Hemiascomycetous yeasts at the forefront of comparative genomics

Bernard Dujon

With more than a dozen species fully sequenced, as many as this partially sequenced, and more in development, yeasts are now used to explore the frontlines of comparative genomics of eukaryotes. Innovative procedures have been developed to compare and annotate genomes at various evolutionary distances, to identify short cis-acting regulatory elements, to map duplications, or to align syntenic blocks. Human and plant pathogens, in addition to yeasts that show a variety of interesting physiological properties, are included in this multidimensional comparative survey, which encompasses a very broad evolutionary range. As major steps of the evolutionary history of hemiascomycetous genomes emerge, precise questions on the general mechanisms of their evolution can be addressed, using both experimental and in silico methods.

Addresses
Unité de Génétique moléculaire des levures (associated with CNRS and University Pierre and Marie Curie), Institut Pasteur, 25, rue du Docteur Roux, F-75724, Paris Cedex 15, France

Corresponding author: Dujon, Bernard (bdujon@pasteur.fr)

Introduction

Less than ten years have now passed since the first DNA sequence of a eukaryotic organism — that of the baker's yeast, Saccharomyces cerevisiae — was entirely unveiled [1]. This remarkable achievement quickly contributed to the emergence of functional genomics. But rare were those at this time who anticipated that, a few years later, the genome sequences of many other yeast species would also become available, promoting these unicellular fungi to the forefront of comparative genomics. Presently, the complete, near complete or partial genome sequences of more than two dozen yeast species have been reported, offering a collection of genomic information without equal among other eukaryotic groups (Figure 1). The significance of this novel situation, made possible by the progress in sequencing techniques, emerges from the fact that, despite their similar morphology and common life styles, yeasts form a much diversified group. Furthermore, several of them, none more so than S. cerevisiae, are favoured organisms for genetic experiments. Most yeasts sequenced to date are members of the Hemiascomycete class, the group of fungi to which budding yeasts belong and which, from genome analysis, was recently discovered to cover an evolutionary range larger than that of the entire phylum of Chordates [2]. Other yeasts belonging to the Archiascomycetes or the Basidiomycetes have also been sequenced but will not be discussed here, because the phylogenetic distances among those fungal groups are so considerable that it is difficult to compare genomes in any detail. By contrast, comparisons within the Hemiascomycetes can be performed at various phylogenetic distances, depending on the type of question examined.

The large-scale comparative exploration of hemiascomycetous genomes started five years ago. Thirteen yeast species, selected to sample various branches of the known phylogenetic tree, were sequenced at low coverage, and each was compared with S. cerevisiae [3]. The results indicated the power of rapid genome survey to identify conserved or specific genes, to examine the evolution of functional categories or to compare genetic maps in search of the mechanisms of genome evolution. But yeast comparative genomics has considerably accelerated over the past two years, with the successive publications of the complete or high-coverage sequences of a large panel of yeast species, selected on the basis of their intrinsic interest and/or for their phylogenetic position. Some species are major human pathogens; others are used in food processing. Some are able to propagate on a variety of natural substrates; others show specific niche adaptation. The novel genomic data were used to examine questions of general significance regarding eukaryotic genome evolution, but they also served to explore and develop novel methods and strategies of general applicability for comparative genomics. Using the yeast sequences, a large variety of biological questions can now be addressed by experimental and/or in silico analyses. This short review only focuses on a limited number of prominent results obtained during the past two years.

Comparative genomics on a short evolutionary range: gene discovery, speciation and identification of conserved regulatory sites

Several species of the Saccharomyces sensu stricto clade have been sequenced and compared [4,5]. Their sequence divergence is significant but they share very high map-synteny (see Glossary), interrupted only by a limited...
number of chromosomal translocations and a higher number of single gene-deletions [6]. Species definition is number of chromosomal translocations and a higher number of chromosomal translocations and a higher number of the different chromosomes. This status can be caused, for example, by the loss of one chromosome from a complete diploid set, or by the addition of a supernumerary chromosome copy to a complete chromosome set.

**Autotetraploidy:** The status of a cell or an organism having four full sets of chromosome complements derived from the duplication of an originally diploid set.

**Collinearity:** Relates to objects (for example, a gene and its corresponding protein, or two chromosomes) having corresponding parts arranged in the same linear order.

**Gene conversion:** The process by which a gene sequence (acceptor) is partially replaced by a copy of another gene sequence (donor) from the same genome. In general, the donor and acceptor sequences must share a sufficient degree of sequence similarity (for example, the two alleles in a diploid, or two paralogs).

**Paralogy:** Homology between two non-allelic genes of the same genome, derived by duplication from a common ancestor.

**Synonymous substitution:** A nucleotide substitution, in a gene sequence encoding a protein, that does not result in an amino-acid change.

**Synteny:** The common presence of genes along a given chromosome or chromosomal segment. The notion generally also implies the order of those genes. Hence, conservation of synteny indicates the conservation of the order of homologous genes between two chromosomes or between chromosomal segments of different species.

---

The broad evolutionary range covered by Hemiascomycetes: synten, genome content, pathway conservation and niche adaptation

Estimated to have separated from the fission yeast, *Schizosaccharomyces pombe*, between 350 and 1000 million years ago [16], Hemiascomycetes cover a broad evolutionary range. Judging by the general distributions of conserved amino-acid identities between orthologous proteins, *Candida glabrata* and *S. cerevisiae*, for example, are as distant from each other as are man and fishes [2**]. And much broader distances from *S. cerevisiae* exist for other clades (for example, from *Candida albicans* or *Yarrowia lipolytica*; see Figure 1). The considerable reshuffling of genetics maps is congruent with these large evolutionary distances between clades [2**,17**,18**]. When comparing species of different clades, mosaics of conserved syntenic blocks, separated by numerous breakpoints and often containing internal inversions of a few genes, are found between all chromosomes.

Despite the evolutionary distances, there exists a large set of protein families that are common among these yeasts, and most of these families are also common to other groups of fungi or are universally conserved [2**,17**,18**,19,20]. Within some families, specific expansion or contraction of gene numbers occurs in the various yeasts and can be related to their known physiological properties or used to suggest novel ones. Against this common heritage, each species harbours several specific genes the function and origin of which is often unclear but probably contributes to its originality. Specificity is also obtained by the loss of certain genes that are common to other species. This phenomenon is frequent...
The yeast species presently sequenced, and a chart of their evolutionary history. All species, except for *S. pombe* [45] and *Cryptococcus neoformans* [46] (used here as outgroups), belong to the Hemiascomycete class, the general phylogenetic topology of which is indicated [47,48]. Closely related species are defined as clades (grey triangles). The extensively studied *Saccharomyces sensu stricto* clade is shown by a black triangle. Completed or essentially completed sequences are bold and underlined; high coverage (greater than six genome equivalents) shotgun sequences are bold; others are medium- (approximately 3X) or low-coverage shotgun sequences and/or work in progress. Only publicly available sequences are indicated. References are in square brackets. URLs of specialized sites where data can be accessed:
and has happened in every yeast lineage. The same pathway can be lost independently in several evolutionary branches, as in the case of galactose utilization [21]. Reductive evolution by gene loss is particularly striking in the case of the pathogenic yeast *C. glabrata*, which has specifically lost several functional pathways that are present in related species [2**].

The evolutionary conservation of functional pathways has been studied at a large scale using yeast genome sequences: novel transporters have been identified [22]; carefully measured metabolic-fluxes can be compared between yeasts and related to their gene content, as was done for the glucose utilization pathways [23]; the conservation of genes involved in replication, recombination and repair has been systematically examined [24]; proteins involved in gene silencing have been shown to evolve rapidly [25], indicating that several independent solutions to this problem have been explored throughout evolution; the evolution of genes involved in mating-type and sexual cycle has been studied in detail in relation to the multiple, independent loss of sexuality in the various Hemiastcomycete lineages [25–27]. Finally, the structure and gene content of subtelomeric regions has been compared between *Kluyveromyces lactis* and *S. cerevisiae* [28]: these regions appear as highly dynamic structures, offering a preferred location for genes involved in rapid adaptive evolution, and they contribute to a significant degree of the global genome redundancy.

**A whole genome duplication in the ancestry of some Hemiastcomycete yeasts**

The ancient whole-genome duplication in the ancestry of *S. cerevisiae*, postulated several years ago [29] on the basis of the numerous pairs of chromosomal homologous regions, has been recently confirmed by two independent criteria. As expected from this hypothesis, the genomes of *Kluyveromyces waltii* [17**] and *Ashbya gossypii* [18**], which have not inherited this duplication, appear as a succession of segments, covering nearly their entire lengths, which show conserved synteny, simultaneously, with two distinct segments of the genome of *S. cerevisiae*. Comparison with *C. glabrata* [2**] shows an extensive coincidence of the chromosomal homologous regions of the two species, indicating that they have inherited the same ancestral duplication event, which can be more precisely located on the phylogenetic tree (Figure 1). But the precise nature of this ancient event remains uncertain. In *S. cerevisiae*, autotetraploids (see Glossary) show a highly elevated rate of chromosomal instability and fail to arrest in glucose-limited stationary phase, resulting in low rates of survival [30]. By opposition, allotetraploids (see Glossary) appear healthy and relatively stable in mitotic growth but tend to be meiotically impaired [31]. In agreement with commonly held views on evolution, the majority of the newly duplicated genes have been lost. Deletions appear to be random and essentially concern single genes, not segments, resulting in the observed mosaic nature of the chromosomal homologous regions [17**,18**]. In *C. glabrata*, deletions have been so numerous as to leave only approximately 2% of the postulated ancestral pairs of paralogs, compared with the approximately 8% that remain in *S. cerevisiae* [2**]. Relics of genes lost by massive accumulation of deleterious mutations are also visible in the genome [6,32]. Cases of functional specialization between the duplicated copies have been mentioned [17**], but, in general, a rapid divergence of expression after duplication seems to have occurred, causing important functional asymmetry between the copies [33]. Synonymous substitutions (see Glossary) between the remaining active pairs of paralogs are not uniform, suggesting a concerted evolution by gene conversion (see Glossary) [34*.35].

**Segmental duplications, tandem gene arrays, and single gene duplication**

Comparative genomics also illustrates the role of other duplication processes in the evolution of yeast genomes. Traces of a few segmental duplications were recognized in the genome of *S. cerevisiae*, taking into account the presence of gene relics [32]. Segmental duplications are also regularly observed in subtelomeric regions [28] and were recognized in the genomes of several yeast species [2**]. The spontaneous formation of large segmental duplications, in which dozens or hundreds of neighboring genes are simultaneously duplicated, was recently demonstrated experimentally using a gene dosage recovery assay in *S. cerevisiae* [36**]. These events are observed at a frequency of between approximately 10⁻⁹ and 10⁻¹⁰ per mitosis in haploid cell cultures, suggesting that, given
the size of natural yeast populations, they must occur very frequently over time. Intrachromosomal direct-tandem duplications are the most frequent events, but tend to be unstable at meiosis, hence limiting their possible evolutionary role. But interchromosomal duplications also occur frequently and might occasionally generate a supernumerary chromosome. The mechanism at the origin of the spontaneous segmental duplications has not yet been elucidated, but indirect evidences suggest accidental secondary firing of some replicons during S phase.

Short tandem gene arrays are also observed in all yeasts. Globally, they are more numerous in some species, in which a few, specific, larger arrays are also observed [2]. The absence of coincidence of tandem arrays between species, and the dynamics of expansion and contraction of some large arrays within populations [37] suggest that such gene arrays are the sites of rapid adaptive evolution.

Dispersed copies of paralogous genes are also observed in all yeast genomes and are generally in higher numbers than those identifiable to all above mechanisms; however, their origin remains uncertain. The duplication of single genes at ectopic locations seems improbable. Dispersed paralogs might be the remnants of ancient segmental duplications after deletion of all other genes. But the retrotransposon-mediated duplication of partial gene copies that has been recently demonstrated in S. cerevisiae, using a genetic selection system [38*], offers an attractive alternative hypothesis. An important consequence of this mechanism, along with the segmental duplication mechanism, is the formation of chimeric genes at junctions. Although probably non-functional in the majority of cases, chimeric proteins with two distinct functional domains are likely to emerge over time, and several interesting examples are observed in yeast genomes [39**].

Accidental horizontal gene transfers
Contrary to its important role in bacteria, horizontal gene transfer is numerically limited in yeast genomes, for which only a few cases (less than 0.2% of the total gene number) have been recorded [2,40,41]. But the contribution of these rare events might become significant for niche specialization over time. When functionally identified, yeast genes originating from horizontal gene transfer almost always correspond to enzymatic functions, and, in several cases, they are duplicated in the species in which they reside, suggesting a selective advantage. A "prokaryotic-type" gene encoding a dihydroorotate dehydrogenase in S. cerevisiae and other related yeasts, has been proposed to be at the origin of those yeast species able to grow in complete anaerobiosis, because the corresponding enzyme is active in the absence of oxygen, contrary to the case for the common "eukaryotic-type" enzyme [40,41]. However, other differences between strictly aerobic yeasts and facultative anaerobes exist, in particular in the large-scale modulation of the transcriptional network [42**]. Another example of horizontal gene transfer concerns an alkylsulfatase-encoding gene of S. cerevisiae, which is believed to have been horizontally transferred from α-Proteobacteria, and which confers to its new yeast "host" the ability to grow on sulfur-free minimal medium [41].

Conclusions
The multiple genome comparisons now possible among a large and rapidly increasing number of yeast species gradually reveal with ever increasing detail the evolutionary history of this diversified group of eukaryotes, at the same time as they unveil novel dimensions in our understanding of gene and genome evolution and offer multiple tools to explore them. The active evolutionary dynamics encountered, illustrated by the various modes of duplication, numerous chromosomal rearrangements, extensive gene loss, rewiring of transcriptional networks as briefly summarized above, is such that novel surprises are likely in the future exploration of novel, carefully selected yeast genomes. The formation of novel genes that, for lack of homologs, seem to have occurred in every yeast lineage remains puzzling. Several other exciting aspects that could not be addressed in this short review concern the non-coding RNA genes, introns, transposable elements, repeated DNA and protein segments. Yeasts are now a favoured case for fundamental studies on phylogenies [43] and, with S. cerevisiae in particular, will soon enable us to explore population genomics. At the same time, the data collected have important consequences for applications in biotechnology (with the discovery of novel enzymes or the efficient manipulation of industrial strains), in medicine and in agronomy (with the complete genetic characterization of important human and plant pathogens, and the possibility of identifying novel drug targets). If the variety of known yeast species is large, their hidden variety is probably much larger, because many more species remain to be isolated and identified from the variety of natural environments, as can be judged from recent explorations [44].

Acknowledgements
I wish to acknowledge my colleagues from the Génolevures Consortium (GDR 2354 Centre National de la Recherche Scientifique [CNRS]) in addition to the members of my own laboratory for efficient collaboration and fruitful discussions. Sequences analyzed by the Génolevures Consortium were produced by Génoscope (Jean Weissenbach and Patrick Wincker) and the Génopole Institut Pasteur-Ile de France (Christiane Bouchier). Work in my laboratory is supported by the Institut Pasteur, CNRS, Université Pierre-et-Marie-Curie and grants from Association pour la Recherche sur le Cancer and Action Coordonnée Incitative. BD is a member of Institut Universitaire de France.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

Hemiascomycetous yeasts at the forefront of comparative genomics

Dujon 619


17. Kells M, Birren BB, Lander ES: Proof and evolutionary analysis **of ancient genome duplication in the yeast Saccharomyces cerevisiae.** Nature 2004, 428:617-624. This article reports the high-coverage sequencing of K. waltii, and subsequent comparison of its genome with that of S. cerevisiae. The comparison reveals a global one-to-two map coincidence as predicted for an ancestral whole genome duplication. The authors continue with an analysis of duplicated gene loss and of the evolution of gene pairs.


34. Gao L-Z, Innan H: Very low gene duplication rate in the yeast genome. Science 2004, 306:1367-1370. Using a method distinct from the classical molecular clock, the authors show that the gene duplication rates in yeast are two orders of magnitude lower than was previously suggested.


36. Koszul R, Caboret S, Dujon B, Fischer G: Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. EMBO J 2004, 23:234-243. This work demonstrates the spontaneous formation of segmental duplications in S. cerevisiae and discusses the impact of this mechanism on genome evolution.


Using functionally characterized protein complexes from S. cerevisiae, the authors demonstrate that they evolved by step-wise partial duplications. They illustrate how module duplication is associated with functional specialization.


The authors describe the large-scale modulation of the transcription programs between two distantly related yeasts and show the fundamental differences between a strict aerobe and a facultative anaerobe. This study demonstrates that the changes in gene expression are connected with the loss of numerous cis-regulatory elements following the apparent whole-genome duplication event.


