



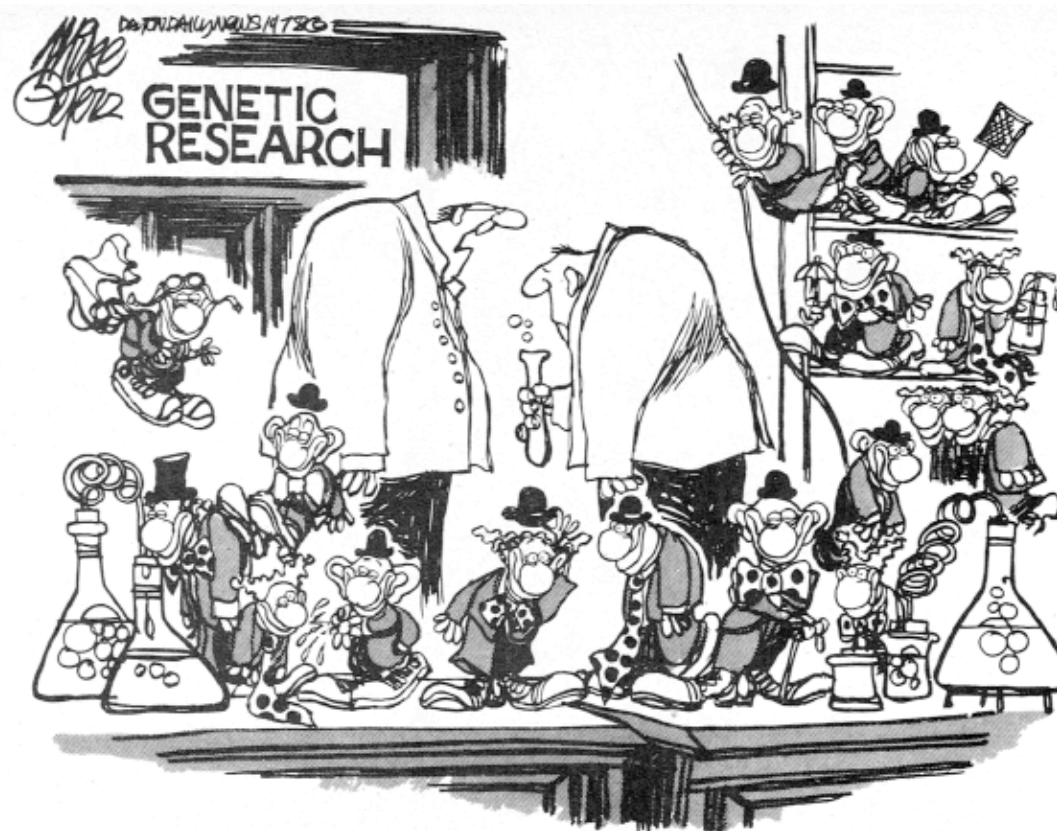
Genomic Exploration of the Hemiascomycetous Yeasts  
3rd Workshop on Algorithms in Bioinformatics  
Moscow 2008-10-08

David J. Sherman

U. Bordeaux, France

LaBRI CNRS & INRIA team "MAGNOME"

# Comparative genomics



CLONES, YOU IDIOT ... I SAID CLONES.

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
- hormones and vaccines

Medical interest

Biodiversity

Systems

Understand mechanisms of **molecular evolution**

Genome redundancy

Ortho-/para- log divergence

Expansion and contraction of universal families

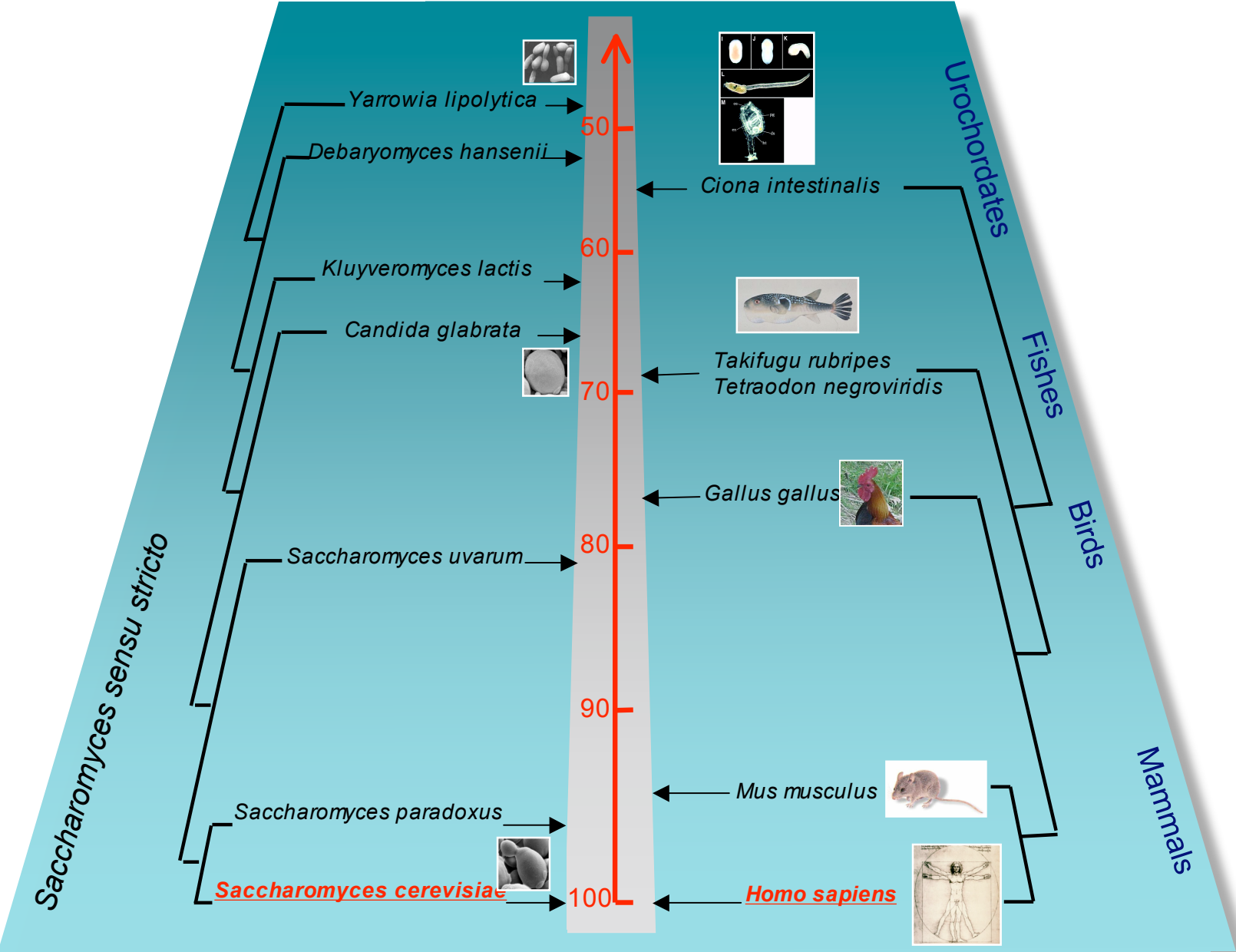
Tandem duplications

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

# Génolevures Sequencing Projects

## Génolevures 1

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

## Génolevures 2

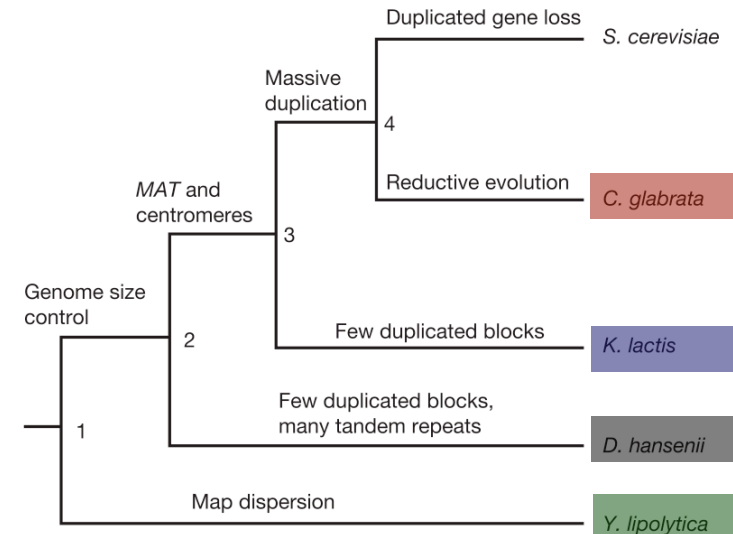
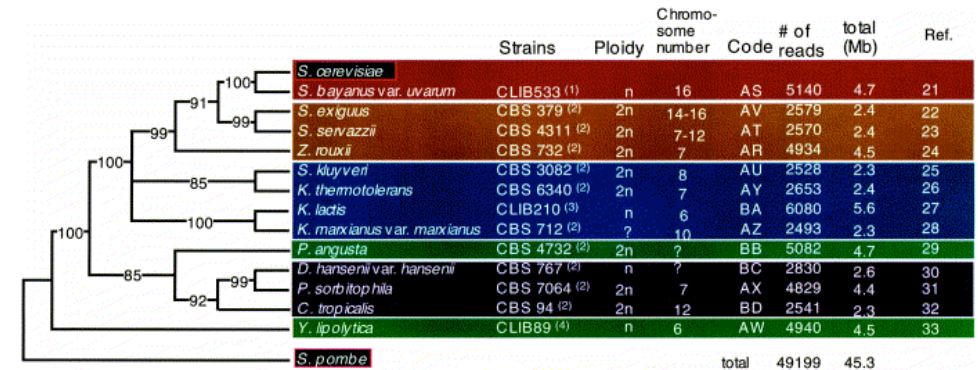
- 4 species complete 12X
- Dujon, Sherman *et al* 2004 Nature 430
- Sherman *et al* 2006 NAR 34

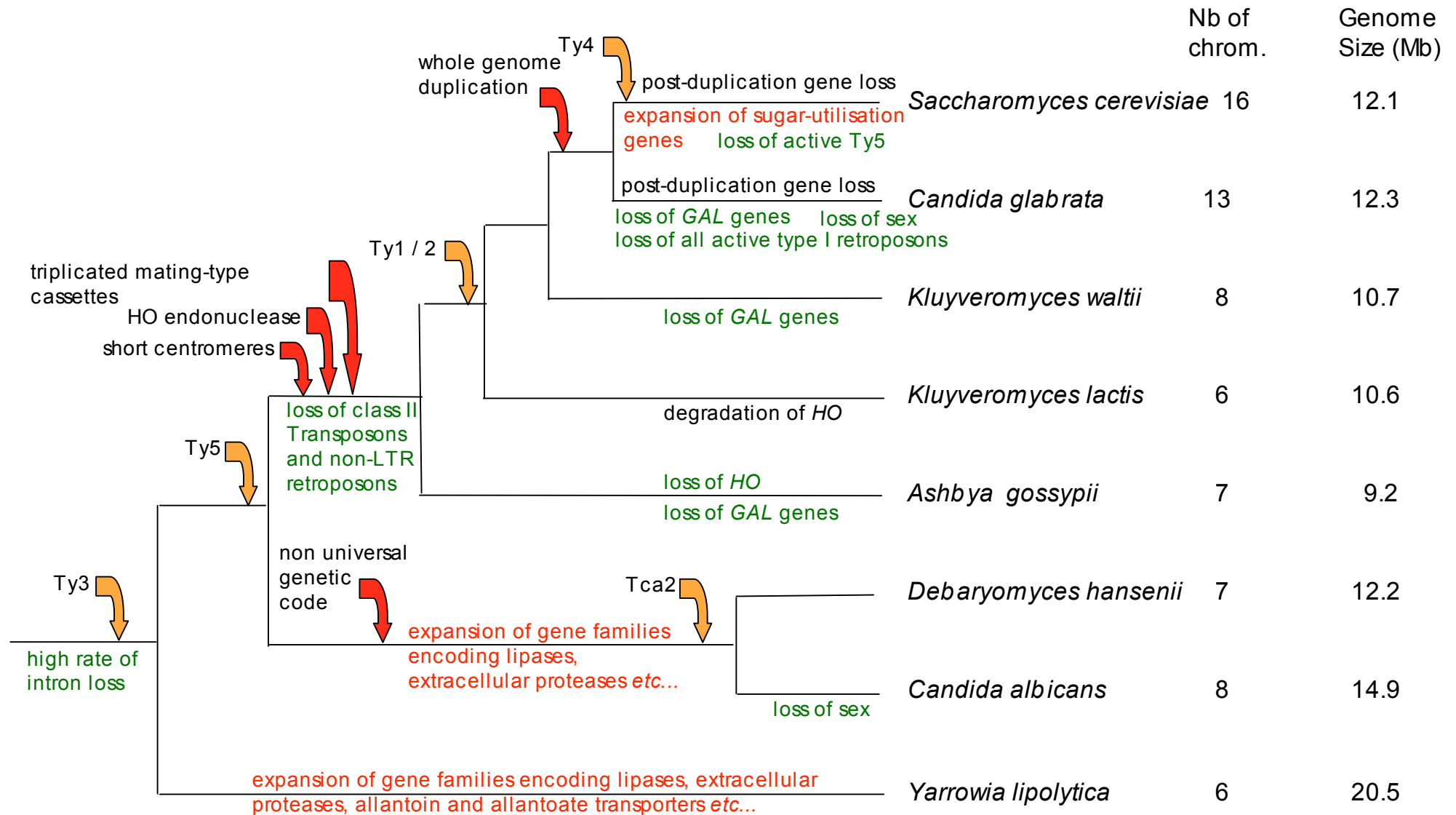
## Génolevures 3

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

## Génolevures 4

- 4 + 5 + 5 close species, NGS





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



And what  
do we do  
after that?



# 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

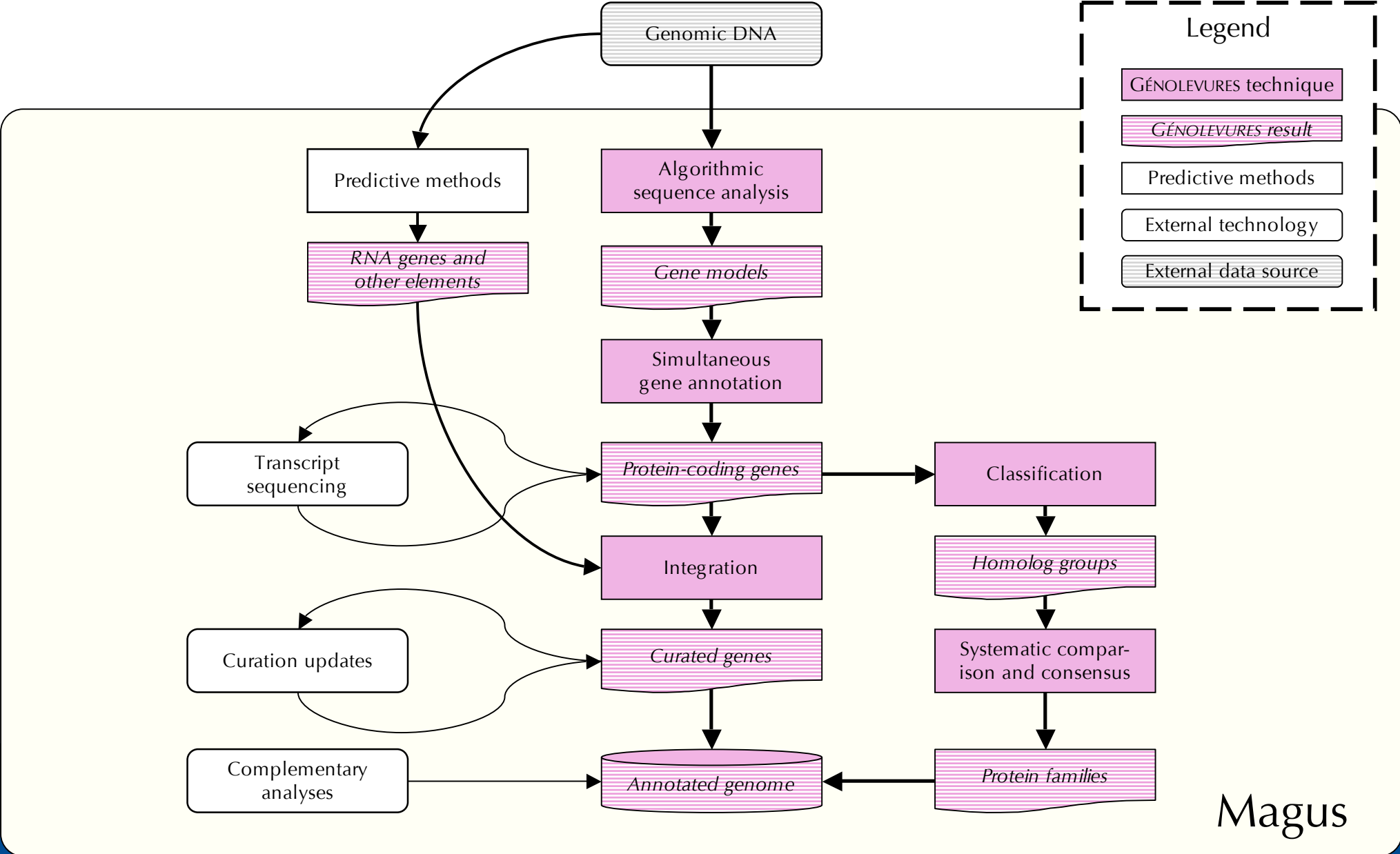
It is hard, painstaking work

It is not fully automatic

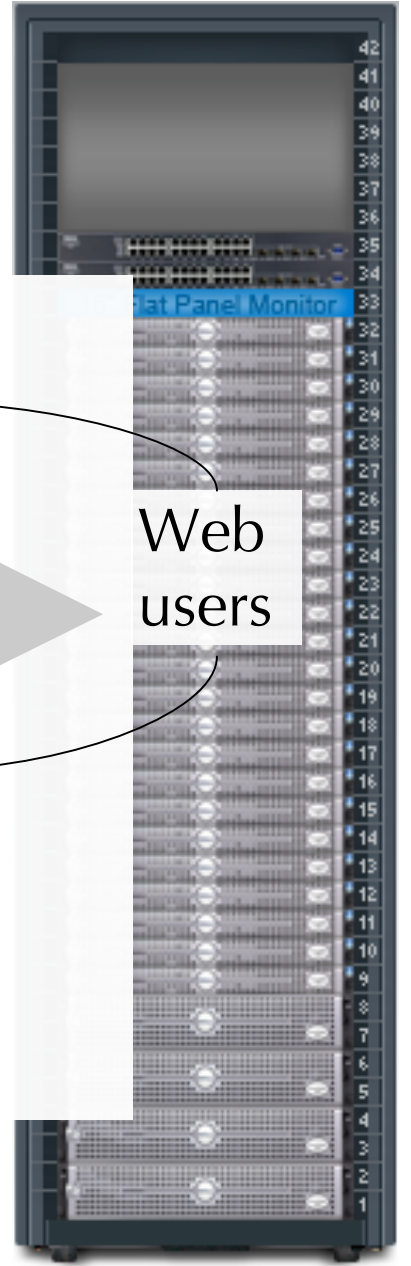
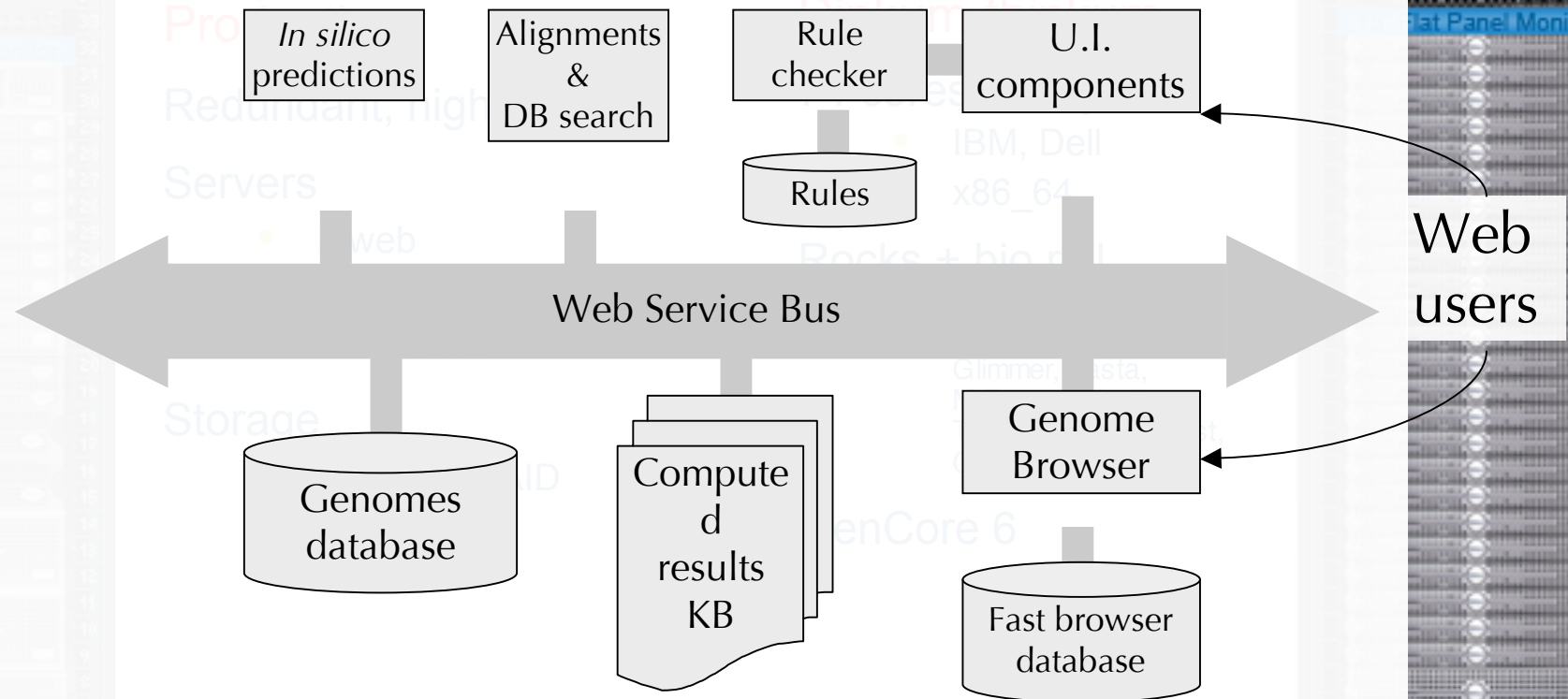
Good tools help



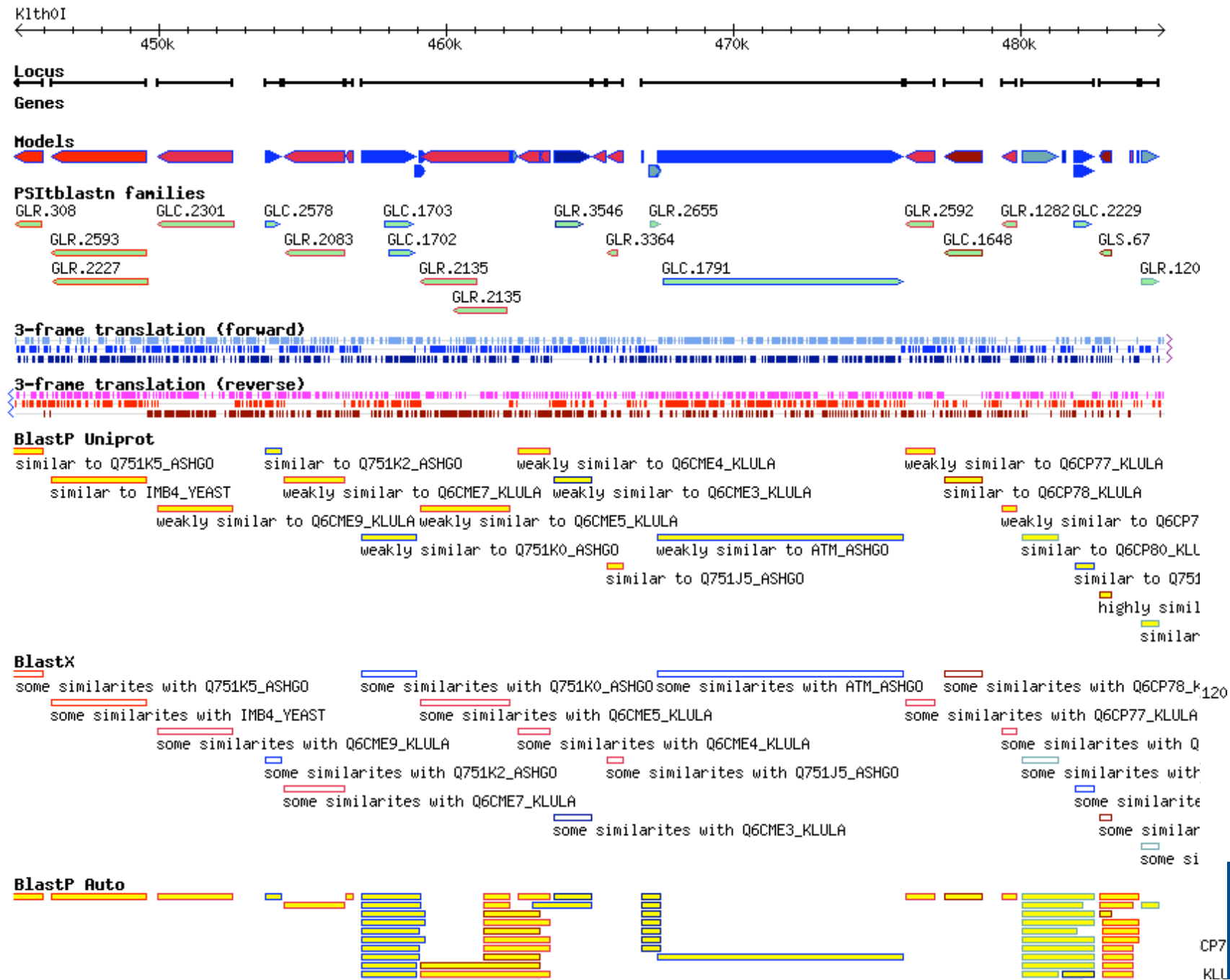
# The Annotation Process



# The "big iron"



# Browsing a genome region



# Viewing a Locus on a Genome

Genolevures locus locus.Sak10G.17298

http://cbl.labri.fr/Genolevures/magus/locus?id=locus.Sak10G.17298

locus.Sak10G.17297 ← locus.Sak10G.17299

**locus.Sak10G.17298**

Locus Sak10G from 1292755 to 1294627  
 Found 18 mRNA genes overlapping locus.Sak10G.17298.  
**1 Validated genes**

- vSak10G.mRNA.1321.p 252 aa

**0 Invalidated genes**

**17 Gene models**

• Sak10G.mRNA.1321.p	252 aa
• Sak10G.mRNA.1318.p	230 aa
• Sak10G.mRNA.1317.p	226 aa
• Sak10G.mRNA.1313.p	221 aa
• Sak10G.mRNA.1314.p	220 aa
• Sak10G.mRNA.1312.p	211 aa
• Sak10G.mRNA.1311.p	200 aa
• Sak10G.mRNA.1316.p	194 aa
• Sak10G.mRNA.1309.p	183 aa
• Sak10G.mRNA.1310.p	182 aa
• Sak10G.mRNA.1308.p	178 aa
• SAKL-ORF3148	142 aa
• Sak10G.mRNA.1320.p	83 aa
• Sak10G.mRNA.1319.p	74 aa
• Sak10G.mRNA.1315.p	66 aa
• SAKL-ORF3147	55 aa
• SAKL-ORF3149	55 aa

74%

**Actions**

Mark locus.Sak10G.17298 as **DONE**

Jump:

Sak10G  
 1291k 1292k 1293k 1294k 1295k 1296k

**Locus**

**Genes**

vSAKL-ORF3150 vSak10G.mRNA.1321.p vSAKL-ORF3146 vSAKL-ORF3145

**Models**

**codon>>>**

**codon<<<**

**GeneMark**

**PSItblastn families**

GLC.1409 GLS.8 GLR.1105 GLR.1106

**ncRNA genes**

**BlastP Uniprot**

63.11111111111111 YLR047C Saccharomyces cerevisiae 73.1481481481482 gn|GLV  
 58.8145896656535 gn|GLV|CAGL0M07942g Candida glabrata 52.8455284552846 tr|Q75C  
 60.7871720116618 sp|Q75C08 Ashbya gossypii 67.3228346456693 gn|GLV|KLLA0E1f  
 27.1532846715328 gn|GLV|YALIOB17292g Yarrowia lipolytica 55.8245083207262 t  
 69.4793536804309 gn|GLV|KLLA0E16247g Kluyveromyces lactis 31.383737517  
 22.6548672566372 gn|GLV|DEHA0B12122g Debaryomyces hansenii  
 9.38053097345133 gn|GLV|DEHA0B12122g Debaryomyces hansenii  
 68.5258964143426 gn|GLV|CAGL0M0z

# Validating a Gene Model

Genolevures gene Klth0C.mRNA.2174.m

http://cbl.labri.fr/Genolevures/magus/gene?id=Klth0C.mRNA.2174.m

KLTH-ORF14155 ← Jump: → KLTH-ORF14151

***Klth0C.mRNA.2174.m***

protein length is 252 aa  
 Klth0C from 189741 to 190838 (antisense (-) strand)  
 CDS sequence is 756 nt,  
 join(complement(189741..190406),complement(190749..190838))  
 wide nucleotide sequence 189541 to 191338  
 GC% = , GC3% =  
 Protein MW 27726.9 Da, IP 4.46, Gravy -0.218

This locus could contain a **protein-coding gene**. If this is the best predicted mRNA transcript,  
 Choose Choose this mRNA using this V\_NOTE:  
 highly similar to sp|P32905 Saccharomyces cerevisiae YGR214W RPS0A  
 Protein component of the small (40S) ribosomal subunit

**Quick links**

Results Homolog groups Best-Blastp Comments  
 SEQ NT SEQ mRNA & start SEQ AA History

**Results**

Auto blastp Auto blastp UniProtKB blastp UniProtKB blastp  
 Hemiasc blastx GeneMark img GeneMark lst Interpro scan  
 Hemiasc tblastn T-Coffee TMHMM spans

**Homolog groups**  
 Curated homolog groups  
 (no curated groups)  
 Alignments with families  
 PSSM for [GLS.8](#) 7e-109 unknown witness

**GeneMark**

**PSITblastn families**  
 GLR.1105  
 GLS.8

**ncRNA genes**

**BlastP Uniprot**

64.8148148148148	gnl GLV KLLA0E16170g	Kluyveromyces lactis
52.8455284552846	tr Q75CR0	Ashbya gossypii
58.8785046728972	gnl GLV CAGL0B02255g	Candida glabrata
61.4678899082569	YGR215W	Saccharomyces cerevisiae
46.4285714285714	gnl GLV YALI0E15312g	Yarrowia lipolytica
42.7184466019417	gnl GLV DEHA0D15840g	Debaryomyces hansenii
69.6850393700787	gnl GLV KLLA0E16214g	Kluyveromyces
75.098814229249	tr Q75CQ9	Ashbya gossypii
70.1195219123506	gnl GLV CAGL0M02849g	Candida glabrata
70.1612903225807	YGR214W	Saccharomyces cerevisiae
60.3703703703704	gnl GLV YALI0A18205g	Yarrowia lipolytica
64.367816091954	gnl GLV DEHA0B14702g	Debaryomyces

# Annotating Homolog Groups

Genolevures group vGLR.641

http://cbl.labri.fr/Genolevures/magus/group?id=vGLR.641

PSITblastn families

Klla0B  
199k 198k 197k 196k 195k 194k 193k 192k 191k 190k 189k 188k 187k 186k 185k 184k 183k 182k 181k

Locus

Genes  
 vKLLA-ORF9717 vKLLA-ORF9721 vKLLA-ORF9724 vKLLA-ORF9730 vKLLA-ORF9734  
 vKLLA-ORF9718 vKLLA-ORF9722 vKLLA-ORF9727

PSITtblastn families

Validate these annotations // Select all Clear all Reset // Copy group define Copy GO terms \*

Homolog group annotation vGLR.641 (change name here)

Define (for the family as a whole)  
 homolog group vGLR.641 derived from GLR.641

GO terms (add terms here)  
 kinase activity  
 catalytic activity

Gene 1 – YPL214C in groups vGLR.641, vC.7284\_1, vC.7284, GLR.641.  
 p|P41835 Saccharomyces cerevisiae YPL214c THI6 thiamin-phosphate pyrophosphorylase, member vGLR.641  
 kinase activity  
 catalytic activity

Gene 2 – vCAGL0E05808g in groups vGLR.641, vC.7284\_1, vC.7284, GLR.641, GLR.641.  
 highly similar to sp|P41835 Saccharomyces cerevisiae YPL214c THI6 thiamin-phosphate pyrophosphorylase start by similarity, member vGLR.641  
 kinase activity  
 catalytic activity

Gene 3 – vKLLA-ORF9724 in groups vGLR.641, vC.7284\_1, vC.7284, GLR.641.  
 similar to sp|P41835 Saccharomyces cerevisiae YPL214c THI6 bifunctional enzyme with thiamine-phosphate pyrophosphorylase and 4-methyl-5-beta-  
 kinase activity  
 catalytic activity

Gene 4 – vKLTN-ORF1764 in groups vGLR.641, vC.7284\_1, vC.7284.  
 enzyme with thiamine-phosphate pyrophosphorylase and 4-methyl-5-beta-hydroxyethylthiazole kinase activities, member vGLR.641  
 kinase activity  
 catalytic activity

Gene 5 – vSAKL-ORF15337 in groups vGLR.641, vC.7284\_1, vC.7284.  
 enzyme with thiamine-phosphate pyrophosphorylase and 4-methyl-5-beta-hydroxyethylthiazole kinase activities, member vGLR.641  
 kinase activity  
 catalytic activity

Gene 6 – vYALI0C15554g in groups vGLR.641, vC.7284, GLR.641, GLR.641.  
 phosphate pyrophosphorylase and hydroxyethylthiazole kinase start by similarity, member vGLR.641  
 kinase activity  
 catalytic activity

Validate these annotations // Select all Clear all Reset // Copy group define Copy GO terms \*

2190k 2180k

Locus

Genes  
 vYALI0C15554g

PSITtblastn families



# Protein families

Multi-species groups of related proteins

Phylogenetic relationship → functional similarity

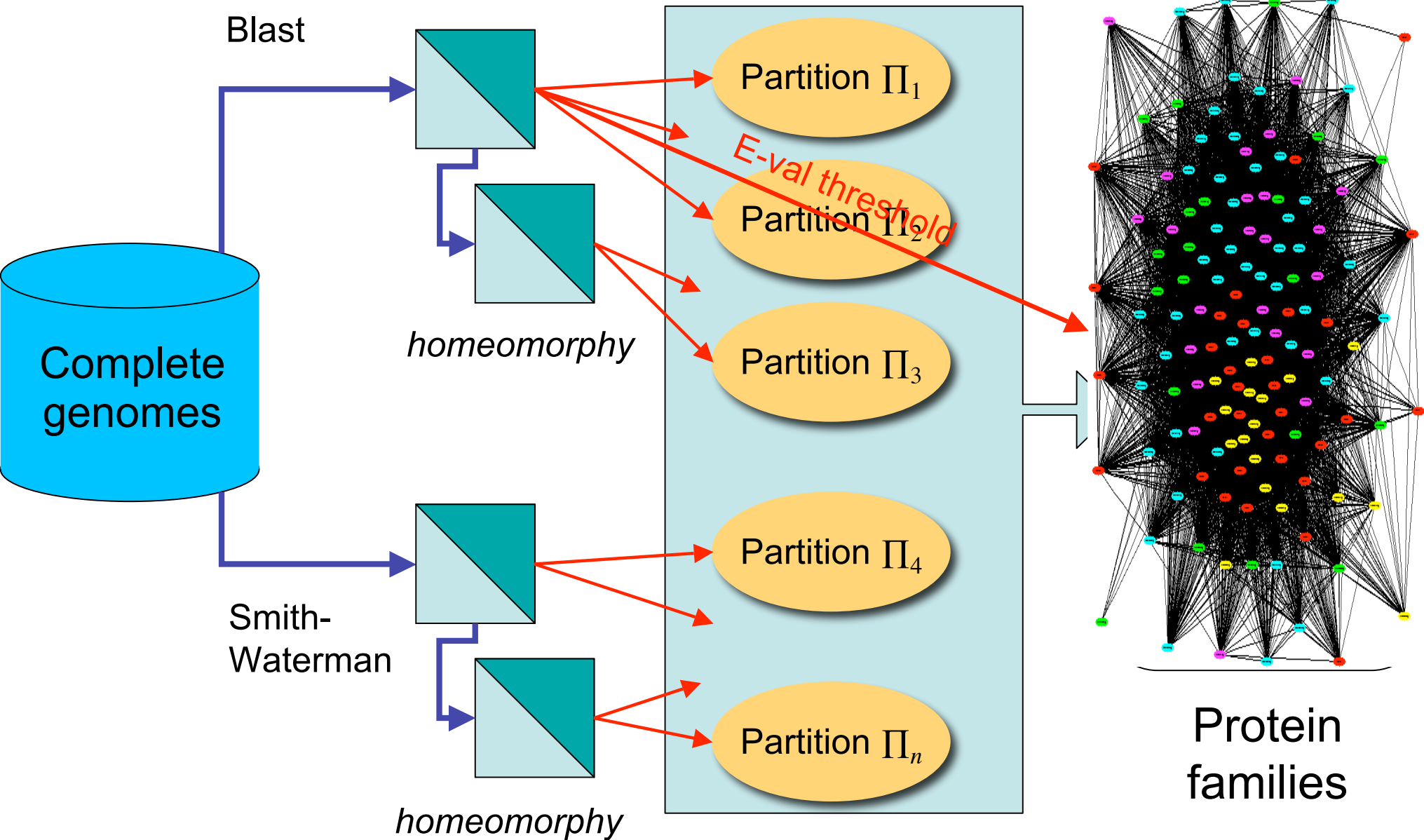
Diversity of *in silico* results

Need to calibrate or train methods for different phylogenetic groups

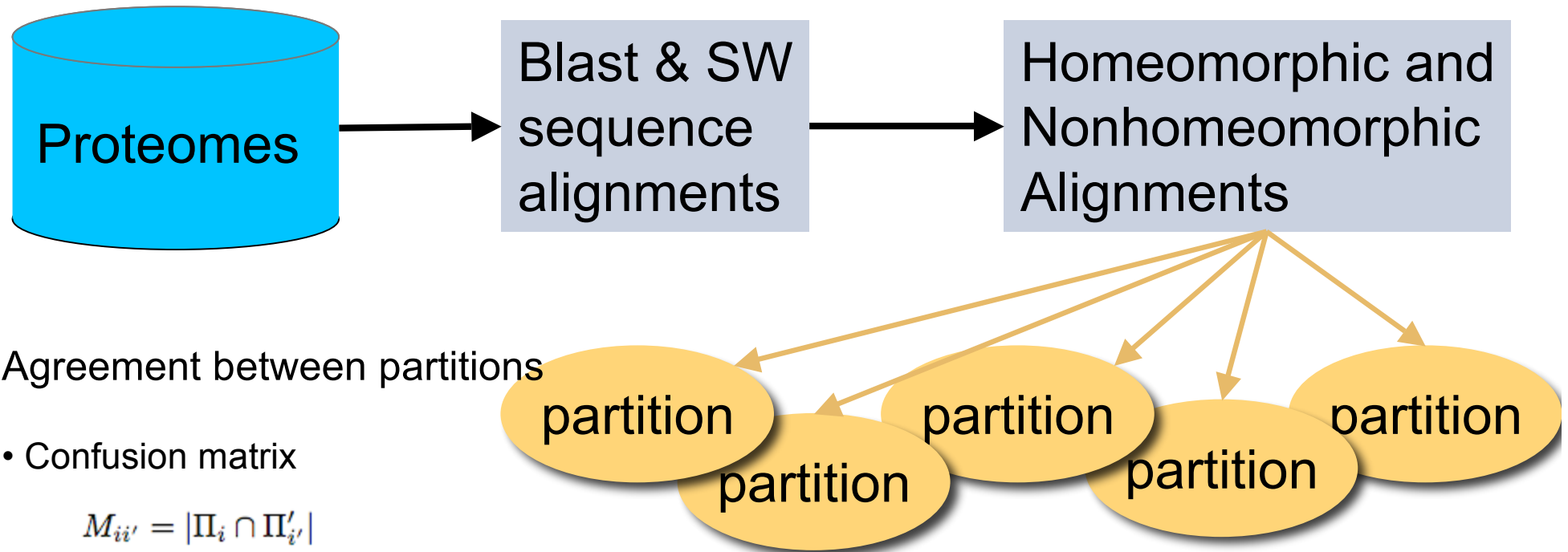
New algorithm for **consensus clustering** that is efficient in practice



# What's the goal?



# Reconciling different *in silico* predictions



Agreement between partitions

- Confusion matrix

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

- Distance between partitions

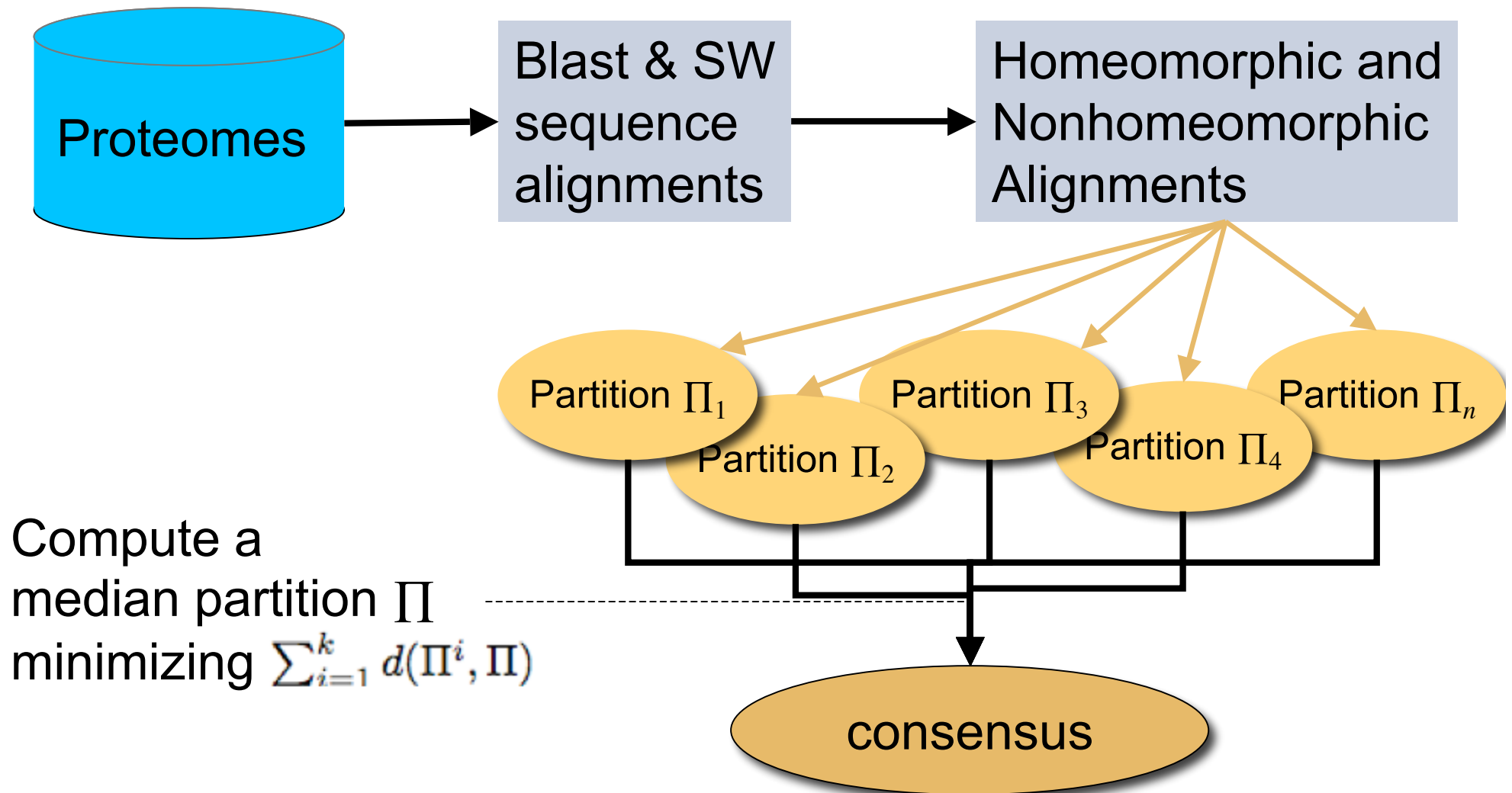
$$d(\Pi, \Pi') = 2n - \pi_{\Pi}(\Pi') - \pi_{\Pi'}(\Pi)$$

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

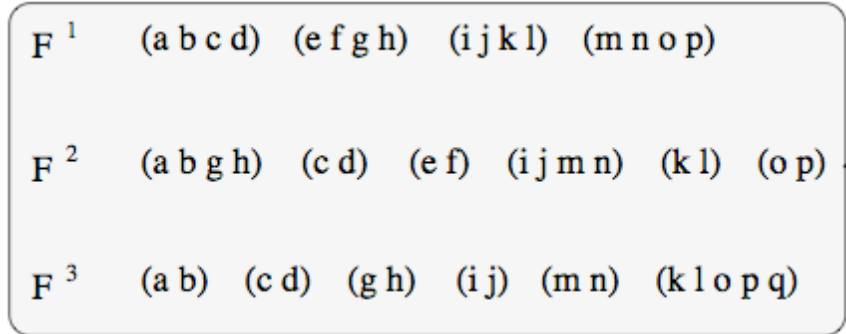
**Consensus clustering (CC):** Given  $k$  partitions,  $\Pi^1, \dots, \Pi^k$ , find a consensus partition  $\Pi$  that minimizes  $S = \sum_{i=1}^k d(\Pi^i, \Pi)$  where  $d$  is a distance.

NP-complete

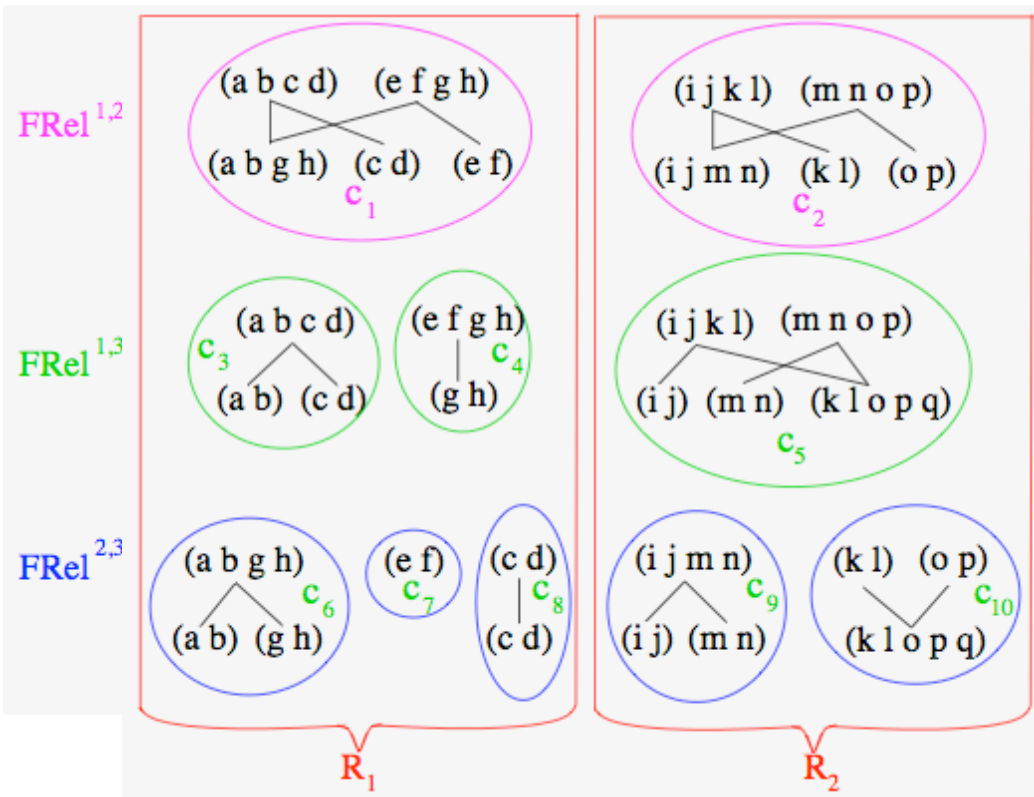
# Median partitions by consensus clustering



# Construction and algorithm



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



Define a similarity measure based on the components  $c_i$

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

$R_k$  maximal conflict regions

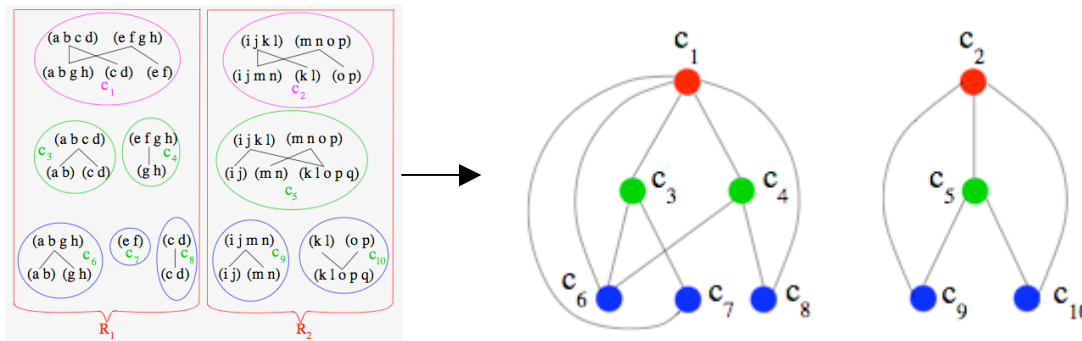
NP-complete

# Efficient heuristic

Relaxation: admit *inexact* cover

(Not all proteins are in families)

Resolve conflicts by *election* + *policy*



Conflict regions  $\longrightarrow$  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  $p$

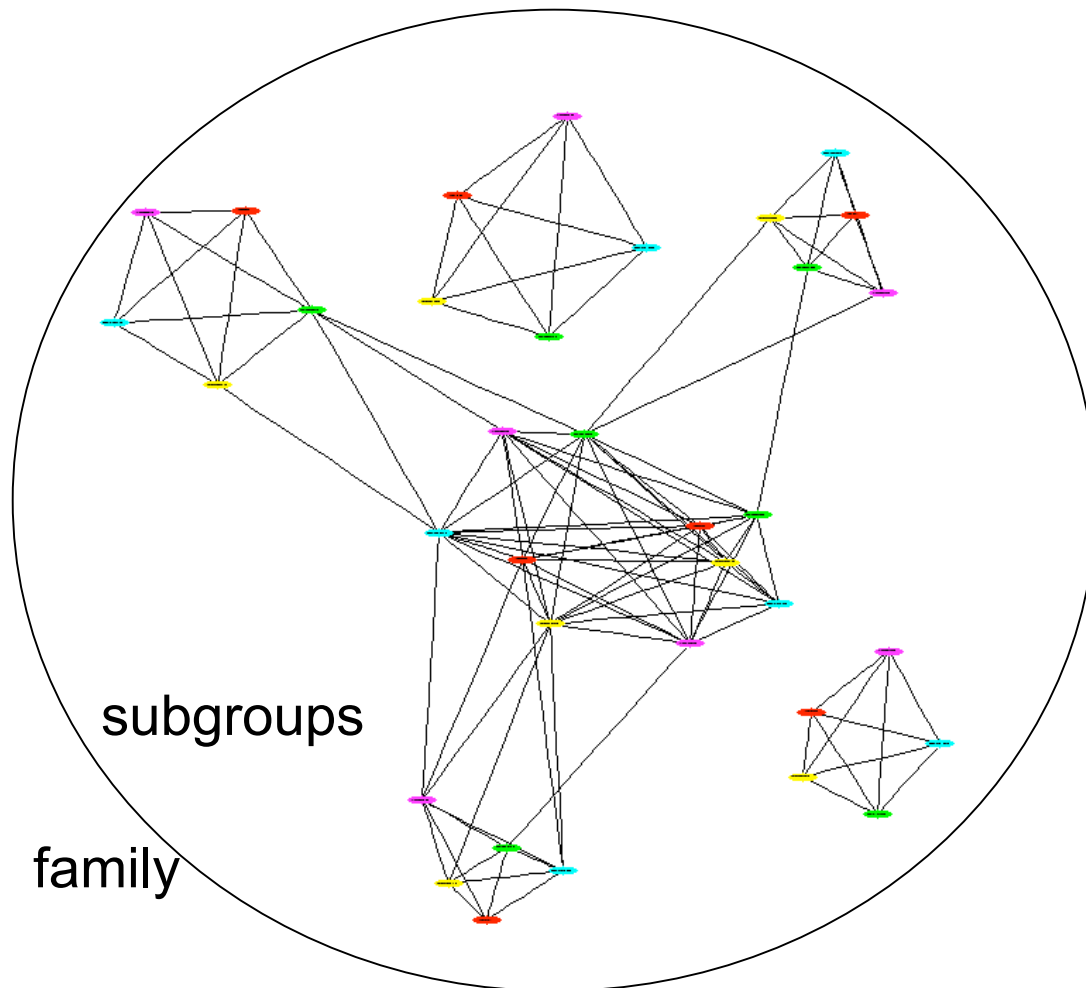
$$n^2 O(m^2) + O(|\mathbb{P}|^2)$$

## GLC.1720 consensus family

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 YALIOA00264g YALIOC10780g YALIOD03619g YALIOE24607g YALIOF02299g YALIOF20218g YALIOF29843g ([fasta](#))

[index](#) -- [GLC.1720](#) -- [cartoon](#) -- [defines](#) -- [GO terms](#) -- [comparisons](#)



### Defines

- 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 spIP00958 *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 spIP00958 *Saccharomyces cerevisiae* YGR264C MES1 methionyl-tRNA synthetase singleton, start by similarity [*Kluyveromyces lactis*] Complete CDS. KLLA0A09647g|KLLA-IPF8320|KLLA-CDS0813.1
- YLR382c spIP11325 *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

# MOLECULAR BIOLOGY AND EVOLUTION

[ABOUT THIS JOURNAL](#) [CONTACT THIS JOURNAL](#) [SUBSCRIPTIONS](#)

[CURRENT ISSUE](#) [ARCHIVE](#) [SEARCH](#)

[Oxford Journals](#) > [Life Sciences](#) > [Molecular Biology and Evolution](#) > [Volume 22, Number 4](#) > Pp. 1011-1023

**MBE Advance Access originally published online on January 12, 2005**

Molecular Biology and Evolution 2005 22(4):1011-1023; doi:10.1093/molbev/msi083

[Molecular Biology and Evolution vol. 22 no. 4 © Society for Molecular Biology and Evolution 2005; all rights reserved.](#)

## 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](mailto:gfrichar@pasteur.fr).

### This Article

- ▶ [Abstract](#) **FREE**
- ▶ [FREE Full Text \(PDF\)](#) **FREE**
- ▶ [Supplementary Material](#)
- ▶ **All Versions of this Article:**  
[22/4/1011](#) *most recent*  
[msi083v1](#)
- ▶ [Alert me when this article is cited](#)
- ▶ [Alert me if a correction is posted](#)

### Services

- ▶ [Email this article to a friend](#)
- ▶ [Similar articles in this journal](#)
- ▶ [Similar articles in ISI Web of Science](#)
- ▶ [Similar articles in PubMed](#)
- ▶ [Alert me to new issues of the journal](#)
- ▶ [Add to My Personal Archive](#)

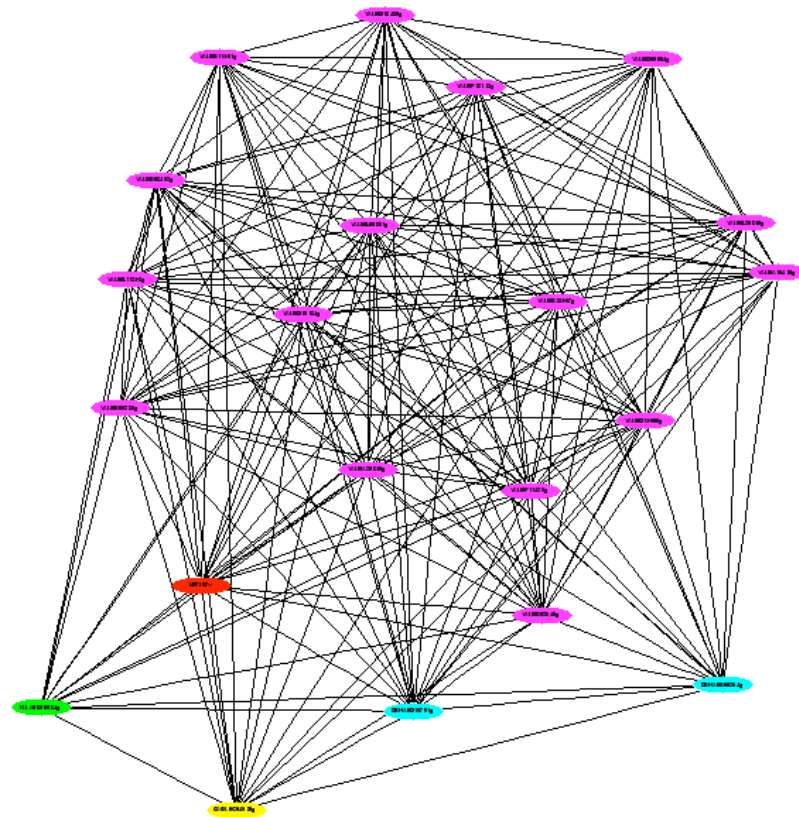


# Correlated gain and loss and in networks and metabolic pathways

## GLS.94 standard family

Family GLS.94  
 Phyletic pattern sckdy  
 Phylogen. profile 1 1 1 2 16  
 Proteins YJR107w CAGL0C04939g KLLA0D10934g DEHA0D19701g DEHA0E00264g YALIOA10439g YALIOA20350g YALIOB09361g YALIOB11858g YALIOB20350g YALIOD09064g YALIOD15906g YALIOD18480g YALIOD19184g YALIOE00286g YALIOE02640g YALIOE08492g YALIOE11561g YALIOE34507g YALIOF11429g YALIOF32131g ([fasta](#))

[index](#) -- [GLS.94](#) -- [cartoon](#) -- [defines](#) -- [GO terms](#) -- [comparisons](#)

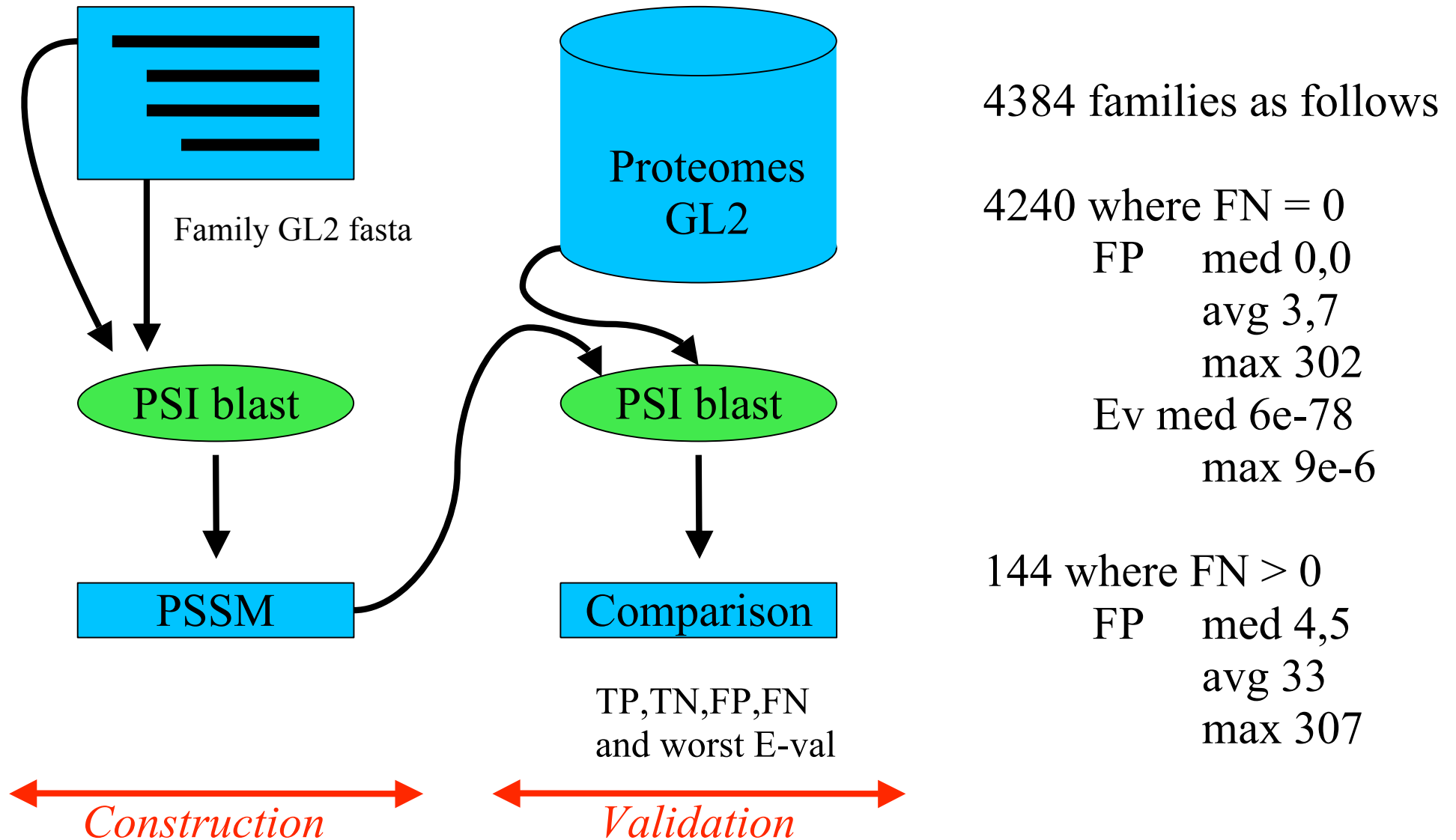


- YALIOE11561g similar to spICAD70713 *Yarrowia lipolytica* lipase LIP4, hypothetical CDS. YALIOE11561g|YALI-IPF2705|YALI-CDS3171.1
- YALIOE34507g weakly similar to trIQ9P8F7 *Yarrowia lipolytica* Triacylglycerol lipase start [*Yarrowia lipolytica*] Complete CDS. YALIOE34507g|YALI-IPF180|YALI-CDS4484.1
- YALIOF11429g weakly similar to trIQ9P8F7 *Yarrowia lipolytica* Triacylglycerol lipase spIP47145 *Saccharomyces cerevisiae* YJR107w, hypothetical start [YARWOLIP1] Complete CDS. YALIOF11429g|YALI-IPF180|YALI-CDS4484.1
- YALIOF32131g similar to trICAD70715 *Yarrowia lipolytica* lipase, start by similarity [YARWOLIP1] Complete CDS. YALIOF32131g|YALI-IPF1763|YALI-CDS3658.1
- YJR107w spIP47145 *Saccharomyces cerevisiae* YJR107W Hypothetical ORF

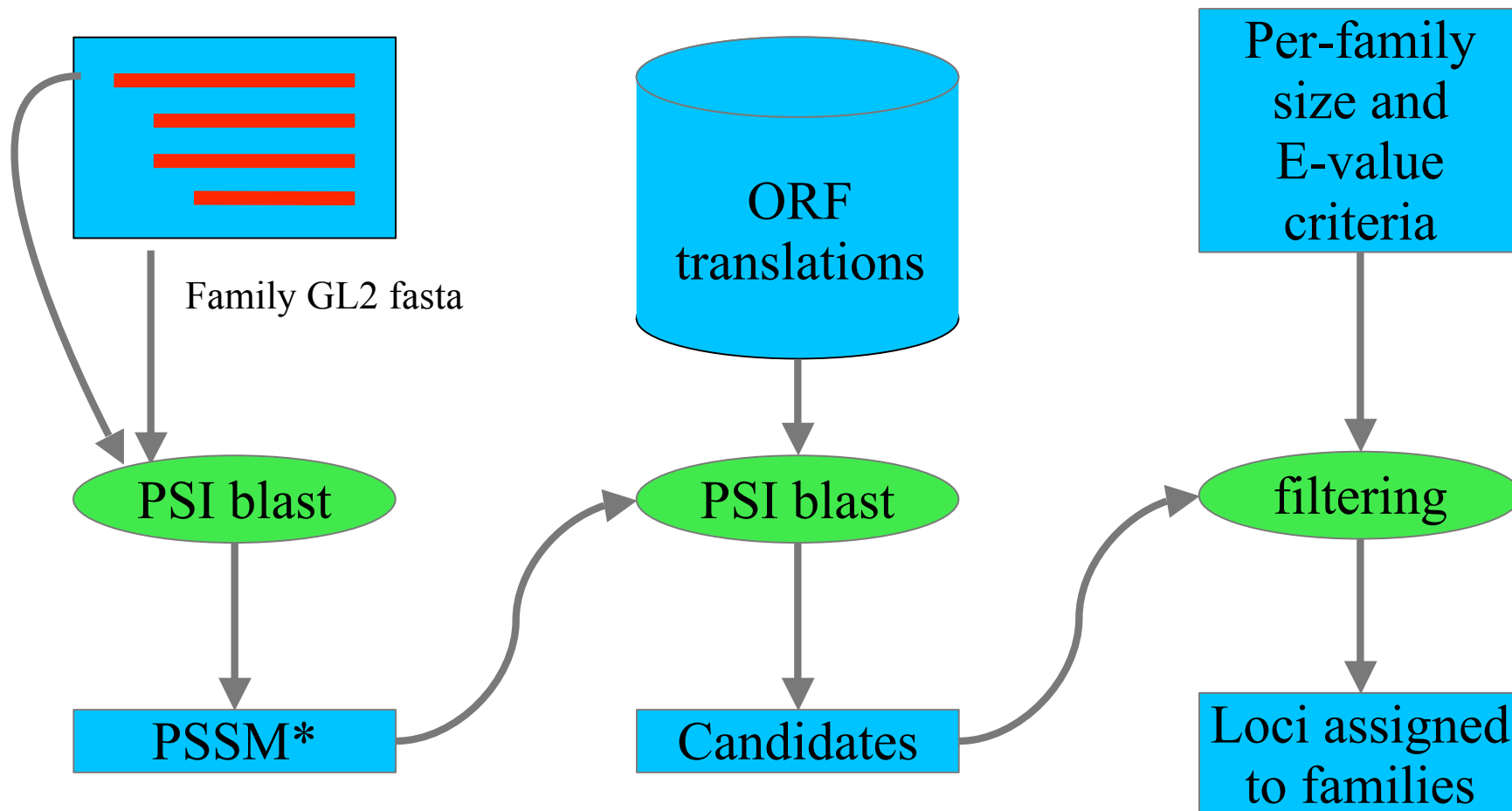
### Related GO terms (from *S. cerevisiae* genes)

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

# Construct a PSSM for each family



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



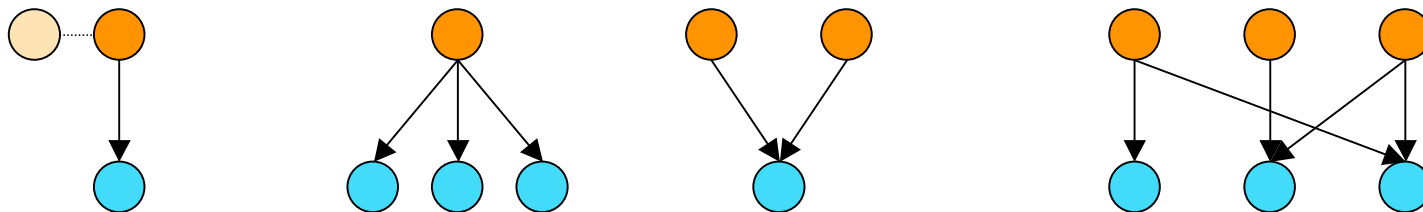
*\*PSSM: position-specific scoring matrix for PSIBLAST*

# Comparison with KOGs

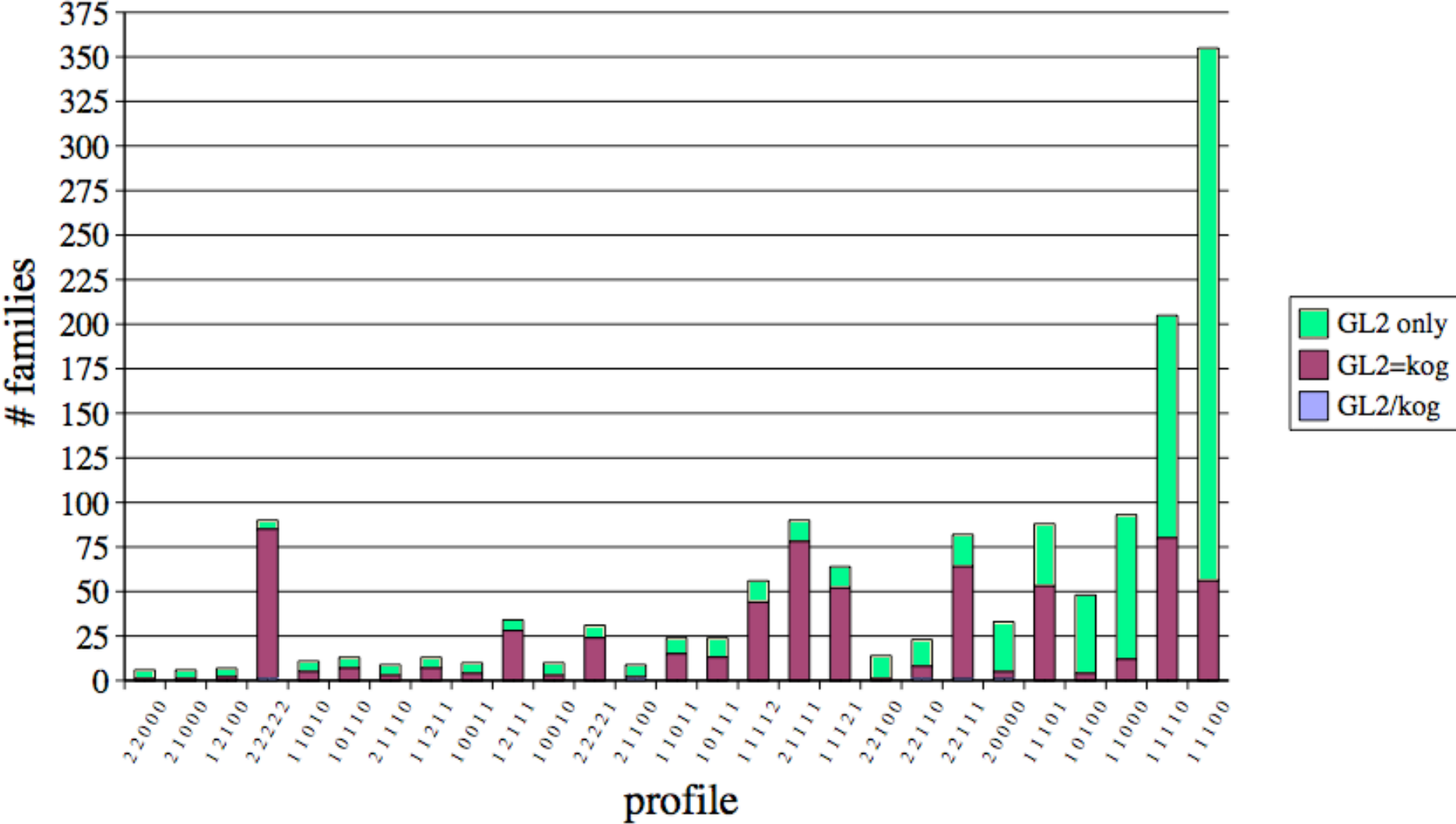
Project families on *S. cerevisiae*

Select intersection and compare  
3625 proteins (~2500 families)

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

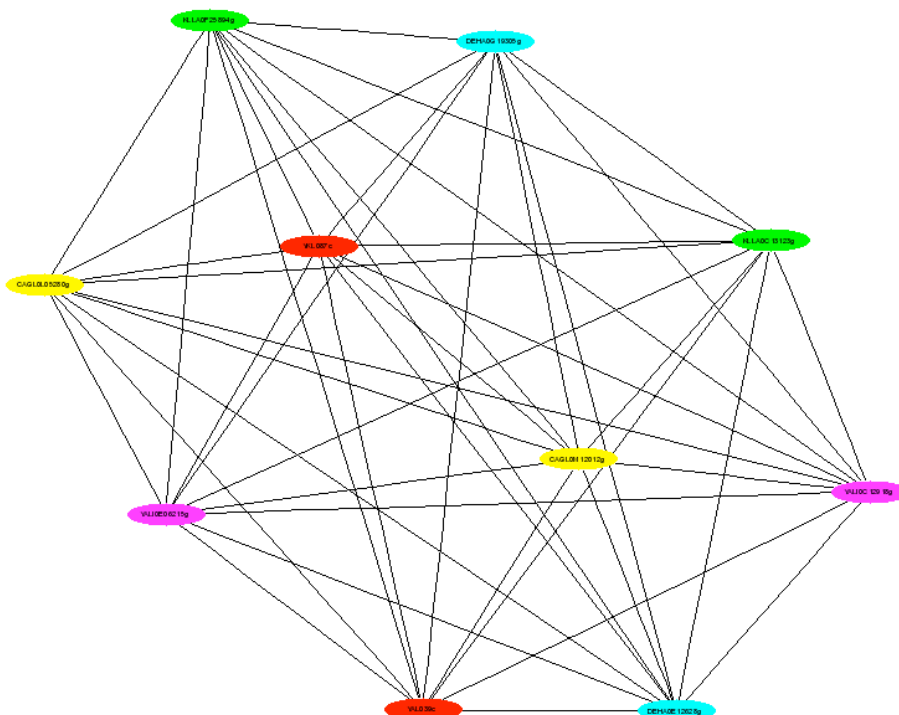


# Comparison with KOGs





# Comparison of GLR.3292 with PIRSF 017297 and 016767



```

YAL039c      MDR-----NPLNMLLAK--KPPGQRDL-P
YALI0C12918g RSWFSW-----KPSGPDAMNKLNPLMHP-ILSSQAPGQVLDL-P
DEHA0E12628g LVS-----ECPIKMDLVLNPLMHPHAISSKAPGQKLSL-S
CRGL0H12012g SKV-----DVKSPFPLMHP-MLSQTSQHGKTL-P
KLLA0C13123g KTIKPT-----GTEGGCPVLSQCGHGLMPLNPIALITVVKQSGKLDL-P
KLLA0F25894g QDVMLKHLQTVNGSAGGARARAPAPIISIIIDVDCSSSDIPEHFKYITVAVL-P
YKL087c      KKLMLK-----NEMERHPGATAPGNQLKCSANPQMDKPTPYITVDS
DEHA0G19305g RADWLSK-----VSUVVQTIKQEPALVSSITTCSSDKLDTSAHHSDSMSKL-P
CRGL0L05280g KDVWLK-----SGKKRGRANDVGL-R
YALI0E06215g EDVWVKR-----QNVASHELRASVDVWFYGNALSCPNTGL-D
  
```

Cons ■ ■ ■ ■ ■ ■ ■ ■ ■

```

YAL039c      VERTISSIPKSPD-SNMFWEYSPQQHYNARHVEKGIK-GSGEVAID-VESHVQVH
YALI0C12918g LERTHSTIPKKESSAGVWEYSPQQHYNARHLEKGG-----GEIPEDAVESHVDVH
DEHA0E12628g IERTISSIPRGLDDQGLWEYSPQQHYNARHLEKGG-----SQGIPEDAVESHVDVH
CRGL0H12012g KERTVSSIPKSG--SNMFWEYSPQQHYNARHLEKGIKDPNTEELIPEDAVESHVDVH
KLLA0C13123g IERTVSSIPKGTIRDDDFWEYSPQQHYNARHVEKGIKDFETELIPEDAVESHVDVH
KLLA0F25894g IERTKSSIPRT--GTSDNWYPSKQEFDAHKKRW-----DPLAUD-NKAVVPIH
YKL087c      QSEVVSIPRT--NSDRNWIYPSKQEFYEAHKKRW-----DPSDD-NKVVVPLH
DEHA0G19305g IERTISSIPRT--SGQSNWIYPSKQEFYEAHKKRW-----EPLAQD-NKTVVPIH
CRGL0L05280g SKRDVSSIPRT--GSEGNWYPSKQEFDAHKKRW-----DPPQPD-NKTIVPLH
YALI0E06215g QQRKLSIPRA--GSDSNWYPSQQQEFNAHKKRW-----DPPQAD-NQSIVPIH
  
```

Cons \* \*\* \*\*\* \*\*\*\* \*\*\*\*\* \*\* \* \* \*

```

YAL039c      NFLNCGQWQVLEWKEKPHIDESHV---QPKLKEHGKPGVLSPRARWHLCC-LLPS
YALI0C12918g NFLNCGAWWICDWEKQTERIDV---EPFLHGFQGRPNDSPPRAQHIQALGRVFA
DEHA0E12628g NFLNCGAWQIILEWQKYTEGTEKI---EPFLLEFTEKPHDLSPRASHYWLKGIKFPD
CRGL0H12012g NFLNCGWQWILLWKEKPYTEKSKI---EPKLLKETEKPKLSPRARNYHLKELKFPZ
KLLA0C13123g NFLNCGQWQIILDWKPKYIDKIDHT---YPKLLQPHKPPQLSPRACENIETGLFPG
KLLA0F25894g NSVNRKVVNYIKIWDGQGGDV---CGGKILTSEKGDSEKLTFRAMESSI
YKL087c      NSINRQVWV-NYSWEDKQGGCA---CGGKILTSEKGDSEKLTFRAMESSTI
DEHA0G19305g NAWNKEAWLHILNWEKSHYQQSLAQCGGKILTSEKGDSEKLTFRAMENSTI
CRGL0L05280g NYVNRKVVWYIHWENGLGDS---CGGKILTSEKGDSEKLTFRAMESSTI
YALI0E06215g NAWNKEAWWICIQWEGQHADK---CGGKLVSEQSDSEKLTFRAMENWE
  
```

Cons \*\* \* \* \* \* \* \* \* \*

```

YAL039c      HFSQMLPFEDHDWVLRGC-NKAKQPPTEKCVRYVLDYGGPD---DGN--GHPTE
YALI0C12918g IEGSAAPFEDHDWVLRVVEARTEKQTVWQCVRYVIDYSGDD---EGSSGDTPE
DEHA0E12628g IENTYPPFEDHDWVLRSLGR---DQGEQVRYVIDYSGAPD---DEE-DGHPAE
CRGL0H12012g IYHGLPFDHDWVLRSM---PEGKCVRYVLDYGGPD---DAM--GLPTE
KLLA0C13123g YFSQMLPFEDHDWVLRPDPDTSDDTEMPGYKRVYIIDYFGPD---DEE--GLPTE
KLLA0F25894g -LGHKPFEDHDWVDRCG-----KQIDYVIDEYSMPD---PEKHLFPI
YKL087c      -LHLAKPFEDHDWVDRCG-----KTVV-VIDEYSIDLD--AMSQQPLI
DEHA0G19305g -LGYKPFEDHDWVDRCG-----KVYVVIDEYGGNG-----EGGAE
CRGL0L05280g -LGHAKPFEDHDWVDRCG-----KTVVYVIDEYSKQKSKOMIATEPQI
YALI0E06215g -LGYKPFEDHDWVDRCG-----IKIDYVIDEYEGKQ---LPGHI GHPSE
  
```

Cons \*\*\*\*\* \* \* \* \* \* \* \* \*

```

YAL039c      HUDVEF-ILSLNKAEDTREFLDRIHSGFSSSSSN-----
YALI0C12918g VLDVEPALDSPGAVIDEVGKWSSETVWAGNGEPLPKYVPSMLEDIEE
DEHA0E12628g ILDVEPALDNPIDRYDFDHWKPLVWAGNGEPLD-----
CRGL0H12012g NVDVEPALDSYENCRDREIRYVTSPIHDYVFSK-----
KLLA0C13123g NLDVEPALDNGMAKDRFTKWHAPTLKRYFNKRDQ-----
YKL087c      YLDVEPKVNTFECKLELVHKLGF-----
DEHA0G19305g YLDVEPKLNSFEGLKLEFGRAFGE-----
CRGL0L05280g YLDVEPKLNSFEGLKRLIKSVGL-----
YALI0E06215g YLDVEPKVNSLEGIHRITANIFGE-----
  
```

Cons \*\* \* \* \* \* \* \* \* \*

# Comparative maps

Despite similarity in size, gene content, ecological niche,

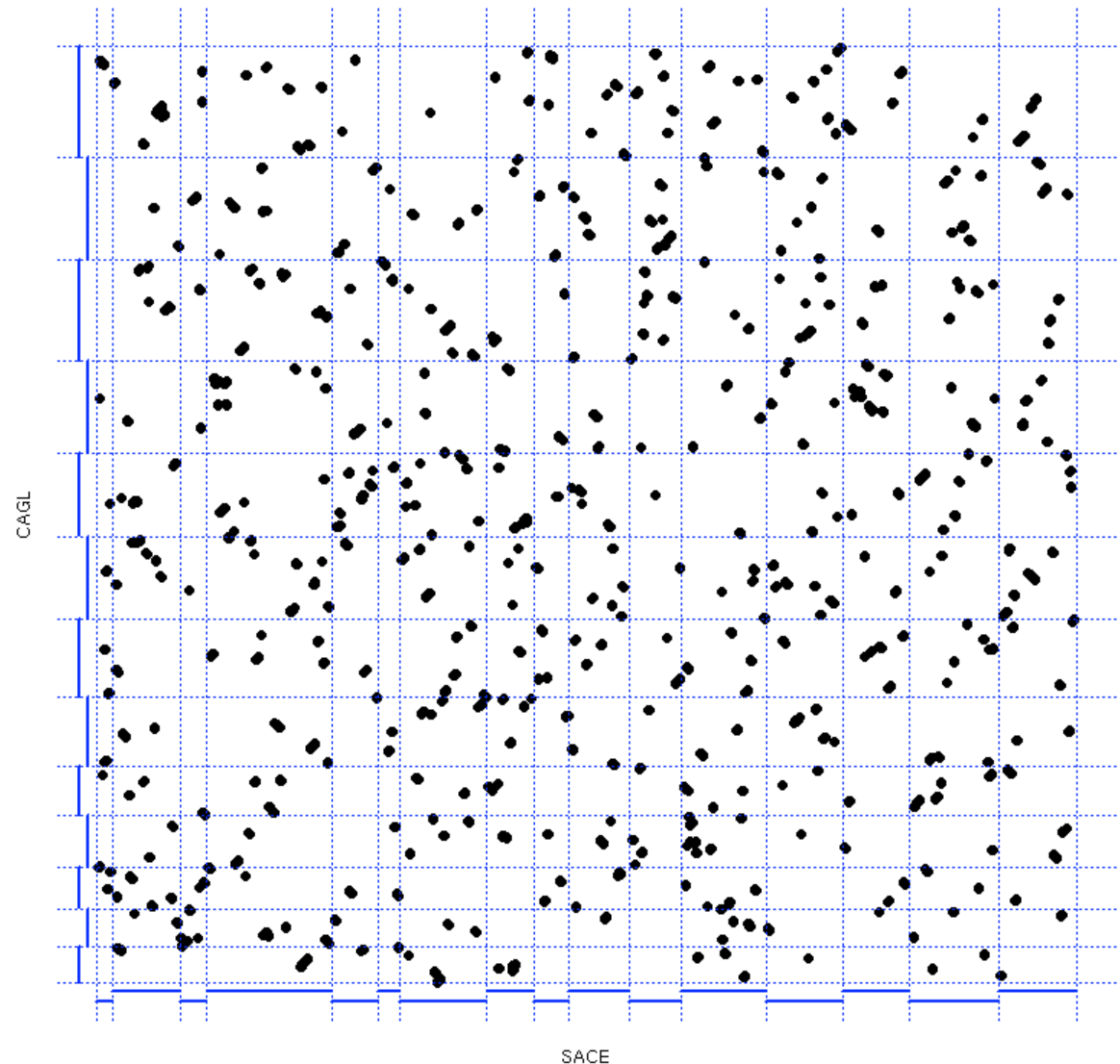
yeast genomes are highly rearranged.

In general, synteny is poorly conserved.

In part:

- evolutionary distance
- artifact of WGD

Synteny dotplot SACE CAGL



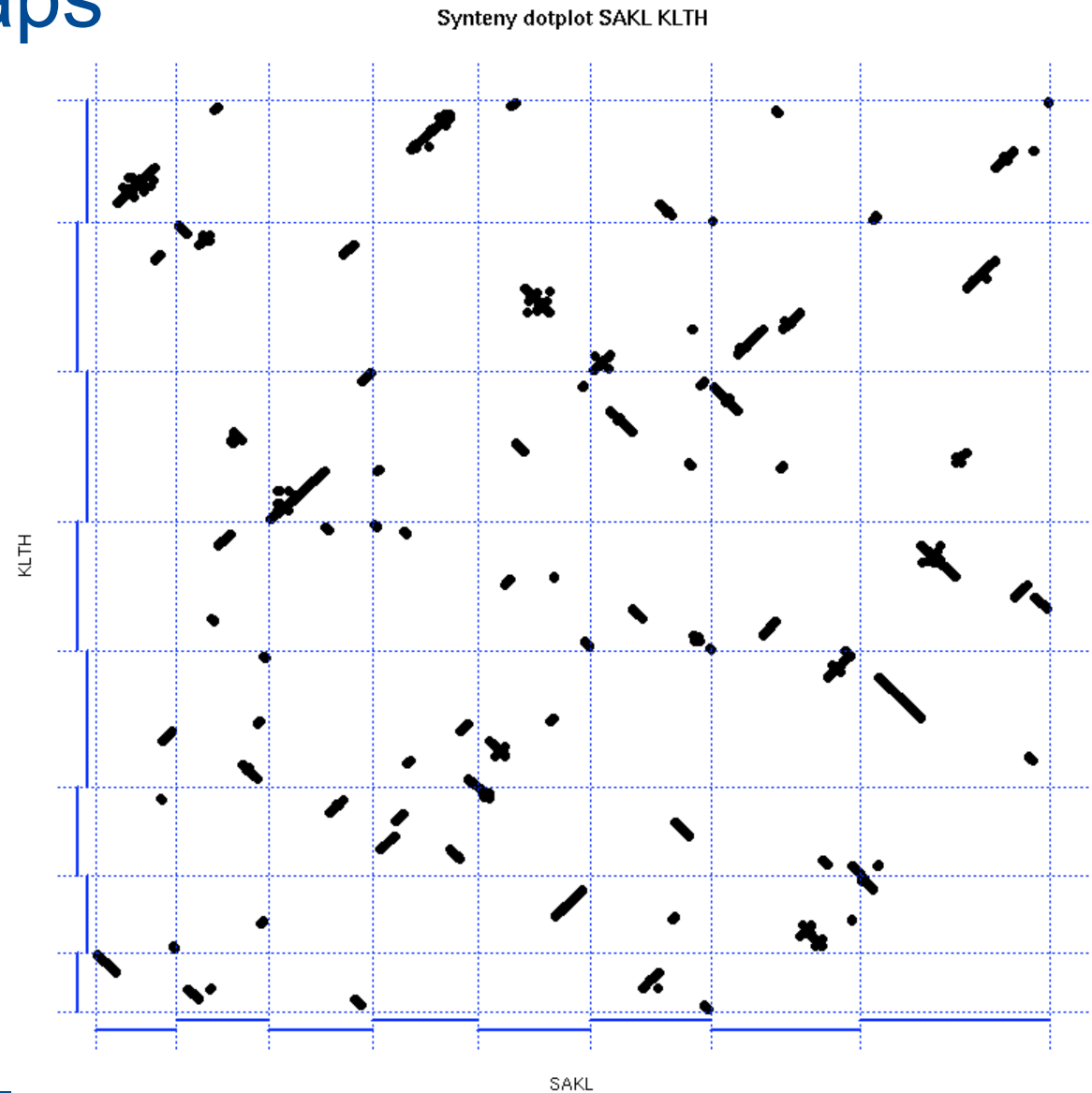


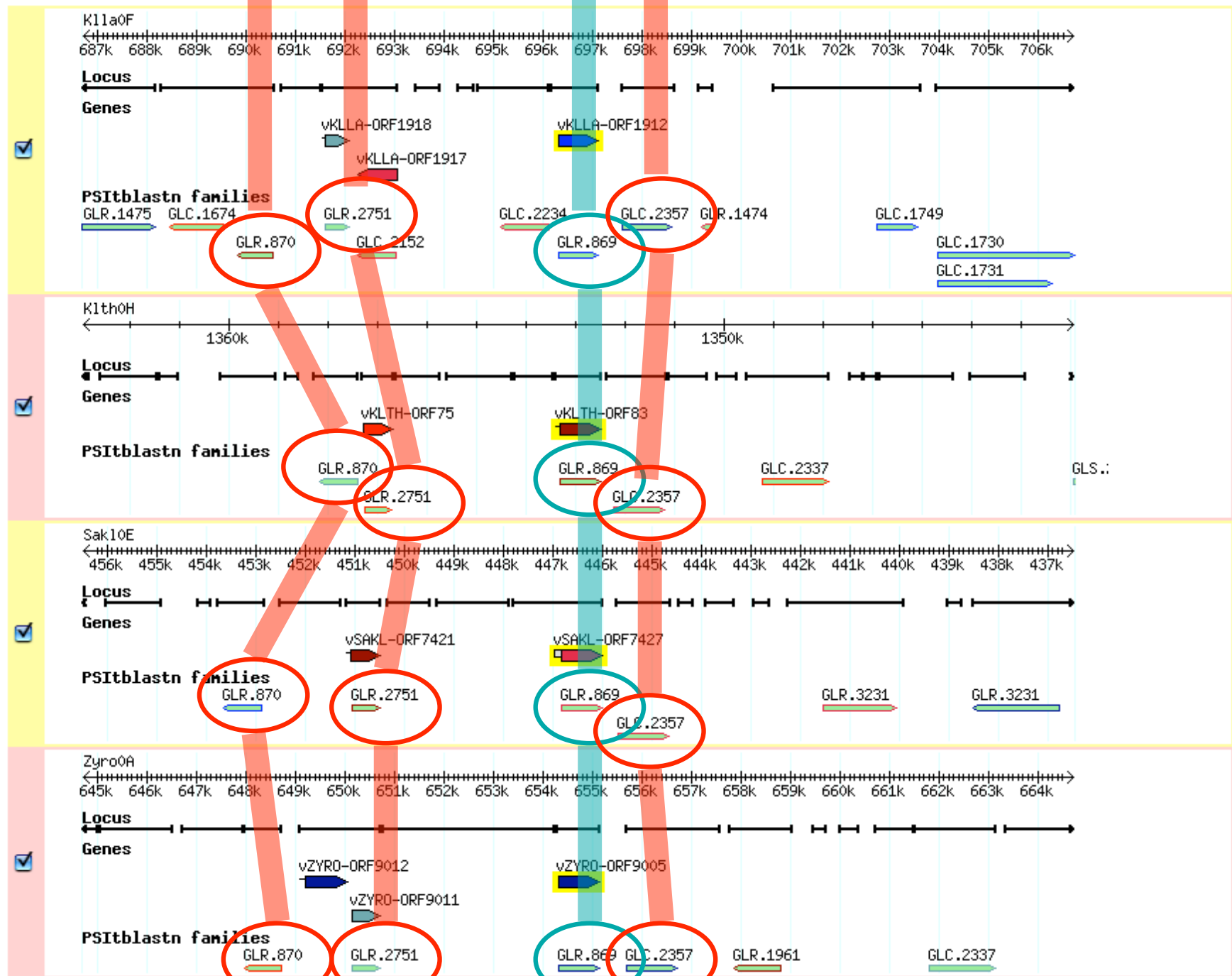
# Comparative maps

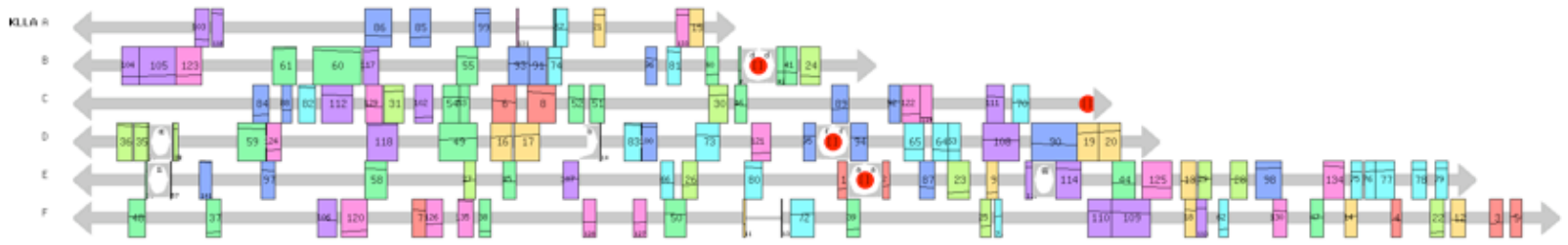
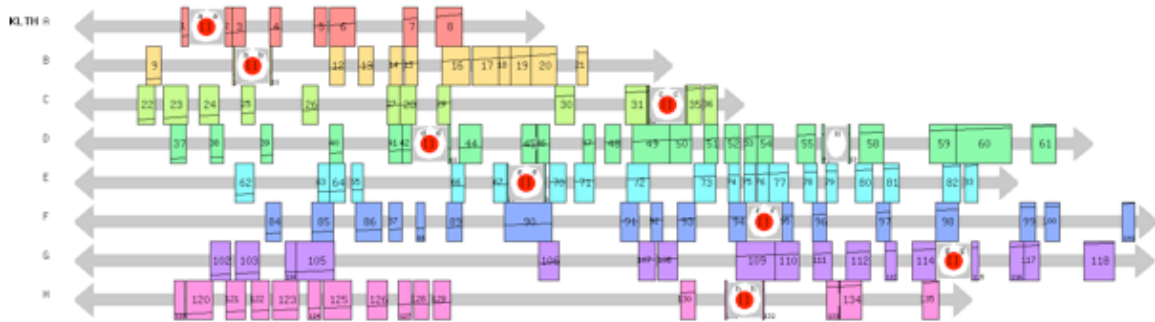
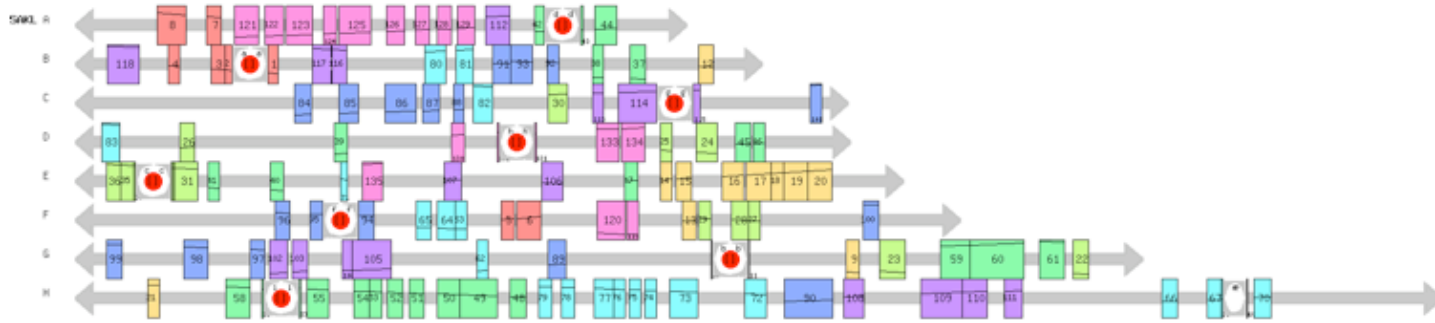
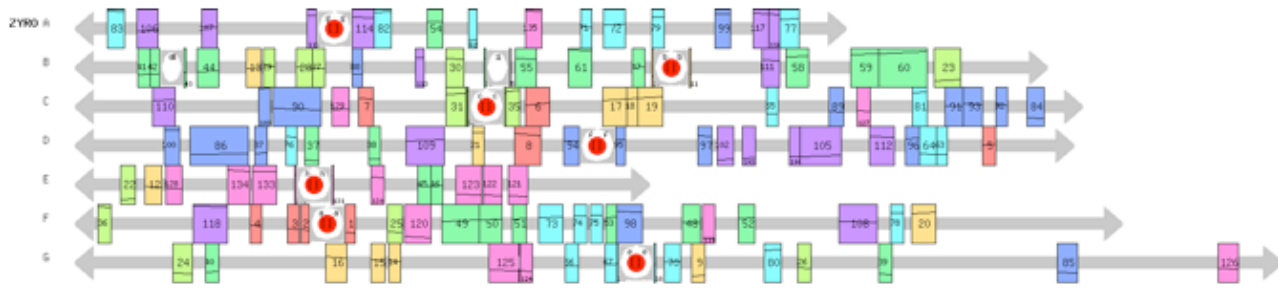
But, if we focus on the *Saccharomycetaceae* that did not undergo a recent whole genome duplication

## Protoploid genomes

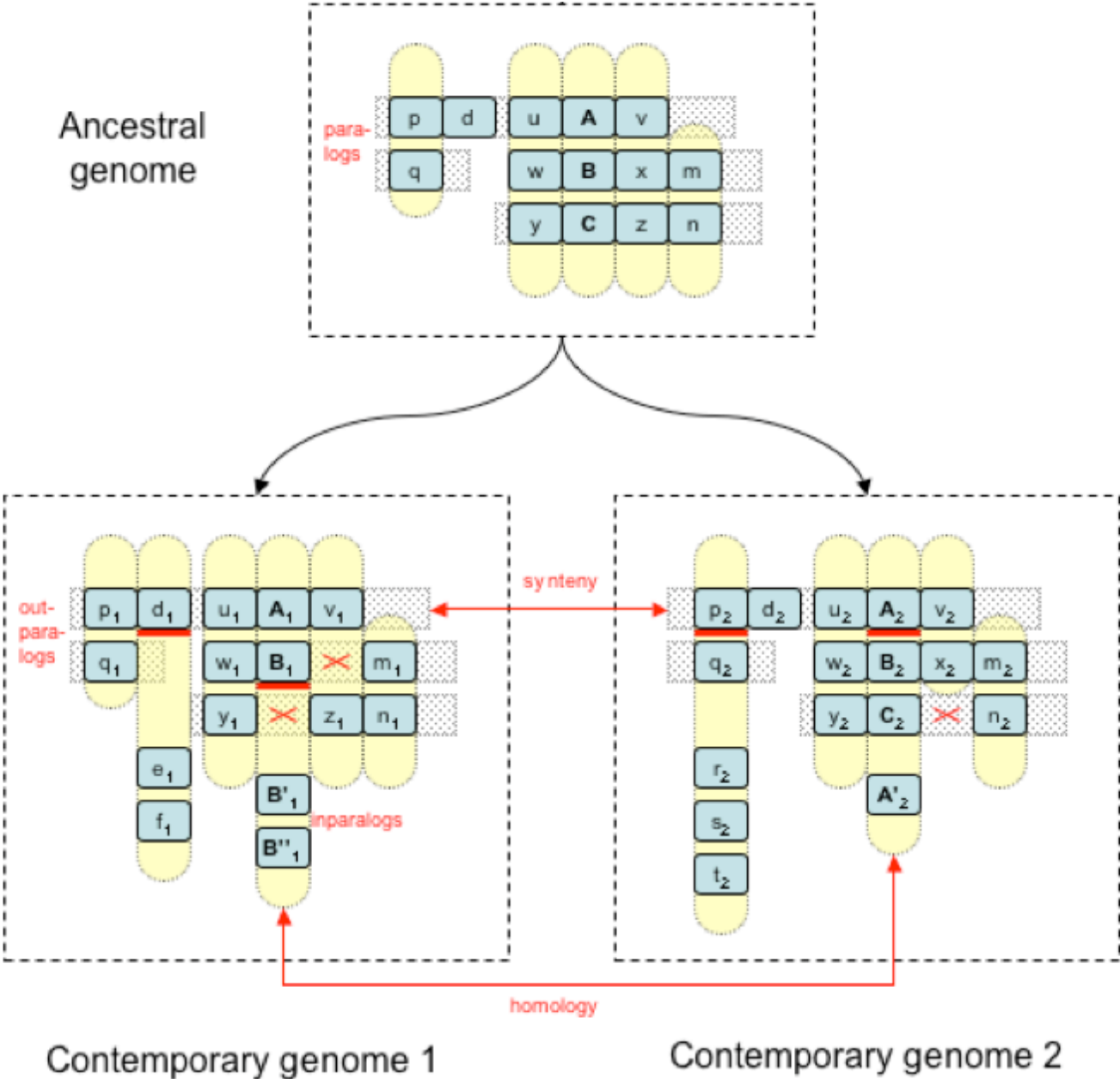
- homogeneous
- low redundancy
- less reshuffling







# 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

