Eukaryotic genome evolution, an approach with the hemiascomycetous yeasts.
Evolution des génomes de microorganismes, GMGM, Université de Strasbourg / CNRS

Nicolas Jauniaux
Véronique Leh-louis
Laurence Despons
Claudine Bleykasten
Emilie Fritsch
Anne Friedrich
Paul Jung
Serge Potier
Jacky de Montigny
Jean-Luc Souciet

and

GDR CNRS Génolevures (sequencing consortium)

Bernard Dujon, Fredj Tekaia (Institut Pasteur)
Philippe Baret (Louvain-la-neuve)
Pascal Durrens (Bordeaux)
Claude Gaillardin (Agro-Paris -Tech)
Patrick Wincker (Génoscope, Evry)
Jean-Luc Souciet (Strasbourg)

2010, July 16th
The basic property of life is:

reproduction
(sexual or asexual)

History of living species:

Identical reproduction

and

Evolution

contradiction?
Genetics

= heredity

= information is transmitted from generation to generation

Heredity laws: Mendel 1864,
1903, « re-discovery » of the heredity laws
Genetics : one example
Bioinformatics and Comparative Genome Analysis course » Institut Pasteur
Genetics
= heredity
  = information is transmitted from generation to generation

Heredity laws: Mendel 1864,
  1903, « re-discovery » of the heredity laws

Genetic analysis is based on statistical analysis,
until 1970/75: one gene = one function; one protein (purification)

Starting:
-1970/1975 = molecular studies
-1977 beginning of DNA sequencing = big transformation, the quantity of genetic data (molecular data) is now VERY HUGE BUT
- gene definition is more and more difficult
Gene: an evolving concept

Gene: An Evolving Concept

Gene Research 2007, ENCODE Project

2010, July 16th

"Bioinformatics and Comparative Genome Analysis course » Institut Pasteur"
Old view of genetics ……Babylone period
Evolution is possible and only possible if at the beginning there is duplication.

Susumu Ohno postulated that gene duplication plays a major role in evolution.

(His classic book Evolution by Gene Duplication (1970))

Ohnologs is often used for genes created by a whole genome duplication event (WGD).
Sequencing of numerous genomes, GOLD
(http://www.genomesonline.org/cgibin/GOLD/bin/gold.cgi)

We observe that the level of redundancy (in CDSs) is always important
Redundancy

How can gene arise?

- single gene duplication  X
- tandem duplications  X
- segmental duplications  X
- duplication by aneuploidy  X
- duplication by polyploidy  X (WGD)
- in frame deletions  X
- more…

Conservation or divergence in similarity between paralogous gene copies
A tandem duplication, 190 kb large, is the genetic determinant of the “white” phenotype for sheep (dominant white/tan(A<sup>Wt</sup>) agouti sheep)

Genome Research, 2008, 18, 1282-93
However

Genomic analysis indicates that the gene number (in CDSs) within eucaryotes is relatively low (ex: ca. 21,000 genes in H.s.) and mostly independant:

- of the genome size

- of phylogenetic position of the studied organisms

How can gene disappear?
General view

Gene arise

Gene disappear
Starting with this:
2010, July 16th

"Bioinformatics and Comparative Genome Analysis course » Institut Pasteur"
We expect to reconstitute this:
Musée de Delphes, Grèce, Janvier 2007
Once it is working well... the history of genome evolution will be better understood!

How can we apply this strategy to study genomes evolution?

One essential key of molecular evolution:
Chromosomal dynamics leading to genes dynamic
This chromosomal dynamics leads to reconstitute the successive steps of genome modifications:

- direct link with identification of ortholog and paralog genes

- synteny conservation
Additional questions
Question 1:
Identification of all the genetic objects

Empty?
Unidentified genetic objects? How to analyse this problem?
Degenerated genes or relics? Conservation between species?...
Question 2: Biodiversity among the eukaryotes

levures

you are here
Eubacteria

Tree of life

Eukaryotes

Archaea
Model organisms (eukaryotes)

Saccharomyces cerevisiae

Drosophila melanogaster

Caenorhabditis elegans

Mus musculus

Arabidopsis thaliana
YEASTS
The Génolevures strategy  
(different aspects) 

Biology, what to study and why: 
- phylogenetically related species  
- species or clades choice (Kurtzman)  
  - criteria ? 

Analysis: 
- complete sequencing ? telomere to telomere…  
  (link to new emerging sequencing technologies, 454, Illumina, Solid, and soon Pacific)  
- annotation and annotation by experts?  
  (proteome of high quality and most of the detectable ncARN are identified)  
- just working by similarity, no functional analysis  

Information for the community:  
- database devoted to comparative genomic analysis  
  (http://www.genolevures.org/)
Bioinformatics and Comparative Genome Analysis course » Institut Pasteur

**Basidiomycota**

- Basidiomycota; Ustilaginomycotina; Ustilaginomycetes; Exobasidiomycetes;
- Agaricomycotina; Tremellomycetes;
- Pucciniomycotina; Cystobasidiomycetes; Microbotryomycetes; Agaricostilbomycetes; Mixiomyces;

**Ascomycota**

- Ascomycota; Pezizomycotina; Sordariomycetes; Leotiomycetes; Lecanoromycetes; Eurotiomycetes; Lichinomycetes; Eurotiomycetes; Arthoniomycetes; Pezizomycetes; Orbiliomycetes;
- Saccharomycotina; Saccharomycetes; (Eurascomycetes)
- Saccharomycotina; (Eurascomycetes)
- Taphrinomycotina; Schizosaccharomycetes; (Archiascomycetes)
A least 3 major groups within *Saccharomycotina*, (previously hemiascomycetes):

- *Saccharomyces* complex or *Saccharomycetaceae*
  14 clades (Kurtzman 2003), point centromeres

- reassignment of the genetic code: the CTG codons are translated in serine rather than leucine ex. *Candida albicans*
  or
  *Debaryomyces hansenii*

- GC rich genome ex. *Yarrowia lipolytica*
Kurtzman 2003

14 clades for the Saccharomyces complex or Saccharomycetaceae (compact centromere)

2010, July 16th

Bioinformatics and Comparative Genome Analysis course » Institut Pasteur
Is WGD a frequent event?

This kind of event appears as kingdom dependent:

- one time for a part of the Hemiascomycetous lineage (Fungi)
- two times in the lineage conducting to *Homo sapiens*
- three times in the lineage conducting to *Paramecium tetraurelia*
- common in the plant lineages (also in few amphibians, reptiles, and insects)

Ohnologs is often used for genes created by a whole genome duplication event (WGD).
Not discuss during this presentation.
Present study: focus on protoplasts

Saccharomyces complex or Saccharomycetaceae after WGD
  Saccharomyces cerevisiae or SACE
  Candida glabrata or CAGL
protoploids (to distinguish them from “duplicated” yeasts)
  Zygosaccharomyces rouxii or ZYRO
  Kluyveromyces thermotolerans or KLTH
  Saccharomyces kluyveri or SAKL
  Kluyveromyces lactis or KLLA
  Ashbya gossypii or ERGO

Reassignment of the genetic code
  Debaryomyces hansenii or DEHA
GC rich genome
  Yarrowia lipolytica or YALI
But more complicated in the future!

groups

Kurtzman and Boekhout in prep. 2010/11 : 21 “groups”

…but only for the future.
Brief description of the protoploid genomes

protoploids = species not affected by WGD
Figure 1

Kurtzman’s clades:
1: Saccharomyces
4: Nakaseomyces
7: Zygosaccharomyces
10: Lachancea
11: Kluveromyces
12: Eremothecium
Ploidy refers to the number of sets of chromosomes in a biological cell:
- \( n \) = haploid number is the number of chromosomes in a gamete.
- \( 2n \) = diploid number a somatic cell has twice that many chromosomes

Link to the diploid level is the state of the studied cell:
- homozygote or heterozygote
- it is a key point for annotation and genomic analysis
Few points on genome anatomy:

- nucleotide composition of the yeast genomes

- centromere organisations
Nucleotide composition of the yeast genomes
Nucleotide composition of the yeast genomes, chromosome per chromosome

**SAKL**
- Mean: 43.1% GC
- Variance: 22.4
- Var. coef.: 0.11

**KLTH**
- Mean: 49.2% GC
- Variance: 33.8
- Var. coef.: 0.12

**ZYRO**
- Mean: 40.2% GC
- Variance: 8.1
- Var. coef.: 0.07

**KLLA**
- Mean: 40.5% GC
- Variance: 6.9
- Var. coef.: 0.07

**ERGO**
- Mean: 54.9% GC
- Variance: 50.0
- Var. coef.: 0.13

2010, July 16th
Centromere organisations
Supplementary Figure 2

A:  

<table>
<thead>
<tr>
<th>CDE I</th>
<th>CDE II</th>
<th>CDE III</th>
</tr>
</thead>
<tbody>
<tr>
<td>SACE</td>
<td>KLTH</td>
<td>ZYRO</td>
</tr>
<tr>
<td>SAKL</td>
<td>ERGO</td>
<td>KLLA</td>
</tr>
</tbody>
</table>

100 bp

B:  

```
ZYRO  
TCA  TG  TGT  TTCCGAA

SAKL  
CAC  TG  GTT  TCCGAA

KLTH  
CA   TG  TTT  TCCGAA

ERGO  
CAC  TG  TGT  TCCGAA

KLLA  
CACGTG  TTT  TCCGAAA

Consensus  
CAC  TG  TTT  TTCCGAA
```
As in *S. cerevisiae*, centromeres of *Z. rouxii, K. thermodotolerans* and *S. kluyveri* are short, suggesting that all hemiascomycetous yeasts have kinetochores with a single microtubule-binding site.

In comparison in the distant yeast species *Yarrowia lipolytica*, the 6 centromeres for each of the 6 chromosomes (A to F) are totally different in size and in sequence:

<table>
<thead>
<tr>
<th>Cen YALI A</th>
<th>60 nt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cen YALI B</td>
<td>620 nt</td>
</tr>
<tr>
<td>Cen YALI C</td>
<td>809 nt</td>
</tr>
<tr>
<td>Cen YALI D</td>
<td>1013 nt</td>
</tr>
<tr>
<td>Cen YALI E</td>
<td>200 nt</td>
</tr>
<tr>
<td>Cen YALI F</td>
<td>986 nt</td>
</tr>
</tbody>
</table>

No similarity, identification by functional analysis, no conservation within proximal clades;
Protein families and functional repertoire
Protein families and functional repertoire

The predicted proteomes were classified in protein family (Nikolski and Sherman 2007); The 48,889 proteins of the 9 species (5 protoplasts + 4 ) were partitioned in 7,927 families.
Examples of protein families

S C Z K S K E D Y
-1 1 1 1 1 1 1 1 1
-1 3 0 0 2
- etc

- Results in Table 1
Table 3: Numerical distribution of protein families and corresponding numbers of protein-coding genes

<table>
<thead>
<tr>
<th></th>
<th>Families present in all 9 yeast species</th>
<th>Families present in a subset of species (2-8)</th>
<th>Families present in only one species</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Families with ( \leq 1 ) gene per species</td>
<td>1,689</td>
<td>1,416 *</td>
<td>3,343</td>
<td>6,448</td>
</tr>
<tr>
<td></td>
<td>15,201</td>
<td>7,640</td>
<td>3,343</td>
<td>26,184</td>
</tr>
<tr>
<td>Families with ( &gt; 1 ) genes per species</td>
<td>902</td>
<td>362 **</td>
<td>215</td>
<td>1,479</td>
</tr>
<tr>
<td></td>
<td>17,518</td>
<td>3,378</td>
<td>794</td>
<td>21,960</td>
</tr>
<tr>
<td>Total</td>
<td>2,591</td>
<td>1,778</td>
<td>3,558</td>
<td>7,927</td>
</tr>
<tr>
<td></td>
<td>32,719</td>
<td>11,018</td>
<td>4,137</td>
<td>47,874</td>
</tr>
</tbody>
</table>

* including 498 families (1,562 genes) absent from \( S.\ cervisiae \), ** including 99 families (607 genes) absent from \( S.\ cervisiae \).
A 30% (2,591) are common to 9 yeasts:
   - a single gene per species (1,689)
   - several paralogous per species (902)
A 45% (3,558) are species specific
A 22% (1,778) subset of species

The 5 protoploid proteomes are quite homogenous, with only small proportions of species specific-families.

To the common set of the 9 species the 5 species share a total of 3,295 families; this core repertoire represent 81 to 88% of the protein families present in each of these 5 protoploids.
Figure 1

Kurtzman’s clades: 1: Saccharomyces, 4: Nakaseomyces, 7: Zygosaccharomyces, 10: Lachancea, 11: Kluyveromyces, 12: Eremothecium

S. cerevisiae (4072)
C. glabrata (3829)
Z. rouxii (3875)
K. thermotolerans (3945)
K. lactis (3971)
D. hansenii (4584)
S. kluyveri (4066)
A. gossypii (3764)
Y. lipolytica (4552)

Families ubiquitous to all species posterior to the node
Families ubiquitous to all five protoploid species
Species-specific families

"Bioinformatics and Comparative Genome Analysis course » Institut Pasteur"
For 32 informative GO terms recovered (SACE) the proportion of the core repertoire families was computed
Paralogous genes, genome redundancy, TGA
Paralogous genes, genome redundancy, TGA

1,479 protein families are represented by more than one protein in a given yeast species.

Distributions of protein family sizes:
- 300 to 350 families of 2 paralogs
- 95 to 115 families of 3 paralogs
- and a few larger families
Figure 2

A

B

Global genome redundancy

Paralogs per family

Nb of protein families

“Bioinformatics and Comparative Genome Analysis course” Institut Pasteur
30 to 34 % of the CDSs are members of a multigene families. For SACE the same calculation = 44%

Conclusion without WGD duplication 30 % redundancy

For YALI and DEHA the total gene number increase without a corresponding increase in the protein repertoire.

Redundancy with and without TGA
Paralogous genes:
- dispersed
- included in Tandem Gene Arrays (TGA)

Ignoring the TGA paralogs appear randomly dispersed:

- amino-acid identity *ca. 27 %* ancient duplication (*major*)

- amino-acid identity *ca. 35 to 75 %* (slow decreasing frequency)

- amino-acid identity *ca. 90 %* recent duplication (*low*)
Figure 3

![Graph showing kernel density estimates for protein identity distribution.](image_url)

- **Paral. TGA**
  - KLTH
  - SAKL
  - ZYRO
  - KLLA
  - ERGO

Protein Identity (%) on the x-axis, Kernel Density Estimates on the y-axis.
Paralogous genes:
- dispersed
- included in Tandem Gene Arrays (TGA)

- 2 or 3 gene copies (but up to 31 in one case in ERGO)
- generally less diverged in sequence indicating a more recent event or adaptative type (several appears species-specific, as TGA occurred independently in phylogenetic branches)
- but other TGAs are conserved within protoploids (using synteny conservation 18 TGAs were formed and conserved prior to species divergence)
Conservation or loss of paralogous gene copies:

A/ Conservation of dispersed paralogous gene copies:
- families of 2 or 3 members conserved among the 5 protoplasts:
  - histones H4 and H2B, …
  - some of unknown function
(analysis performed only when synteny conservation is obvious)

B/ Variation of dispersed paralogous gene copies:
- 1 copy in 4 species and 2 or more copies in species 5
- 2 or more copies in 1 species absent in the 4 others
Orthologous genes and synteny conservation

5 protoploid species: identification of orthologous genes based on conservation of their genetic location.

The SONS (Subset of Orthologs by Neighborhood and Similarity) method (Baret et al. in preparation).

Objectif 1:
- families ou pas par la synténie

Objectif 2:
- orthologs
Supplementary Figure 7

1) GL family content

2) Neighborhood identification

3) Clustering by transitivity

4) SONS

GL3C0025: YOL075c, CAGL0I08019g, ZYRO0B05588g, KLTH0E15796g, SAKL0C08074, KLLA0C04477g, ...

Queries belonging to the GL3C0025 family

Neighbors
Orthologous genes within the 5 protoploids
Using all alignments of orthologous proteins the distributions of sequence identities for all pairwise comparisons were computed (5 protoplasts):
- distribution monomodal
- even species from the same clade SAKL et KLTH show an low average amino-acid identity (58,2%). Other pairwise comparisons show lower similarity but consistent with the proposed phylogeny.

Conclusion: an homogeneous phylogenetic group as such *Saccharomycetaceae* spans a very broad evolutionary range
Identification of synteny blocks and utilisation
Synteny conservation

The maximum synteny conservation was computed among the 5 protoploids:
- map reshuffling is important
- however block size 14 to 26 genes and block number 180-291 vary slightly between pairs of species
  (except KLTH and SAKL 59 genes and 84 blocks, same clade)
Sequence divergence and synteny conservation

Relationship between rate of divergence and chromosomal rearrangements during the evolution of hemiascomycetes:
Figure 5

Conclusion: sequence divergence occurs before chromosomal maps become extensively rearranged.
Utilisation of the synteny conservation
- detection of intervening gene
- detection of HGT

Rolland et al., PLoS ONE, 2009, e6515
Novo et al., PNAS, 2009,
Horizontal gene transfer (HGT)
(lateral gene transfer)
- prokaryotes to eukaryotes
- eukaryotes to eukaryotes

Identification of HGT:
- nucleotide composition
- bias in codon usage
- phylogeny incongruence

Additional criteria:
- group of phylogenetically related species (hemiascomycetous yeasts)
- synteny conservation
Prokaryotes to eukaryotes
Prokaryotes to eukaryotes
Thermoanaerobacter ethanolicus, T. tengcongencis

2010, July 16th

"Bioinformatics and Comparative Genome Analysis course » Institut Pasteur"
A tandem of 4 copies

Single (solo) copy
Protein similarities between the 5 copies
Eukaryotes to eukaryotes

Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118

Maite Novo¹, Frédéric Bigey¹, Emmanuelle Beyne², Virginie Galeote², Frédéric Gavory³, Sandrine Mallet⁴, Brigitte Cambon⁵, Jean-Luc Legras⁶, Patrick Wincker⁷, Serge Casaregola⁸, and Sylvie Dequin⁹,²
Zygosaccharomyces bailii

Fig. 1. Chromosomal distribution of the 3 unique EC1118 regions. The alignment of EC1118 contigs with S288c chromosomes led to the identification of 3 genomic regions unique to EC1118. The localization and length of these 3 regions are indicated by colored chromosomal segments. The insertion into chromosome VI of a 12-kb fragment from chromosome VIII is also shown.
Function of the EC1118 ORFs Encompassed by the Unique Regions. Within the three unique EC1118 regions, 34 ORFs predicted to code for proteins of >150 aa in length and with homologs in other species were identified (Table S5). These genes were classified according to the Munich Information Center for Protein Sequences (MIPS) functional catalog and were found to be involved mostly in key functions of the winemaking process, such as carbon and nitrogen metabolism, cellular transport, and the stress response (Fig. 2).

![Diagram of functional classification of unique genes of EC1118](image)

**Fig. 2.** Functional classification of the unique genes of EC1118. The potential functions of the 34 unique genes of EC1118 were deduced from their S. cerevisiae orthology. EC1118 genes were clustered according to the MIPS functional catalog. Each category is represented in the chart by a color and a description of function.
2010, July 16

(B) The genomic localization and orientation of orthologs of region B–specific genes are represented for the following species: Z. rouxii (ZYRO, pale yellow arrows), Lachancea kluyveri (SAKL, gold arrows), K. thermotolerans (KLTH, pink arrows), Candida guilliermondii (CAGU, turquoise arrows), and Pichia sorbitophila (PISO, blue arrows). S288c orthologs for the EC1118 genes flanking the region B are shown in light blue. Arrows represent ORFs and their orientation. Genes are identified by their name, and the chromosome or the scaffold to which they belong is shown by a letter or a number within the arrow. In EC1118, N refers to the scaffold N26, and in P. sorbitophila the numbers refer to the gene coordinates on the chromosome. Genes syntenic with those of EC1118 are shown as fully colored arrows. Gene order was analyzed with a genome browser for all species except P. sorbitophila, for which tblastn was used, because this genome is not yet annotated.
Conclusions

- comparisons of 5 protoploids species reveals common signatures: short centromeres, single rDNA locus, low level of spliceosomal introns, same genetic code

- variation of the total number of CDSs (7%), chromosome number from 6 to 8, GC variation 39% to 52%, significant amino-acid sequence divergence even between SAKL and KLTH (57% similarity) species belonging to the same clade

- ZYRO branches separately for the 4 others and separates from them before the WGD, however it is the most closely related to the putative ancestral genome of SACE
- the 5 protoploids show a high degree of synteny conservation even with numerous rearrangement breakages they leave relatively long synteny blocks covering up to 80% of the genome

- the non linear relationship between sequence divergence and synteny conservation indicates that genetic maps are more robusts than DNA sequences over evolutionary periods leading to entire families

- identification of conserved synteny blocks leads to the characterisation of “intervening” genes are the landmarks for chromosomal events as : ectopic gene relocation, retroposition, or HGT (6 copies of an bacterial IS element are present in SAKL)
- the 5 protoploids contains very few segmental duplications (except in telomeric area) but the level of paralogous genes is important (ca. 30%)

- families of paralogs issue from ancient duplications events leave diverge and dispersed copies but conservation is higher in TGA

- despite this important dynamics the level of global redundancy is elevated (ca. 1.25) and reflects and equilibrium between gene duplication and gene loss over long evolutionary periods
Thank you for your interest
If necessary for few questions
Graphical View of Protein Coding Genes (as of Jun 29, 2010)

Verified: 4679 ORFs, 73.85%
Uncharacterized: 916 ORFs, 13.89%
Dubious: 810 ORFs, 12.26%
- non linear relationship between average amino-acid identity of orthologous proteins and the portion of genes remaining in synteny

- 5 protoplasts are grouped in a sector in which sequence identity decreases rapidly while synteny conservation decreases slowly

Conclusion : sequence divergence occurs before chromosomal maps become extensively rearranged
Animals are a major group of mostly multicellular, heterotrophs, eukaryotic organisms of the kingdom Animalia or Metazoa. Origin: Cambrian explosion 542 M years ago.
2010, July 16th

"Bioinformatics and Comparative Genome Analysis course » Institut Pasteur"
Conservation of Centromere organizations of *Z. rouxii*, *K. thermotolerans* and *S. kluveri*

For each centromere, the *S. cerevisiae* homologues of the centromere-proximal CDSs are indicated.

<table>
<thead>
<tr>
<th>ZYRO</th>
<th>size</th>
<th><em>S. cerevisiae</em> homologs(left)</th>
<th>dist. (kb)</th>
<th>Centromere coord.</th>
<th>dist. (kb)</th>
<th><em>S. cerevisiae</em> homologs (right)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zyro0A</td>
<td>167</td>
<td>YBL001-X-X</td>
<td>2.1</td>
<td>369077-369243</td>
<td>0.2</td>
<td>YBR001c-YDR002w-YBR002c</td>
</tr>
<tr>
<td>Zyro0B</td>
<td>167</td>
<td>YJR002w-YJR001w-YLR001c</td>
<td>0.3</td>
<td>788730-788896</td>
<td>0.4</td>
<td>YLL001w-YLL002w-YLL001w</td>
</tr>
<tr>
<td>Zyro0C</td>
<td>167</td>
<td>YOL003c-YOL002c-YOL001w</td>
<td>0.2</td>
<td>581961-581795</td>
<td>0.2</td>
<td>YOR001w-YML001w-X</td>
</tr>
<tr>
<td>Zyro0D</td>
<td>167</td>
<td>YIL002c-YIL001w-YEL001c</td>
<td>0.2</td>
<td>807719-807885</td>
<td>0.3</td>
<td>YIR001c-YER002w-YER003c</td>
</tr>
<tr>
<td>Zyro0E</td>
<td>169</td>
<td>X-YCR003w-YCR002c</td>
<td>0.9</td>
<td>335012-334844</td>
<td>0.1</td>
<td>YNL001w-YCL001w-YCL002c</td>
</tr>
<tr>
<td>Zyro0F</td>
<td>167</td>
<td>YFL001w-YPL002c-YPL001w</td>
<td>0.1</td>
<td>372701-372867</td>
<td>0.3</td>
<td>YFR001w-YPR001w-YPR002w</td>
</tr>
<tr>
<td>Zyro0G</td>
<td>167</td>
<td>YGL002w-YAR002w-YGL001c</td>
<td>0.4</td>
<td>852551-852385</td>
<td>0.1</td>
<td>YGR001c-YAL001c-YAL002w</td>
</tr>
</tbody>
</table>
### Equivalence between centromeres

<table>
<thead>
<tr>
<th></th>
<th>Z. rouxii 7 chromosomes</th>
<th>K. thermotolerans 8 chromosomes</th>
<th>S. kluveri 8 chromosomes</th>
<th>K. lactis 6 chromosomes</th>
<th>A. gossypii 7 chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (w)</td>
<td>A (w)</td>
<td>G (c)</td>
<td>C (w)</td>
<td>(1)</td>
<td>E (w)</td>
</tr>
<tr>
<td>B (w)</td>
<td>B (w)</td>
<td>B (w)</td>
<td>G (w)</td>
<td>(2)</td>
<td>G (w)</td>
</tr>
<tr>
<td>C (c)</td>
<td>C (c)</td>
<td>C (c)</td>
<td>E (c)</td>
<td>(4)</td>
<td>C (c)</td>
</tr>
<tr>
<td>D (w)</td>
<td>F (w)</td>
<td>F (w)</td>
<td>F (w)</td>
<td>(1)</td>
<td>D (w)</td>
</tr>
<tr>
<td>E (w)</td>
<td>H (w)</td>
<td>D (c)</td>
<td>A (c)</td>
<td>(5)</td>
<td>A (w)</td>
</tr>
<tr>
<td>F (w)</td>
<td>A (w)</td>
<td>B (w)</td>
<td>E (c)</td>
<td>(6)</td>
<td>G (w)</td>
</tr>
<tr>
<td>G (c)</td>
<td>E (w)</td>
<td>H (w)</td>
<td>C (w)</td>
<td>F (c)</td>
<td>(7)</td>
</tr>
</tbody>
</table>

*(Images of the table cells are not provided in the text)*
## Summary of annotated features in the 9 yeast genomes of interest

<table>
<thead>
<tr>
<th>Species</th>
<th>strain</th>
<th>Nb of chromosomes</th>
<th>Genome size (Mb)</th>
<th>Average G+C content (%)</th>
<th>Total nb of CDS</th>
<th>Genome coding coverage (%)</th>
<th>Gene density (nb. of CDS per 10 kb)</th>
<th>Average G+C in CDS (%)</th>
<th>Average CDS size (codons)</th>
<th>Total tRNA genes (a)</th>
<th>Total snoRNA genes (b)</th>
<th>Total tRNA clusters (c)</th>
<th>Nb of tRNA clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae</td>
<td>S288c</td>
<td>16</td>
<td>12.1</td>
<td>38.3</td>
<td>5,769</td>
<td>70.0</td>
<td>4.8</td>
<td>40.3</td>
<td>488</td>
<td>274</td>
<td>6</td>
<td>77</td>
<td>1 (I)</td>
</tr>
<tr>
<td>C. glabrata</td>
<td>CBS138</td>
<td>13</td>
<td>12.3</td>
<td>38.8</td>
<td>5,204</td>
<td>64.2</td>
<td>4.2</td>
<td>41.7</td>
<td>507</td>
<td>207</td>
<td>6</td>
<td>ND</td>
<td>2 (S)</td>
</tr>
<tr>
<td>Z. rouxii</td>
<td>CBS732</td>
<td>7</td>
<td>9.8</td>
<td>39.1</td>
<td>4,992</td>
<td>76.1</td>
<td>5.1</td>
<td>40.2</td>
<td>497</td>
<td>272</td>
<td>5</td>
<td>44</td>
<td>1 (I)</td>
</tr>
<tr>
<td>K. thermotolerans</td>
<td>CBS6340</td>
<td>8</td>
<td>10.4</td>
<td>47.3</td>
<td>5,094</td>
<td>72.3</td>
<td>4.9</td>
<td>49.2</td>
<td>492</td>
<td>229</td>
<td>5</td>
<td>43</td>
<td>1 (I)</td>
</tr>
<tr>
<td>S. kluveri</td>
<td>CBS3082</td>
<td>8</td>
<td>11.3</td>
<td>41.5</td>
<td>5,320</td>
<td>69.6</td>
<td>4.7</td>
<td>43.1</td>
<td>497</td>
<td>257</td>
<td>5</td>
<td>43</td>
<td>1 (I)</td>
</tr>
<tr>
<td>K. lactis</td>
<td>CBS2359</td>
<td>6</td>
<td>10.7</td>
<td>38.8</td>
<td>5,076</td>
<td>69.2</td>
<td>4.8</td>
<td>40.5</td>
<td>485</td>
<td>163</td>
<td>5</td>
<td>43</td>
<td>1 (I)</td>
</tr>
<tr>
<td>A. gossypii</td>
<td>ATCC10895</td>
<td>7</td>
<td>8.7</td>
<td>52.0</td>
<td>4,715</td>
<td>79.6</td>
<td>5.4</td>
<td>52.0</td>
<td>491</td>
<td>190</td>
<td>5</td>
<td>79</td>
<td>1 (I)</td>
</tr>
<tr>
<td>D. hansenii</td>
<td>CBS767</td>
<td>7</td>
<td>12.2</td>
<td>36.3</td>
<td>6,395</td>
<td>74.2</td>
<td>5.2</td>
<td>38.0</td>
<td>479</td>
<td>200</td>
<td>5</td>
<td>ND</td>
<td>3 (I)</td>
</tr>
<tr>
<td>Y. lipolytica</td>
<td>CBS7504</td>
<td>6</td>
<td>20.5</td>
<td>49.0</td>
<td>6,580</td>
<td>46.0</td>
<td>3.1</td>
<td>53.8</td>
<td>489</td>
<td>510</td>
<td>6</td>
<td>ND</td>
<td>6 (S)</td>
</tr>
</tbody>
</table>