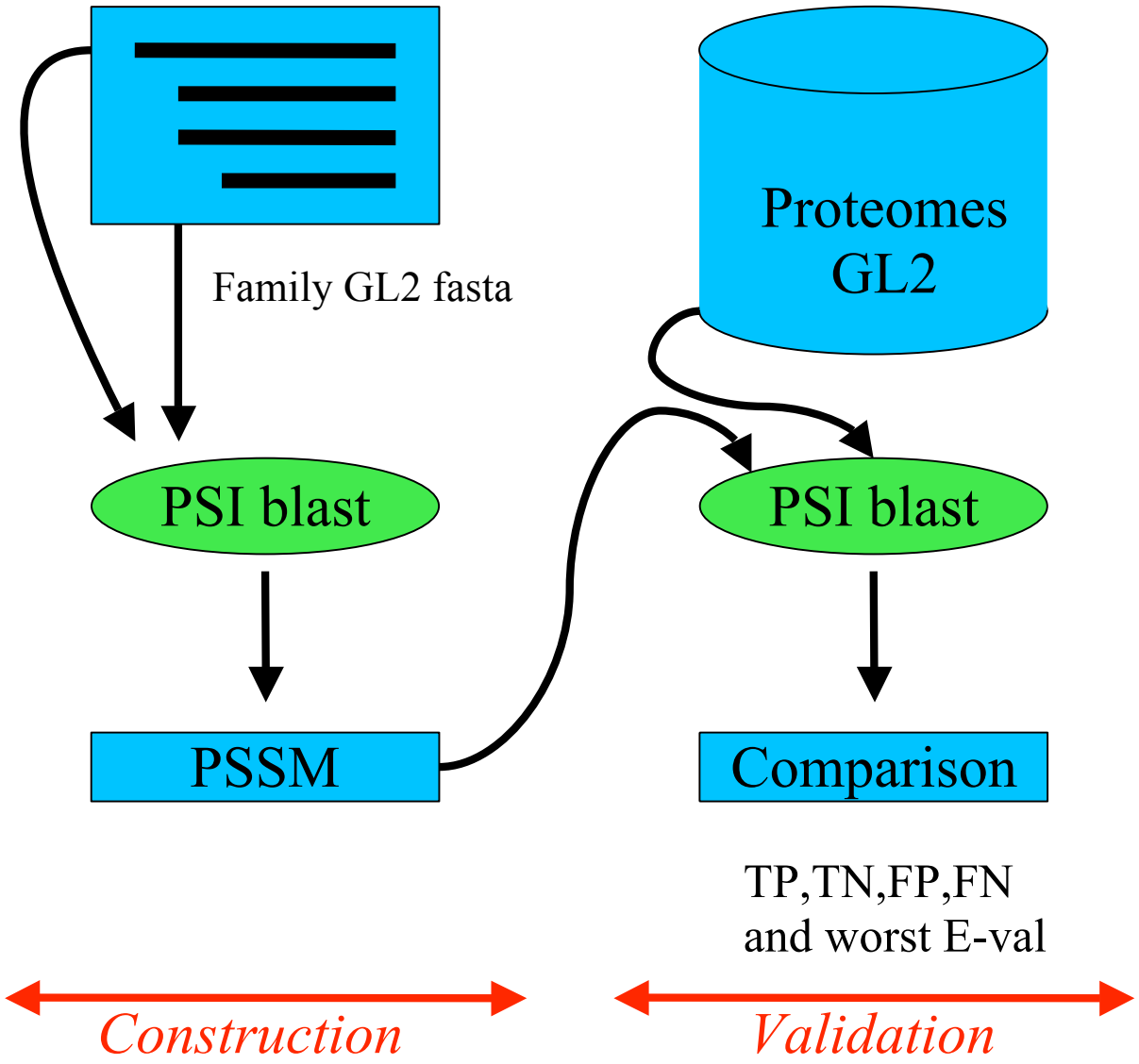


YALI0E11561g	similar to sp|CAD70713 Yarrowia lipolytica lipase LIP4, hypothetical CDS. YALI0E11561g|YALI-IPF2705|YALI-CDS3171.1
YALI0E34507g	weakly similar to tr|Q9P8F7 Yarrowia lipolytica Triacylglycerol lipase start [Yarrowia lipolytica] Complete CDS. YALI0E34507g|YALI-IPF
YALI0F11429g	weakly similar to tr|Q9P8F7 Yarrowia lipolytica Triacylglycerol lipase sp|P47145 Saccharomyces cerevisiae YJR107w, hypothetical start [Ya YALI0F11429g|YALI-IPF180|YALI-CDS4484.1
YALI0F32131g	similar to tr|CAD70715 Yarrowia lipolytica lipase, start by similarity [YALI0F32131g|YALI-IPF1763|YALI-CDS3658.1
YJR107w	sp|P47145 Saccharomyces cerevisiae YJR107W Hypothetical ORF

Related GO terms (from *S. cerevisiae* genes)

GO:0000004 biological process unknown
GO:0008372 cellular component unknown
GO:0016298 lipase activity

Construct a PSSM for each family

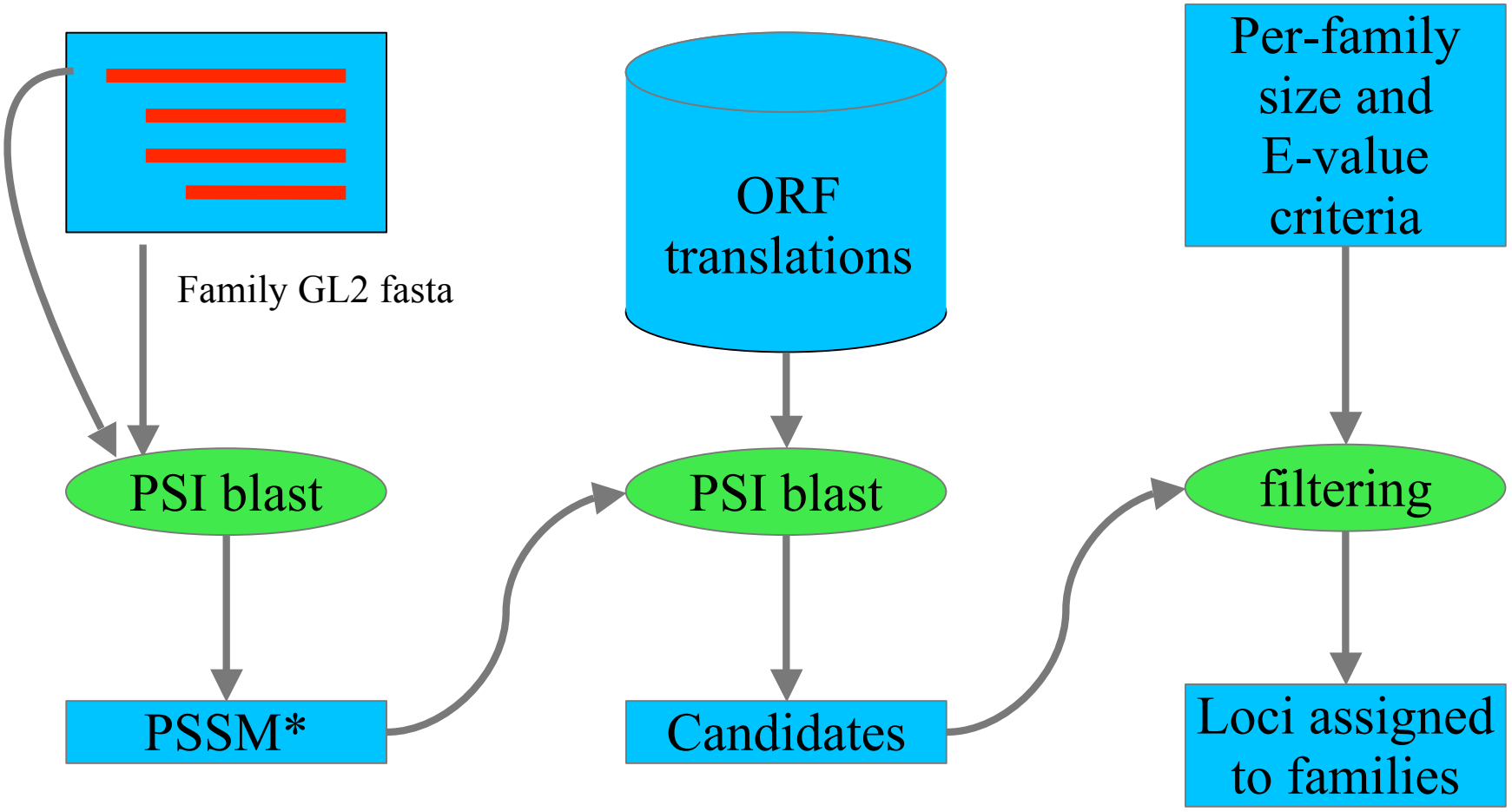


4384 families as follows

4240 where FN = 0
FP med 0,0
avg 3,7
max 302
Ev med 6e-78
max 9e-6

144 where FN > 0
FP med 4,5
avg 33
max 307

Build a PSSM for each family and use to improve gene prediction



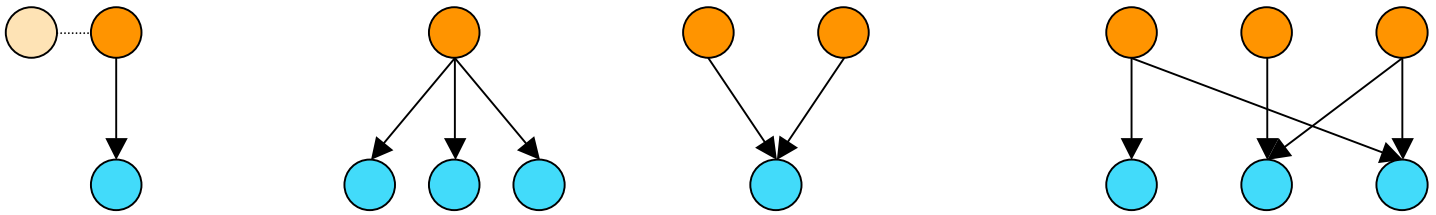
**PSSM: position-specific scoring matrix for PSIBLAST*

Comparison with KOGs

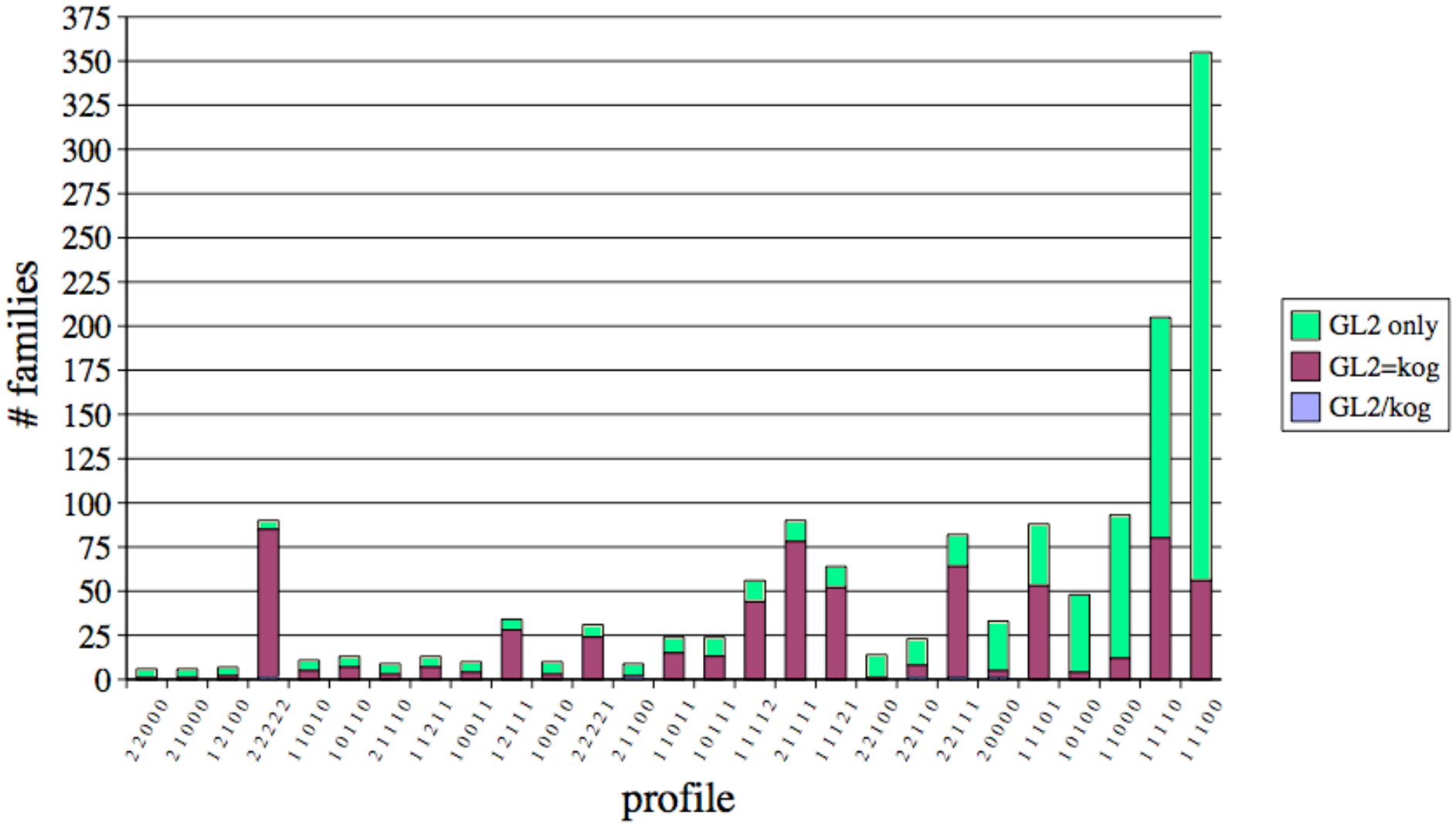
Project families on *S. cerevisiae*

Select intersection and compare
3625 proteins (~2500 families)

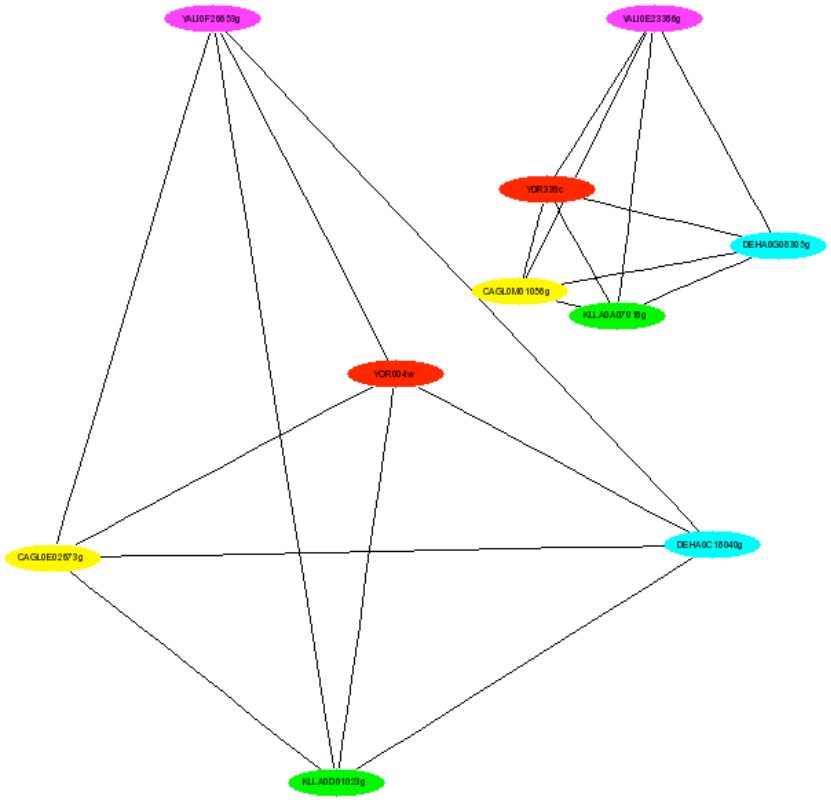
identities:	1901	
split:	159	(4 GLS, 42 GLR, 113 GLC)
merge:	117	(6 GLS, 70 GLR, 79 GLC)
messy:	25	(2 GLS, 13 GLR, 23 GLC)



Comparison with KOGs

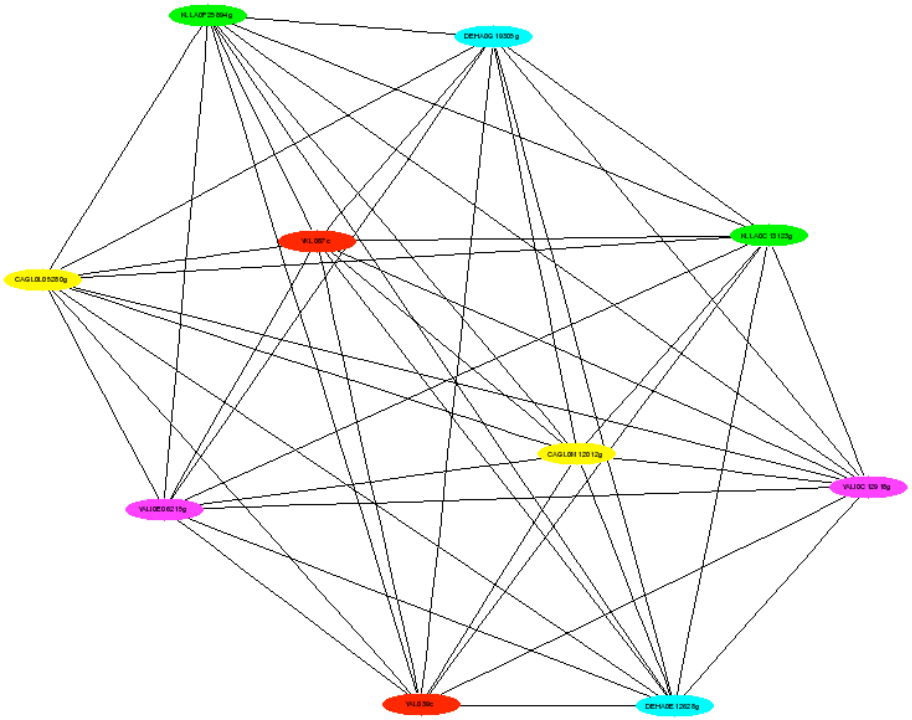


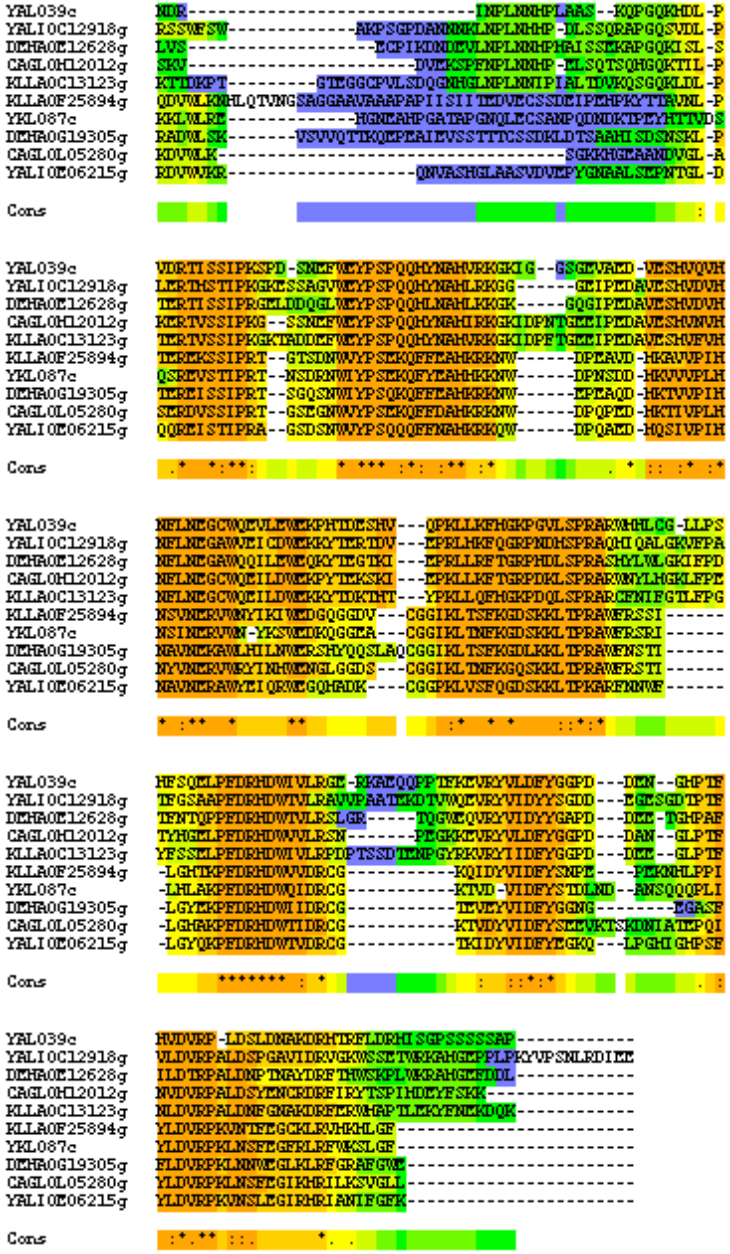
Comparison GLR.3294 with PIRSF (UniProt) 017196 and 017667




```
KLLA0D01023g  --YNHTEKFRQPYQVLVDDQLVLETFNSSFDFVKGLQRTLQA-EVKPHITQCCHQSL
CAGL0E02673g  --YNHTEKFREPYQVLLDDQIVHDSTTSKYDLVKGLKRTLQA-EVKPHITQCCHQAL
YOR004w       --YSHTEKFREPYQVLVDNQLVLECNNSNFNLPSGLKRTLQA-DVK-VITQCCIQAL
DEHA0C18040g  --YLHTEKFREPFQTIVDDEIILNCCKASFDIAKGLNRTIQG-ETKPHITQCSIEAL
YALI0F26653g  --YHLNFKFREPYQVLVDHDLIIECVEHKYDLLKGLQRTLQTKEVKPHITQHSLMKL
CAGL0H01056g  LFFQYNEAIKPPYQVLIDTNFINFSIQKKIDIVKGHHDCLLA-KCNPLITDCVIAEL
KLLA0A07018g  LFFQYNEAIKPPYQVLIDTNFINFSIQKKIDIVKGHHDCLLA-KCNILITDCVHAEL
YDR339c       LFFQYNQAI-PPYQVLIDTNFINFSIQKKVDIVRGHHDCLLA-KCNPLITDCVHAEL
DEHA0G08305g  LFFQFNKSIKPPYQILIDTNFINFSIQKKIDITKGHHDTLFA-KCIPHITDCVHAEL
YALI0E23386g  LFFQYNEAIKPPYQVLIDTNFINFSIQKKLDIHRAHHDCLYA-KANPIVTDCVHAEL

Cons          :  .  :  *:* ::* :::     . ::  .:  :   .  ::*:  :  *

KLLA0D01023g  YQTNNQGAI--DAGKQFERRRCNHDPKEP----------------KSVLECLTSVV
CAGL0E02673g  YETKNEHAI--DLGKEFERRRCGHHPGEA----------------VSPKECILNVV
YOR004w       YETRNDGAI--NLAKQFERRRCNHSFKDP----------------KSPAECIESVV
DEHA0C18040g  YKTNNQDAI--SIAKLFERRRCNHPPSNP----------------IPSSECIRSIV
YALI0F26653g  FESNNEEAI--QLGKTEERRRTLHFENSTETQKKKLKYNGSNAARKHHPSLAIYKSV
CAGL0H01056g  EKLGPKYRIALKLARDPRIKELTCTHKG----------------TYADDCIVNRV
KLLA0A07018g  EKLGPKYRIALKLARDPRIKELSCSHKG----------------TYADDCIVHRV
YDR339c       EKLGPKYRIALKL-RDPRIKELSCSHKG----------------TYADDCLVHRV
DEHA0G08305g  EKLGPKYRIALKLAKDPRIQRLSCSHKG----------------TYADDCLVNRV
YALI0E23386g  EKLGPKYRIALKLARDPRIKELTCTHKG----------------TYADDCLVNRV

Cons          :  .  *  .  :  . :*                          :.:  *

KLLA0D01023g  DVNGKNKHR-YVVATQDVEIRRRLRKIPGVPLVYHNRSVHVHEPLSNASEKVSREVE
CAGL0E02673g  DVKGKNKHR-YVVACQDVEIRRLLRKVPGVPLLHISRSVHIHEPLSDASAKVSREEE
YOR004w       NISGANKHR-YVV-SQDIDLRRKLRTVPGVPLIHLTRSVHVHEPLSTASAKASKITE
DEHA0C18040g  DINGENKHR-YLVATQDKLLRNKLSRVPGVPLIYHNRSVHVHEPHSEATTTYSNSVE
YALI0F26653g  VVDGSNVHR-YVVATQDPSLRGRLRSIPGVPLIHHNRAIHLLEPFSPASERARQLIE
CAGL0H01056g  -----LQHKCYIVATNDAGLKQRVRKIPGIPLHSVGGHAYVIEKLPDVF--------
KLLA0A07018g  -----LQHKCYIVATNDAGLKQRVRKVPGIPLHSVGGHSYVIEKLPGVF--------
YDR339c       -----LQHKCYIVATNDAGLKQRIRKIPGIPLHSVGGHA-VIEKLPDVF--------
DEHA0G08305g  -----IQHKCYIVATNDADLKRRIRKVPGVPIHSVGGHAYVIEKLPDVF--------
YALI0E23386g  -----LQHKCYIVATNDQDLKRRLRKVPGIPLHGVGNKGYVIEKLPDVF--------

Cons          :  *: *:*  :*  ::  :  :**:*:: :     ::* :. .

KLLA0D01023g  EQKLYKGLNDPKFAGIARDE-NDEAGAENQENK-PKKKK-GPKEPNPLSHKKKKTTE
CAGL0E02673g  ESKLFRGLNDPKCYAGLKGEH-EEEEQSK-EQTV-AKKKKIGPKQPNPLSHKKKKKEN
YOR004w       EQKLYKGLNDPNIEKL------ESGDGSGKESI-TKKKKLGPKAPNPLSVKKKKKVN
DEHA0C18040g  RKKLTGGLNDLKVGKVDKHKLKTEGIEGTEGAPTKKKK-GPKEPNPLSHKKKKATN
YALI0F26653g  ASKLTSGLNAVKR----------KAPEGDEEQ-AKKKK-GPKAPNPLSVKKKKVSK
CAGL0H01056g  ------------------------------------------------------
KLLA0A07018g  ------------------------------------------------------
YDR339c       ------------------------------------------------------
DEHA0G08305g  ------------------------------------------------------
YALI0E23386g  ------------------------------------------------------

Cons

KLLA0D01023g  STGSISSLSASSNNISTPTEAGDVSGTSKKKRRRKHKKSNDKPNDSADSIQ------
CAGL0E02673g  TKK----------EQQTETSADQSQPTKRRRSKKHKHGSKNS---IDTNENGNLSE
YOR004w       SPS----------DE-VKDKKDTSKKKKKRRRRKHKNTNVPVSNGTTRAQ------
DEHA0C18040g  TDT----------KSNQDSKD-NLKPKTKRRRKHKSSSERTAVSLLPNGNSSNSS
YALI0F26653g  EDQ---------GKRERPEKKDDSQSVKKVRRKRAKRKDGEG---GESNEGGETDE
CAGL0H01056g  ------------------------------------------------------
KLLA0A07018g  ------------------------------------------------------
YDR339c       ------------------------------------------------------
DEHA0G08305g  ------------------------------------------------------
YALI0E23386g  ------------------------------------------------------
```

Comparison of GLR.3292 with PIRSF 017297 and 016767





Comparative maps

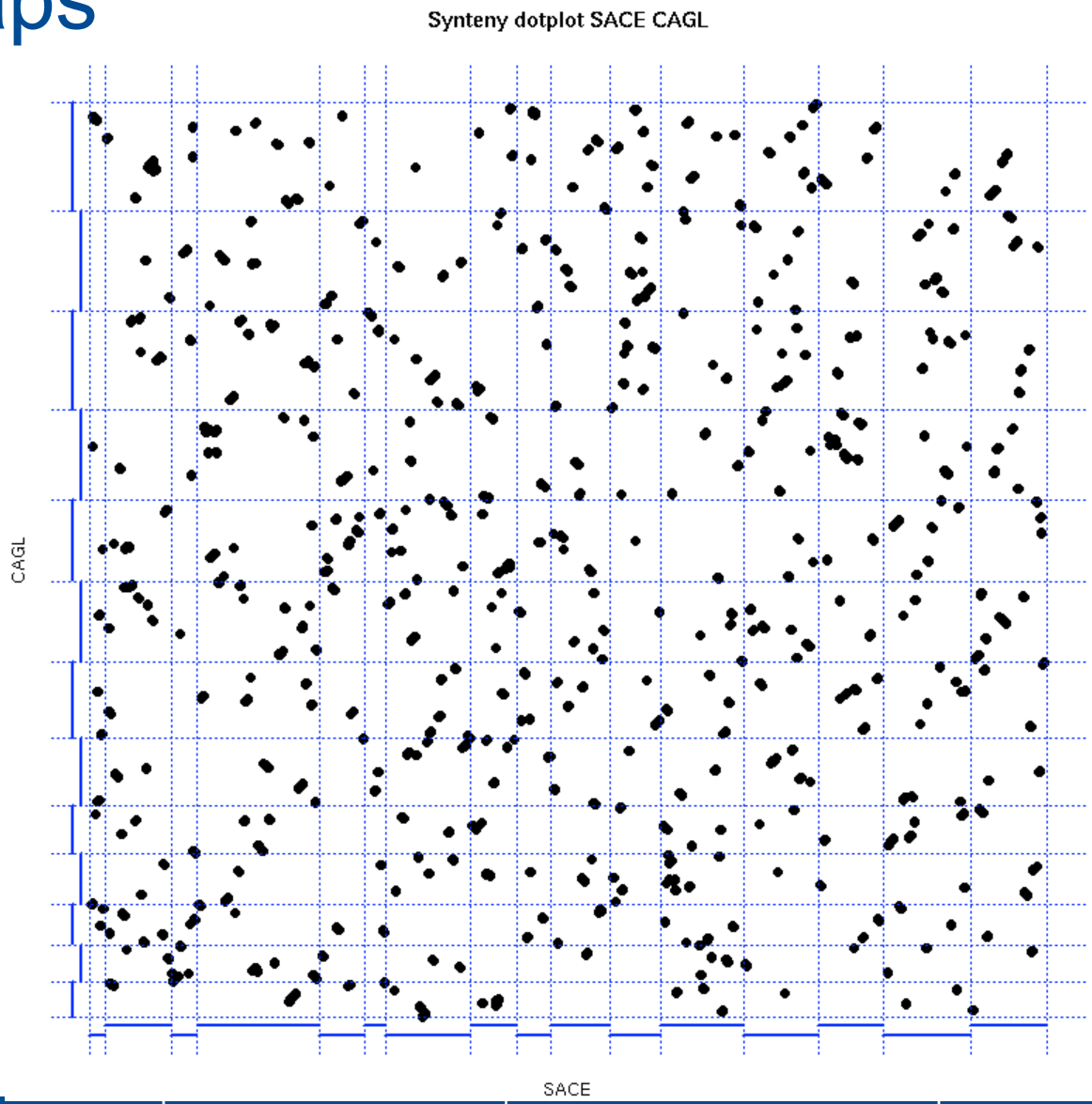
Despite similarity in size, gene content, ecological niche,

yeast genomes are highly rearranged.

In general, synteny is poorly conserved.

In part:

- evolutionary distance
- artifact of WGD

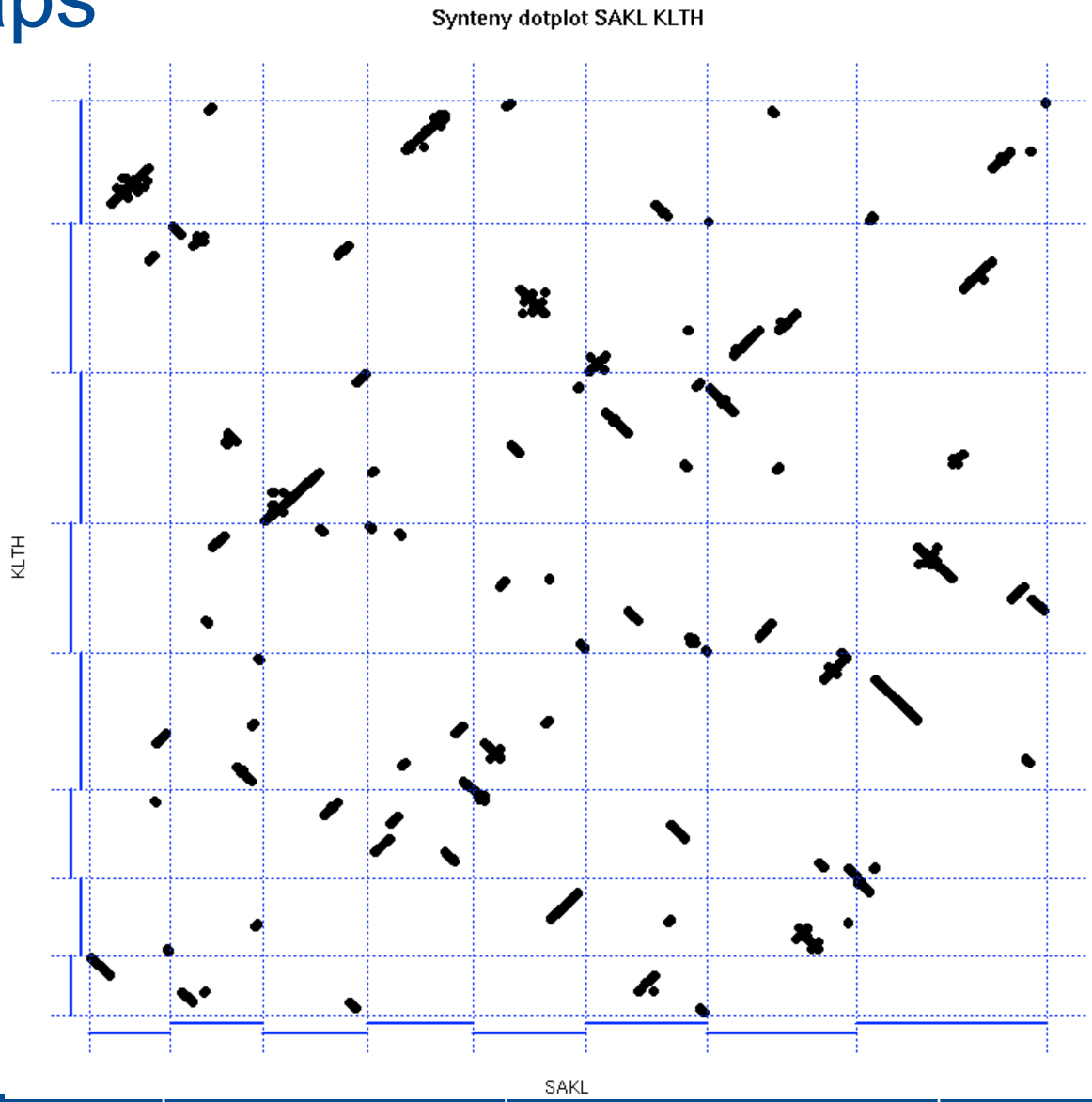


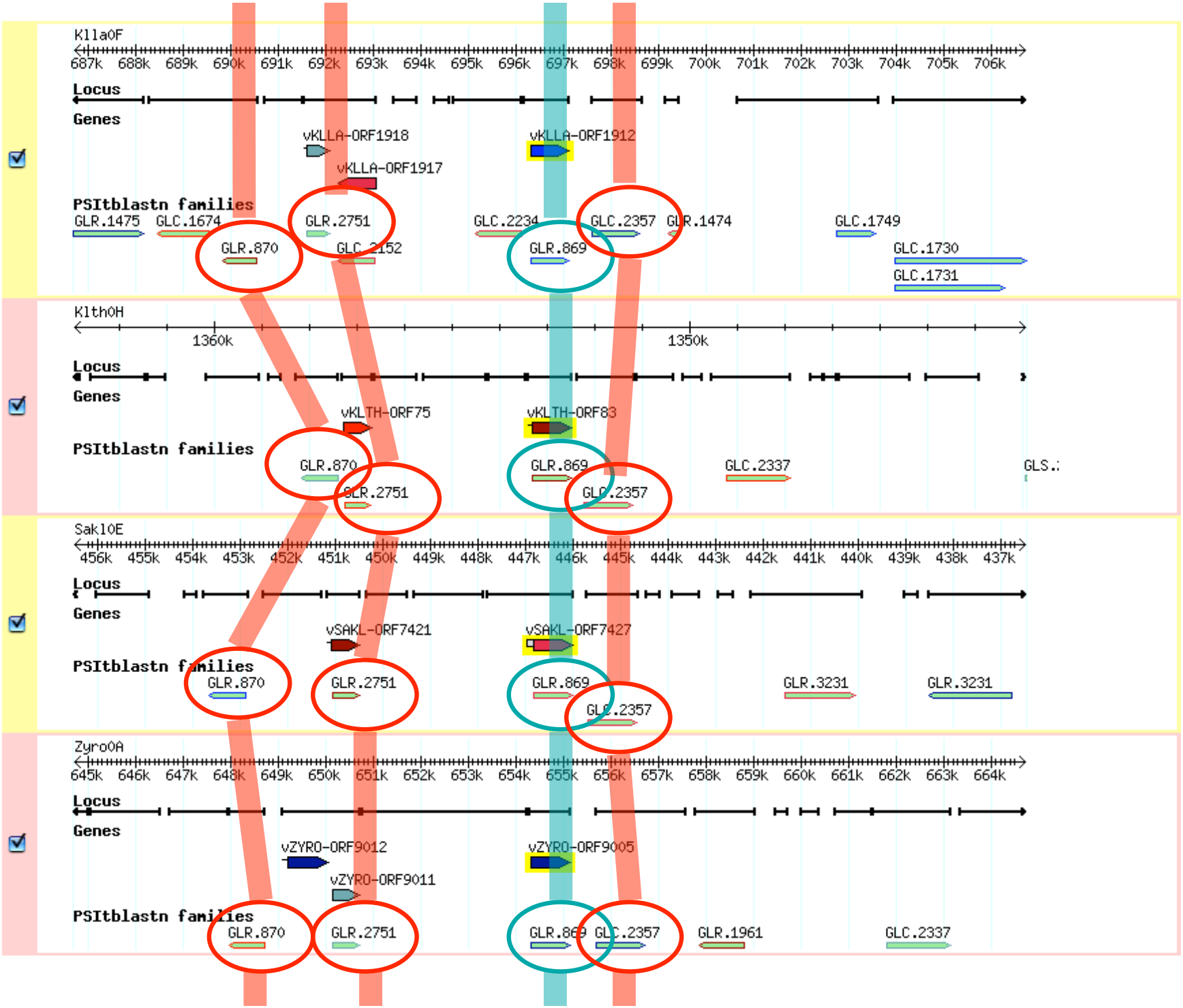
Comparative maps

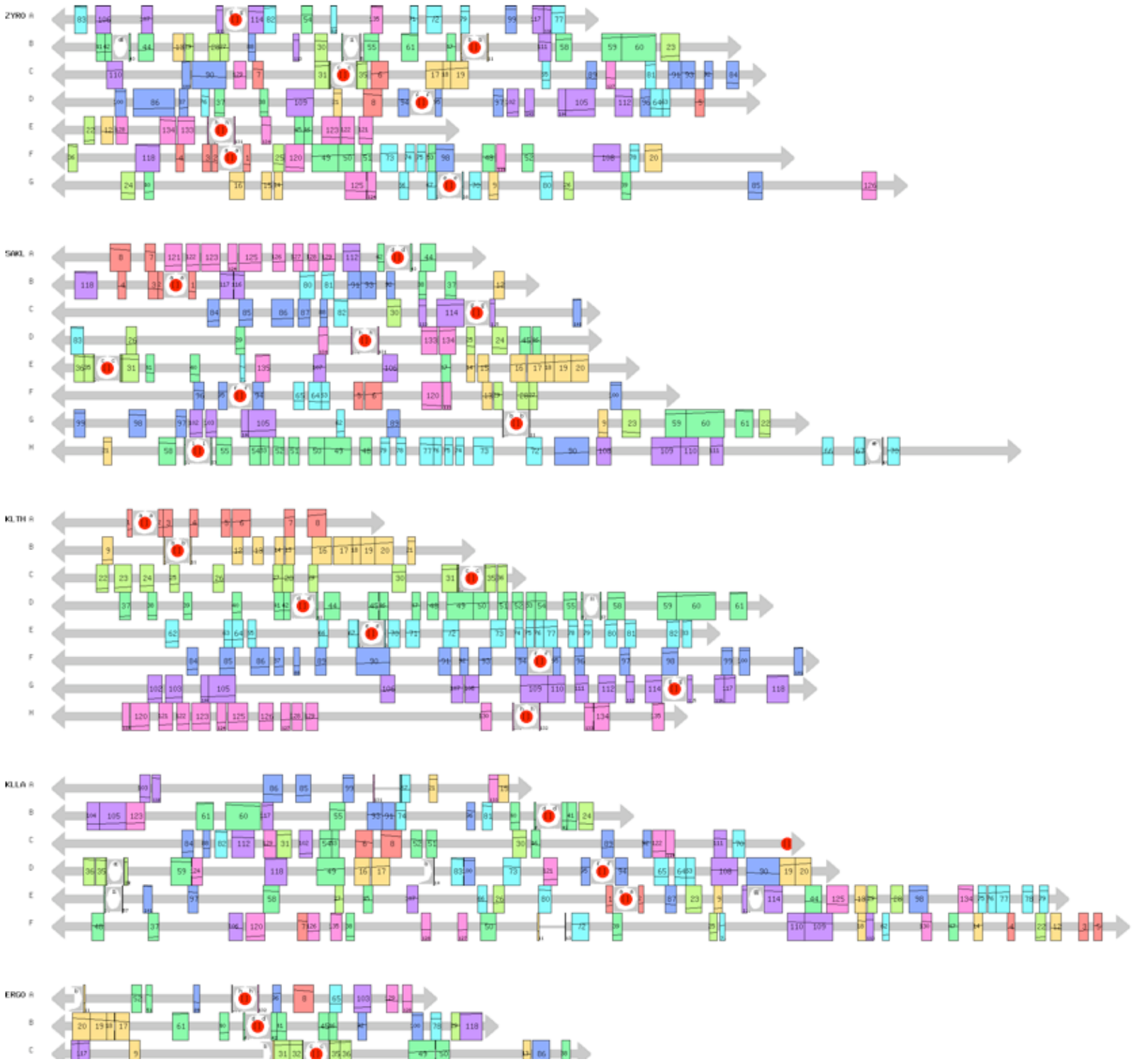
But, if we focus on the *Saccharomycetacae* that did not undergo a recent whole genome duplication

Protoploid genomes

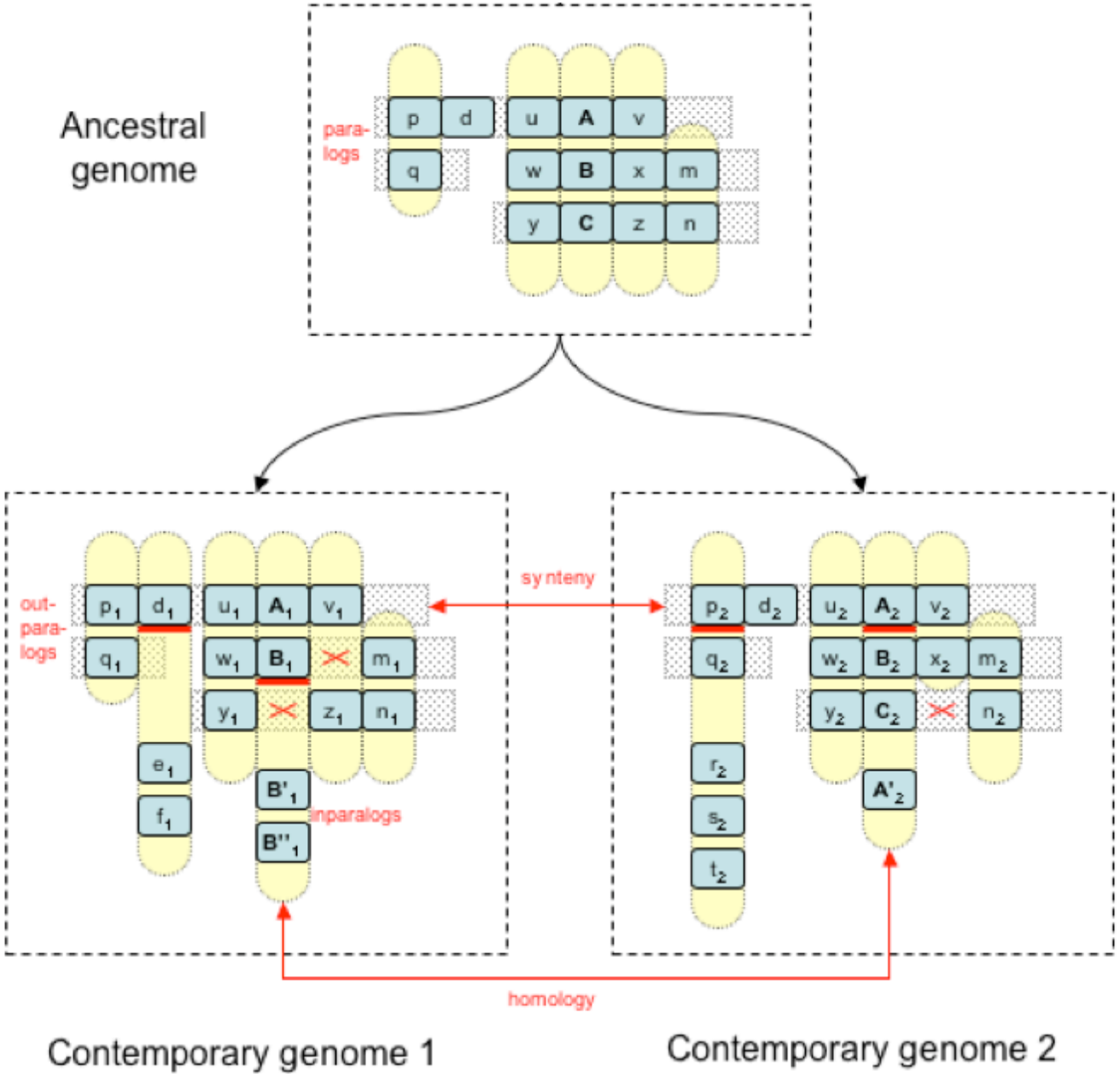
- homogeneous
- low redundancy
- less reshuffling




Klla0F
687k 688k 689k 690k 691k 692k 693k 694k 695k 696k 697k 698k 699k 700k 701k 702k 703k 704k 705k 706k
Locus
Genes
vKLLA-ORF1918
vKLLA-ORF1912
vKLLA-ORF1917
PSItblastn families
GLR.1475
GLC.1674
GLR.870
GLR.2751
GLC.2152
GLC.2234
GLR.869
GLC.2357
GLR.1474
GLC.1749
GLC.1730
GLC.1731
Klth0H
1360k
1350k
Locus
Genes
vKLTH-ORF75
vKLTH-ORF83
PSItblastn families
GLR.870
GLR.2751
GLR.869
GLC.2337
GLC.2357
Sakl0E
456k 455k 454k 453k 452k 451k 450k 449k 448k 447k 446k 445k 444k 443k 442k 441k 440k 439k 438k 437k
Locus
Genes
vSAKL-ORF7421
vSAKL-ORF7427
PSItblastn families
GLR.870
GLR.2751
GLR.869
GLR.3231
GLR.3231
GLC.2357
Zyro0A
645k 646k 647k 648k 649k 650k 651k 652k 653k 654k 655k 656k 657k 658k 659k 660k 661k 662k 663k 664k
Locus
Genes
vZYRO-ORF9012
vZYRO-ORF9005
vZYRO-ORF9011
PSItblastn families
GLR.870
GLR.2751
GLR.869
GLC.2357
GLR.1961
GLC.2337


ZYRO
SWKL
KLTH
KLLA
ERGO

Syntenic homologs are orthologs



So, in conclusion

Comparative genomics works if you pay attention to the data

- High-quality, complete genomes
- Chosen from interesting phylogenetic groups

Building tools and analyses works if you have a plan

- Genome annotation
- Protein families and subgroups
- Syntenic blocks and common markers

Many opportunities for further work

http://genolevures.org/ ≡ http://cbi.labri.fr/Genolevures/

Acknowledgments and support

Bordeaux

- Macha Nikolski CNRS
- Tiphaine Martin CNRS
- Pascal Durrens CNRS
- David Sherman INRIA
- Géraldine Jean
- Hayssam Soueidan
- Nicolás Loira
- Adrien Goëffon
- Julie Bourbeillon
- Rodrigo Assar

Génolevures

- Jean-Luc Souciet
- Bernard Dujon
- Claude Gaillardin
- Christian Marck
- Eric Westhof
- Cécile Neuvéglise
- Cécile Fairhead
- André Goffeau
- Philippe Baret
- Ed Louis
- Mark Johnston

CNRS GDR 2354 Génolevures
CNRS UMR 5800 LaBRI
INRIA team MAGNOME