Extensive loss of cell cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts

Jacob L. Steenwyk[1], Dana A. Opulente[2], Jacek Kominek[2], Xing-Xing Shen[1], Xiaofan Zhou[3], Abigail L. Labella[1], Noah P. Bradley[1], Brandt F. Eichman[1], Neža Čadež[5], Diego Libkind[6], Jeremy DeVirgilio[7], Amanda Beth Hulfachor[4], Cletus P. Kurtzman[7], Chris Todd Hittinger[2]*, and Antonis Rokas[1]*

1. Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA

2. Laboratory of Genetics, Genome Center of Wisconsin, DOE Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, University of Wisconsin–Madison, Wisconsin 53706, USA

3. Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Centre, South China Agricultural University, 510642 Guangzhou, China

4. Laboratory of Genetics, Genome Center of Wisconsin, Wisconsin Energy Institute, J.F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI 53706, USA

5. University of Ljubljana, Biotechnical Faculty, Department of Food Science and Technology, Jamnikarjeva 101, 1000 Ljubljana, Slovenia

6. Laboratorio de Microbiología Aplicada, Biotecnología y Bioinformática, Instituto Andino Patagónico de Tecnologías Biológicas y Geoambientales (IPATEC), Universidad Nacional del Comahue - CONICET, San Carlos de Bariloche, 8400, Río Negro, Argentina

7. Mycotoxin Prevention and Applied Microbiology Research Unit, National Center for Agricultural Utilization Research, Agricultural Research Service, US Department of Agriculture, Peoria, Illinois 61604, USA

*Correspondence: cthittinger@wisc.edu and antonis.rokas@vanderbilt.edu

Running title: Gene loss among *Hanseniaspora*

Keywords: phylogenomics, Saccharomycotina, biodiversity, carbohydrate metabolism, genomics, DNA repair, cell cycle, genome stability, DNA damage response

Abstract

Cell cycle checkpoints and DNA repair processes protect organisms from potentially lethal mutational damage. Compared to other budding yeasts in the subphylum Saccharomycotina, we noticed that a lineage in the genus *Hanseniaspora* exhibited very high evolutionary rates, low GC content, small genome sizes, and lower gene numbers. To better understand *Hanseniaspora* evolution, we analyzed 25 genomes, including 11 newly sequenced, representing 18 / 21 known species in the genus. Our phylogenomic analyses identify two *Hanseniaspora* lineages, the fast-evolving lineage (FEL), which began diversifying ~87 million years ago (mya), and the slow-evolving lineage (SEL), which began diversifying ~54 mya. Remarkably, both lineages lost genes associated with the cell cycle and genome integrity, but these losses were greater in the FEL. For example, all species lost the cell cycle regulator *WHI5*, and the FEL lost components of the spindle checkpoint pathway (e.g., *MAD1*, *MAD2*) and DNA damage checkpoint pathway (e.g., *MEC3*, *RAD9*). Similarly, both lineages lost genes involved in DNA repair pathways, including the DNA glycosylase gene *MAG1*, which is part of the base excision repair pathway, and the DNA photolyase gene *PHR1*, which is involved in pyrimidine dimer repair. Strikingly, the FEL lost 33 additional genes, including polymerases (i.e., *POL4* and *POL32*) and telomere-associated genes (e.g., *RIF1*, *RFA3, CDC13, PBP2*). Echoing these losses, molecular evolutionary analyses reveal that, compared to the SEL, the FEL stem lineage underwent a burst of accelerated evolution, which resulted in greater mutational loads, homopolymer instabilities, and higher fractions of mutations associated with the common endogenously damaged base, 8-oxoguanine. We conclude that *Hanseniaspora* is an ancient lineage that has diversified and thrived, despite lacking many otherwise highly conserved cell cycle and genome integrity genes and pathways, and may represent a novel system for studying cellular life without them.

Introduction

Genome maintenance is largely attributed to the fidelity of cell cycle checkpoints, DNA repair pathways, and their interaction [1]. Dysregulation of these processes often leads to the loss of genomic integrity [2] and hypermutation, or the acceleration of mutation rates [3]. For example, improper control of cell cycle and DNA repair processes can lead to 10- to 100-fold increases in mutation rate [4]. Furthermore, deletions of single genes can have profound effects on genome stability. For example, the deletion of *MEC3*, which is involved in sensing DNA damage in the G1 and G2/M cell cycle phases, can lead to a 54-fold increase in the gross chromosomal rearrangement rate [5]. Similarly, nonsense mutations in mismatch repair proteins account for the emergence of hypermutator strains in the yeast pathogens *Cryptococcus deuterogattii* [6] and *Cryptococcus neoformans* [7,8]. Due to their importance in ensuring genomic integrity, most genome maintenance-associated processes are thought to be evolutionarily ancient and broadly conserved [9].

One such ancient and highly conserved process in eukaryotes is the cell cycle [10,11]. Landmark features of cell cycle control include cell size control, the mitotic spindle checkpoint, the DNA damage response checkpoint, and DNA replication [9]. Cell size is controlled, in part, through the activity of *WHI5*, which represses the G1/S transition by inhibiting G1/S transcription [12]. Similarly, when kinetochores are improperly attached or are not attached to microtubules, the mitotic spindle checkpoint helps to prevent activation of the anaphase-promoting complex (APC), which controls the G1/S and G2/M transitions [9,13]. Additional key regulators in this process are Mad1 and Mad2, which dimerize at unattached kinetochores and delay anaphase. Failure of Mad1:Mad2 recruitment to unattached kinetochores results in failed checkpoint activity [14]. Importantly, many regulators, including but not limited to those mentioned here, are highly similar in structure and function between fungi and animals and are thought to have a shared ancestry [10]. Interestingly, cell cycle initiation in certain fungi (including *Hanseniaspora*) is achieved through SBF, a transcription factor that is functionally equivalent but evolutionarily unrelated to E2F, the transcription factor that that initiates the cycle in animals, plants, and certain early-diverging fungal lineages [11]. SBF is postulated to have been acquired via a viral infection, suggesting that evolutionary changes in this otherwise highly conserved process can and do rarely occur [11,15].

DNA damage checkpoints can arrest the cell cycle and influence the activation of DNA repair pathways, the recruitment of DNA repair proteins to damaged sites, and the composition and length of telomeres [16]. For example, *MEC3* and *RAD9*, function as checkpoint genes required for arrest in the G2 phase after DNA damage has occurred [17]. Additionally, the deletions of DNA damage and checkpoint genes have been known to cause hypermutator phenotypes in the baker's yeast *Saccharomyces cerevisiae* [18]. Similarly, hypermutator phenotypes are associated with loss-of-function mutations in DNA polymerase genes [19]. For example, deletion of the DNA polymerase δ subunit gene, *POL32*, which participates in multiple DNA repair processes, causes an increased mutational load and hypermutation in *S. cerevisiae*, in part, through the increase of genomic deletions and small indels [18,20]. Likewise, the deletion of *MAG1*, a gene encoding a DNA glycosylase that removes damaged bases via the multi-step base excision repair pathway, can cause a 2,500-fold increased sensitivity to the DNA alkylating agent methyl methanesulfonate [21].

In contrast to genes in multi-step DNA repair pathways, other DNA repair genes function individually or are parts of simpler regulatory processes. For example, *PHR1*, a gene that encodes a photolyase, is activated in response to and repairs pyrimidine dimers, one of the most frequent types of lesions caused by damaging UV light [22,23]. Other DNA repair genes do not interact with DNA but function to prevent the misincorporation of damaged bases. For example, *PCD1* encodes a 8-oxo-dGTP diphosphatase [24], which suppresses G → T or C → A transversions by removing 8-oxo-dGTP, thereby preventing the incorporation of the base 8-oxo-dG, one of the most abundant endogenous forms of an oxidatively damaged base [24–26]. Collectively, these studies demonstrate that the loss of DNA repair genes can lead to hypermutation and increased sensitivity to DNA damaging agents.

Hypermutation phenotypes are generally short-lived because most mutations are deleterious and are generally adaptive only in highly stressful or rapidly fluctuating environments [27]. For example, in *Pseudomonas aeruginosa* infections of cystic fibrosis patients [28] and mouse gut-colonizing *Escherichia coli* [29], hypermutation is thought to facilitate adaptation to the host environment and the evolution of drug resistance. Similarly, in the fungal pathogens *C. deuterogattii* [6] and *C. neoformans* [7,8], hypermutation is thought to contribute to within-host adaptation, which may involve modulating traits such as drug resistance [6]. However, as adaptation to a new environment nears completion, hypermutator alleles are expected to decrease in frequency due to the accumulation of deleterious mutations that result as a consequence of the high mutation rate [30,31]. In agreement with this prediction, half of experimentally evolved hypermutating lines of *S. cerevisiae* had reduced mutation rates after a few thousand generations [32], suggesting hypermutation is a short-lived phenotype and that compensatory mutations can restore or lower the mutation rate. Additionally, this experiment also provided insights to how strains may cope with hypermutation; for example, all *S. cerevisiae* hypermutating lines increased their ploidy to presumably reduce the impact of higher mutation rates [32]. Altogether, hypermutation can produce short-term advantages but causes long-term disadvantages, which may explain its repeated but short-term occurrence in clinical environments [29] and its sparseness in natural ones. While these theoretical and experimental studies have provided seminal insights to the evolution of mutation rate and hypermutation, we still lack understanding of the long-term, macroevolutionary effects of increased mutation rates.

Recently, multiple genome-scale phylogenies of species in the budding yeast subphylum Saccharomycotina showed that certain species in the bipolar budding yeast genus *Hanseniaspora* are characterized by very long branches [33–35], which are reminiscent of the very long branches of fungal hypermutator strains [6–8]. Most of what is known about these cosmopolitan apiculate yeasts relates to their high abundance on mature fruits and in fermented beverages [36], especially on grapes and in wine must [37,38]. As a result, *Hanseniaspora* plays a significant role in the early stages of fermentation and can modify wine color and flavor through the production of enzymes and aroma compounds [39]. Surprisingly, even with the use of *S. cerevisiae* starter cultures, *Hanseniaspora* species, particularly *Hanseniaspora uvarum,* can achieve very high cell densities , in certain cases comprising greater than 80% of the total yeast population, during early stages of fermentation [40], suggesting exceptional growth capabilities in this environment.

To gain insight into the long branches and the observed fast growth of *Hanseniaspora*, we sequenced and extensively characterized gene content and patterns of evolution in 25 genomes, including 11 newly sequenced for this study, from 18 / 21 known species in the genus. Our analyses delineated two lineages, the fast-evolving lineage (FEL), which has a strong signature of acceleration in evolutionary rate at its stem branch, and the slow-evolving lineage (SEL), which has a weaker signature of evolutionary rate acceleration at its stem branch. Relaxed molecular clock analyses estimate that the FEL and SEL split ~95 million years ago (mya). The degree of evolutionary rate acceleration is commensurate with the preponderance of loss of genes associated with metabolic, cell cycle, and DNA repair processes. Specifically, compared to *S. cerevisiae*, there are 748 genes that were lost from two-thirds of *Hanseniaspora* genomes with FEL yeasts having lost an additional 661 genes and SEL yeasts having lost only an additional 23. Both lineages have lost major cell cycle regulators, including *WHI5* and components of the APC, while FEL species additionally lost numerous genes associated with the spindle checkpoint (e.g., *MAD1* and *MAD2*) and DNA damage checkpoint (e.g., *MEC3* and *RAD9*). Similar patterns are observed among DNA repair-related genes; *Hanseniaspora* species have lost 14 genes, while the FEL yeasts have lost an additional 33 genes. For example, both lineages have lost *MAG1* and *PHR1*, while the FEL has lost additional genes including polymerases (i.e., *POL32* and *POL4*) and multiple telomere-associated genes (e.g., *RIF1*, *RFA3, CDC13, PBP2*). Compared to the SEL, analyses of substitution patterns in the FEL show higher levels of sequence substitutions, greater instability of homopolymers, and a greater mutational signature associated with the commonly damaged base, 8-oxo-dG [26]. Furthermore, we find that the transition to transversion (or transition / transversion) ratios of the FEL and the SEL are both very close to the ratio expected if transitions and transversions occur neutrally. These results are consistent with the hypothesis that species in the FEL represent a novel example of diversification and long-term evolutionary survival of a hypermutator lineage, which highlights the potential of *Hanseniaspora* for understanding the long-term effects of hypermutation on genome function and evolution.

Results

An exceptionally high evolutionary rate in the FEL stem branch

Concatenation and coalescence analyses of a data matrix of 1,034 single-copy OGs (522,832 sites; 100% taxon-occupancy) yielded a robust phylogeny of the genus *Hanseniaspora* (Fig 1A, Fig S2, Fig S3). Consistent with previous analyses [34,35,41], our phylogeny revealed the presence of two major lineages, each of which was characterized by long stem branches; we hereafter refer to the lineage with a longer stem branch as the fast-evolving lineage (FEL) and to the other as the slow-evolving lineage (SEL). Relaxed molecular clock analysis suggests that the FEL and SEL split 95.34 (95% credible interval (CI): 117.38 – 75.36) mya, with the origin of their crown groups estimated at 87.16 (95% CI: 112.75 – 61.38) and 53.59 (95% CI: 80.21 – 33.17) mya, respectively (Fig 1A, Fig S4, File S2).

The FEL stem branch is much longer than the SEL stem branch in the *Hanseniaspora* phylogeny (Fig 1) (see also phylogenies in: Shen et al., 2016, 2018). To determine whether this difference in branch length was a property of some or all single-gene phylogenies, we compared the difference in length of the FEL and SEL stem branches among all single-gene trees where each lineage was recovered monophyletic ($n = 946$). We found that the FEL stem branch was nearly four times longer (0.62 ± 0.38 substitutions / site) than the SEL stem branch (0.17 ± 0.11 substitutions / site) (Fig 1B; $p < 0.001$; Paired Wilcoxon Rank Sum test). Furthermore, of the 946 gene trees examined, 932 had a much longer FEL stem branch (0.46 ± 0.33 Δ substitutions / site), whereas only 14 had a slightly longer SEL stem branch (0.06 ± 0.05 Δ substitutions / site).

The genomes of FEL species have lost substantial numbers of genes

Examination of GC content, genome size, and gene number revealed that the some of the lowest GC content values, as well as the smallest genomes and lowest gene numbers, across the subphylum Saccharomycotina are primarily observed in FEL yeasts (Fig S1). Specifically, the average GC contents for FEL yeasts (33.10 ± 3.53%), SEL yeasts (37.28 ± 2.05%), and all other Saccharomycotina yeasts (40.77 ± 5.58%) are significantly different ($\chi^2(2) = 30.00$, $p < 0.001$; Kruskal-Wallis rank sum test). Further examination revealed only the FEL was significantly different from other Saccharomycotina yeasts ($p < 0.001$; Dunn's test for multiple comparisons with Benjamini-Hochberg multi-test correction). For genome size and gene number, FEL yeast genomes have average sizes of 9.71 ± 1.32 Mb and contain 4,707.89 ± 633.56 genes, respectively, while SEL yeast genomes have average sizes of 10.99 ± 1.66 Mb and contain 4,932.43 ± 289.71 genes. In contrast, all other Saccharomycotina have average genome sizes and gene numbers of 13.01 ± 3.20 Mb and 5,726.10 ± 1,042.60, respectively. Statistically significant differences were observed between the FEL, SEL, and all other Saccharomycotina (genome size: $\chi^2(2) = 33.47$, $p < 0.001$ and gene number: $\chi^2(2) = 31.52$, $p < 0.001$; Kruskal-Wallis rank sum test for both). Further examination revealed the only significant difference for genome size was between FEL and other Saccharomycotina yeasts ($p < 0.001$; Dunn's test for multiple comparisons with Benjamini-Hochberg multi-test correction), while both the FEL and SEL had smaller gene sets compared to other Saccharomycotina yeasts ($p < 0.001$ and $p = 0.008$, respectively; Dunn's test for multiple comparisons with Benjamini-Hochberg multi-test correction). The lower numbers of genes in the FEL (especially) and SEL lineages were also supported by gene content completeness analyses using orthologous sets of genes constructed from sets of genomes representing multiple taxonomic levels across eukaryotes (Fig S5) from the ORTHODB database [43].

To further examine which genes have been lost in the genomes of FEL and SEL species relative to other representative Saccharomycotina genomes, we conducted HMM-based sequence similarity searches using annotated *S. cerevisiae* genes as queries in HMM construction (see *Methods*) (Fig S6). Because we were most interested in identifying genes absent from the FEL and SEL, we focused our analyses on genes lost in at least two-thirds of each lineage (i.e., ≥ 11 FEL taxa or ≥ 5 SEL taxa). Using this criterion, we found that 1,409 and 771 genes have been lost in the FEL and SEL, respectively (Fig 2A). Among the genes lost in each lineage, 748 genes were lost across both lineages, 661 genes have been uniquely lost in the FEL, and 23 genes have been uniquely lost in the SEL (File S3).

To identify the likely functions of genes lost from each lineage, we conducted GO enrichment analyses. Examination of significantly over-represented GO terms for the sets of genes that have been lost in *Hanseniaspora* genomes revealed numerous categories related to metabolism (e.g., MALTOSE METABOLIC PROCESS, GO:0000023, $p = 0.006$; SUCROSE ALPHA-GLUCOSIDASE ACTIVITY, GO:0004575, $p = 0.003$) and genome-maintenance processes (e.g., MEIOTIC CELL CYCLE, GO:0051321, $p < 0.001$) (File S4). Additional terms, such as CELL CYCLE, GO:0007049 ($p < 0.001$), CHROMOSOME SEGREGATION, GO:0007059 ($p < 0.001$), CHROMOSOME ORGANIZATION, GO:0051276 ($p = 0.009$), and DNA-DIRECTED DNA POLYMERASE ACTIVITY, GO:0003887 ($p < 0.001$), were significantly over-represented among genes absent only in the FEL. Next, we examined in more detail the identities and likely functional consequences of extensive gene losses across *Hanseniaspora* associated with metabolism, cell cycle, and DNA repair.

Metabolism-associated gene losses. Examination of the genes causing over-representation of metabolism-associated GO terms revealed gene losses in the *IMA* gene family and the *MAL* loci, both of which are associated with growth primarily on maltose but can also facilitate growth on sucrose, raffinose, and melezitose [44,45]. All *IMA* genes have been lost in *Hanseniaspora*, whereas *MALx3*, which encodes the *MAL*-activator protein [46] has been lost in all but one species (*Hanseniaspora jakobsenii*; Fig 2B). Consistent with these losses, *Hanseniaspora* species cannot grow on the carbon substrates associated with these genes (i.e., maltose, raffinose, and melezitose) with the exception of *H. jakobsenii*, which has weak/delayed growth on maltose (Fig 2B; File S5). The growth of *H. jakobsenii* on maltose may be due to a cryptic α-glucosidase gene or represent a false positive, as *MALx2* encodes the required enzyme for growth on maltose and is absent in *H. jakobsenii*. Because these genes are also associated with growth on sucrose in some species [44], we also examined their ability to grow on this substrate. In addition to the *MAL* loci conferring growth on sucrose, the invertase Suc2 can also break down sucrose into glucose and fructose [47]. We found that FEL yeasts have lost *SUC2* and are unable to grow on sucrose, while SEL yeasts have *SUC2* and are able to grow on this substrate (Fig 2B; File S5). Altogether, patterns of gene loss are consistent with known metabolic traits.

Examination of gene sets associated with growth on other carbon substrates revealed that *Hanseniaspora* species also cannot grow on galactose, consistent with the loss of one or more of the three genes involved in galactose assimilation (*GAL1, GAL7,* and *GAL10*) from their genomes (Fig 2C; File S5). Additionally, all *Hanseniaspora* genomes appear to have lost two key genes, *PCK1* and *FBP1*, encoding enzymes in the gluconeogenesis pathway (Fig S7A and S7C); in contrast, all *Hanseniaspora* have an intact glycolysis pathway (Fig S7B and S7D).

Manual examination of other metabolic pathways revealed that *Hanseniaspora* genomes are also missing some of their key genes. For example, we found that THIAMINE BIOSYNTHETIC PROCESS, GO:0009228 ($p = 0.003$), was an over-represented GO term among genes missing in both the FEL and SEL due to the absence of *THI* and *SNO* family genes. Further examination of genes present in the thiamine biosynthesis pathway revealed extensive gene loss (Fig 2D), which is consistent with their inability to grow on vitamin-free media [45] (File S5). Notably, *Hanseniaspora* are still predicted to be able to import extracellular thiamine via Thi73 and convert it to its active cofactor via Thi80, which may explain why they can rapidly consume thiamine [39]. Similarly, examination of amino acid biosynthesis pathways revealed the methionine salvage pathway was also largely disrupted by gene losses across all *Hanseniaspora* (Fig 2E). Lastly, we found that *GDH1* and *GDH3* from the glutamate biosynthesis pathway from ammonium are missing in FEL yeasts (File S3). However, *Hanseniaspora* have *GLT1,* which enables glutamate biosynthesis from glutamine.

Cell cycle and genome integrity-associated gene losses. Many genes involved in cell cycle and genome integrity, including cell cycle checkpoint genes, have been lost across

Hanseniaspora (Fig 3). For example, *WHI5* and *DSE2*, which are responsible for repressing the Start (i.e., an event that determines cells have reached a critical size before beginning division) [48] and help facilitate daughter-mother cell separation through cell wall degradation [49], have been lost in both lineages. Additionally, the FEL has lost the entirety of the DASH complex (i.e., *ASK1*, *DAD1*, *DAD2*, *DAD3*, *DAD4*, *DUO1, DAM1*, *HSK3, SPC19*, and *SPC34*), which forms part of the kinetochore and functions in spindle attachment and stability, as well as chromosome segregation, and the MIND complex (i.e., *MTW1*, *NNF1*, *NSL1*, and *DSN1*), which is required for kinetochore bi-orientation and accurate chromosome segregation (File S3 and S4). Similarly, FEL species have lost *MAD1* and *MAD2*, which are associated with spindle checkpoint processes and have abolished checkpoint activity when their encoded proteins are unable to dimerize [14]. Lastly, components of the anaphase-promoting complex, a major multi-subunit regulator of the cell cycle, are lost in both lineages (i.e., *CDC26* and *MND2*) or just the FEL (i.e., *APC2, APC4, APC5,* and *SWM1*).

Another group of genes that have been lost in *Hanseniaspora* are genes associated with the DNA damage checkpoint and DNA damage sensing. For example, both lineages have lost *RFX1*, which controls a late point in the DNA damage checkpoint pathway [50], whereas the FEL has lost *MEC3* and *RAD9*, which encode checkpoint proteins required for arrest in the G2 phase after DNA damage has occurred [17]. Since losses in DNA damage checkpoints and dysregulation of spindle checkpoint processes are associated with genomic instability, we next evaluated the ploidy of *Hanseniaspora* genomes [51]. Using base frequency plots, we found that the ploidy of genomes of FEL species ranges between 1 and 3, with evidence suggesting that certain species, such as *H. singularis, H. pseudoguilliermondii*, and *H. jakobsenii,* are potentially aneuploid (Fig

S8). In contrast, the genomes of SEL species have ploidies of 1-2 with evidence of potential aneuploidy observed only in *H. occidentalis* var. *citrica.* Greater variance in ploidy and aneuploidy in the FEL compared to the SEL may be due to the FEL's loss of a greater number of components of the anaphase-promoting complex (APC), whose dysregulation is thought to increase instances of aneuploidy [52].

Pronounced losses of DNA repair genes in the FEL. Examination of other GO-enriched terms revealed numerous genes associated with diverse DNA repair processes that have been lost among *Hanseniaspora* species, and especially the FEL (Fig 4). We noted 14 lost DNA repair genes across all *Hanseniaspora*, including the DNA glycosylase gene *MAG1* [53], the photolyase gene *PHR1* that exclusively repairs pyrimidine dimers [23], and the diphosphatase gene *PCD1*, a key contributor to the purging of mutagenic nucleotides, such as 8-oxo-dGTP, from the cell [24]. An additional 33 genes were lost specifically in the FEL such as *TDP1*, which repairs damage caused by topoisomerase activity [54]; the DNA polymerase gene *POL32* that participates in base-excision and nucleotide-excision repair and whose null mutants have increased genomic deletions [20]; and the *CDC13* gene that encodes a telomere-capping protein [55].

FEL gene losses are associated with accelerated sequence evolution

Loss of DNA repair genes is associated with a burst of sequence evolution. To examine the mutational signatures of losing numerous DNA repair genes on *Hanseniaspora* substitution rates, we tested several different hypotheses that postulated changes in the ratio of the rate of nonsynonymous (dN) to the rate of synonymous substitutions (dS) (dN/dS or ω) along the phylogeny (Table 1; Fig 5). For each hypothesis tested, the null was that the ω value remained constant across all branches of the phylogeny. Examination of the hypothesis that the ω values of both the FEL and SEL stem branches were distinct from the background ω value ($H_{FE\text{-}SE\ branch}$; Fig 5B), revealed that 678 genes (68.55% of examined genes) significantly rejected the null hypothesis (Table 1; α = 0.01; LRT; median FEL stem branch ω = 0.57, median SEL stem branch ω = 0.29, and median background ω = 0.060). Examination of the hypothesis that the ω value of the FEL stem branch and the ω value of the FEL crown branches were distinct from the background ω value (H_{FE}; Fig 5C) revealed 743 individual genes (75.13% of examined genes) that significantly rejected the null hypothesis (Table 1; α = 0.01; LRT; median FEL stem branch ω = 0.71, median FEL crown branches ω = 0.06, median background ω = 0.063). Testing the same hypothesis for the SEL (H_{SE}; Fig 5D) revealed 528 individual genes (53.7% of examined genes) that significantly rejected the null hypothesis (Table 1; α = 0.01; LRT; median SEL stem branch ω = 0.267, median SEL crown branches ω = 0.074, median background ω = 0.059). Finally, testing of the hypothesis that the FEL and SEL crown branches have ω values distinct from each other and the background ($H_{FE\text{-}SE\ crown}$; Fig 5E) revealed 717 genes (72.5% of examined genes) that significantly rejected the null hypothesis (Table 1; α = 0.01; LRT; median FEL crown branches ω = 0.062, median SEL crown branches ω = 0.074, median background ω = 0.010). These results suggest a dramatic, genome-wide increase in evolutionary rate in the FEL stem branch (Fig 5B and 5C), which coincided with the loss of a large number of genes involved in DNA repair.

The FEL has a greater number of base substitutions and indels. To better understand the mutational landscape in the FEL and SEL, we characterized patterns of base substitutions across the 1,034 OGs. Focusing on first (*n* = 240,565), second (*n* = 318,987), and third (*n* =

58,151) codon positions that had the same character state in all outgroup taxa, we first examined how many of these sites had experienced base substitutions in FEL and SEL species (Fig 6A). We found significant differences between the proportions of base substitutions in the FEL and SEL (F(1) = 196.88, $p < 0.001$; Multi-factor ANOVA) at each codon position (first: $p < 0.001$; second: $p < 0.001$; and third: $p = 0.02$; Tukey Honest Significance Differences post-hoc test).

Examination of whether the observed base substitutions were AT- (i.e., G|C → A|T) or GC- (i.e., A|T → G|C) biased revealed differences between the FEL and SEL (F(1) = 447.1, $p < 0.001$; Multi-factor ANOVA), as well as between AT- and GC-bias (F(1) = 914.5, $p < 0.001$; Multi-factor ANOVA) among sites with G|C ($n = 232{,}546$) and A|T ($n = 385{,}157$) pairs (Fig 6B). Specifically, we observed significantly more base substitutions in the FEL compared to the SEL and a significant bias toward A|T across both lineages ($p < 0.001$ for both tests; Tukey Honest Significance Differences post-hoc test). Examination of transition / transversion ratios revealed a lower transition / transversion ratio in the FEL (0.67 ± 0.02) compared to the SEL (0.76 ± 0.01) (Fig 6C; $p < 0.001$; Wilcoxon Rank Sum test); this finding is in contrast to the transition / transversion ratios found in most known organisms, whose values are substantially above 1.00 [56–59]. Altogether, these analyses reveal more base substitutions in the FEL and SEL across all codon positions and a significant AT-bias in base substitutions across all *Hanseniaspora*.

Examination of indels revealed that the total number of insertions or deletions was significantly greater in the FEL ($\text{mean}_{\text{insertions}} = 7521.11 \pm 405.34$; $\text{mean}_{\text{deletions}} = 3894.11 \pm 208.16$) compared to the SEL ($\text{mean}_{\text{insertions}} = 6049.571 \pm 155.85$; $\text{mean}_{\text{deletions}} = 2346.71 \pm 326.22$) (Fig 6D; $p < 0.001$ for both tests; Wilcoxon Rank Sum test). The difference in number of indels between the

FEL and SEL remained significant after taking into account indel size (F(1) = 2102.87, $p < 0.001$; Multi-factor ANOVA). Further analyses revealed there are significantly more insertions in the FEL compared to the SEL for insertion sizes 3-18 bp ($p < 0.001$ for all comparisons between each lineage for each insertion size; Tukey Honest Significance Differences post-hoc test), while there were significantly more deletions in the FEL compared to the SEL for deletion sizes 3-21 bp ($p < 0.001$ for all comparisons between each lineage for each deletion size; Tukey Honest Significance Differences post-hoc test). These analyses suggest that there are significantly more indels in the FEL compared to the SEL and that this pattern is primarily driven by short indels.

Greater sequence instability in the FEL and signatures of endogenous and exogenous DNA damage

The FEL has greater instability of homopolymers. Examination of the total proportion of mutated bases among homopolymers (i.e., (substituted bases + deleted bases + inserted bases) / total homopolymer bases) revealed significant differences between the FEL and SEL (Fig 6G; F(1) = 27.68, $p < 0.001$; Multi-factor ANOVA). Although the FEL had a higher proportion of mutations among homopolymers across all sizes of two (n = 17,391), three (n = 1,062), four (n = 104), and five (n = 5), significant differences were observed for homopolymers of length two and three (p = 0.02 and p = 0.003, respectively; Tukey Honest Significance Differences post-hoc). To gain more insight into the drivers differentiating mutational load in homopolymers, we considered the additional factors of homopolymer sequence type (i.e., A|T or C|G) and mutation type (i.e., base substitution, insertion, or deletion) (Fig S9). In addition to recapitulating differences between the types of mutations that occur at homopolymers (F(2) = 1686.70, $p <$

0.001; Multi-factor ANOVA), we observed that base substitutions occurred more frequently than insertions and deletions ($p < 0.001$ for both tests; Tukey Honest Significance Differences post-hoc test). For example, among A|T and C|G homopolymers of length two and C|G homopolymers of length three, base substitutions were higher in the FEL compared to the SEL ($p = 0.009$, $p < 0.001$, and $p < 0.001$, respectively; Tukey Honest Significance Differences post-hoc test). Additionally, there were significantly more base substitutions in A|T homopolymers of length five in the FEL compared to the SEL ($p < 0.001$; Tukey Honest Significance Differences post-hoc test). Altogether, these analyses reveal greater instability of homopolymers in the FEL compared to the SEL due to more base substitutions.

The FEL has a stronger signature of endogenous DNA damage from 8-oxo-dG. Examination of mutational signatures associated with common endogenous and exogenous mutagens revealed greater signatures of mutational load in the FEL compared to the SEL, as well as in both FEL and SEL compared to the outgroup taxa. The oxidatively damaged guanine base, 8-oxo-dG, is a commonly observed endogenous form of DNA damage that causes the transversion mutation of G → T or C → A [26]. Examination of the direction of base substitutions among all sites with a G base in all outgroup taxa revealed differences in the direction of base substitutions (F(2) = 5,682, $p < 0.001$; Multi-factor ANOVA). Moreover, there are significantly more base substitutions at G sites associated with 8-oxo-dG damage in the FEL compared to the SEL (Fig 6H; $p < 0.001$; Tukey Honest Significance Differences post-hoc test). These analyses reveal that FEL genomes have higher proportions of G site substitutions associated with the mutational signature of a common endogenous mutagen.

Hanseniaspora *have a greater genomic signature of UV-damage.* Both the FEL and SEL have lost *PHR1*, a gene encoding a DNA photolyase that repairs pyrimidine dimers, so we next examined the genomes for evidence of a CC → TT dinucleotide substitution bias, an indirect molecular signature of UV radiation damage (Fig 6I). To do so, we used a CC|GG and TT|AA score, which quantifies the abundance of CC|GG and TT|AA dinucleotides in a genome and corrects for the total number of dinucleotides and GC content in the same genome. When comparing CC|GG scores between the FEL, SEL, and outgroup taxa, there were no significant differences ($\chi^2(2) = 5.96$, $p = 0.051$; Kruskal-Wallis rank sum test). When comparing all *Hanseniaspora* to the outgroup, we found that the CC|GG score was significantly lower in *Hanseniaspora* ($p = 0.03$; Wilcoxon Rank Sum test). Examination of TT|AA scores revealed significant differences between the three groups ($\chi2(2) = 8.84$, $p = 0.012$; Kruskal-Wallis rank sum test), which was driven by differences between the FEL and SEL compared to the outgroup ($p = 0.011$ and 0.016, respectively; Dunn's test for multiple comparisons with Benjamini-Hochberg multi-test correction). The same result was observed when comparing all *Hanseniaspora* to the outgroup ($p < 0.001$; Wilcoxon Rank Sum test). Altogether, these analyses suggest *Hanseniaspora* have a greater signal of UV damage compared to other budding yeasts.

Lastly, we examined if all of these mutations were associated with more radical amino acid changes in the FEL compared to the SEL using two measures of amino acid change: Sneath's index [60] and Epstein's coefficient of difference [61]. For both measures, we observed significantly more radical amino acid substitutions in the FEL compared to the SEL (Fig S10; $p < 0.001$; Wilcoxon Rank Sum test for both metrics). Altogether, these analyses reveal greater

DNA sequence instability in the FEL compared to the SEL, which is also associated with more radical amino acid substitutions.

Discussion

The genus *Hanseniaspora* has been recently observed to exhibit the longest branches among budding yeasts (Fig 1) [33–35], and their genomes have some of the lowest numbers of genes, lowest GC contents, and smallest assembly sizes in the subphylum (Fig S1). Through the analysis of the genomes of nearly every known *Hanseniaspora* species this study presents multiple lines of evidence suggesting that one lineage of *Hanseniaspora*, which we have named FEL, is a lineage of long-term, hypermutator species that have undergone extensive gene loss (Figs. 1-4).

Evolution by gene loss is gaining increasing attention as a major mode of genome evolution [34,62] and is mainly possible due to the dispensability of the majority of genes. For example, 90% of *E. coli* [63], 80% of *S. cerevisiae* [64], and 73% of *Candida albicans* [65] genes are dispensable in laboratory conditions. The loss of dispensable genes can be selected for [66] and is common in lineages of obligate parasites or symbionts, such as in the microsporidia, intracellular fungi which have lost key metabolic pathways such as amino acid biosynthesis pathways [67,68], and myxozoa, a group of cnidarian obligate parasites that infect vertebrates and invertebrates [69]. Similar losses are also increasingly appreciated in free-living organisms, such as the budding yeasts (this study; Hittinger et al., 2004; Riley et al., 2016; Shen et al., 2018; Slot and Rokas, 2010; Wolfe et al. 2015) and animals [62]. For example, a gene known to enable sucrose utilization, *SUC2* [47], is lost in the FEL and reflects an inability to grow on sucrose, while the *SUC2* is present in the SEL and reflects an ability to grow on sucrose (Fig 2).

However, *Hanseniaspora* species have experienced not just the typically observed losses of metabolic genes (Figs. 2A and 2B), but more strikingly, the atypical loss of dozens of cell cycle and DNA damage, response, and repair genes (Figs. 3 and 4). Losses of cell cycle genes are extremely rare [11], and most such losses are known in the context of cancers [73]. Losses of individual or a few DNA repair genes have also been observed in individual hypermutator fungal isolates [6–8]. In contrast, the *Hanseniaspora* losses of cell cycle and DNA repair genes are not only unprecedented in terms of the numbers of genes lost and their striking impact on genome sequence evolution, but also in terms of the evolutionary longevity of the lineage.

Missing checkpoint processes are associated with fast growth and bipolar budding.

Hanseniaspora species lost numerous components of the cell cycle (Fig 3), such as *WHI5*, which causes accelerated G1/S transitions in knock-out *S. cerevisiae* strains [12,48], as well as components of APC (i.e., *CDC26* and *MND2*), which may accelerate the transition to anaphase [13]. These and other cell cycle gene losses are suggestive of rapid cell division and growth and consistent with the known ability of *Hanseniaspora* yeast of rapid growth in the wine fermentation environment [40].

One of the distinguishing characteristics of the *Hanseniaspora* cell cycle is bipolar budding, which is known only in the genera *Wickerhamia* (Debaryomycetaceae) and *Nadsonia* (Dipodascaceae), as well as in *Hanseniaspora* and its sister genus *Saccharomycodes* (both in the family Saccharomycodaceae) [45][74]. These three lineages are distantly related to one another on the budding yeast phylogeny [34], so bipolar budding likely evolved three times independently in Saccharomycotina, including in the last common ancestor of *Hanseniaspora* and *Saccharomycodes*. Currently, there is only one genome available for *Saccharomycodes* [74], making robust inferences of ancestral states challenging. Interestingly, examination of cell cycle gene presence and absence in the only representative genome from the genus, *Saccharomycodes ludwigii* [74], reveals that *CDC26*, *PCL1*, *PDS1*, *RFX1*, *SIC1*, *SPO12*, and *WHI5* are absent (File S6), most of which are either absent from all *Hanseniaspora* (i.e., *CDC26*, *RFX1*, *SPO12*, and *WHI5*) or just from the FEL (i.e., *PDS1* and *SIC1*). This evidence raises the hypothesis that bipolar budding is linked to the dysregulation of cell cycle processes due to the absence of cell cycle genes and in particular cell cycle checkpoints (Fig 3).

Some gene losses may be compensatory. Deletion of many of the genes associated with DNA maintenance that have been lost in *Hanseniaspora* lead to dramatic increases of mutation rates and gross genome instability [12,13,20], raising the question of how these gene losses were tolerated in the first place. Examination of the functions of the genes lost in *Hanseniaspora* suggests that at least some of these gene losses may have been compensatory. For example, *POL4* knock-out strains of *S. cerevisiae* can be rescued by the deletion of *YKU70* [75], both of which were lost in the FEL. Similarly, the loss of genes responsible for key cell cycle functions (e.g., kinetochore functionality and chromosome segregation) appears to have co-occurred with the loss of checkpoint genes responsible for delaying the cell cycle if its functions fail to complete, which may have allowed *Hanseniaspora* cells to bypass otherwise detrimental cell cycle arrest. Specifically, *MAD1* and *MAD2*, which help delay anaphase when kinetochores are unattached [14]; the 10-gene DASH complex, which participates in spindle attachment, stability, and chromosome segregation [76]; and the 4-gene MIND complex, which is required for kinetochore bi-orientation and accurate chromosome segregation [77], were all lost in the FEL.

Long-term hypermutation and the subsequent slowing of sequence evolution. Estimates of ω suggest the FEL and SEL, albeit to a much lower degree in the latter, underwent a burst of accelerated sequence evolution in their stem lineages, followed by a reduction in the pace of sequence evolution (Fig 5). This pattern is consistent with theoretical predictions that selection against mutator phenotypes will reduce the overall rate of sequence evolution [27], as well as with evidence from experimental evolution of hypermutator lines of *S. cerevisiae* that showed that their mutation rates were quickly reduced [32]. Although we do not know the catalyst for this burst of sequence evolution, hypermutators may be favored in maladapted populations or in conditions where environmental parameters frequently change [27,32]. Although the environment occupied by the *Hanseniaspora* last common ancestor is unknown, it is plausible that environmental instability or other stressors favored hypermutators in *Hanseniaspora*. Extant *Hanseniaspora* species are well known to be associated with the grape environment [39,78,79]. Interestingly, grapes appear to have originated [80] around the same time window that *Hanseniaspora* did (Fig 1B), leading us to speculate that the evolutionary trigger of *Hanseniaspora* hypermutation could have been adaptation to the grape environment.

Losses of DNA repair genes are reflected in patterns of sequence evolution. Although the relationship between genotype and phenotype is complex, the loss of genes involved in DNA repair can have predictable outcomes on patterns of sequence evolution in genomes. In the case of the observed losses of DNA repair genes in *Hanseniaspora*, the mutational signatures of this loss and the consequent hypermutation can be both general (i.e., the sum total of many gene losses), as well as specific (i.e., can be putatively linked to the losses of specific genes or pathways). Arguably the most notable general mutational signature is that *Hanseniaspora* genome sequence evolution is largely driven by random (i.e., neutral) mutagenic processes with a strong AT-bias. For example, whereas the transition / transversion ratios of eukaryotic genomes are typically within the 1.7 and 4 range [56–59], *Hanseniaspora* ratios are ~0.66-0.75 (Fig 6C), which are values on par with estimates of transition / transversion caused by neutral mutations alone (e.g., 0.6-0.95 in *S. cerevisiae* [56,81], 0.92 in *E. coli* [82], 0.98 in *Drosophila melanogaster* [83], and 1.70 in humans [84]). Similarly, base substitutions across *Hanseniaspora* genomes are strongly AT-biased*,* especially in the FEL (Fig 6), an observation consistent with the general AT-bias of mutations observed in diverse organisms, including numerous bacteria [85], the fruit fly [83], *S. cerevisiae* [56], and humans [84].

In addition to these general mutational signatures, examination of *Hanseniaspora* sequence evolution also reveals mutational signatures that can be linked to the loss of specific DNA repair genes. For example, we found a higher proportion of base substitutions associated with the most abundant oxidatively damaged base, 8-oxo-dG, which causes G → T or C → A transversions [26], in the FEL compared to the SEL, which reflects specific gene losses. Specifically, *Hanseniaspora* yeasts have lost *PCD1*, which encodes a diphosphatase that contributes to the removal of 8-oxo-dGTP [24] and thereby reduces the chance of misincorporating this damaged base. Once 8-oxo-dG damage has occurred, it is primarily repaired by the base excision repair pathway [26]. Notably, the FEL is missing a key component of the base excision repair pathway, a DNA polymerase δ subunit, encoded by *POL32*, which aids in filling the gap after excision [86]. Accordingly, the proportion of G|C sites with substitutions indicative of 8-oxo-dG damage (i.e., G → T or C → A transversions) is significantly greater in the FEL compared to the SEL (Fig 5H). Similarly, the numbers of dinucleotide substitutions of CC → TT associated with UV-induced pyrimidine dimers [87] are higher across *Hanseniaspora* compared to other yeasts due to the loss of *PHR1*, which encodes a DNA photolyase that repairs pyrimidine dimers (Fig 5I) [23].

Our analyses provide the first major effort to characterize the genome function and evolution of the enigmatic genus *Hanseniaspora* and identify major and extensive losses of genes associated with metabolism, cell cycle, and DNA repair processes. These extensive losses and the concomitant acceleration of evolutionary rate mean that levels of amino acid sequence divergence within each of the two *Hanseniaspora* lineages alone, but especially within the FEL, are similar to those observed within plant classes and animal subphyla (Fig S11). These discoveries set the stage for further fundamental molecular and evolutionary investigations among *Hanseniaspora*, such as potential novel rewiring of cell cycle and DNA repair processes.

Methods

DNA sequencing For each species, genomic DNA (gDNA) was isolated using a two-step phenol:chloroform extraction previously described to remove additional proteins from the gDNA [34]. The gDNA was sonicated and ligated to Illumina sequencing adaptors as previously described [88], and the libraries were submitted for paired-end sequencing (2 x 250) on an Illumina HiSeq 2500 instrument.

Phenotyping We qualitatively measured growth of species on five carbon sources (maltose, raffinose, sucrose, melezitose, and galactose) as previously described in [34]. We used a minimal media base with ammonium sulfate and all carbon sources were at a 2% concentration. Yeast were initially grown in YPD and transferred to carbon treatments. Species were visually scored for growth for about a week on each carbon source in three independent replicates over multiple days. A species was considered to utilize a carbon source if it showed growth across ≥ 50% of biological replicates. Growth data for *Hanseniaspora gamundiae* were obtained from Čadež et al., 2019.

Genome assembly and annotation To generate *de novo* genome assemblies, we used paired-end DNA sequence reads as input to iWGS, version 1.1 [89], a pipeline which uses multiple assemblers and identifies the "best" assembly according to largest genome size and N50 (i.e., the shortest contig length among the set of the longest contigs that account for 50% of the genome assembly's length) [90] as described in [34]. More specifically, sequenced reads were first quality-trimmed, and adapter sequences were removed used TRIMMOMATIC, version 0.33 [91], and LIGHTER, version 1.1.1 [92]. Subsequently, KMERGENIE, version 1.6982 [93], was used to determine the optimal *k*-mer length for each genome individually. Thereafter, six *de novo* assembly tools (i.e., ABYSS, version 1.5.2 [94]; DISCOVAR, release 51885 [95]; MASURCA, version 2.3.2 [96]; SGA, version 0.10.13 [97]; SOAPDENOVO2, version 2.04 [98]; and SPADES, version 3.7.0 [99]) were used to generate genome assemblies from the processed reads. Using QUAST, version 4.4 [100], the best assembly was chosen according to the assembly that provided the largest genome size and best N50.

Annotations for eight of the *Hanseniaspora* genomes (i.e., *H. clermontiae*, *H. osmophila* CBS 313, *H. pseudoguilliermondii*, *H. singularis*, *H. uvarum* DSM2768, *H. valbyensis*, *H. vineae* T02 19AF, and *K. hatyaiensis*) and the four outgroup species (i.e., *Cy. jadinii*, *K. marxianus*, *S. cerevisiae*, and *W. anomalus*) were generated in a recent comparative genomic study of the budding yeast subphylum [34]. The other 11 *Hanseniaspora* genomes examined here were annotated by following the same protocol as in [34].

In brief, the genomes were annotated using the MAKER pipeline, version 2.31.8 [101]. The homology evidence used for MAKER consists of fungal protein sequences in the SwissProt database (release 2016_11) and annotated protein sequences of select yeast species from MYCOCOSM [102], a web portal developed by the US Department of Energy Joint Genome Institute for fungal genomic analyses. Three *ab initio* gene predictors were used with the MAKER pipeline, including GENEMARK-ES, version 4.32 [103]; SNAP, version 2013-11-29 [104]; and AUGUSTUS, version 3.2.2 [105], each of which was trained for each individual genome. GENEMARK-ES was self-trained on the repeat-masked genome sequence with the fungal-specific option ("–fugus"), while SNAP and AUGUSTUS were trained through three iterative MAKER runs. Once all three *ab initio* predictors were trained, they were used together with homology evidence to conduct a final MAKER analysis in which all gene models were reported ("keep_preds" set to 1), and these comprise the final set of annotations for the genome.

Data acquisition All publicly available *Hanseniaspora* genomes, including multiple strains from a single species, were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/ File S1).

These species and strains include *H. guilliermondii* UTAD222 [78], *H. opuntiae* AWRI3578, *H. osmophila* AWRI3579, *H. uvarum* AWRI3580 [106], *H. uvarum* 34-9, *H. vineae* T02-19AF [107], *H. valbyensis* NRRL Y-1626 [33], and *H. gamundiae* [41]. We also included *Saccharomyces cerevisiae* S288C, *Kluyveromyces marxianus* DMKU3-1042*, Wickerhamomyces anomalus* NRRL Y-366-8*,* and *Cyberlindnera jadinii* NRRL Y-1542, four representative budding yeast species that are all outside the genus *Hanseniaspora* [34], which we used as outgroups. Together with publicly available genomes, our sampling of *Hanseniaspora* encompasses all known species in the genus (or its anamorphic counterpart, *Kloeckera*), except *Hanseniaspora lindneri*, which likely belongs to the FEL based on a four-locus phylogenetic study [108], and *Hanseniaspora taiwanica*, which likely belongs to the SEL based on neighbor-joining analyses of the LSU rRNA gene sequence [109].

Assembly assessment and identification of orthologs To determine genome assembly completeness, we calculated contig N50 [90] and assessed gene content completeness using multiple databases of curated orthologs from BUSCO, version 3 [110]. More specifically, we determined gene content completeness using orthologous sets of genes constructed from sets of genomes representing multiple taxonomic levels, including Eukaryota (superkingdom; 100 species; 303 BUSCOs), Fungi (kingdom; 85 species; 290 BUSCOs), Dikarya (subkingdom; 75 species; 1,312 BUSCOs), Ascomycota (phylum; 75 species; 1,315 BUSCOs), Saccharomyceta (no rank; 70 species; 1,759 BUSCOs), and Saccharomycetales (order; 30 species; 1,711 BUSCOs).

Genomes sequenced in the present project were sequenced at an average depth of 63.49 ± 52.57 (File S1). Among all *Hanseniaspora*, the average scaffold N50 was 269.03 ± 385.28 kb, the average total number of scaffolds was 980.36 ± 835.20 (398.32 ± 397.97 when imposing a 1kb scaffold filter), and the average genome assembly size was 10.13 ± 1.38 Mb (9.93 ± 1.35 Mb when imposing a 1kb scaffold filter). Notably, the genome assemblies and gene annotations created in the present project were comparable to publicly available ones. For example, the genome size of publicly available *Hanseniaspora vineae* T02 19AF is 11.38 Mb with 4,661 genes, while our assembly of *Hanseniaspora vineae* NRRL Y-1626 was 11.15 Mb with 5,193 genes.

We found that our assemblies were of comparable quality to those from publicly available genomes. For example, *Hanseniaspora uvarum* NRRL Y-1614 (N50 = 267.64 kb; genome size = 8.82 Mb; number of scaffolds = 258; gene number = 4,227), which was sequenced in the present study, and *H. uvarum* AWRI3580 (N50 = 1,289.09 kb; genome size = 8.81 Mb; number of scaffolds = 18; gene number = 4,061), which is publicly available [106] had similar single-copy BUSCO genes present in the highest and lowest ORTHODB [43] taxonomic ranks (Eukaryota and Saccharomycetales, respectively). Specifically, *H. uvarum* NRRL Y-1614 and *H. uvarum* AWRI3580 had 80.20% (243 / 303) and 79.87% (242 / 303) of universally single-copy orthologs in Eukaryota present in each genome respectively, and 52.31% (895 / 1,711) and 51.49% (881 / 1,711) of universally single-copy orthologs in Saccharomycetales present in each genome, respectively.

To identify single-copy orthologous genes (OGs) among all protein coding sequences for all 29 taxa, we used ORTHOMCL, version 1.4 [111]. ORTHOMCL clusters genes into OGs using a Markov clustering algorithm (van Dongen, 2000; https://micans.org/mcl/) from gene similarity information acquired from a blastp 'all-vs-all' using NCBI's BLAST+, version 2.3.0 (Fig S2; Madden, 2013) and the proteomes of species of interest as input. The key parameters used in blastp 'all-vs-all' were: e-value = $1e^{-10}$, percent identity cut-off = 30%, percent match cutoff = 70%, and a maximum weight value = 180. To conservatively identify OGs, we used a strict ORTHOMCL inflation parameter of 4.

To identify additional OGs suitable for use in phylogenomic and molecular sequence analyses, we identified the single best putatively orthologous gene from OGs with full species representation and a maximum of two species with multiple copies using PHYLOTREEPRUNER, version 1.0 [114]. To do so, we first aligned and trimmed sequences in 1,143 OGs out a total of 11,877 that fit the criterion of full representation and a maximum of two species with duplicate sequences. More specifically, we used MAFFT, version 7.294b [115], with the BLOSUM62 matrix of substitutions [116], a gap penalty of 1.0, 1,000 maximum iterations, the 'genafpair' parameter, and TRIMAL, version 1.4 [117], with the 'automated1' parameter to align and trim individual sequences, respectively. The resulting OG multiple sequence alignments were then used to infer gene phylogenies using FASTTREE, version 2.1.9 [118], with 4 and 2 rounds of subtree-prune-regraft and optimization of all 5 branches at nearest-neighbor interchanges, respectively, as well as the 'slownni' parameter to refine the inferred topology. Internal branches with support lower than 0.9 Shimodaira-Hasegawa-like support implemented in FASTTREE [118] were collapsed using PHYLOTREEPRUNER, version 1.0 [114], and the longest sequence for species with multiple sequences per OG were retained, resulting a robust set of OGs with every taxon being represented by a single sequence. OGs were realigned (MAFFT) and trimmed (TRIMAL) using the same parameters as above.

Phylogenomic analyses To infer the *Hanseniaspora* phylogeny, we performed phylogenetic inference using maximum likelihood [119] with concatenation [120,121] and coalescence [122] approaches. To determine the best-fit phylogenetic model for concatenation and generate single-gene trees for coalescence, we constructed trees per single-copy OG using RAXML, version 8.2.8. [123], where each topology was determined using 5 starting trees. Single-gene trees that did not recover all outgroup species as the earliest diverging taxa when serially rooted on outgroup taxa were discarded. Individual OG alignments or trees were used for species tree estimation with RAXML (i.e., concatenation) using the LG [124] model of substitution, which is the most commonly supported model of substitution (874 / 1,034; 84.53% genes), or ASTRAL-II, version 4.10.12 (i.e., coalescence; Mirarab and Warnow, 2015). Branch support for the concatenation and coalescence phylogenies was determined using 100 rapid bootstrap replicates [126] and local posterior support [122], respectively.

Several previous phylogenomic studies have shown that the internal branches preceding the *Hanseniaspora* FEL and SEL are long [33,35]. To examine whether the relationship between the length of the internal branch preceding the FEL and the length of the internal branch preceding the SEL was consistent across genes in our phylogeny, we used NEWICK UTILITIES, version 1.6 [127] to remove the 88 single-gene trees where either lineage was not recovered as monophyletic and calculated their difference for the remaining 946 genes.

Estimating divergence times To estimate divergence times among the 25 *Hanseniaspora* genomes, we used the Bayesian method MCMCTree in the PAML, version 4.9 [128], and the concatenated 1,034-gene matrix. The input tree was derived from the concatenation-based ML analysis under a single LG+G4 [124] model (Figure 1A). The in-group root (i.e., the split between the FEL and SEL) age was set between 0.756 and 1.177 time units (1 time unit = 100 million years ago [mya]), which was adopted from a recent study [34].

To infer the *Hanseniaspora* timetree, we first estimated branch lengths under a single LG+G4 [124] model with codeml in the PAML, version 4.9 [128], package and obtained a rough mean of the overall mutation rate. Next, we applied the approximate likelihood method [129,130] to estimate the gradient vector and Hessian matrix with Taylor expansion (option usedata = 3). Last, we assigned (a) the gamma-Dirichlet prior for the overall substitution rate (option rgene_gamma) as G(1, 1.55), with a mean of 0.64, (b) the gamma-Dirichlet prior for the rate-drift parameter (option sigma2 gamma) as G(1, 10), and (c) the parameters for the birth-death sampling process with birth and death rates λ=μ=1 and sampling fraction ρ=0. We employed the independent-rate model (option clock=2) to account for the rate variation across different lineages and used soft bounds (left and right tail probabilities equal 0.025) to set minimum and maximum values for the in-group root mentioned above. The MCMC run was first run for 1,000,000 iterations as burn-in and then sampled every 1,000 iterations until a total of 30,000 samples was collected. Two separate MCMC runs were compared for convergence, and similar results were observed.

Gene presence and absence analysis To determine the presence and absence of genes in *Hanseniaspora* genomes, we built hidden Markov models (HMMs) for each gene present in *Saccharomyces cerevisiae* and used the resulting HMM profile to search for the corresponding homolog in each *Hanseniaspora* genome, as well as outgroup taxa. More specifically, for each of the 5,917 verified open reading frames from *S. cerevisiae* [131] (downloaded Oct 2018 from the *Saccharomyces* genome database), we searched for putative homologs in NCBI's Reference Sequence Database for Fungi (downloaded June 2018) using NCBI's BLAST+, version 2.3.0 [113], blastp function, and an e-value cut-off of $1e^{-3}$ as recommended for homology searches [132]. We used the top 100 hits for the gene of interest and aligned them using MAFFT, version 7.294b [115], with the same parameters described above. The resulting gene alignment was then used to create an HMM profile for the gene using the hmmbuild function in HMMER, version 3.1b2 [133]. The resulting HMM profile was then used to search for each individual gene in each *Hanseniaspora* genome and outgroup taxa using the hmmsearch function with an expectation value cutoff of 0.01 and a score cutoff of 50. This analysis was done for the 5,735 genes with multiple blast hits allowing for the creation of a HMM profile. To evaluate the validity of constructed HMMs, we examined their ability to recall genes in *S. cerevisiae* and found that we recovered all nuclear genes. Altogether, our ability to recall 99.63% of genes demonstrates the validity of our pipeline for the vast majority of genes and for nuclear genes in particular.

To determine if any functional categories were over- or under-represented among genes present or absent among *Hanseniaspora* species, we conducted gene ontology (GO) [134] enrichment analyses using GOATOOLS, version 0.7.9 [135]. We used a background of all *S. cerevisiae* genes and a *p*-value cut-off of 0.05 after multiple-test correction using the Holm method [136].

Plotting gene presence and absence among pathways was done by examining depicted pathways available through the KEGG project [137] and the *Saccharomyces* Genome Database [131].

We examined the validity of the gene presence and absence pipeline by examining under-represented terms and the presence or absence of essential genes in *S. cerevisiae* [138]. We hypothesized that under-represented GO terms will be associated with basic molecular processes and that essential genes will be under-represented among the set of absent genes. In agreement with these expectations, GO terms associated with basic biological processes and essential *S. cerevisiae* genes are under-represented among genes that are absent across *Hanseniaspora* genomes. For example, among all genes absent in the FEL and SEL, the molecular functions BASE PAIRING, GO:0000496 ($p < 0.001$); GTP BINDING, GO:0005525 ($p < 0.001$); and ATPASE ACTIVITY, COUPLED TO MOVEMENT OF SUBSTANCES, GO:0043492 ($p < 0.001$), are significantly under-represented (File S4). Similarly, *S. cerevisiae* essential genes are significantly under-represented ($p < 0.001$; Fischer's exact test for both lineages) among lost genes with only 3 and 2 *S. cerevisiae* essential genes having been lost from the FEL and SEL genomes, respectively.

Ploidy estimation To determine ploidy, we leveraged base frequency distributions at variable sites, which we generated by mapping each genome's reads to its assembly. To ensure high-quality read mapping, we first quality-trimmed reads suing TRIMMOMATIC, version 0.36 [91], using the parameters leading:10, trailing:10, slidingwindow:4:20, and minlen:50. Reads were subsequently mapped to their respective genome using BOWTIE2, version 1.1.2 [139], with the "sensitive" parameter and converted the resulting file to a sorted bam format using SAMTOOLS, version 1.3.1 [140]. We next used NQUIRE [141], which extracts base frequency information at segregating sites with a minimum frequency of 0.2. Prior to visualization, we removed background noise by utilizing the Gaussian Mixture Model with Uniform noise component [141].

Molecular evolution and mutation analysis *Molecular sequence rate analysis along the phylogeny.* To determine the rate of sequence evolution over the course of *Hanseniaspora* evolution, we examined variation in the rate of nonsynonymous (dN) to the rate of synonymous (dS) substitutions (dN/dS or ω) across the species phylogeny. We first obtained codon-based alignments of the protein sequences used during phylogenomic inference by threading nucleotides on top of the amino acid sequence using PAL2NAL, version 14 [142], and calculated ω values under the different hypotheses using the CODEML module in PAML, version 4.9 [128]. For each gene tested, we set the null hypothesis (H_o) where all internal branches exhibit the same ω (model = 0) and compared it to four different alternative hypotheses. Under the $H_{FE\text{-}SE\ branch}$ hypothesis, the branches immediately preceding the FEL and SEL were assumed to exhibit distinct ω values from the background (model = 2) (Fig 5Bi). Under the H_{FE} hypothesis, the branch immediately preceding the FEL was assumed to have a distinct ω value, all FEL crown branches were assumed to have their own collective ω value, and all background branches were assumed to have their own collective ω value (model = 2) (Fig 5Ci). The H_{SE} hypothesis assumed the branch preceding the lineage had its own ω value, all SEL crown branches had their own collective ω value, and all background branches were assumed to have their own collective ω value (model = 2) (Fig 5Di). Lastly, the $H_{FE\text{-}SE\ crown}$ hypothesis assumed that all FEL crown branches had their own collective ω value, all SEL crown branches had their own collective ω value, and the rest of the branches were assumed to have their own collective ω value (model = 2) (Fig 5Ei). To determine if each of the alternative hypotheses was significantly different from the null hypothesis, we used the likelihood ratio test (LRT) ($\alpha = 0.01$). A few genes could not be analyzed due to fatal interruptions or errors during use in PAML, version 4.9 [128], which have been reported by other users [143]; these genes were removed from the analysis. Thus, this analysis was conducted for 989 genes for three tests ($H_{\text{FE-SE branch}}$, H_{FE}, and H_{SE} hypotheses) and 983 genes for one test ($H_{\text{FE-SE crown}}$ hypothesis).

Examination of mutational signatures To conservatively identify base substitutions, insertions, and deletions found in taxa in the FEL or SEL, we examined the status of each nucleotide at each position in codon-based and amino acid-based OG alignments. We examined base substitutions, insertions, and deletions at sites that are conserved in the outgroup (i.e., all outgroup taxa have the same character state for a given position in an alignment). For base substitutions, we determined if the nucleotide or amino acid residue in a given *Hanseniaspora* species differed from the conserved outgroup nucleotide or amino acid residue at the same position. To measure if amino acid substitutions in each lineage were conservative or radical (i.e., a substitution to a similar amino acid residue versus a substitution to an amino acid residue with different properties), we used Sneath's index of dissimilarity, which considers 134 categories of biological activity and chemical change to quantify dissimilarity of amino acid substitutions, and Epstein's coefficient of difference, which considers differences in polarity and size of amino acids to quantify dissimilarity. Notably, Sneath's index is symmetric (i.e., isoleucine to leucine is equivalent to leucine to isoleucine), whereas Epstein's coefficient is not (i.e., isoleucine to leucine is not equivalent to leucine to isoleucine). For indels, we used a sliding window approach with a step size of one nucleotide. We considered positions where a nucleotide was present in all outgroup taxa but a gap was present in *Hanseniaspora* as deletions, and positions where a gap was present in all outgroup taxa and a nucleotide was present in *Hanseniaspora* species as insertions. Analyses were conducted using custom PYTHON, version 3.5.2 (https://www.python.org/), scripts, which use the BIOPYTHON, version 1.70 [144], and NUMPY, version 1.13.1 [145], modules.

We discovered that all *Hanseniaspora* species lack the *PHR1* gene, which is associated with the repair of UV radiation damage. UV exposure induces high levels of CC → TT dinucleotide substitutions [87]. If *Hanseniaspora* have a reduced capacity to repair UV radiation damage, they would be expected to contain fewer CC|GG dinucleotides and more TT|AA ones. To test whether this was the case, we created a CC or GG (hereby denoted as CC|GG) score, which was calculated using the following formula:

$$CC|GG\ score = \frac{CC|GG}{D} \times \frac{1}{G|C} \quad \text{where } D = \frac{GS}{2}$$

where CC|GG is the number of observed CC or GG dinucleotides in a genome, D is the number of dinucleotides in the genome, GS is the genome size, and G|C is GC-content. Similarly, we created a TT|AA score calculated the following formula:

$$TT|AA\ score = \frac{TT|AA}{D} \times \frac{1}{A|T} \quad \text{where } D = \frac{GS}{2}$$

where TT|AA is the number of TT or AA dinucleotides in a genome, D is the number of dinucleotides in the genome, GS is the genome size, and A|T is AT-content.

Data Availability

Data matrices, species-level and single-gene phylogenies, dN/dS results, and HMMs will be made available through the figshare repository upon publication.

Acknowledgements

We thank members of the Rokas and Hittinger laboratories for helpful suggestions and discussion.

Financial Statement

This work was supported in part by the National Science Foundation (https://www.nsf.gov) (DEB-1442113 to AR, DEB-1442148 to CTH and CPK, and DGE-1445197 to NPB) and the DOE Great Lakes Bioenergy Research Center (https://www.glbrc.org) (DOE Office of Science BER DE-FC02-07ER64494 and DE-SC0018409 to Timothy J. Donohue). CTH is a Pew Scholar in the Biomedical Sciences and Vilas Faculty Early Career Investigator, supported by the Pew Charitable Trusts (https://www.pewtrusts.org) and Vilas Trust Estate (https://www.rsp.wisc.edu/Vilas), respectively. AR is supported by a Guggenheim fellowship (https://www.gf.org/about/fellowship), JLS by Vanderbilt's Biological Sciences graduate program (https://as.vanderbilt.edu/biosci), and XZ in part by the National Key Project for Basic Research of China (http://www.most.gov.cn) (973 Program, No. 2015CB150600). NC was supported by funding from the Slovenian Research Agency (https://www.arrs.gov.si) (P4-0116). DL was supported by CONICET (https://www.conicet.gov.ar) (PIP 392), FONCyT (https://www.argentina.gob.ar/ciencia/agencia/fondo-para-la-investigacion-cientifica-y-tecnologica-foncyt) (PICT 2542), and Universidad Nacional del Comahue (https://www.uncoma.edu.ar) (B199). This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University (https://www.vanderbilt.edu/accre), the Center for High-Throughput Computing at UW-Madison (http://chtc.cs.wisc.edu), and the UW Biotechnology Center DNA Sequencing Facility (https://www.biotech.wisc.edu/services/dnaseq). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture (https://www.usda.gov/). USDA is an equal opportunity provider and employer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Lindahl T. Quality Control by DNA Repair. Science (80-). 1999;286: 1897–1905. doi:10.1126/science.286.5446.1897

2. Hakem R. DNA-damage repair; the good, the bad, and the ugly. EMBO J. 2008;27: 589–605. doi:10.1038/emboj.2008.15

3. Broustas CG, Lieberman HB. DNA Damage Response Genes and the Development of Cancer Metastasis. Radiat Res. 2014;181: 111–130. doi:10.1667/RR13515.1

4. Pal C, Maciá MD, Oliver A, Schachar I, Buckling A. Coevolution with viruses drives the evolution of bacterial mutation rates. Nature. 2007;450: 1079–1081. doi:10.1038/nature06350

5. Myung K, Datta A, Kolodner RD. Suppression of Spontaneous Chromosomal Rearrangements by S Phase Checkpoint Functions in Saccharomyces cerevisiae. Cell. 2001;104: 397–408. doi:10.1016/S0092-8674(01)00227-6

6. Billmyre RB, Clancey SA, Heitman J. Natural mismatch repair mutations mediate phenotypic diversity and drug resistance in Cryptococcus deuterogattii. Elife. 2017;6. doi:10.7554/eLife.28802

7. Boyce KJ, Wang Y, Verma S, Shakya VPS, Xue C, Idnurm A. Mismatch Repair of DNA Replication Errors Contributes to Microevolution in the Pathogenic Fungus Cryptococcus neoformans. Alspaugh JA, editor. MBio. 2017;8. doi:10.1128/mBio.00595-17

8. Rhodes J, Beale MA, Vanhove M, Jarvis JN, Kannambath S, Simpson JA, et al. A Population Genomics Approach to Assessing the Genetic Basis of Within-Host Microevolution Underlying Recurrent Cryptococcal Meningitis Infection. G3 Genes|Genomes|Genetics. 2017;7: 1165–1176. doi:10.1534/g3.116.037499

9. Barnum KJ, O'Connell MJ. Cell Cycle Regulation by Checkpoints. 2014. pp. 29–40. doi:10.1007/978-1-4939-0888-2_2

10. Cross FR, Buchler NE, Skotheim JM. Evolution of networks and sequences in eukaryotic cell cycle control. Philos Trans R Soc Lond B Biol Sci. 2011;366: 3532–44. doi:10.1098/rstb.2011.0078

11. Medina EM, Turner JJ, Gordân R, Skotheim JM, Buchler NE. Punctuated evolution and transitional hybrid network in an ancestral cell cycle of fungi. Elife. 2016;5. doi:10.7554/eLife.09492

12. Costanzo M, Nishikawa JL, Tang X, Millman JS, Schub O, Breitkreuz K, et al. CDK Activity Antagonizes Whi5, an Inhibitor of G1/S Transcription in Yeast. Cell. 2004;117: 899–913. doi:10.1016/j.cell.2004.05.024

13. Castro A, Bernis C, Vigneron S, Labbé J-C, Lorca T. The anaphase-promoting complex: a key factor in the regulation of cell cycle. Oncogene. 2005;24: 314–325. doi:10.1038/sj.onc.1207973

14. Heinrich S, Sewart K, Windecker H, Langegger M, Schmidt N, Hustedt N, et al. Mad1 contribution to spindle assembly checkpoint signalling goes beyond presenting Mad2 at kinetochores. EMBO Rep. 2014;15: 291–298. doi:10.1002/embr.201338114

15. Hendler A, Medina EM, Kishkevich A, Abu-Qarn M, Klier S, Buchler NE, et al. Gene duplication and co-evolution of G1/S transcription factor specificity in fungi are essential for optimizing cell fitness. Snyder M, editor. PLOS Genet. 2017;13: e1006778. doi:10.1371/journal.pgen.1006778

16. Zhou B-BS, Elledge SJ. The DNA damage response: putting checkpoints in perspective. Nature. 2000;408: 433–439. doi:10.1038/35044005

17. Weinert TA, Kiser GL, Hartwell LH. Mitotic checkpoint genes in budding yeast and the dependence of mitosis on DNA replication and repair. Genes Dev. 1994;8: 652–665. doi:10.1101/gad.8.6.652

18. Serero A, Jubin C, Loeillet S, Legoix-Né P, Nicolas AG. Mutational landscape of yeast mutator strains. Proc Natl Acad Sci. 2014;111: 1897–1902. doi:10.1073/pnas.1314423111

19. Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R, et al. Comprehensive Analysis of Hypermutation in Human Cancer. Cell. 2017;171: 1042–1056.e10. doi:10.1016/j.cell.2017.09.048

20. Huang M-E, de Calignon A, Nicolas A, Galibert F. POL32 , a subunit of the Saccharomyces cerevisiae DNA polymerase δ, defines a link between DNA replication and the mutagenic bypass repair pathway. Curr Genet. 2000;38: 178–187. doi:10.1007/s002940000149

21. Xiao W, Chow BL, Hanna M, Doetsch PW. Deletion of the MAG1 DNA glycosylase gene suppresses alkylation-induced killing and mutagenesis in yeast cells lacking AP endonucleases. Mutat Res - DNA Repair. 2001; doi:10.1016/S0921-8777(01)00113-6

22. Sebastian J, Sancar GB. A damage-responsive DNA binding protein regulates transcription of the yeast DNA repair gene PHR1. Proc Natl Acad Sci. 1991;88: 11251–11255. doi:10.1073/pnas.88.24.11251

23. Sebastian J, Kraus B, Sancar GB. Expression of the yeast PHR1 gene is induced by DNA-damaging agents. Mol Cell Biol. 1990;10: 4630–7.

24. Nunoshiba T. A novel Nudix hydrolase for oxidized purine nucleoside triphosphates encoded by ORFYLR151c (PCD1 gene) in Saccharomyces cerevisiae. Nucleic Acids Res. 2004;32: 5339–5348. doi:10.1093/nar/gkh868

25. Cartwright JL, Gasmi L, Spiller DG, McLennan AG. The Saccharomyces cerevisiae PCD1 gene encodes a peroxisomal nudix hydrolase active toward coenzyme A and its derivatives. J Biol Chem. 2000; doi:10.1074/jbc.M005015200

26. De Bont R. Endogenous DNA damage in humans: a review of quantitative data. Mutagenesis. 2004;19: 169–185. doi:10.1093/mutage/geh025

27. Ram Y, Hadany L. The evolution of stress-induced hypermutation in asexual populations. Evolution (N Y). 2012;66: 2315–2328. doi:10.1111/j.1558-5646.2012.01576.x

28. Oliver A. High Frequency of Hypermutable Pseudomonas aeruginosa in Cystic Fibrosis Lung Infection. Science (80-). 2000;288: 1251–1253. doi:10.1126/science.288.5469.1251

29. Giraud A, Matic I, Tenaillon O, Clara A, Radman M, Fons M, et al. Costs and Benefits of High Mutation Rates: Adaptive Evolution of Bacteria in the Mouse Gut. Science (80-). 2001;291: 2606–2608. doi:10.1126/science.1056421

30. Taddei F, Radman M, Maynard-Smith J, Toupance B, Gouyon PH, Godelle B. Role of mutator alleles in adaptive evolution. Nature. 1997;387: 700–702. doi:10.1038/42696

31. Sniegowski PD, Gerrish PJ, Lenski RE. Evolution of high mutation rates in experimental populations of E. coli. Nature. 1997;387: 703–705. doi:10.1038/42701

32. McDonald MJ, Hsieh Y-Y, Yu Y-H, Chang S-L, Leu J-Y. The Evolution of Low Mutation Rates in Experimental Mutator Populations of Saccharomyces cerevisiae. Curr Biol. 2012;22: 1235–1240. doi:10.1016/j.cub.2012.04.056

33. Riley R, Haridas S, Wolfe KH, Lopes MR, Hittinger CT, Göker M, et al. Comparative genomics of biotechnologically important yeasts. Proc Natl Acad Sci. 2016;113: 9882–9887. doi:10.1073/pnas.1603941113

34. Shen X-X, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh K V., et al. Tempo and

Mode of Genome Evolution in the Budding Yeast Subphylum. Cell. 2018; doi:10.1016/j.cell.2018.10.023

35. Shen X-X, Zhou X, Kominek J, Kurtzman CP, Hittinger CT, Rokas A. Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. G3 Genes|Genomes|Genetics. Genetics Society of America; 2016;6: 3927–3939. doi:10.1534/g3.116.034744

36. Albertin W, Setati ME, Miot-Sertier C, Mostert TT, Colonna-Ceccaldi B, Coulon J, et al. Hanseniaspora uvarum from Winemaking Environments Show Spatial and Temporal Genetic Clustering. Front Microbiol. 2016;6. doi:10.3389/fmicb.2015.01569

37. Jordão A, Vilela A, Cosme F. From Sugar of Grape to Alcohol of Wine: Sensorial Impact of Alcohol in Wine. Beverages. 2015;1: 292–310. doi:10.3390/beverages1040292

38. Montero CM, Dodero MCR, Sanchez DAG, Barroso CG. Analysis of low molecular weight carbohydrates in food and beverages: A review. Chromatographia. 2004; doi:10.1365/s10337-003-0134-3

39. Martin V, Valera M, Medina K, Boido E, Carrau F. Oenological Impact of the Hanseniaspora/Kloeckera Yeast Genus on Wines—A Review. Fermentation. 2018;4: 76. doi:10.3390/fermentation4030076

40. Langenberg A-K, Bink FJ, Wolff L, Walter S, von Wallbrunn C, Grossmann M, et al. Glycolytic Functions Are Conserved in the Genome of the Wine Yeast Hanseniaspora uvarum, and Pyruvate Kinase Limits Its Capacity for Alcoholic Fermentation. Dudley EG, editor. Appl Environ Microbiol. 2017;83. doi:10.1128/AEM.01580-17

41. Čadež N, Bellora N, Ulloa R, Hittinger CT, Libkind D. Genomic content of a novel yeast species Hanseniaspora gamundiae sp. nov. from fungal stromata (Cyttaria) associated with

a unique fermented beverage in Andean Patagonia, Argentina. Yurkov AM, editor. PLoS One. 2019;14: e0210792. doi:10.1371/journal.pone.0210792

42. Zhou X, Shen X-X, Hittinger CT, Rokas A. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. Mol Biol Evol. 2018;35: 486–503. doi:10.1093/molbev/msx302

43. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva E V. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 2013;41: D358–D365. doi:10.1093/nar/gks1116

44. Opulente DA, Rollinson EJ, Bernick-Roehr C, Hulfachor AB, Rokas A, Kurtzman CP, et al. Factors driving metabolic diversity in the budding yeast subphylum. BMC Biol. 2018;16: 26. doi:10.1186/s12915-018-0498-3

45. Kurtzman CP, Fell JW. The Yeasts - A Taxonomic Study. 4th ed. Kurtzman CP, Fell JW, editors. Elsevier Science; 1998.

46. Charron MJ, Read E, Haut SR, Michels CA. Molecular evolution of the telomere-associated MAL loci of Saccharomyces. Genetics. 1989;122: 307–16.

47. Koschwanez JH, Foster KR, Murray AW. Sucrose Utilization in Budding Yeast as a Model for the Origin of Undifferentiated Multicellularity. Keller L, editor. PLoS Biol. 2011;9: e1001122. doi:10.1371/journal.pbio.1001122

48. Jorgensen P. Systematic Identification of Pathways That Couple Cell Growth and Division in Yeast. Science (80-). 2002;297: 395–400. doi:10.1126/science.1070850

49. Colman-Lerner A, Chin TE, Brent R. Yeast Cbk1 and Mob2 Activate Daughter-Specific Genetic Programs to Induce Asymmetric Cell Fates. Cell. 2001;107: 739–750. doi:10.1016/S0092-8674(01)00596-7

50. Lubelsky Y, Reuven N, Shaul Y. Autorepression of Rfx1 Gene Expression: Functional Conservation from Yeast to Humans in Response to DNA Replication Arrest. Mol Cell Biol. 2005;25: 10665–10673. doi:10.1128/MCB.25.23.10665-10673.2005

51. Galgoczy DJ, Toczyski DP. Checkpoint Adaptation Precedes Spontaneous and Damage-Induced Genomic Instability in Yeast. Mol Cell Biol. 2001;21: 1710–1718. doi:10.1128/MCB.21.5.1710-1718.2001

52. Kim IY, Kwon HY, Park KH, Kim DS. Anaphase-Promoting Complex 7 is a Prognostic Factor in Human Colorectal Cancer. Ann Coloproctol. 2017;33: 139–145. doi:10.3393/ac.2017.33.4.139

53. Xiao W, Chow BL. Synergism between yeast nucleotide and base excision repair pathways in the protection against DNA methylation damage. Curr Genet. 1998; doi:10.1007/s002940050313

54. Nitiss KC, Malik M, He X, White SW, Nitiss JL. Tyrosyl-DNA phosphodiesterase (Tdp1) participates in the repair of Top2-mediated DNA damage. Proc Natl Acad Sci. 2006;103: 8953–8958. doi:10.1073/pnas.0603455103

55. Lustig AJ. Cdc13 subcomplexes regulate multiple telomere functions. Nat Struct Biol. 2001;8: 297–9. doi:10.1038/86157

56. Zhu YO, Siegal ML, Hall DW, Petrov DA. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci. 2014;111: E2310–E2318. doi:10.1073/pnas.1323011111

57. Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong W, et al. The functional spectrum of low-frequency coding variation. Genome Biol. 2011;12: R84. doi:10.1186/gb-2011-12-9-r84

58. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics. 2015;31: 318–23. doi:10.1093/bioinformatics/btu668

59. Vignal A, Milan D, SanCristobal M, Eggen A. A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol. 2002;34: 275–305. doi:10.1051/gse:2002009

60. Sneath PH. Relations between chemical structure and biological activity in peptides. J Theor Biol. 1966;12: 157–95.

61. Epstein CJ. Non-randomness of Ammo-acid Changes in the Evolution of Homologous Proteins. Nature. 1967;215: 355–359. doi:10.1038/215355a0

62. Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet. 2016;17: 379–391. doi:10.1038/nrg.2016.39

63. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006;2. doi:10.1038/msb4100050

64. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature. 2002;418: 387–391. doi:10.1038/nature00935

65. Segal ES, Gritsenko V, Levitan A, Yadav B, Dror N, Steenwyk JL, et al. Gene Essentiality Analyzed by In Vivo Transposon Mutagenesis and Machine Learning in a Stable Haploid Isolate of Candida albicans. Di Pietro A, editor. MBio. 2018;9. doi:10.1128/mBio.02048-18

66. Koskiniemi S, Sun S, Berg OG, Andersson DI. Selection-Driven Gene Loss in Bacteria.

Casadesús J, editor. PLoS Genet. 2012;8: e1002787. doi:10.1371/journal.pgen.1002787

67. Keeling PJ, Slamovits CH. Simplicity and complexity of microsporidian genomes. Eukaryot Cell. 2004;3: 1363–9. doi:10.1128/EC.3.6.1363-1369.2004

68. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, et al. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature. 2001;414: 450–453. doi:10.1038/35106579

69. Chang ES, Neuhof M, Rubinstein ND, Diamant A, Philippe H, Huchon D, et al. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc Natl Acad Sci U S A. 2015;112: 14912–7. doi:10.1073/pnas.1511468112

70. Hittinger CT, Rokas A, Carroll SB. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. Proc Natl Acad Sci. 2004;101: 14144–14149. doi:10.1073/pnas.0404319101

71. Slot JC, Rokas A. Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. Proc Natl Acad Sci. 2010;107: 10136–10141. doi:10.1073/pnas.0914418107

72. Wolfe KH, Armisén D, Proux-Wera E, ÓhÉigeartaigh SS, Azam H, Gordon JL, et al. Clade- and species-specific features of genome evolution in the Saccharomycetaceae. Nielsen J, editor. FEMS Yeast Res. 2015;15: fov035. doi:10.1093/femsyr/fov035

73. Hartwell L. Defects in a cell cycle checkpoint may be responsible for the genomic instability of cancer cells. Cell. 1992. doi:10.1016/0092-8674(92)90586-2

74. Tavares MJ, Güldener U, Esteves M, Mendes-Faia A, Mendes-Ferreira A, Mira NP. Genome Sequence of the Wine Yeast Saccharomycodes ludwigii UTAD17. Cuomo CA, editor. Microbiol Resour Announc. 2018;7. doi:10.1128/MRA.01195-18

75. Sterling CH. DNA Polymerase 4 of Saccharomyces cerevisiae Is Important for Accurate Repair of Methyl-Methanesulfonate-Induced DNA Damage. Genetics. 2005;172: 89–98. doi:10.1534/genetics.105.049254

76. Jenni S, Harrison SC. Structure of the DASH/Dam1 complex shows its role at the yeast kinetochore-microtubule interface. Science (80-). 2018;360: 552–558. doi:10.1126/science.aar6436

77. Dimitrova YN, Jenni S, Valverde R, Khin Y, Harrison SC. Structure of the MIND Complex Defines a Regulatory Focus for Yeast Kinetochore Assembly. Cell. 2016;167: 1014–1027.e12. doi:10.1016/j.cell.2016.10.011

78. Seixas I, Barbosa C, Salazar SB, Mendes-Faia A, Wang Y, Güldener U, et al. Genome Sequence of the Nonconventional Wine Yeast Hanseniaspora guilliermondii UTAD222. Genome Announc. American Society for Microbiology; 2017;5: e01515-16. doi:10.1128/genomeA.01515-16

79. Chavan P, Mane S, Kulkarni G, Shaikh S, Ghormade V, Nerkar DP, et al. Natural yeast flora of different varieties of grapes used for wine making in India. Food Microbiol. 2009;26: 801–808. doi:10.1016/j.fm.2009.05.005

80. Wikstrom N, Savolainen V, Chase MW. Evolution of the angiosperms: calibrating the family tree. Proc R Soc B Biol Sci. 2001;268: 2211–2220. doi:10.1098/rspb.2001.1782

81. Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, et al. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci. 2008; doi:10.1073/pnas.0803466105

82. Lynch M. The Origins of Genome Architecture. Journal of Heredity. 2007. doi:10.1093/jhered/esm073

83. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. Genome Res. 2009;19: 1195–1201. doi:10.1101/gr.091231.109

84. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci. 2010;107: 961–968. doi:10.1073/pnas.0912629107

85. Hershberg R, Petrov DA. Evidence That Mutation Is Universally Biased towards AT in Bacteria. Nachman MW, editor. PLoS Genet. 2010;6: e1001115. doi:10.1371/journal.pgen.1001115

86. Seeberg E, Eide L, Bjørås M. The base excision repair pathway. Trends Biochem Sci. 1995;20: 391–397. doi:10.1016/S0968-0004(00)89086-6

87. Huang MN, Yu W, Teoh WW, Ardin M, Jusakul A, Ng AWT, et al. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. Genome Res. 2017;27: 1475–1486. doi:10.1101/gr.220038.116

88. Hittinger CT, Gonçalves P, Sampaio JP, Dover J, Johnston M, Rokas A. Remarkably ancient balanced polymorphisms in a multi-locus gene network. Nature. 2010;464: 54–58. doi:10.1038/nature08791

89. Zhou X, Peris D, Kominek J, Kurtzman CP, Hittinger CT, Rokas A. in silico Whole Genome Sequencer &amp; Analyzer (iWGS): A Computational Pipeline to Guide the Design and Analysis of de novo Genome Sequencing Studies. G3 Genes|Genomes|Genetics. 2016; doi:10.1534/g3.116.034249

90. Yandell M, Ence D. A beginner’s guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13: 329–42. doi:10.1038/nrg3174

91. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence

data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

92. Song L, Florea L, Langmead B. Lighter: fast and memory-efficient sequencing error correction without counting. Genome Biol. 2014;15: 509. doi:10.1186/s13059-014-0509-9

93. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;30: 31–37. doi:10.1093/bioinformatics/btt310

94. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: A parallel assembler for short read sequence data. Genome Res. 2009;19: 1117–1123. doi:10.1101/gr.089532.108

95. Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, et al. Comprehensive variation discovery in single human genomes. Nat Genet. 2014;46: 1350–1355. doi:10.1038/ng.3121

96. Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29: 2669–2677. doi:10.1093/bioinformatics/btt476

97. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012;22: 549–556. doi:10.1101/gr.126953.111

98. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1: 18. doi:10.1186/2047-217X-1-18

99. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19: 455–477. doi:10.1089/cmb.2012.0021

100. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for

genome assemblies. Bioinformatics. 2013;29: 1072–1075. doi:10.1093/bioinformatics/btt086

101. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12: 491. doi:10.1186/1471-2105-12-491

102. Grigoriev I V., Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, et al. MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 2014;42: D699–D704. doi:10.1093/nar/gkt1183

103. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 2008;18: 1979–1990. doi:10.1101/gr.081612.108

104. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5: 59. doi:10.1186/1471-2105-5-59

105. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19: ii215-ii225. doi:10.1093/bioinformatics/btg1080

106. Sternes PR, Lee D, Kutyna DR, Borneman AR. Genome Sequences of Three Species of Hanseniaspora Isolated from Spontaneous Wine Fermentations. Genome Announc. 2016;4: e01287-16. doi:10.1128/genomeA.01287-16

107. Giorello FM, Berná L, Greif G, Camesasca L, Salzman V, Medina K, et al. Genome Sequence of the Native Apiculate Wine Yeast Hanseniaspora vineae T02/19AF. Genome Announc. 2014;2: e00530-14. doi:10.1128/genomeA.00530-14

108. Cadez N. Phylogenetic placement of Hanseniaspora-Kloeckera species using multigene sequence analysis with taxonomic implications: descriptions of Hanseniaspora

pseudoguilliermondii sp. nov. and Hanseniaspora occidentalis var. citrica var. nov. Int J Syst Evol Microbiol. 2006;56: 1157–1165. doi:10.1099/ijs.0.64052-0

109. Chang C-F, Huang L-Y, Chen S-F, Lee C-F. Kloeckera taiwanica sp. nov., an ascomycetous apiculate yeast species isolated from mushroom fruiting bodies. Int J Syst Evol Microbiol. 2012;62: 1434–1437. doi:10.1099/ijs.0.034231-0

110. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol. 2018;35: 543–548. doi:10.1093/molbev/msx319

111. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13: 2178–2189. doi:10.1101/gr.1224503

112. van Dongen S. Graph clustering by flow simulation. Graph Stimul by flow Clust. 2000;PhD thesis: University of Utrecht. doi:10.1016/j.cosrev.2007.05.001

113. Madden T. The BLAST sequence analysis tool. BLAST Seq Anal Tool. 2013; 1–17.

114. Kocot KM, Citarella MR, Moroz LL, Halanych KM. PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. Evol Bioinforma. 2013;2013: 429–435. doi:10.4137/EBO.S12813

115. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010

116. Mount DW. Using BLOSUM in Sequence Alignments. Cold Spring Harb Protoc. 2008;2008: pdb.top39-pdb.top39. doi:10.1101/pdb.top39

117. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25: 1972–

1973. doi:10.1093/bioinformatics/btp348

118. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5. doi:10.1371/journal.pone.0009490

119. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol. 1981;17: 368–376. doi:10.1007/BF01734359

120. Philippe H, Delsuc F, Brinkmann H, Lartillot N. Phylogenomics. Annu Rev Ecol Evol Syst. 2005;36: 541–562. doi:10.1146/annurev.ecolsys.35.112202.130205

121. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 2003;425: 798–804. doi:10.1038/nature02053

122. Edwards SV. Is a new and general theory of molecular systematics emerging? Evolution (N Y). 2009;63: 1–19. doi:10.1111/j.1558-5646.2008.00549.x

123. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033

124. Le SQ, Gascuel O. An Improved General Amino Acid Replacement Matrix. Mol Biol Evol. 2008;25: 1307–1320. doi:10.1093/molbev/msn067

125. Mirarab S, Warnow T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 2015;31: i44–i52. doi:10.1093/bioinformatics/btv234

126. Stamatakis A, Hoover P, Rougemont J. A Rapid Bootstrap Algorithm for the RAxML Web Servers. Renner S, editor. Syst Biol. 2008;57: 758–771. doi:10.1080/10635150802429642

127. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010;26: 1669–1670. doi:10.1093/bioinformatics/btq243

128. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24: 1586–1591. doi:10.1093/molbev/msm088

129. dos Reis M, Donoghue PCJ, Yang Z. Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet. 2016;17: 71–80. doi:10.1038/nrg.2015.8

130. Reis M d., Yang Z. Approximate Likelihood Calculation on a Phylogeny for Bayesian Estimation of Divergence Times. Mol Biol Evol. 2011;28: 2161–2172. doi:10.1093/molbev/msr045

131. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 2012;40: D700–D705. doi:10.1093/nar/gkr1029

132. Pearson WR. An Introduction to Sequence Similarity (“Homology”) Searching. Curr Protoc Bioinforma. 2013;42: 3.1.1-3.1.8. doi:10.1002/0471250953.bi0301s42

133. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Comput Biol. 2011;7: e1002195. doi:10.1371/journal.pcbi.1002195

134. GeneOntologyConsortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32: 258D–261. doi:10.1093/nar/gkh036

135. Klopfenstein D V., Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, et al. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep. 2018;8: 10872. doi:10.1038/s41598-018-28948-z

136. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6: 65–

70.

137. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44: D457–D462. doi:10.1093/nar/gkv1070

138. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, et al. Functional characterization of the S-cerevisiae genome by gene deletion and parallel analysis. Science (80-). 1999;285: 901–906. doi:10.1126/science.285.5429.901

139. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923

140. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

141. Weiß CL, Pais M, Cano LM, Kamoun S, Burbano HA. nQuire: a statistical framework for ploidy estimation using next generation sequencing. BMC Bioinformatics. 2018;19: 122. doi:10.1186/s12859-018-2128-z

142. Suyama M, Torrents D, Bork P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34. doi:10.1093/nar/gkl315

143. Liu L, Zhang J, Rheindt FE, Lei F, Qu Y, Wang Y, et al. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. Proc Natl Acad Sci. 2017;114: E7282–E7290. doi:10.1073/pnas.1616744114

144. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics.

Bioinformatics. 2009;25: 1422–1423. doi:10.1093/bioinformatics/btp163

145. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: A structure for efficient numerical computation. Comput Sci Eng. 2011;13: 22–30. doi:10.1109/MCSE.2011.37

Tables

Table 1. Rate of sequence evolution hypotheses and results.

Hypotheses for inter-lineage comparisons	Parameters	Fraction of genes significantly different than H_O	**Median ω values**		
			$\omega_{background}$	ω_1	ω_2
H_O: Uniform rate for all branches **Figure S9A**	Single ω value	N/A	N/A	N/A	N/A
$H_{FE\text{-}SE\ branch}$: Unique rates for FEL and SEL stem **Figure S9B**	$\omega_{background} \neq \omega_1 \neq \omega_2$ ω_1 = FEL stem branch ω_2 = SEL stem branch	678 genes (68.55% of examined genes)	0.060	0.566	0.293
H_{FE}: Unique rates for FEL stem and FEL crown **Figure S9C**	$\omega_{background} \neq \omega_1 \neq \omega_2$ ω_1 = FEL stem branch ω_2 = FEL crown branches	743 genes (75.13% of examined genes)	0.063	0.711	0.061
H_{SE}: Unique rates for SEL stem and SEL crown **Figure S9D**	$\omega_{background} \neq \omega_1 \neq \omega_2$ ω_1 = SEL stem branch ω_2 = SEL crown branches	528 genes (53.7% of examined genes)	0.059	0.267	0.074
$H_{FE\text{-}SE\ crown}$: Unique rates for FEL crown and SEL crown **Figure S9E**	$\omega_{background} \neq \omega_1 \neq \omega_2$ ω_1 = FEL crown branches ω_2 = SEL crown branches	717 genes (72.5% of examined genes)	0.010	0.062	0.074

Main Figure Legends

Fig 1. The evolutionary history and timeline of *Hanseniaspora* diversification and the stability of a long internode branch. (A) Using 1,034 single-copy orthologous genes (SCOG), the evolutionary history of *Hanseniaspora* in geologic time revealed two well-support lineages termed the fast evolving and slow evolving lineages (FEL and SEL, respectively), which began diversifying around 87.2 and 53.6 million years ago (mya) after diverging 95.3 mya. (B) Among single-gene phylogenies where the FEL and SEL were monophyletic (*n* = 946), internode branch lengths leading up to each lineage revealed significantly longer internode branches leading up to the FEL (0.62 ± 0.38 base substitutions / site) compared to the SEL (0.17 ± 0.11 base substitutions / site) ($p < 0.001$; Paired Wilcoxon Rank Sum test). (C) Examination of the difference between internode branch lengths per single-gene tree revealed 932 single-gene phylogenies had a longer branch length in the FEL compared to the SEL (depicted in orange with values greater than 0), while the converse was only observed in 14 single-gene phylogenies (depicted in blue with values less than 0). Across all single-gene phylogenies, the average difference between the internode branch length leading up to the two lineages was 0.45.

Fig 2. Gene presence and absence analyses reflect phenotype and reveal disrupted pathways. (A) Examination of gene presence and absence (see *Methods*) revealed numerous genes that had been lost across *Hanseniaspora*. Specifically, 1,409 have been lost in the FEL, and 771 genes have been lost in the SEL. A Euler diagram represents the overlap of these gene sets. Both lineages have lost 748 genes, the FEL has lost an additional 661, and the SEL has lost an additional 23. (B) The *IMA* gene family (*IMA1-5*) encoding α-glucosidases,

MAL (*MALx1-3*) loci, and *SUC2* are associated with growth on maltose, sucrose, raffinose, and melezitose. The *IMA* and *MAL* loci are largely missing among *Hanseniaspora* with the exception of homologs *MALx1*, which encode diverse transporters of the major facilitator superfamily whose functions are difficult to predict from sequence; as expected, *Hanseniaspora* spp. cannot grow on maltose, raffinose, and melezitose with the sole exception of *Hanseniaspora jakobsenii*, which has delayed/weak growth on maltose and is the only *Hanseniaspora* species with (*MALx3*), which encodes a homolog of the *MAL*-activator protein. (C) The genes involved with galactose degradation are largely missing among *Hanseniaspora* species, which correlates with their inability to grow on galactose. Genes that are present are depicted in white, and genes that are absent are depicted in black. The ability to grow, grow with delayed/weak growth on a given substrate, or the inability to grow is specified using white, grey, and black circles, respectively; dashes indicate no data. (D) Most genes involved in the thiamine biosynthesis pathway are absent among all *Hanseniaspora*. (E) Many genes involved in the methionine salvage pathway are absent among all *Hanseniaspora*. Absent genes are depicted in purple.

Fig 3. Gene presence and absence in the budding yeast cell cycle. Examination of genes present and absent in the cell cycle of budding yeasts revealed numerous missing genes. Many genes are key regulators, such as *WHI5*; participate in spindle checkpoint processes and segregation, such as *MAD1* and *MAD2*; or DNA damage checkpoint processes, such as *MEC3*, *RAD9*, and *RFX1*. Genes missing in both lineages, the FEL, or the SEL are colored purple, orange, or blue, respectively. The "e" in the PHO cascade represents expression of Pho4:Pho2. Dotted lines with arrows indicate indirect links or unknown reactions. Lines with arrows indicate molecular interactions or relations. Circles indicate chemical compounds such as DNA.

Fig 4. A panoply of genome maintenance and DNA repair genes are missing among *Hanseniaspora*, especially in the FEL. Genes annotated as DNA repair genes according to gene ontology (GO:0006281) and child terms were examined for presence and absence in at least two-thirds of each lineage, respectively (268 total genes). 47 genes are missing among the FEL species, and 14 genes are missing among the SEL. Presence and absence of genes was clustered using hierarchical clustering (cladogram on the left) where each gene's ontology is provided as well. Genes with multiple gene annotations are denoted as such using the 'multiple' term.

Fig 5. dN/dS (ω) analyses supports a historical burst of accelerated evolution in the FEL. (A) The null hypothesis (H_O) that all branches in the phylogeny have the same ω value. Alternative hypotheses (B-E) evaluate ω along three sets of branches. (Bi) The alternative hypothesis ($H_{FE\text{-}SE}$ branch) examined ω values along the branch leading up the FEL and the SEL. (Bii) 311 supported H_O and 678 genes supported $H_{FE\text{-}SE}$ branch. (Biii) Among the genes that supported $H_{FE\text{-}SE}$ branch, we examined the distribution of the difference between ω_1 and ω_2 as specified in part Bi. Here, a range of ω_1 - ω_2 of -3.5 to 3.5 is shown in the histogram. Additionally, we report the median ω_1 and ω_2 values, which are 0.57 and 0.29, respectively. (Biv) Among all genes examined, 0.39 genes significantly rejected H_O and were faster in the FEL than the SEL, and 0.30 genes were faster in the SEL than the FEL. (Ci) The alternative hypothesis (H_{FE}) examined ω values along the branch leading up to the FEL and all branches thereafter (FEL_{crown}). (Cii) 246 genes supported H_O, and 743 genes supported H_{FE}. (Ciii) Among the genes that supported H_{FE}, we examined the distribution of the difference between ω_1 and ω_2

as specified in part Ci. The median ω_1 and ω_2 values were 0.71 and 0.06, respectively. (Civ) Among all genes, 0.73 genes significantly rejected H_O and were faster in the FEL than the FEL_{crown}, and 0.02 genes were faster in the FEL_{crown} than the FEL. (Di) The alternative hypothesis (H_{SE}) examined ω values along the branch leading up to the SEL and all branches thereafter (SEL_{crown}). (Dii) 455 genes supported H_O, and 528 genes supported H_{SE}. (Diii) Among the genes that supported H_{SE}, we examined the distribution of the difference between ω_1 and ω_2 as specified in part Di. The median ω_1 and ω_2 values were 0.27 and 0.07, respectively. (Div) Among all genes, 0.49 genes significantly rejected H_O and were faster in the SEL than the SEL_{crown}, and 0.05 genes were faster in the SEL_{crown} than the SEL. (Ei) The alternative hypothesis ($H_{FE\text{-}SE\ crown}$) examined ω values in the crown of the FEL_{crown} and SEL_{crown}. (Eii) 272 genes supported H_O, and 717 genes supported $H_{FE\text{-}SE\ crown}$. (Eiii) Among the genes that supported $H_{FE\text{-}SE\ crown}$, we examined the distribution of the difference between ω_1 and ω_2 as specified in part Di. The median ω_1 and ω_2 values were 0.06 and 0.07, respectively. (Eiv) Among all genes, 0.22 genes significantly rejected H_O and were faster in the FEL_{crown} compared to the SEL_{crown}, and 0.51 genes were faster in the SEL_{crown} than the FEL_{crown}.

Fig 6. Analyses of base substitutions and indels reveal a higher mutational load in the FEL compared to the SEL. (A) Analyses of substitutions at evolutionarily tractable sites among codon-based alignments revealed a higher number of base substitutions in the FEL compared to the SEL (F(1) = 196.88, $p < 0.001$; Multi-factor ANOVA) and an asymmetric distribution of base substitutions at codon sites (F(2) = 1691.60, $p < 0.001$; Multi-factor ANOVA). A Tukey Honest Significance Differences post-hoc test revealed a higher proportion of substitutions in the FEL compared to the SEL at evolutionarily tractable sites at the first (*n* =

240,565; $p < 0.001$), second ($n = 318{,}987$; $p < 0.001$), and third ($n = 58{,}151$; $p = 0.02$) codon positions. (B) Analyses of the direction of base substitutions (i.e., G|C → A|T or A|T → G|C) reveals significant differences between the FEL and SEL (F(1) = 447.1, $p < 0.001$; Multi-factor ANOVA) and differences between the directionality of base substitutions (F(1) = 914.5, $p < 0.001$; Multi-factor ANOVA). A Tukey Honest Significance Differences post-hoc test revealed a significantly higher proportion of substitutions were G|C → A|T compared to A|T → G|C among evolutionarily tractable sites that are G|C ($n = 232{,}546$) and A|T ($n = 385{,}157$) ($p < 0.001$), suggesting a general AT-bias of base substitutions. Additionally, there was a significantly higher proportion of evolutionary tractable sites with base substitutions in the FEL compared to the SEL ($p < 0.001$). More specifically, a higher number of base substitutions were observed in the FEL compared to the SEL for both G|C → A|T ($p < 0.001$) and A|T → G|C mutations ($p < 0.001$), but the bias toward AT was greater in the FEL. (C) Examinations of transition / transversion ratios revealed a lower transition / transversion ratio in the FEL compared to the SEL ($p < 0.001$; Wilcoxon Rank Sum test). (D) Comparisons of insertions and deletions revealed a significantly greater number of insertions ($p < 0.001$; Wilcoxon Rank Sum test) and deletions ($p < 0.001$; Wilcoxon Rank Sum test) in the FEL ($\overline{x}_{\text{insertions}} = 7521.11 \pm 405.34$; $\overline{x}_{\text{deletions}} = 3894.11 \pm 208.16$) compared to the SEL ($\overline{x}_{\text{insertions}} = 6049.571 \pm 155.85$; $\overline{x}_{\text{deletions}} = 2346.71 \pm 326.22$). (E and F) When adding the factor of size per insertion or deletion, significant differences were still observed between the lineages (F(1) = 2102.87, $p < 0.001$; Multi-factor ANOVA). A Tukey Honest Significance Differences post-hoc test revealed that most differences were caused by significantly more small insertions and deletions in the FEL compared to the SEL. More specifically, there were significantly more insertions in the FEL compared to the SEL for sizes 3-18 ($p < 0.001$ for all comparisons between each lineage for each insertion size), and there were significantly more deletions in the FEL compared to the SEL for sizes 3-21 ($p < 0.001$ for all comparisons between each lineage for each deletion size). Black lines at the top of each bar show the 95% confidence interval for the number of insertions or deletions for a given size. (G) Evolutionarily conserved homopolymers of sequence length two (n = 17,391), three (n = 1,062), four (n = 104), and five (n = 5) were examined for substitutions and indels. Statistically significant differences of the proportion mutated bases (i.e., (base substitutions + deleted bases + inserted bases) / total homopolymer bases) were observed between the FEL and SEL (F(1) = 27.68, $p < 0.001$; Multi-factor ANOVA). Although the FEL had more mutations than the SEL for all homopolymers, a Tukey Honest Significance Differences post-hoc test revealed differences were statistically significant for homopolymers of two ($p = 0.02$) and three ($p = 0.003$). Analyses of homopolymers using additional factors of mutation type (i.e., base substitution, insertion, deletion) and homopolymer sequence type (i.e., A|T and C|G homopolymers) can be seen in Fig S9. (H) G → T or C → A mutations are associated with the common and abundant oxidatively damaged base, 8-oxo-dG. When examining all substituted G positions for each species and their substitution direction, we found significant differences between different substitution directions (F(2) = 5682, $p < 0.001$; Multi-factor ANOVA). More importantly, a Tukey Honest Significance Differences post-hoc test revealed an over-representation of G → T or C → A in the FEL compared to the SEL ($p < 0.001$). (I) CC → TT dinucleotide substitutions are associated with UV damage. Using a CC|GG (left) and TT|AA (right) score, which is an indirect proxy for UV mutation damage where less UV damage would result in a higher CC|GG score and more UV damage would result in a higher TT|AA score, we found no significant differences when comparing CC|GG scores between the FEL, SEL, and outgroup taxa ($\chi 2(2) = 5.964$, $p = 0.05$; Kruskal-Wallis rank sum test); however, when comparing the outgroup taxa to all *Hanseniaspora*, a significant difference was observed (p = 0.03; Wilcoxon Rank Sum test). When examining TT|AA scores, we found significant differences between the FEL, SEL, and outgroup taxa ($\chi2(2) = 8.84$, $p = 0.01$; Kruskal-Wallis rank sum test). A post-hoc Dunn's test using the Benjamini-Hochberg method for multi-test correction revealed significant differences between the FEL and SEL compared to the outgroup taxa (p = 0.01 and 0.02, respectively). A significant difference between all *Hanseniaspora* and the outgroup taxa were also observed ($p < 0.001$; Wilcoxon Rank Sum test). Results from the Kruskal-Wallis rank sum test and the Wilcoxon Rank Sum test are differentiated using lines and asterisks that are red and black, respectively.

Supplementary Figure Legends

Fig S1. *Hanseniaspora* have among the smallest genome sizes, lowest number of genes, and lowest percent GC content in the budding yeast subphylum Saccharomycotina.

(A) The genus *Hanseniaspora* (family Saccharomycodaceae) includes the smallest budding yeast genome. The FEL, SEL, and all of Saccharomycotina have an average genome size of 9.71 ± 1.32 Mb (min: 8.10; max: 14.05), 10.99 ± 1.66 Mb (min: 7.34; max: 12.17), 12.80 ± 3.20 Mb (min: 7.34; max: 25.83), respectively. (B) The genus *Hanseniaspora* includes the budding yeast genome with the fewest genes. The FEL, SEL, and all of Saccharomycotina have an average number of genes per genome of 4,707.89 ± 633.56 (min: 3,923; max: 6,380), 4,932.43 ± 289.71 (min: 4,624; max: 5,349), and 5,657.66 ± 1,044.78 (min: 3,923; max: 12,786), respectively. (C) The genus *Hanseniaspora* has among the lowest GC-content values in budding yeast genomes. The FEL, SEL, and all of Saccharomycotina GC-content values were 33.10 ± 3.53% (min: 26.32; max: 37.17), 37.28 ± 2.05% (min: 34.82; max: 39.93), and 40.30 ± 5.71% (min: 25.2; max: 53.98), respectively. Families of Saccharomycotina are depicted on the y-axis. Median values are depicted with a line, and dashed lines indicate plus or minus one standard deviation from the median. To the right of each figure, boxplots depict the median and standard deviations of each grouping. The grey represents all of Saccharomycotina. Blue represents the SEL, and orange represents the FEL.

Fig S2. Phylogenomics method pipeline. Using 25 *Hanseniaspora* proteomes and the proteomes of 4 outgroup taxa, 11,877 orthologous groups (OGs) of genes were identified. 1,143 OGs with few paralogs were identified has having few paralogs – that is, ≥ 90% of species do not have paralogs and have one gene in the OG. The sequences of the 1,143 OGs were

individually aligned, trimmed, had their evolutionary history inferred, and paralogs were trimmed based on tree topology. Using the resulting 1,142 OGs with paralogs trimmed, sequences were realigned and trimmed and had their evolutionary history inferred. If the outgroup taxa were not the earliest diverging taxa after serially rooting on the outgroup taxa, the OG was removed resulting in 1,034 OGs. Among these 1,034 OGs of genes, a concatenated 1,034-gene matrix was constructed and used for reconstructing evolutionary history. Similarly, evolutionary history was inferred using coalescence of the 1,034 OG single-gene phylogenies.

Fig S3. Concatenation and coalescence produce nearly identical and well-supported phylogenies that support two distinct lineages. (Left) Concatenation supports one lineage with a long internode branch leading to the clade, which we term the fast-evolving lineage (FEL) and another lineage with a much shorter internode branch length leading to the clade (SEL). (Right) Coalescence supports monophyly of the FEL and SEL. Minor discrepancies are observed between the topologies. Bipartitions without full support have their support values depicted. Support for concatenation and coalescence was determined using 100 rapid bootstrap replicates and local posterior support, respectively.

Fig S4. Internode key to accompany divergence time estimate file per internode.

Internode identifiers for timetree analysis in Fig 1B. Associated mean divergence time and credible intervals can be found in File S2.

Fig S5. BUSCO analyses reveals extensive gene 'missingness' across various taxonomic ranks. BUSCO analyses of *Hanseniaspora* proteomes using the Eukaryota (n_{BUSCOs} =

303), Fungi (n_{BUSCOs} = 290), Dikarya (n_{BUSCOs} = 1,312), Ascomycota (n_{BUSCOs} = 1,315), Saccharomyceta (n_{BUSCOs} =1,759), and Saccharomycetales (n_{BUSCOs} = 1,711) orthoDB databases revealed numerous BUSCO genes are missing among *Hanseniaspora* genomes, in particular the FEL.

Fig S6. A liberal targeted gene searching pipeline and the number of missing genes in at least two-thirds of FEL and SEL taxa. (A) A FASTA file for gene *X*, where gene *X* is the FASTA entry of a verified ORF in the *Saccharomyces cerevisiae* proteome, is used as a query to search for putative homologs in the Fungal reference sequence (refseq) database. The top 100 putative homologs were subsequently aligned. From the alignment, a Hidden Markov Model (HMM) was made. Using the HMM, gene *X* was searched for in the genome of each species from the FEL, SEL, and outgroup individually using a liberal e-value cut-off of 0.01 and a score of > 50. This pipeline yields presence and absence information of gene *X* among FEL, SEL, and outgroup taxa. This method was subsequently applied to all verified ORF in the *S. cerevisiae* proteome.

Fig S7. Gene presence and absence reveals a putatively diminished gluconeogenesis pathway. Gene presence and absence analysis of genes that participate in the gluconeogenesis (A) and glycolysis (B) pathway reveal key missing genes in the gluconeogenesis pathway, suggestive of a diminished capacity for gluconeogenesis. More specifically, *PCK1*, which encodes the enzyme that converts oxaloacetic acid to phosphoenolpyruvate, and *FBP1*, which encodes the enzyme that converts fructose-1,6-bisphosphate to fructose-6-phospbate, are missing among all *Hanseniaspora* species.

Fig S8. Base frequency plots reveal diversity in ploidy of *Hanseniaspora* species.

(A) A lack of Gaussian distributions suggests *H. occidentalis* var. *occidentalis*, *H. uvarum* CBS 314, and *H. guilliermondii* CBS 465 are haploid. (B) A single Gaussian distribution suggests *H. occidentalis* var. *citrica*, *H. osmophila* CBS 313, *H. meyeri*, *H. clermontiae*, *H. nectarophila*, *H. thailandica*, *H. pseudoguilliermondii*, *H. singularis*, and *K. hatyaiensis* are diploids. (C) Two Gaussian distributions suggest *H. lachancei* and *H. jakobsenii* are triploid. (D) Analyses of *H. vineae* CBS 2171, *H. valbyensis*, *Hanseniaspora* sp. CRUB 1602, and *H. opuntiae* base frequency distributions were ambiguous. Certain FEL species, such as *H. singularis, H. pseudoguilliermondii*, and *H. jakobsenii,* are potentially aneuploid, while evidence of aneuploidy in the SEL is observed in only *H. occidentalis* var. *citrica.*

Fig S9. Analyses of homopolymers by sequence length, type, and type of mutation.

Significant differences among the proportion of mutated bases among homopolymers of various lengths were observed (Figure 5). Addition of variables (i.e., sequence type (A|T or C|G) and mutation type (base substitution, insertion, and deletion)) allowed for further determination of what types of mutations caused differences between the FEL and SEL. As shown in Figure 5, we observed significant differences in the numbers of mutations between the FEL and SEL (F = 27.06, p < 0.001; Multi-factor ANOVA) as well as in the type of mutations (F = 1686.70, p < 0.001; Multi-factor ANOVA). A Tukey Honest Significance Differences post-hoc test revealed that the proportion of nucleotides that underwent base substitutions was significantly greater than insertions (p < 0.001) and deletions (p < 0.001). We next focused on significant differences observed between the FEL and SEL when considering all factors. We observed significant differences between the FEL and SEL at A|T and C|G homopolymers with a length of 2 (p = 0.009 and $p < 0.001$, respectively), C|G homopolymers of length 3 ($p < 0.001$), and A|T homopolymers of length 5 ($p < 0.001$).

Fig S10. Metrics reveal more radical amino acid substitutions in the FEL compared to SEL. Using Sneath's index and Epstein's coefficient of difference, the average difference among amino acid substitutions were determined among sites where the outgroup taxa had all the same amino acid. Using either metric, amino acid substitutions were significantly more drastic in the FEL compared to the SEL ($p < 0.001$; Wilcoxon Rank Sum test for both metrics).

Fig S11. Mean protein similarity reveals immense diversity in *Hanseniaspora*.

The FEL spans a large amount of mean protein similarity when comparing various species to *H. uvarum*. Similarly, but to a lesser degree, the same is true for the SEL when comparing various species to *H. vineae*. The diversity observed in these lineages is roughly on par with genus-level differences within the family Saccharomycetaceae, humans to zebrafish, and thale cress (*Arabidopsis thaliana*) to Japanese rice.

Figure 1

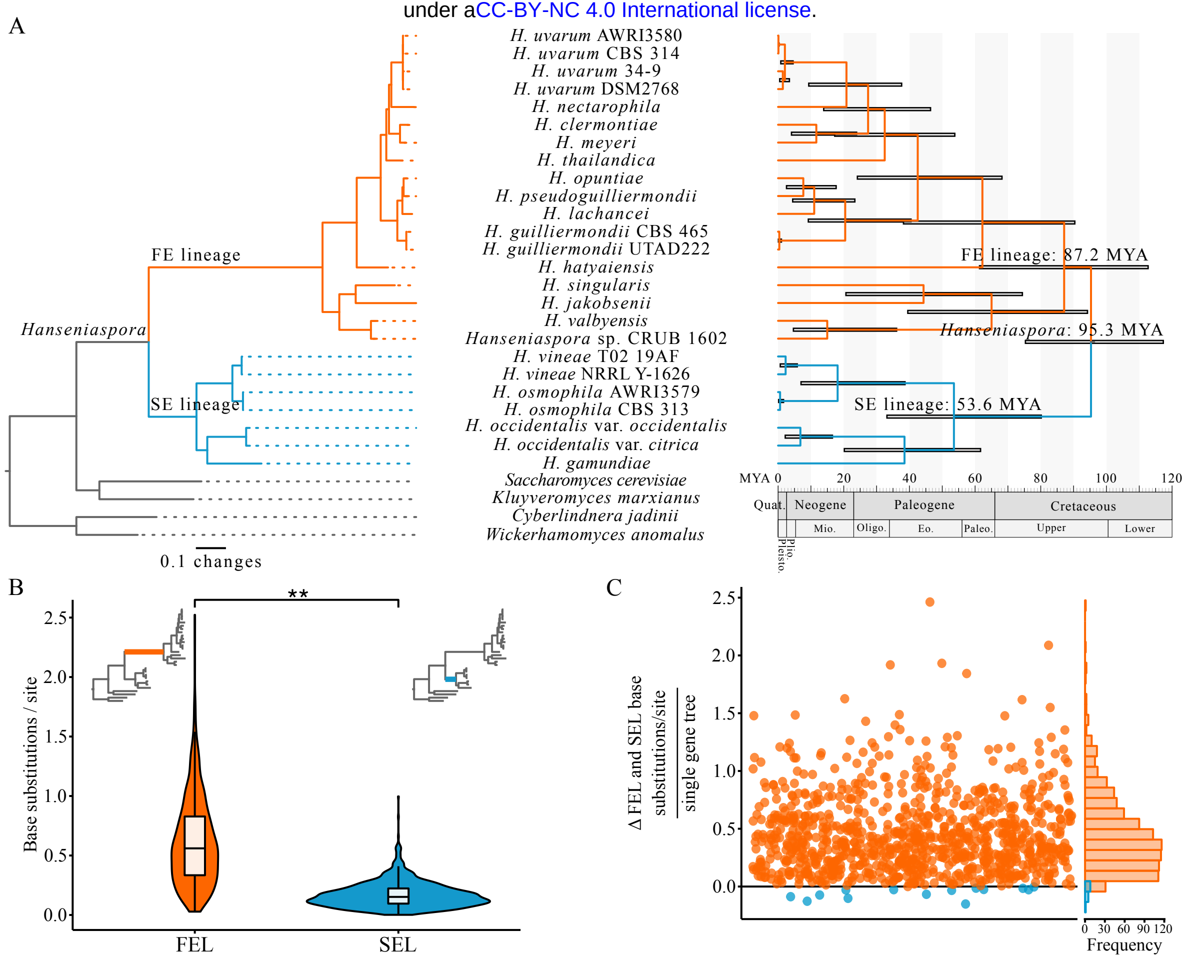

A
Hanseniaspora
FE lineage
SE lineage
0.1 changes
H. uvarum AWRI3580
H. uvarum CBS 314
H. uvarum 34-9
H. uvarum DSM2768
H. nectarophila
H. clermontiae
H. meyeri
H. thailandica
H. opuntiae
H. pseudoguilliermondii
H. lachancei
H. guilliermondii CBS 465
H. guilliermondii UTAD222
H. hatyaiensis
H. singularis
H. jakobsenii
H. valbyensis
Hanseniaspora sp. CRUB 1602
H. vineae T02 19AF
H. vineae NRRL Y-1626
H. osmophila AWRI3579
H. osmophila CBS 313
H. occidentalis var. *occidentalis*
H. occidentalis var. *citrica*
H. gamundiae
Saccharomyces cerevisiae
Kluyveromyces marxianus
Cyberlindnera jadinii
Wickerhamomyces anomalus
FE lineage: 87.2 MYA
Hanseniaspora: 95.3 MYA
SE lineage: 53.6 MYA
MYA 0
20
40
60
80
100
120
Quat.
Neogene
Paleogene
Cretaceous
Plio.
Pleisto.
Mio.
Oligo.
Eo.
Paleo.
Upper
Lower
B
**
Base substitutions / site
0.0
0.5
1.0
1.5
2.0
2.5
FEL
SEL
C
Δ FEL and SEL base substitutions/site / single gene tree
0.0
0.5
1.0
1.5
2.0
2.5
0
30
60
90
120
Frequency

Figure 2

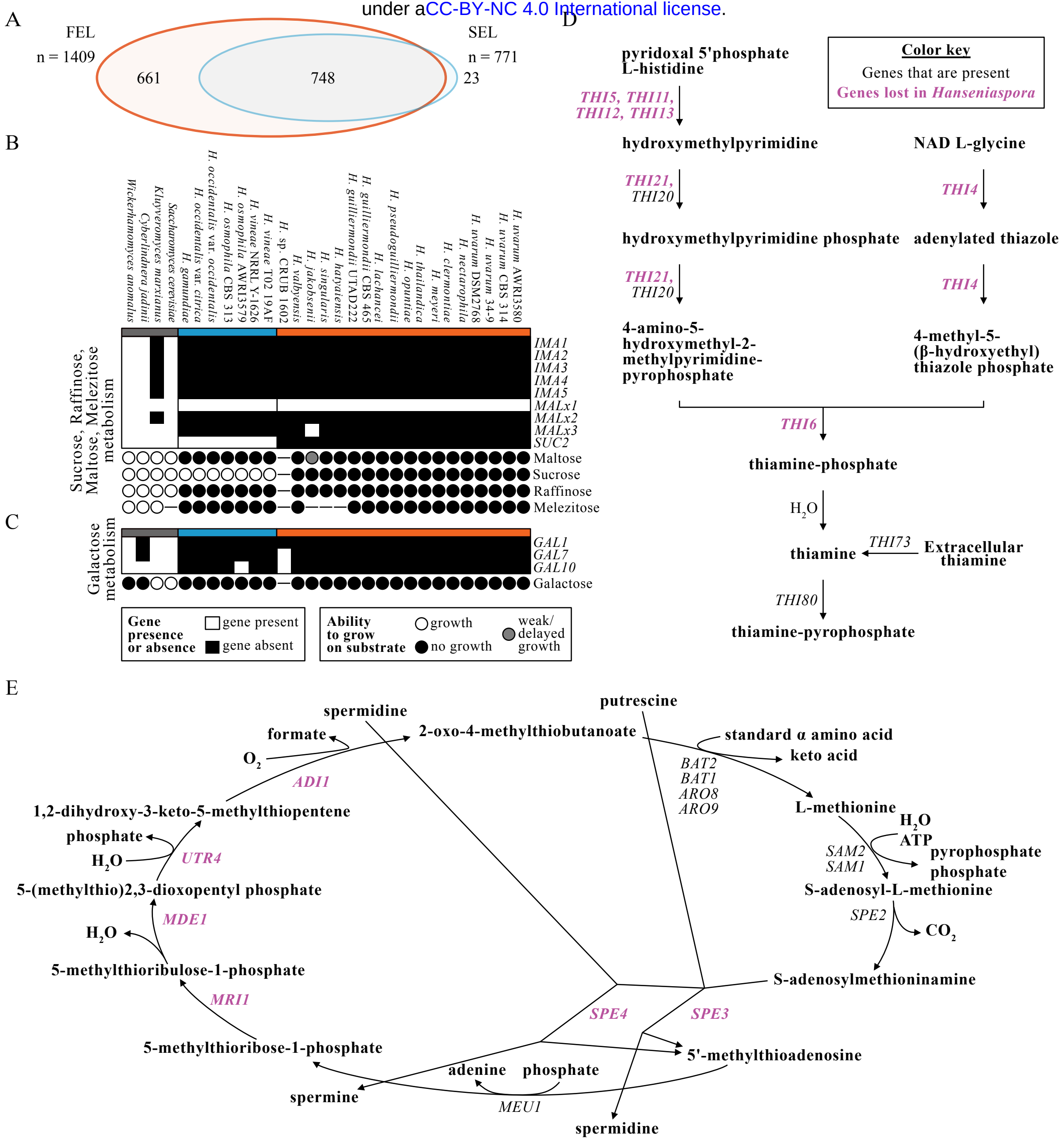

A
FEL
n = 1409
661
748
SEL
n = 771
23
B
Wickerhamomyces anomalus
Cyberlindnera jadinii
Kluyveromyces marxianus
Saccharomyces cerevisiae
H. gamundiae
H. occidentalis var. citrica
H. occidentalis var. occidentalis
H. osmophila CBS 313
H. osmophila AWRI3579
H. vineae NRRL Y-1626
H. vineae T02 19AF
H. sp. CRUB 1602
H. valbyensis
H. jakobsenii
H. singularis
H. hatyaiensis
H. guilliermondii UTAD222
H. guilliermondii CBS 465
H. lachancei
H. pseudoguilliermondii
H. opuntiae
H. thailandica
H. meyeri
H. clermontiae
H. nectarophila
H. uvarum DSM2768
H. uvarum 34-9
H. uvarum CBS 314
H. uvarum AWRI3580
Sucrose, Raffinose, Maltose, Melezitose metabolism
IMA1
IMA2
IMA3
IMA4
IMA5
MALx1
MALx2
MALx3
SUC2
Maltose
Sucrose
Raffinose
Melezitose
C
Galactose metabolism
GAL1
GAL7
GAL10
Galactose
Gene presence or absence
□ gene present
■ gene absent
Ability to grow on substrate
○ growth
● no growth
weak/ delayed growth
D
Color key
Genes that are present
Genes lost in Hanseniaspora
pyridoxal 5'phosphate
L-histidine
THI5, THI11, THI12, THI13
hydroxymethylpyrimidine
THI21, THI20
hydroxymethylpyrimidine phosphate
THI21, THI20
4-amino-5-hydroxymethyl-2-methylpyrimidine-pyrophosphate
NAD L-glycine
THI4
adenylated thiazole
THI4
4-methyl-5-(β-hydroxyethyl) thiazole phosphate
THI6
thiamine-phosphate
H_2O
thiamine
THI73
Extracellular thiamine
THI80
thiamine-pyrophosphate
E
spermidine
putrescine
formate
O_2
2-oxo-4-methylthiobutanoate
standard α amino acid
keto acid
BAT2
BAT1
ARO8
ARO9
ADI1
L-methionine
1,2-dihydroxy-3-keto-5-methylthiopentene
H_2O
ATP
SAM2
SAM1
pyrophosphate
phosphate
phosphate
H_2O
UTR4
5-(methylthio)2,3-dioxopentyl phosphate
S-adenosyl-L-methionine
SPE2
CO_2
MDE1
H_2O
5-methylthioribulose-1-phosphate
S-adenosylmethioninamine
MRI1
SPE4
SPE3
5-methylthioribose-1-phosphate
5'-methylthioadenosine
adenine
phosphate
spermine
MEU1
spermidine

Figure 3

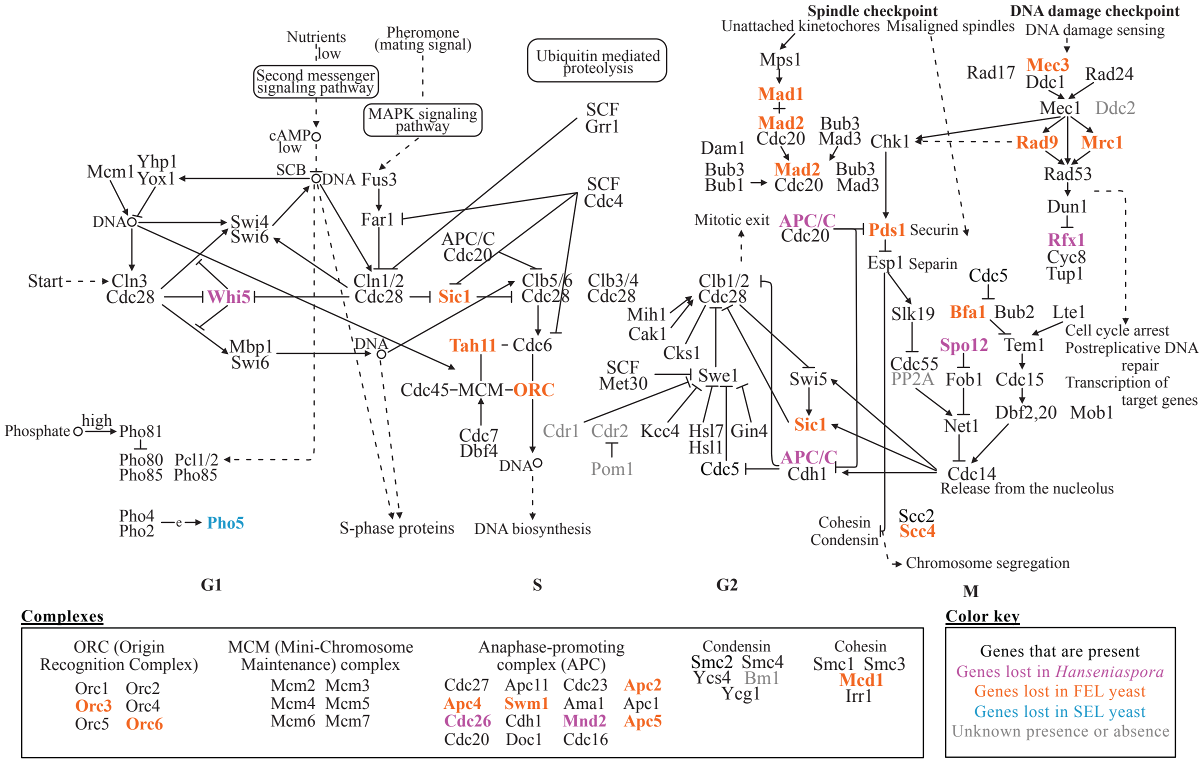

Nutrients
low
Pheromone
(mating signal)
Ubiquitin mediated proteolysis
Second messenger signaling pathway
MAPK signaling pathway
cAMP
low
SCB
DNA
Fus3
Mcm1
Yhp1
Yox1
DNA
Swi4
Swi6
Far1
SCF
Grr1
SCF
Cdc4
APC/C
Cdc20
Start
Cln3
Cdc28
Whi5
Cln1/2
Cdc28
Sic1
Clb5/6
Cdc28
Clb3/4
Cdc28
Mbp1
Swi6
DNA
Tah11
Cdc6
Cdc45
MCM
ORC
Cdc7
Dbf4
DNA
Phosphate
high
Pho81
Pho80
Pho85
Pcl1/2
Pho85
Pho4
Pho2
e
Pho5
S-phase proteins
DNA biosynthesis
Spindle checkpoint
Unattached kinetochores
Misaligned spindles
DNA damage checkpoint
DNA damage sensing
Mps1
Mad1
Mad2
Cdc20
Bub3
Mad3
Dam1
Bub3
Bub1
Mad2
Cdc20
Bub3
Mad3
Chk1
Rad17
Mec3
Ddc1
Rad24
Mec1
Ddc2
Rad9
Mrc1
Rad53
Dun1
Rfx1
Cyc8
Tup1
Mitotic exit
APC/C
Cdc20
Pds1
Securin
Esp1
Separin
Cdc5
Clb1/2
Cdc28
Mih1
Cak1
Cks1
SCF
Met30
Swe1
Swi5
Sic1
Cdr1
Cdr2
Pom1
Kcc4
Hsl7
Hsl1
Gin4
Cdc5
APC/C
Cdh1
Slk19
Cdc55
PP2A
Bfa1
Bub2
Lte1
Spo12
Tem1
Fob1
Cdc15
Net1
Dbf2,20
Mob1
Cdc14
Release from the nucleolus
Cell cycle arrest
Postreplicative DNA repair
Transcription of target genes
Cohesin
Condensin
Scc2
Scc4
Chromosome segregation
G1
S
G2
M
Complexes
ORC (Origin Recognition Complex)
Orc1 Orc2
Orc3 Orc4
Orc5 Orc6
MCM (Mini-Chromosome Maintenance) complex
Mcm2 Mcm3
Mcm4 Mcm5
Mcm6 Mcm7
Anaphase-promoting complex (APC)
Cdc27 Apc11 Cdc23 Apc2
Apc4 Swm1 Ama1 Apc1
Cdc26 Cdh1 Mnd2 Apc5
Cdc20 Doc1 Cdc16
Condensin
Smc2 Smc4
Ycs4 Brn1
Ycg1
Cohesin
Smc1 Smc3
Mcd1
Irr1
Color key
Genes that are present
Genes lost in Hanseniaspora
Genes lost in FEL yeast
Genes lost in SEL yeast
Unknown presence or absence

Figure 4

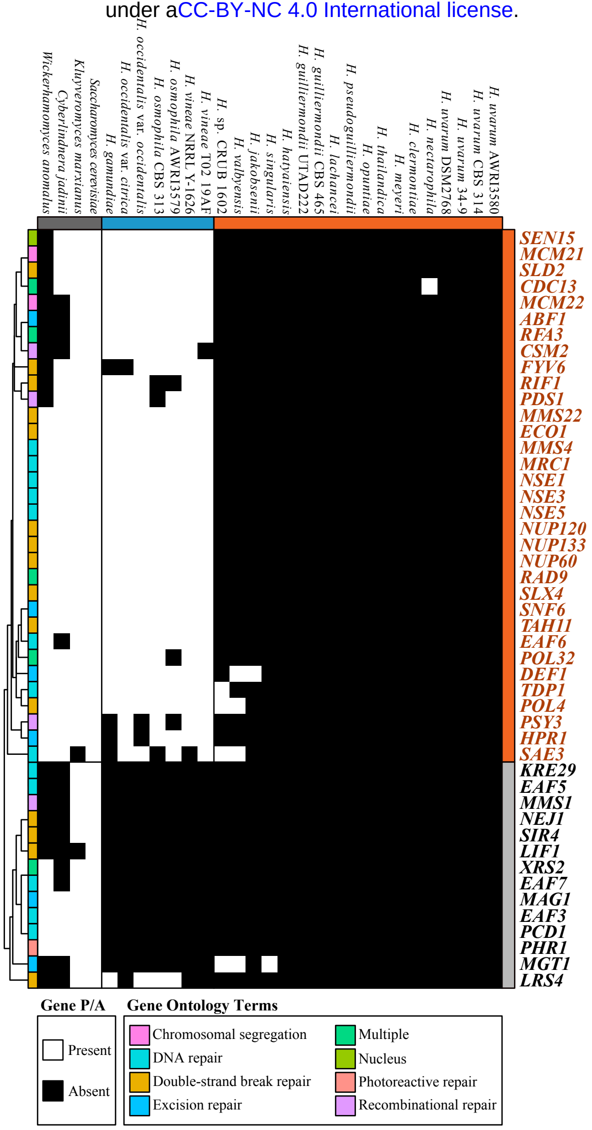
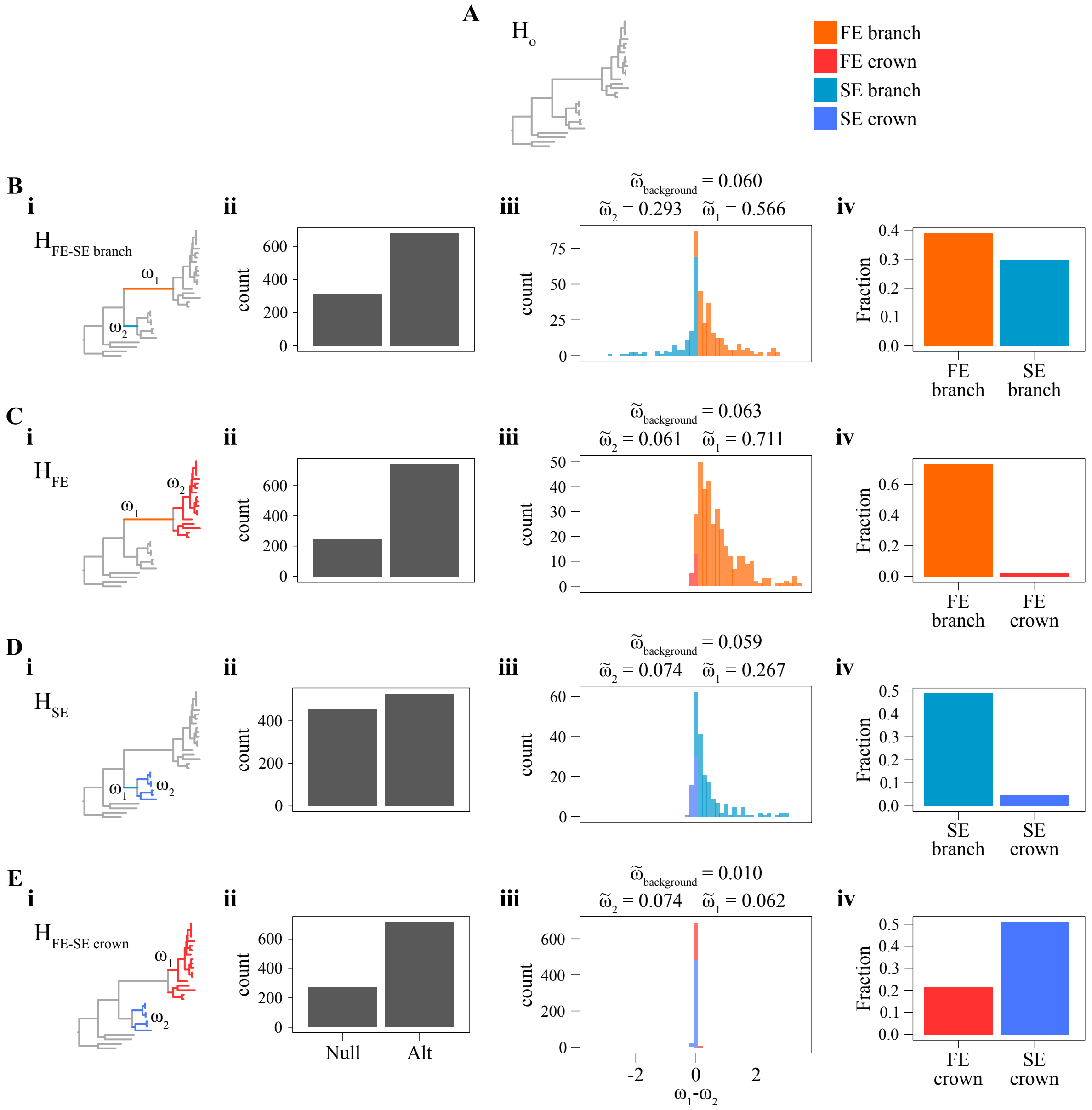
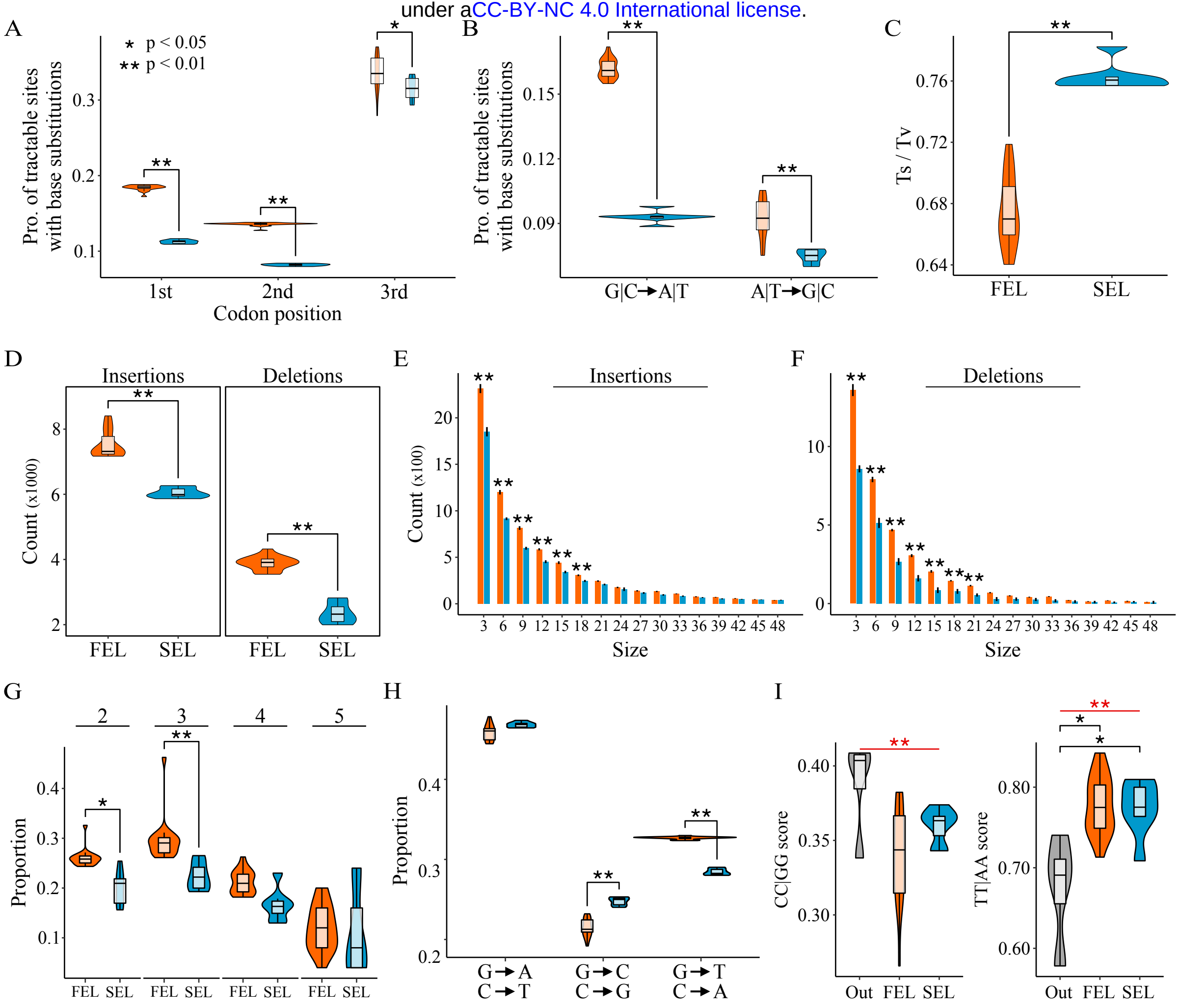

Wickerhamomyces anomalus
Cyberlindnera jadinii
Kluyveromyces marxianus
Saccharomyces cerevisiae
H. gamundiae
H. occidentalis var. citrica
H. occidentalis var. occidentalis
H. osmophila CBS 313
H. osmophila AWRI3579
H. vineae NRRL Y-1626
H. vineae T02 19AF
H. sp. CRUB 1602
H. valbyensis
H. jakobsenii
H. singularis
H. hatyaiensis
H. guilliermondii UTAD222
H. guilliermondii CBS 465
H. pseudoguilliermondii
H. lachancei
H. opuntiae
H. thailandica
H. meyeri
H. clermontiae
H. nectarophila
H. uvarum DSM2768
H. uvarum 34-9
H. uvarum CBS 314
H. uvarum AWRI3580
SEN15
MCM21
SLD2
CDC13
MCM22
ABF1
RFA3
CSM2
FYV6
RIF1
PDS1
MMS22
ECO1
MMS4
MRC1
NSE1
NSE3
NSE5
NUP120
NUP133
NUP60
RAD9
SLX4
SNF6
TAH11
EAF6
POL32
DEF1
TDP1
POL4
PSY3
HPR1
SAE3
KRE29
EAF5
MMS1
NEJ1
SIR4
LIF1
XRS2
EAF7
MAG1
EAF3
PCD1
PHR1
MGT1
LRS4
Gene P/A
Present
Absent
Gene Ontology Terms
Chromosomal segregation
DNA repair
Double-strand break repair
Excision repair
Multiple
Nucleus
Photoreactive repair
Recombinational repair

Figure 5



A
H_o
FE branch
FE crown
SE branch
SE crown
B
i
$H_{FE\text{-}SE\ branch}$
ω_1
ω_2
ii
count
0
200
400
600
iii
$\tilde{\omega}_{background} = 0.060$
$\tilde{\omega}_2 = 0.293$ $\tilde{\omega}_1 = 0.566$
count
0
25
50
75
iv
Fraction
0.0
0.1
0.2
0.3
0.4
FE branch
SE branch
C
i
H_{FE}
ω_1
ω_2
ii
count
0
200
400
600
iii
$\tilde{\omega}_{background} = 0.063$
$\tilde{\omega}_2 = 0.061$ $\tilde{\omega}_1 = 0.711$
count
0
10
20
30
40
50
iv
Fraction
0.0
0.2
0.4
0.6
FE branch
FE crown
D
i
H_{SE}
ω_1
ω_2
ii
count
0
200
400
iii
$\tilde{\omega}_{background} = 0.059$
$\tilde{\omega}_2 = 0.074$ $\tilde{\omega}_1 = 0.267$
count
0
20
40
60
iv
Fraction
0.0
0.1
0.2
0.3
0.4
0.5
SE branch
SE crown
E
i
$H_{FE\text{-}SE\ crown}$
ω_1
ω_2
ii
count
0
200
400
600
Null
Alt
iii
$\tilde{\omega}_{background} = 0.010$
$\tilde{\omega}_2 = 0.074$ $\tilde{\omega}_1 = 0.062$
count
0
200
400
600
-2
0
2
$\omega_1\text{-}\omega_2$
iv
Fraction
0.0
0.1
0.2
0.3
0.4
0.5
FE crown
SE crown

Figure 6


A
* p < 0.05
** p < 0.01
Pro. of tractable sites with base substitutions
0.1
0.2
0.3
1st
2nd
3rd
Codon position
B
Pro. of tractable sites with base substitutions
0.09
0.12
0.15
G|C→A|T
A|T→G|C
C
Ts / Tv
0.64
0.68
0.72
0.76
FEL
SEL
D
Insertions
Deletions
Count (x1000)
2
4
6
8
FEL
SEL
E
Insertions
Count (x100)
0
5
10
15
20
3 6 9 12 15 18 21 24 27 30 33 36 39 42 45 48
Size
F
Deletions
0
5
10
3 6 9 12 15 18 21 24 27 30 33 36 39 42 45 48
Size
G
2
3
4
5
Proportion
0.1
0.2
0.3
0.4
FEL SEL
H
Proportion
0.2
0.3
0.4
G→A C→T
G→C C→G
G→T C→A
I
CC|GG score
0.30
0.35
0.40
TT|AA score
0.60
0.70
0.80
Out FEL SEL
*
**