PRIMER

Origin of the Yeast Whole-Genome Duplication

Kenneth H. Wolfe*

UCD Conway Institute, School of Medicine & Medical Science, University College Dublin, Dublin, Ireland

* kenneth.wolfe@ucd.ie

Abstract

Whole-genome duplications (WGDs) are rare evolutionary events with profound consequences. They double an organism's genetic content, immediately creating a reproductive barrier between it and its ancestors and providing raw material for the divergence of gene functions between paralogs. Almost all eukaryotic genome sequences bear evidence of ancient WGDs, but the causes of these events and the timing of intermediate steps have been difficult to discern. One of the best-characterized WGDs occurred in the lineage leading to the baker's yeast *Saccharomyces cerevisiae*. Marcet-Houben and Gabaldón now show that, rather than simply doubling the DNA of a single ancestor, the yeast WGD likely involved mating between two different ancestral species followed by a doubling of the genome to restore fertility.

Competing Interests: The author has declared that no competing interests exist.

The unicellular baker's yeast *Saccharomyces cerevisiae* was the first eukaryote to have its genome sequenced, using the first generation of automated sequencing machines and before the advent of the whole-genome shotgun approach. The sequencing was done during the period between 1990 and 1996 by an international consortium that included many small European laboratories, one of which was mine. Each laboratory was given a "tranche" of about 30 kb to sequence, and when you had completed that chunk, you could apply for another one. We were paid €2 per base pair. Progress meetings, chaired energetically by André Goffeau [1], were held every six months to ensure that the project remained on track. At these meetings, each group would make a 5-minute presentation about the genes they had found in their current chunk. The presentations were often tedious, enlivened only by the occasional exigency for André to reassign pieces of DNA from the sequencing tortoises to the hares. But as the project progressed, a pattern began to emerge: many of the chunks were similar to other chunks. The first clone that I sequenced happened to contain the centromere of chromosome II, and I noticed that a gene beside it had a paralog beside the centromere of chromosome IV [2]. My second chunk, from chromosome XV, contained four genes that had four paralogs, in the same order, on chromosome I [3].

When the complete genome was released in April 1996, we were able to identify 55 large duplicated blocks of this type, ranging in size from three to 18 duplicated genes (Fig 1) [4]. Two observations indicated that the duplications were quite old: the average amino acid
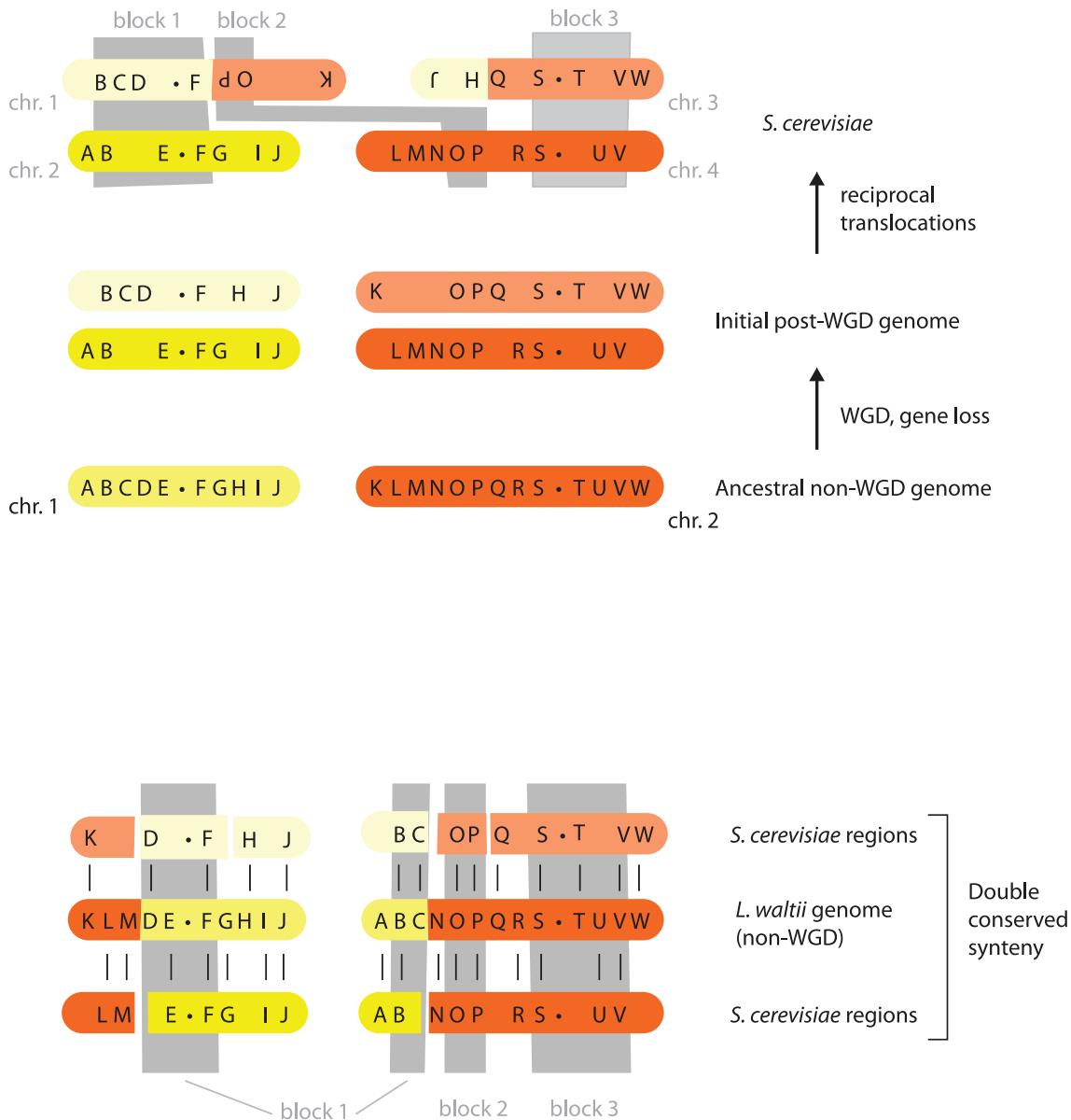
Fig 1. A simple model of WGD, gene loss, and synteny relationships. The upper panel shows how duplicated blocks were initially identified using only genes that remain in duplicate in *S. cerevisiae* [4]. The lower panel shows how additional data from non-WGD yeasts such as *Lachancea waltii* [5] allowed the parts of the genome that were not initially allocated to blocks to be placed into pairs, providing a duplication map that covered the whole *S. cerevisiae* genome. Letters A–W represent genes, and dots represent centromeres. Only two chromosomes (yellow and brown) are shown.

doi:10.1371/journal.pbio.1002221.g001

sequence identity between the gene pairs was only 63%, and within each block only about 25% of the genes were actually duplicated, the others being single copy. This pattern suggested that the whole block was initially duplicated, and then many individual genes were deleted. Two other observations suggested that the blocks were remnants of duplicated chromosomes that had become rearranged during evolution: there were almost no overlaps between the blocks, and the orientation of each pair of blocks was conserved relative to the centromeres and telomeres. This layout of blocks was consistent with duplication of the whole genome followed by both extensive deletion of single genes and genome rearrangement solely by the process of reciprocal translocation between chromosomes [4]. Under this hypothesis, there had been an

ancient whole-genome duplication (WGD), and the 55 blocks that we could identify were simply the most duplicate-dense regions that still survived without evolutionary rearrangement.

The hypothesis of a WGD in *S. cerevisiae* was confirmed in 2004 when three groups sequenced the genomes of species that had branched off from this lineage before the WGD occurred [5–7]. These non-WGD genomes had a "double conserved synteny" relationship with the *S. cerevisiae* genome—that is, instead of each pair of duplicated regions, they had a single region containing all the genes in a merged order (Fig 1). This discovery allowed the entire genome of *S. cerevisiae* to be mapped into pairs of regions via their double conserved synteny with the non-WGD species, even if the pairs retain no duplicated genes, thus filling the gaps between the initial map of 55 duplicated blocks. These analyses proved that the WGD encompassed the entire genome of *S. cerevisiae* and showed that its 16 centromeres fall into eight ancestral pairs that are syntenic with centromeres of the non-WGD species. It therefore appeared that the WGD turned an eight-chromosome ancestor into a 16-chromosome descendant. From this complete map, we now know that among the 5,774 protein-coding genes of *S. cerevisiae*, there are 551 pairs of duplicated genes (ohnologs) that were formed by the WGD and that about 144 chromosomal rearrangements scrambled the genome after the WGD [8,9]. We also know that the WGD is not confined to *Saccharomyces* but occurred in the common ancestor of six genera, some of which diverged from each at an early stage when more than 4,000 genes were still duplicated, leading to later losses of different gene copies in different lineages [10].

What were the molecular events that caused the WGD? It is relatively easy to draw a diagram summarizing the history of each chromosomal region (Fig 2), but it is much more difficult to specify the provenance of the intermediate molecules and the timescales involved. Two alternative scenarios can describe the steps in Fig 2. In both scenarios, event 1 is a DNA replication, and cells W and Z are each capable of mating (they are respectively a non-WGD haploid and a post-WGD haploid). The key question is whether the DNA molecules labeled X and Y existed in (1) two different cells of the same species or (2) two cells of two different species. Scenario 1 is called autopolyploidization, in which case event 1 corresponds to a simple cell division and event 2 is a mating between gametes from the same species or some other form of cell fusion. Scenario 2 is called allopolyploidization or hybridization, in which case event 1 is a speciation and event 2 is an interspecies mating or cell fusion. If event 2 was a mating, then an additional step such as deletion of one allele at the *MAT* locus is necessary to convert cell Z from a nonmating zygote to a mating gamete—but it is not essential that this additional step occurred immediately after event 2. In fact, a long delay in which cell Z replicated mitotically for many generations could be useful because it could allow reproductive isolation from cells of type W to build up. Eventually (event 3), mating between two post-WGD haploid cells of type Z can produce a post-WGD diploid like cell ZZ, which is the state in which *S. cerevisiae* is normally found in nature.

The major difference between these two scenarios is the amount of time (T) that elapsed between events 1 and 2: was it a few generations or millions of years? In scenario 1, molecules X and Y must be identical, whereas in scenario 2 they could have any level of sequence divergence from minimal to extensive, and they could also differ by chromosomal rearrangements. It has been difficult to design tests that could differentiate between these scenarios, but an analysis of the inferred order of genes along molecules X and Y did not find any rearrangements and so did not rule out scenario 1 [9]. However, it has been frustrating that we could not pin down the details of this crucial phase of yeast evolution, which gave birth to many pairs of genes with substantially divergent functions [11–15].

In this issue of *PLOS Biology*, Marcet-Houben and Gabaldón now report strong evidence in support of interspecies hybridization (scenario 2) as the source of the two subgenomes in post-WGD species [16]. By phylogenetic analysis using state-of-the-art methods, they show that
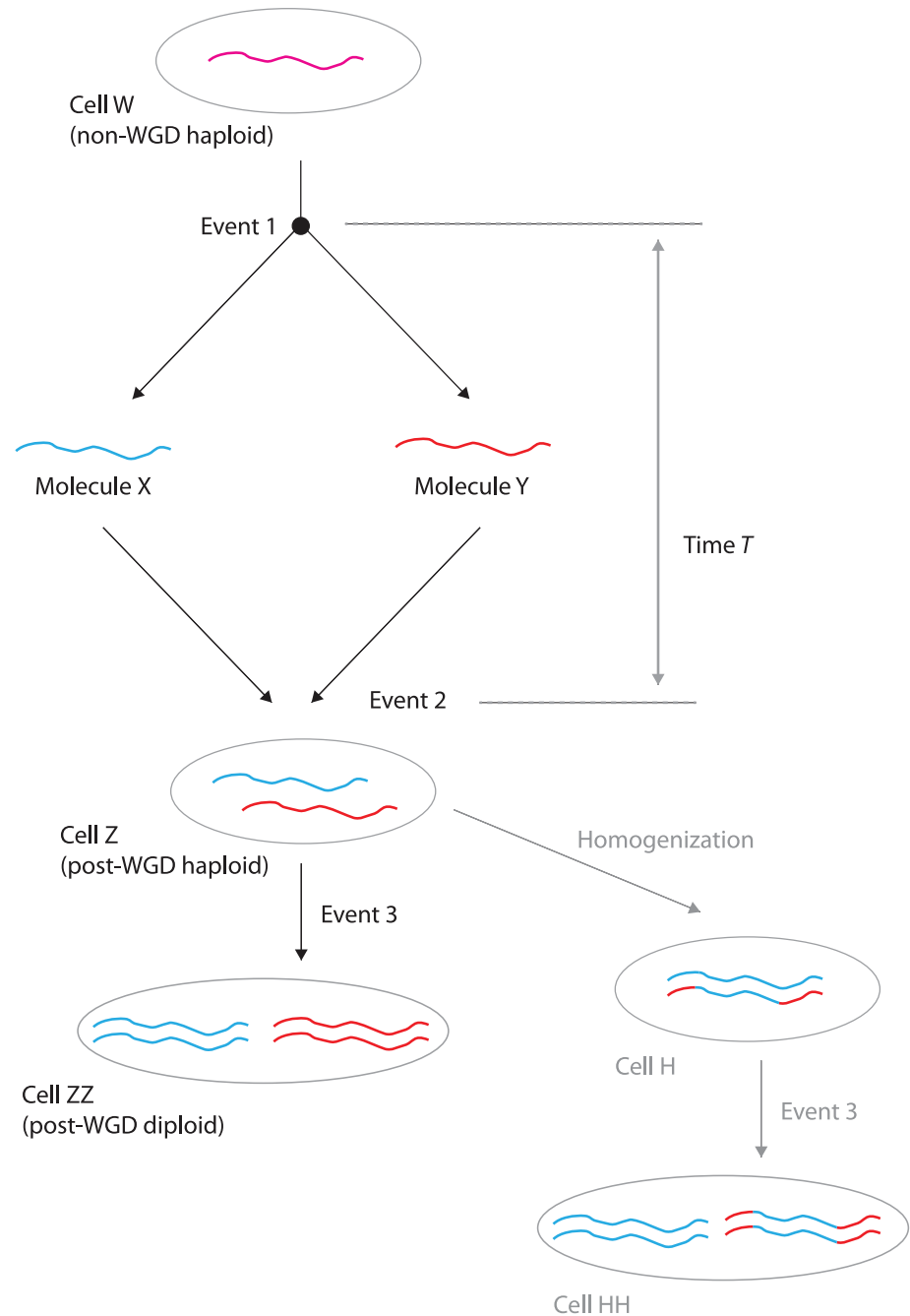
Fig 2. Tracing the history of a single chromosomal region. See text for details. In an allopolyploidization, the red and blue chromosomes are called homeologs.

molecules X and Y have phylogenetic affinities to two different non-WGD lineages that they call the KLE and ZT clades. The KLE clade (*Kluyveromyces*, *Lachancea*, and *Eremothecium*) is the group of non-WGD species that was sequenced in 2004 [5–7]. The ZT clade (*Zygosaccharomyces* and *Torulaspora*) is a separate, more recently studied non-WGD lineage [17,18]. Previous phylogenetic studies using supertrees or concatenated data suggested that the ZT clade is sister to the post-WGD clade, with the KLE clade being an out-group to them both [17,19,20].

The new analysis [16] made trees for each gene individually and found that, although the majority of genes in post-WGD species do cluster phylogenetically with the ZT clade as expected, a significant minority (about 30%) instead either cluster with the KLE clade or form an outgroup to a KLE + ZT clade. This phylogenetic heterogeneity was not noticed before because the KLE signal is only present in a minority of genes, and it is swamped by the ZT signal in methods that try to place the post-WGD clade at a single point on the tree.

Marcet-Houben and Gabaldón interpret this phylogenetic heterogeneity as evidence that the two post-WGD subgenomes have separate origins, one from the ZT clade and the other from an unidentified lineage that is an outgroup to KLE + ZT. Under the simplest hypothesis of hybridization, we might then expect that phylogenetic trees constructed from ohnolog pairs should show one *S. cerevisiae* gene grouping with the ZT clade and the other grouping with the KLE clade, but in fact, most ohnolog pairs group with each other, with the ZT clade as their closest relative [16]. The authors' explanation for these two results—an excess of ZT-like ohnolog pairs, and an excess of ZT-like genes in the whole genome (which is mostly singletons)—is that the post-WGD genomes have been affected by biased gene conversion that preferentially replaced some KLE-derived sequences with copies of the homeologous ZT-derived sequences, homogenizing these regions and obliterating their signal of KLE ancestry.

The hybridization proposed by Marcet-Houben and Gabaldón makes a lot of sense in terms of what we know about the biology of yeast interspecies hybrids. Many yeast strains, most notably those used in commercial settings where stress tolerance is important, have turned out to be interspecies hybrids. For instance, the yeast used to brew lager (*S. pastorianus*) is a hybrid between *S. cerevisiae* and *S. eubayanus* [21,22], and many other combinations of genomes from different species of *Saccharomyces* have been found in nature [23]. These interspecies hybrids are usually infertile (unable to sporulate) because the two copies (homeologs) of each chromosome that they contain are too dissimilar to pair properly during meiosis [24–26]. One simple way to restore fertility is to double the genome, allowing each chromosome to pair with an identical partner instead of trying to pair with the homeolog. In this model, cell Z changes from being a nonmater (effectively diploid) to a mater (effectively haploid—perhaps by deletion of a *MAT* allele), then two cells of type Z mate to produce cell ZZ (diploid), and cell ZZ is able to go through meiosis and make spores with twice the DNA content of cell W. Thus, one hypothesis that Marcet-Houben and Gabaldón propose is that event 2 was an interspecies mating and event 3 was a restoration of fertility by genome doubling, with a possible interval of many mitotic generations between these two events. Alternatively, they hypothesize that event 2 may have been an interspecies fusion of diploid cells, obviating the need for a separate event 3.

The obscuring of the phylogenetic signal of hybridization by subsequent gene conversions [16] is consistent with the known genome structures of some interspecies hybrids. The yeasts *Pichia sorbitophila* [27] and *Candida orthopsilosis* [28] are both interspecies hybrids, but in each case extensive homogenization of parts of the genome has occurred. This process of homogenization has been called overwriting, loss of heterozygosity, or gene conversion by different groups. It leaves the number of chromosomes unchanged (equal to the sum of the numbers of chromosomes in the two incoming subgenomes) but involves the replacement of sequences in one subgenome by sequences copied from the other subgenome (cell H in Fig 2). Homogenization could occur on scales as small as a few hundred base pairs (gene conversion) or as large as whole chromosome arms (break-induced replication). In the latter case, even differences in gene order such as inversions between the parental species could be ironed out.

The discovery that the yeast WGD was an allopolyploidization adds complexity to what initially seemed to be a simple story of duplication. If an interspecies hybrid such as *P. sorbitophila* with a partly homogenized genome developed two mating types and these could mate to form a diploid that could sporulate efficiently, the result would be a species with a genome

resembling the inferred progenitor of the post-WGD clade (cell HH in Fig 2). Allopolyploidy answers some old questions about why genes were retained in duplicate if their sequences were identical (answer: they weren't identical), what the immediate selective advantage of the post-WGD cell was (answer: hybrid vigor), and how the post-WGD lineage became reproductively isolated from the pre-WGD lineage (answer: delay between events 2 and 3). But it also raises new questions about homogenization (how much of the genome? how often? why is it biased?) and about the mechanism of restoration of fertility (why is event 3 so rare, apparently happening only once in the budding yeast family even though event 2 happened quite often?).

Ancient WGDs have been detected right across the eukaryotic tree of life, including in animals, ciliates, fungi, and, most prominently, plants [29–32]. If extensive gene conversion can obscure the traces of allopolyploidization in yeast genomes, one might wonder how many of these other ancient WGDs also began as interspecies hybridizations. In fact, there is evidence from plants that gene conversion acts continually to homogenize ohnolog pairs [32,33] and that hybrid plants can show preferential retention of DNA from one parent over the other [34] similar to the situation in *P. sorbitophila* [27]. Detecting the yeast hybridization in the presence of these obscuring factors required both good luck and good timing: good luck that a reference species closer to one parent than to the other had been sequenced and good timing that the hybrid was sampled before all traces of its hybrid origin had faded away. These fortunate circumstances may not hold for ancient hybridizations in other eukaryotes, but as a famous golfer once said, "The harder I practice, the luckier I get." Detecting that they are hybridizations may become possible with exhaustive sampling of possible parental lineages and the use of sensitive phylogenomic methods of the type introduced by the authors [16].

References

1. Goffeau A. Yeast transport ATPases and the genome sequencing project. Comprehensive Biochemistry. 2004; 43: 493–536.

2. Wolfe KH, Lohan AJ. Sequence around the centromere of *Saccharomyces cerevisiae* chromosome II: similarity of *CEN2* to *CEN4*. Yeast. 1994; 10: S41–46. PMID: 8091860

3. Parle-McDermott AG, Hand NJ, Goulding SE, Wolfe KH. Sequence of 29 kb around the PDR10 locus on the right arm of *Saccharomyces cerevisiae* chromosome XV: similarity to part of chromosome I. Yeast. 1996; 12: 999–1004. PMID: 8896263

4. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. Nature. 1997; 387: 708–713. PMID: 9192896

5. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature. 2004; 428: 617–624. PMID: 15004568

6. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, et al. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science. 2004; 304: 304–307. PMID: 15001715

7. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, et al. Genome evolution in yeasts. Nature. 2004; 430: 35–44. PMID: 15229592

8. Byrne KP, Wolfe KH. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. 2005; 15: 1456–1461. PMID: 16169922

9. Gordon JL, Byrne KP, Wolfe KH. Additions, losses and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. PLoS Genet. 2009; 5: e1000485. doi: 10.1371/journal.pgen.1000485 PMID: 19436716

10. Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci USA. 2007; 104: 8397–8402. PMID: 17494770

11. Hickman MA, Rusche LN. Transcriptional silencing functions of the yeast protein Orc1/Sir3 subfunctionalized after gene duplication. Proc Natl Acad Sci USA. 2010; 107: 19384–19389. doi: 10.1073/pnas.1006436107 PMID: 20974972

12. van Hoof A. Conserved functions of yeast genes support the Duplication, Degeneration and Complementation model for gene duplication. Genetics. 2005; 171: 1455–1461. PMID: 15965245

13. Hittinger CT, Carroll SB. Gene duplication and the adaptive evolution of a classic genetic switch. Nature. 2007; 449: 677–681. PMID: 17928853

14. Marshall AN, Montealegre MC, Jimenez-Lopez C, Lorenz MC, van Hoof A. Alternative splicing and sub-functionalization generates functional diversity in fungal proteomes. PLoS Genet. 2013; 9: e1003376. doi: 10.1371/journal.pgen.1003376 PMID: 23516382

15. Chang CP, Chang CY, Lee YH, Lin YS, Wang CC. Divergent Alanyl-tRNA synthetase genes of *Vanderwaltozyma polyspora* descended from a common ancestor through whole-genome duplication followed by asymmetric evolution. Mol Cell Biol. 2015; 35: 2242–2253. doi: 10.1128/MCB.00018-15 PMID: 25896914

16. Marcet-Houben M, Gabaldón T. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. PLoS Biol. 2015; 13: e1002220. doi: 10.1371/journal.pbio.1002220

17. Souciet JL, Dujon B, Gaillardin C, Johnston M, Baret PV, Cliften P, et al. Comparative genomics of protoploid Saccharomycetaceae. Genome Res. 2009; 19: 1696–1709. doi: 10.1101/gr.091546.109 PMID: 19525356

18. Gordon JL, Armisen D, Proux-Wera E, Oheigeartaigh SS, Byrne KP, Wolfe KH. Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. Proc Natl Acad Sci USA. 2011; 108: 20024–20029. doi: 10.1073/pnas.1112808108 PMID: 22123960

19. Kurtzman CP. Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. FEMS Yeast Res. 2003; 4: 233–245. PMID: 14654427

20. Hedtke SM, Townsend TM, Hillis DM. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst Biol. 2006; 55: 522–529. PMID: 16861214

21. Libkind D, Hittinger CT, Valerio E, Goncalves C, Dover J, Johnston M, et al. Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. Proc Natl Acad Sci USA. 2011; 108: 14539–14544. doi: 10.1073/pnas.1105430108 PMID: 21873232

22. Walther A, Hesselbart A, Wendland J. Genome sequence of *Saccharomyces carlsbergensis*, the world's first pure culture lager yeast. G3 (Bethesda). 2014; 4: 783–793.

23. Hittinger CT. *Saccharomyces* diversity and evolution: a budding model genus. Trends Genet. 2013; 29: 309–317. doi: 10.1016/j.tig.2013.01.002 PMID: 23395329

24. Hunter N, Chambers SR, Louis EJ, Borts RH. The mismatch repair system contributes to meiotic sterility in an interspecific yeast hybrid. EMBO J. 1996; 15: 1726–1733. PMID: 8612597

25. Greig D, Borts RH, Louis EJ, Travisano M. Epistasis and hybrid sterility in *Saccharomyces*. Proc R Soc Lond B Biol Sci. 2002; 269: 1167–1171.

26. Delneri D, Colson I, Grammenoudi S, Roberts IN, Louis EJ, Oliver SG. Engineering evolution to study speciation in yeasts. Nature. 2003; 422: 68–72. PMID: 12621434

27. Louis VL, Despons L, Friedrich A, Martin T, Durrens P, Casarégola S, et al. *Pichia sorbitophila*, an interspecies yeast hybrid, reveals early steps of genome resolution after polyploidization. G3. 2012; 2: 299–311. doi: 10.1534/g3.111.000745 PMID: 22384408

28. Pryszcz LP, Nemeth T, Gacser A, Gabaldon T. Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies. Genome Biol Evol. 2014; 6: 1069–1078. doi: 10.1093/gbe/evu082 PMID: 24747362

29. Sémon M, Wolfe KH. Consequences of genome duplication. Curr Opin Genet Devel. 2007; 17: 505–512.

30. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, et al. Ancestral polyploidy in seed plants and angiosperms. Nature. 2011; 473: 97–100. doi: 10.1038/nature09916 PMID: 21478875

31. Vanneste K, Sterck L, Myburg AA, Van de Peer Y, Mizrachi E. Horsetails Are Ancient Polyploids: Evidence from Equisetum giganteum. Plant Cell. 2015; doi: 10.1105/tpc.15.00157

32. Soltis DE, Visger CJ, Soltis PS. The polyploidy revolution then. . .and now: Stebbins revisited. Amer J Bot. 2014; 101: 1057–1078.

33. Wang X-Y, Paterson AH. Gene conversion in angiosperm genomes with an emphasis on genes duplicated by polyploidization. Genes. 2011; 2: 1–20. doi: 10.3390/genes2010001 PMID: 24710136

34. Freeling M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 2009; 60: 433–453. doi: 10.1146/annurev.arplant.043008.092122 PMID: 19575588