PLOS BIOLOGY

RESEARCH ARTICLE

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[4], Diego Libkind[5], Jeremy DeVirgilio[6], Amanda Beth Hulfachor[7], Cletus P. Kurtzman[6†], Chris Todd Hittinger[2]*, Antonis Rokas[1]*

1 Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, United States of America, **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, United States of America, **3** Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Centre, South China Agricultural University, Guangzhou, China, **4** University of Ljubljana Biotechnical Faculty, Department of Food Science and Technology, University of Ljubljana, Ljubljana, Slovenia, **5** Laboratorio de Microbiología Aplicada, Biotecnología y Bioinformática, Instituto Andino Patagónico de Tecnologías Biológicas y Geoambientales, Universidad Nacional del Comahue-CONICET, San Carlos de Bariloche, Río Negro, Argentina, **6** Mycotoxin Prevention and Applied Microbiology Research Unit, National Center for Agricultural Utilization Research, Agricultural Research Service, United States Department of Agriculture, Peoria, Illinois, United States of America, **7** Laboratory of Genetics, Genome Center of Wisconsin, Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, University of Wisconsin–Madison, Madison, Wisconsin, United States of America

† Deceased.
* cthittinger@wisc.edu (CTH); antonis.rokas@vanderbilt.edu (AR)

Citation: Steenwyk JL, Opulente DA, Kominek J, Shen X-X, Zhou X, Labella AL, et al. (2019) Extensive loss of cell-cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts. PLoS Biol 17(5): e3000255. https://doi.org/10.1371/journal.pbio.3000255

Academic Editor: Sophien Kamoun, The Sainsbury Laboratory, UNITED KINGDOM

Received: February 10, 2019

Accepted: April 18, 2019

Published: May 21, 2019

Data Availability Statement: Phylogenetic and phylogenomic data matrices, single-gene and species-level phylogenies, dN/dS results, individual files used to construct HMMs, as well as the HMMs and a complete gene presence / absence matrix for each yeast species and strain are available through the figshare repository doi: 10.6084/m9.figshare.7670756. Raw sequence read and genome assembly data for all new genomes are available via GenBank BioProject ID PRJNA529215. All sequenced strains have been

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 Guanine–Cytosine (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, a faster-evolving lineage (FEL), which began diversifying approximately 87 million years ago (mya), and a slower-evolving lineage (SEL), which began diversifying approximately 54 mya. Remarkably, both lineages lost genes associated with the cell cycle and genome integrity, but these losses were greater in the FEL. E.g., all species lost the cell-cycle regulator WHIskey 5 (*WHI5*), and the FEL lost components of the spindle checkpoint pathway (e.g., Mitotic Arrest-Deficient 1 [*MAD1*], Mitotic Arrest-Deficient 2 [*MAD2*]) and DNA-damage–checkpoint pathway (e.g., Mitosis Entry Checkpoint 3 [*MEC3*], RADiation sensitive 9 [*RAD9*]). Similarly, both lineages lost genes involved in DNA repair pathways, including the DNA glycosylase gene 3-MethylAdenine DNA Glycosylase 1 (*MAG1*), which is part of the base-excision repair pathway, and the DNA photolyase gene PHotoreactivation Repair deficient 1 (*PHR1*), which is involved in pyrimidine dimer repair. Strikingly, the FEL lost 33 additional genes, including polymerases

publicly deposited in the NRRL, CBS, and/or JCM strain collections.

Funding: 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) (funded by DOE Office of Science BER DE-FC02-07ER64494 and DE-SC0018409 to PI 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.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: *ADI*, Acireductone DIoxygenase; APC, Anaphase-Promoting Complex; *ARO*, AROmatic amino-acid requiring; *ASK1*, Associated with Spindles and Kinetochores 1; AWRI3580, Australian Wine Research Institute 3580; *BAT*, Branched-chain Amino-acid Transaminase; CBS 314, Centraalbureau voor Schimmelcultures 314; *CDC13*, Cell Division Cycle 13; CI, credible interval; *DAD*, Death Upon Overproduction 1 And Death

(i.e., POLymerase 4 [*POL4*] and *POL32*) and telomere-associated genes (e.g., Repressor/activator site binding protein-Interacting Factor 1 [*RIF1*], Replication Factor A 3 [*RFA3*], Cell Division Cycle 13 [*CDC13*], Pbp1p Binding Protein [*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, to our knowledge, 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]. E.g., 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. E.g., the deletion of Mitosis Entry Checkpoint 3 (*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 *C. neoformans* [7,8]. Because of 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 WHIskey 5 (*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 Mitotic Arrest-Deficient 1 (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 SWItching deficient (*SWI*) 4/6 cell-cycle box-binding factor (SBF), a transcription factor that is functionally equivalent but evolutionarily unrelated to E2 promoter binding Factor (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

Upon Overproduction 1 and MonoPolar Spindle 1 interacting; *DAM1*, Duo1 and MonoPolar Spindle 1; DASH, Dam1 Complex; dN, rate of nonsynonymous substitutions; dS, rate of synonymous substitutions; *DSE2*, Daughter-Specific Expression 2; DSM2768, Dutch State Mines 2768; *DSN1*, Dosage Suppressor of Necessary for Nuclear Function 1; *DUO*, Death Upon Overproduction; E2F, E2 promoter binding Factor; Eo, Eocene; *EXO1*, EXOnuclease 1; *FBP1*, Fructose-1,6-BisPhosphatase 1; FEL, faster-evolving lineage; *GAL*, GALactose metabolism; GC, Guanine–Cytosine; *GDH*, Glutamate DeHydrogenase; gDNA, genomic DNA; *GIP1*, Glycogen 7-Interacting Protein 1; *GLT*, GLuTamate synthase; GO, gene ontology; HMM, Hidden Markov Model; *HSK3*, Helper of ASK1 3; *IMA*, IsoMAltase; *IME1*, Inducer of MEiosis 1; LRT, likelihood ratio test; *MAD1*, Mitotic Arrest-Deficient 1; *MAG1*, 3-MethylAdenine DNA Glycosylase 1; *MAL*, MALtose fermentation; *MALx*, MALtose fermentation locus 1 or 3; MCM, Mini-Chromosome Maintenance; *MDE*, Methylthioribulose-1-phosphate DEhydratase; *MEC3*, Mitosis Entry Checkpoint 3; *MEU*, Multicopy Enhancer of Upstream activation site; MIND, Mis TWelve-like 1 protein Including Necessary for Nuclear Function 1 protein,Nnf1 Synthetic Lethal 1 protein, Dosage Suppressor of Necessary for Nuclear Function 1 protein complex; Mio., Miocene; *MND2*, Meiotic Nuclear Divisions 2; *MRC1*, Mediator of the Replication Checkpoint 1; *MRI*, MethylthioRibose-1-phosphate Isomerase; *MTW1*, Mis TWelve-like 1; mya, million years ago; *NNF1*, Necessary for Nuclear Function 1; NRRL, Northern Regional Research Laboratory; *NSL1*, Nnf1 Synthetic Lethal 1; OG, orthologous gene; Oligo., Oligocene; ORC, Origin Recognition complex; Paleo, Paleocene; *PBP2*, Pbp1p Binding Protein 2; *PCD1*, Peroxisomal Coenzyme A Diphosphatase 1; *PCK1*, Phosphoenolpyruvate CarboxyKinase 1; *PCL1*, Pho85 CycLin 1; *PDS1*, Precocious Dissociation of Sisters 1; PHO, PHOsphate metabolism; *PHR1*, PHotoreactivation Repair deficient; Pleisto., Pleistocene; Plio., Pliocene; *POL*, POLymerase; Quat, Quaternary; *RAD9*, RADiation sensitive 9; *RFA3*, Replication Factor A; *RFX1*, Regulatory Factor X 1; *RIF1*, Repressor/activator site binding protein-Interacting Factor 1; *SAM*, S-AdenosylMethionine requiring; SBF, Switching deficient 4/6 cell-cycle box-binding factor; SEL, slower-evolving lineage; *SGS1*, Slow Growth Suppressor 1; *SIC1*, Substrate/Subunit Inhibitor of Cyclin-dependent protein kinase 1; *SNO*, SNooze proximal Open reading frame; *SPC*, Spindle Pole Component; *SPE*, SPErmidine auxotroph; *SPO12*, SPOrulation 12; *SSP1*,

composition and length of telomeres [16]. E.g., *MEC3* and RADiation sensitive 9 (*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]. E.g., deletion of the DNA polymerase δ subunit gene, POLymerase 32 (*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 3-MethylAdenine DNA Glycosylase 1 (*MAG1*), a gene encoding a DNA glycosylase that removes damaged bases via the multistep 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 multistep DNA repair pathways, other DNA repair genes function individually or are parts of simpler regulatory processes. E.g., PHotoreactivation Repair-deficient 1 (*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. E.g., Peroxisomal Coenzyme A Diphosphatase 1 (*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]. E.g., 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], *C. neoformans* [7,8], and *Candida glabrata* [30], hypermutation is thought to contribute to within-host adaptation, which may involve modulating traits such as drug resistance [6,30]. However, as adaptation to a new environment increases, hypermutator alleles are expected to decrease in frequency because of the accumulation of deleterious mutations that result as a consequence of the high mutation rate [31,32]. In agreement with this prediction, half of the experimentally evolved hypermutating lines of *S. cerevisiae* had reduced mutation rates after a few thousand generations [33], 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; e.g., all *S. cerevisiae* hypermutating lines increased their ploidy, presumably to reduce the impact of higher mutation rates [33]. 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 into the evolution of mutation rates 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 [34–36], which are reminiscent of the very long branches of fungal hypermutator strains [6–8]. Most of what is known about these cosmopolitan yeasts relates to their high abundance on mature fruits and in fermented beverages [37],

Sporulation-specific protein 1; *SUC2*, SUCrose 2; *SWM1*, Spore Wall Maturation 1; *TDP1*, Tyrosyl-DNA Phosphodiesterase 1; *THI*, THIamine regulon; UTAD222, University of Trás-os-Montes and Alto Douro 222; *UTR*, Unidentified TRanscript; *WHI5*, WHIskey 5; *YKU70*, Yeast KU protein 70.

especially on grapes and in wine must [38,39]. 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 [40]. 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 [41], 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 showed that species in the genus *Hanseniaspora* lost many genes involved in diverse processes and delineated two lineages within the genus: a faster-evolving lineage (FEL), which has a strong signature of acceleration in evolutionary rate at its stem branch and has lost many additional genes involved in diverse processes, and a slower-evolving lineage (SEL), which has a weaker signature of evolutionary rate acceleration at its stem branch and underwent fewer gene losses. 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. Relaxed molecular clock analyses estimate that the FEL and SEL split approximately 95 million years ago (mya). The degree of evolutionary rate acceleration is commensurate with the preponderance of loss of genes associated with cell-cycle and DNA repair processes. 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 the 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. E.g., 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., Repressor/activator site binding protein-Interacting Factor 1 [*RIF1*], Replication Factor A 3 [*RFA3*], Cell Division Cycle 13 [*CDC13*], Pbp1p Binding Protein 2 [*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, to our knowledge, 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 orthologous genes (OGs) (522,832 sites; 100% taxon-occupancy) yielded a robust phylogeny of the genus *Hanseniaspora* (Fig 1A, S1 and S2 Figs). Consistent with previous analyses [35,36,42], our phylogeny identified two major lineages, each of which had a long stem branch; we hereafter refer to the lineage with the longer stem branch as the FEL and to the other as the SEL. Relaxed molecular clock analysis suggests that the FEL and SEL split 95.34 (95% credible interval [CI]:

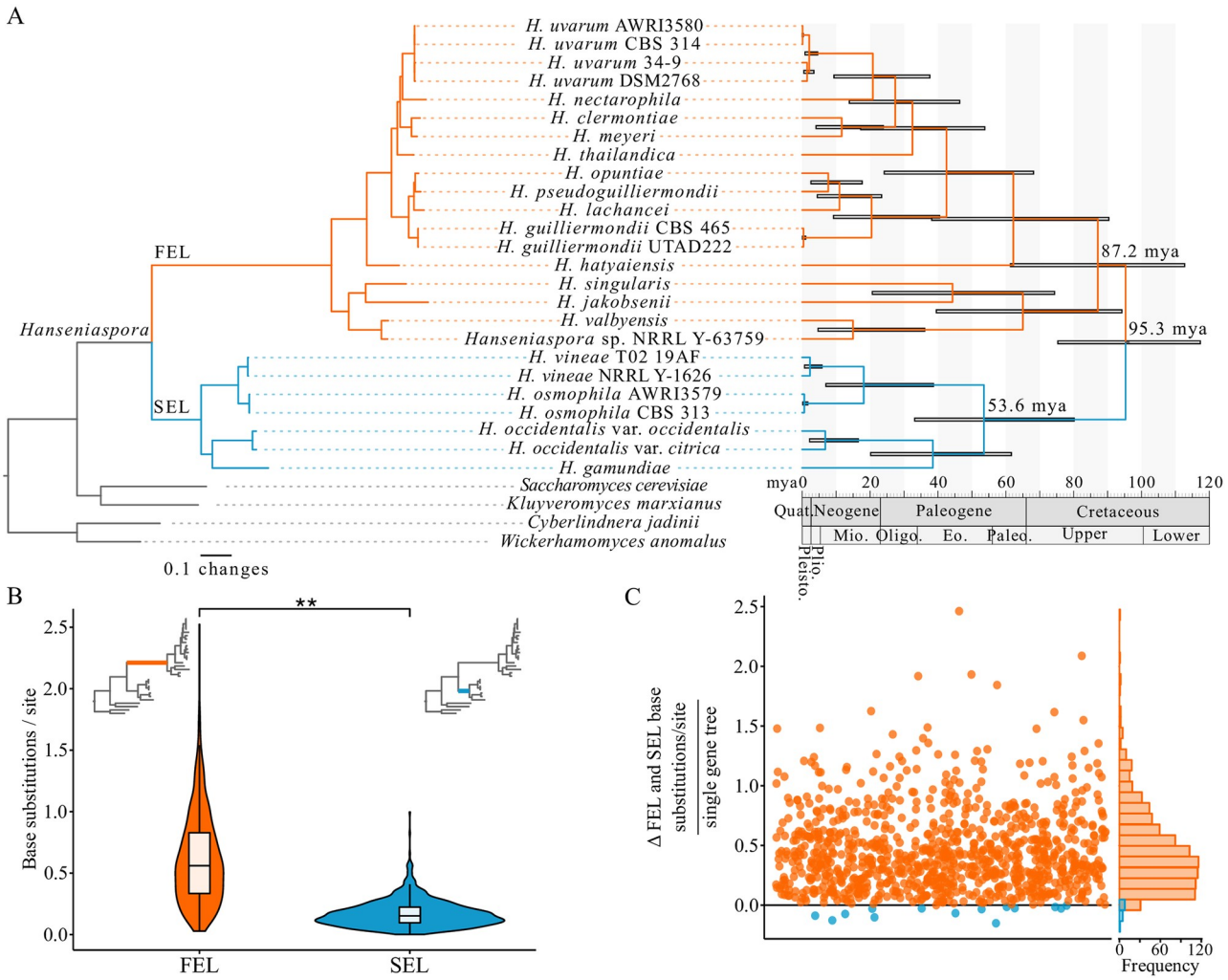

Fig 1. The evolutionary history, rate, and timeline of *Hanseniaspora* diversification. (A) Phylogenomic and relaxed molecular clock analysis of 1,034 single-copy OGs from a near-complete set of *Hanseniaspora* species revealed two well-supported lineages termed the FEL and SEL, which began diversifying around 87.2 and 53.6 mya after diverging 95.3 mya. (B) Among single-gene phylogenies in which the FEL and SEL were monophyletic ($n = 946$), the FEL stem branch was consistently and significantly longer (0.62 ± 0.38 base substitutions/site) than the SEL stem branch (0.17 ± 0.11 base substitutions/site) ($p < 0.001$; paired Wilcoxon rank–sum test). (C) Examination of the difference between FEL and SEL: stem branch lengths per single-gene tree revealed that 932 single-gene phylogenies had a longer FEL stem branch (depicted in orange with values greater than 0), while only 14 single-gene phylogenies had a longer SEL stem branch (depicted in blue with values less than 0). Across all single-gene phylogenies, the average difference in stem branch length between the two lineages was 0.45. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. AWRI, Australian Wine Research Institute; CBS, Centraalbureau voor Schimmelcultures; DSM2768, Dutch State Mines 2768; Eo., Eocene; FEL, faster-evolving lineage; Mio., Miocene; mya, million years ago; NRRL, Northern Regional Research Laboratory; OG, orthologous gene; Oligo., Oligocene; Paleo., Paleocene; Pleisto., Pleistocene; Plio., Pliocene; Quat., Quaternary; SEL, slower-evolving lineage; UTAD222, University of Trás-os-Montes and Alto Douro 222.

https://doi.org/10.1371/journal.pbio.3000255.g001

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, S3 Fig and S2 File).

The FEL stem branch is much longer than the SEL stem branch in the *Hanseniaspora* phylogeny (Fig 1) (see also phylogenies in [35,36]). 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 in which each lineage was inferred to be 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 Guanine–Cytosine (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 (S4 Fig). 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 from one another ($\chi^2(2) = 30.00$, $p < 0.001$; Kruskal–Wallis rank–sum test). Pairwise comparisons of GC contents between the FEL, SEL, and all other Saccharomycotina were not significant, except in the comparison between the FEL and other Saccharomycotina yeasts ($p < 0.001$; Dunn's test for multiple comparisons with Benjamini–Hochberg multitest 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). Pairwise comparisons of genome size and gene number between the FEL, SEL, and all other Saccharomycotina revealed that the only significant difference for genome size was between the FEL and other Saccharomycotina yeasts ($p < 0.001$; Dunn's test for multiple comparisons with Benjamini–Hochberg multitest 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 multitest 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 (S5 Fig) 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 Hidden Markov Model (HMM)-based sequence similarity searches using annotated *S. cerevisiae* genes as queries in HMM construction (see Methods) (S6 Fig). Because we were most interested in broad patterns of gene losses in the FEL and SEL, we focused our analyses on genes lost in at least two-thirds of each lineage (i.e., $\geq$ 11 FEL taxa or $\geq$ 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 were uniquely lost in the FEL, and 23 genes were uniquely lost in the SEL (S3 File).

To identify the likely functions of genes lost from each lineage, we conducted gene ontology (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, $p < 0.001$) (S4 File). Additional terms, such as CELL CYCLE, GO:0007049 ($p < 0.001$), CHROMOSOME SEGREGATION, GO:0007059 ($p < 0.001$), CHROMOSOME

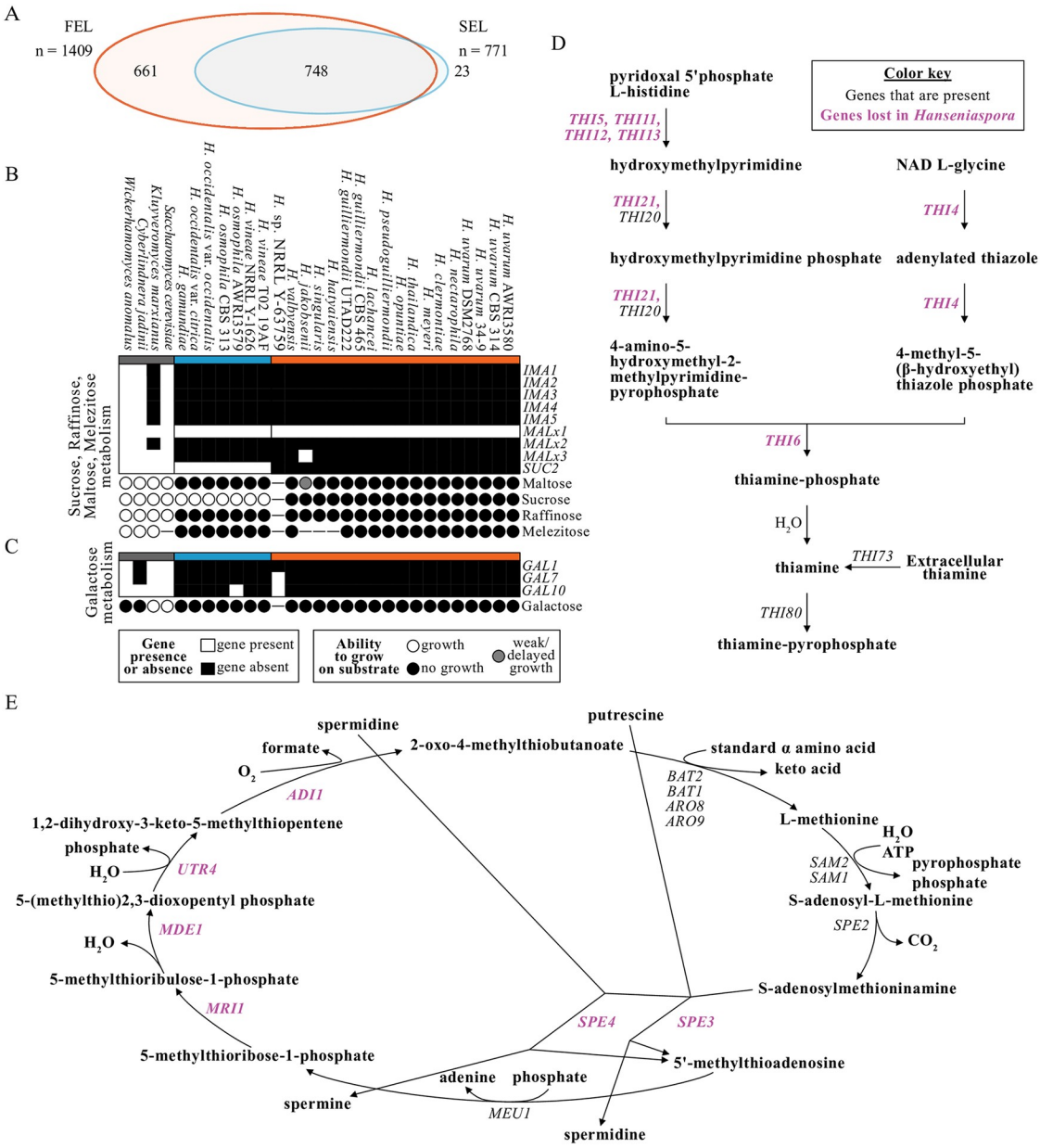

Fig 2. Gene presence and absence analyses reflect phenotype and reveal disrupted pathways. (A) Examination of gene presence and absence (see Methods) revealed numerous genes that were lost across *Hanseniaspora*. Specifically, 1,409 were lost in the FEL, and 771 genes were lost in the SEL. A Euler diagram 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 absent 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 *H. 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 absent 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, the ability to weakly grow/exhibit delayed growth on a given substrate, or the inability to grow is specified using white, gray, 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. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. *ADI*, Acireductone Dioxygenase; *ARO*, AROmatic amino-acid requiring; AWRI, Australian Wine Research Institute; *BAT*, Branched-chain Amino-acid Transaminase; CBS, Centraalbureau voor Schimmelcultures; DSM2768, Dutch State Mines 2768; FEL, faster-evolving lineage; *GAL*, GALactose

metabolism; *IMA*, IsoMAltase; *MAL*, MALtose fermentation; *MDE*, Methylthioribulose-1-phosphate DEhydratase; *MEU*, Multicopy Enhancer of Upstream activation site; *MRI*, MethylthioRibose-1-phosphate Isomerase; NRRL, Northern Regional Research Laboratory; *SAM*, S-AdenosylMethionine requiring; SEL, slower-evolving lineage; *SUC2*, SUCrose; *THI*, THIamine regulon; UTAD222, University of Trás-os-Montes and Alto Douro 222; *UTR*, Unidentified Transcript.

https://doi.org/10.1371/journal.pbio.3000255.g002

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, the 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 IsoMAltase (*IMA*) gene family and the MALtose fermentation (*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 MALtose fermentation locus 3 (*MALx3*), which encodes the *MAL*-activator protein [46], has been lost in all but one species (*H. 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 and S5 File). The growth of *H. jakobsenii* on maltose may be due to a cryptic α-glucosidase gene or represent a false positive because *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 SUCrose 2 (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 and S5 File). 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 (GALactose metabolism 1 [*GAL1*], *GAL7*, and *GAL10*) from their genomes (Fig 2C and S5 File). Additionally, all *Hanseniaspora* genomes appear to have lost two key genes, Phosphoenolpyruvate CarboxyKinase 1 (*PCK1*) and Fructose-1,6-BisPhosphatase 1 (*FBP1*), encoding enzymes in the gluconeogenesis pathway (S7A Fig); in contrast, all *Hanseniaspora* have an intact glycolysis pathway (S7B Fig).

Altogether these metabolism-associated gene losses may reflect *Hanseniaspora* ecology. More specifically, among wine strains of *S. cerevisiae*, genes associated with maltose and thiamine metabolism are frequently absent in their genomes [48,49] and are thought to reflect their ecology in the grape must environment [50]. Interestingly, similar gene losses are observed among *Hanseniaspora* species but are often more pronounced; e.g., *Hanseniaspora* species lack most of the thiamine biosynthesis pathway, while wine strains of *S. cerevisiae* typically lack a single member of the THIamine regulon (*THI*) gene family.

Manual examination of other metabolic pathways revealed that *Hanseniaspora* genomes are also lacking some of their key genes. E.g., we found that THIAMINE BIOSYNTHETIC PROCESS, GO:0009228 ($p = 0.003$), was an over-represented GO term among genes absent in both the FEL and SEL because of the absence of *THI* and SNooze proximal Open reading frame (*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] (S5 File). 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 [40]. 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 Glutamate DeHydrogenase 1 (*GDH1*) and Glutamate DeHydrogenase 3 (*GDH3*) from the glutamate biosynthesis pathway from ammonium are absent in FEL yeasts (S3 File). However, *Hanseniaspora* have GLuTamate synthase 1 (*GLT1*), which enables glutamate biosynthesis from glutamine.

Cell-cycle– and genome-integrity–associated gene losses. Many genes involved in the cell cycle and genome integrity, including cell-cycle checkpoint genes, have been lost across *Hanseniaspora* (Fig 3). E.g., *WHI5* and Daughter-Specific Expression 2 (*DSE2*), which are responsible for repressing the start (i.e., an event that determines cells have reached a critical size before beginning division) [51] and help facilitate daughter–mother cell separation through cell wall degradation [52], have been lost in both lineages. Additionally, the FEL has lost the entirety of the Dam1 complex (or DASH complex) (i.e., Associated with Spindles and Kinetochores 1 [*ASK1*], Death Upon Overproduction 1 [*DUO1*] And Duo1 and MonoPolar Spindle 1 [*MPS1*] [*DAM1*] interacting [DAD] 1, *DAD2*, *DAD3*, *DAD4*, *DUO1*, *DAM1*, Helper of ASK1 3 [*HSK3*], Spindle Pole Component [SPC] 19, and *SPC34*), which forms part of the kinetochore and functions in spindle attachment and stability as well as chromosome segregation, and the Mis TWelve-like 1 (Mtw1) protein Including Necessary for Nuclear Function 1 (Nnf1) protein–Nnf1 Synthetic Lethal 1 (Nsl1) protein–Dosage Suppressor of NNF1 (Dsn1) protein (MIND) complex (i.e., *MTW1*, *NNF1*, *NSL1*, and *DSN1*), which is required for kinetochore biorientation and accurate chromosome segregation (S3 and S4 Files). 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 APC, a major multi-subunit regulator of the cell cycle, are lost in both lineages (i.e., *CDC26* and Meiotic Nuclear Divisions 2 [*MND2*]) or just the FEL (i.e., *APC2*, *APC4*, *APC5*, and Spore Wall Maturation 1 [*SWM1*]).

Another group of genes that have been lost in *Hanseniaspora* are genes associated with the DNA damage checkpoint and DNA damage sensing. E.g., both lineages have lost Regulatory Factor X1 (*RFX1*), which controls a late point in the DNA-damage–checkpoint pathway [53], 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 [54]. 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 (S8 Fig). 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 APC, whose dysregulation is thought to increase instances of aneuploidy [55].

Lastly, we examined losses among genes related to meiosis. Although little is known about meiosis and sexual reproduction in *Hanseniaspora*, recent attempts to induce sporulation and sexual reproduction in different *Hanseniaspora* species have been unsuccessful [37,41,56,57]. In contrast, other species (i.e., *H. thailandica*, *H. singularis*, and *H. gamundiae*) are able to sporulate [42,58]. These inconsistences may be due to the infrequency of sporulation or reduced total number of spores produced, which may be linked to the losses of genes associated with coordinating meiosis such as the major regulator Inducer of MEiosis 1 (*IME1*) [59]

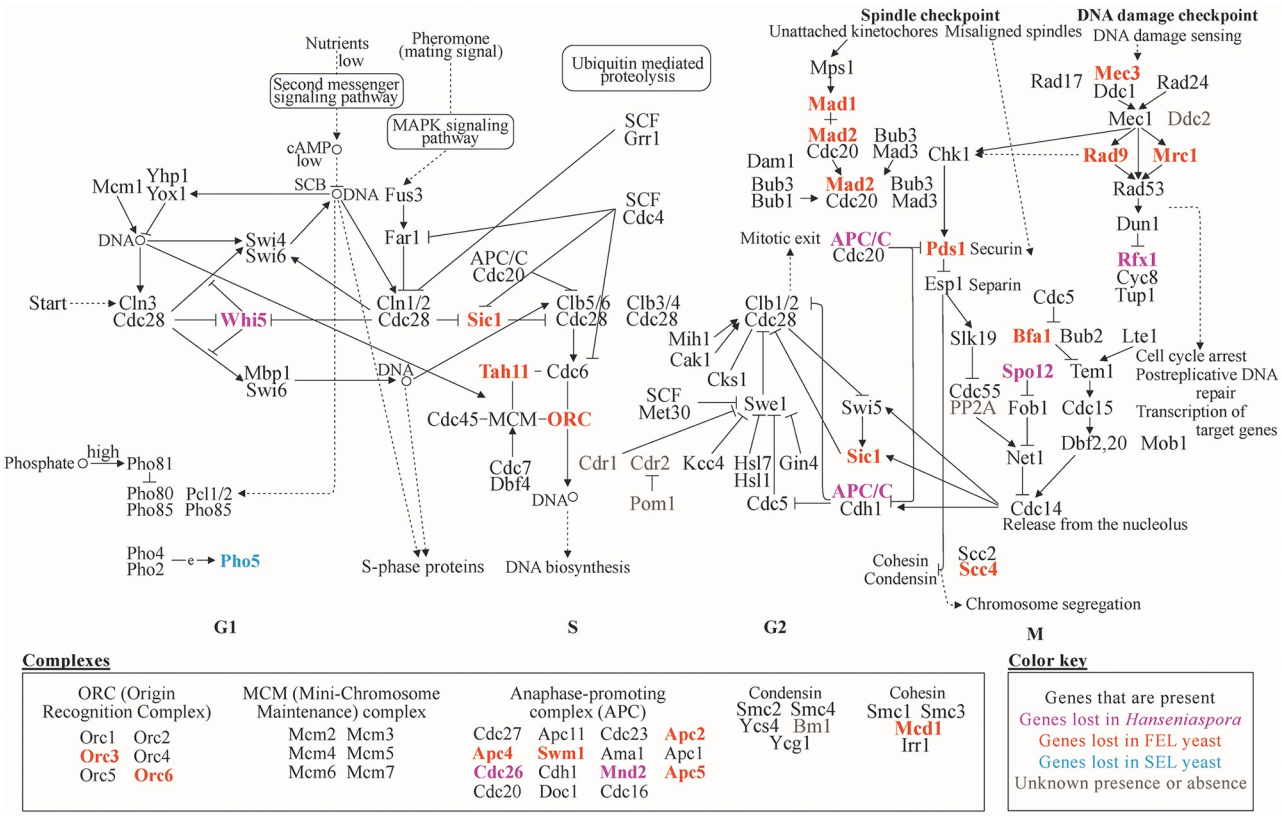

Fig 3. Gene presence and absence in the budding yeast cell cycle. Examination of cell-cycle genes revealed numerous genes that are absent in *Hanseniaspora* genomes. The genes not present in *Hanseniaspora* participate in diverse functions and include key regulators such as *WHI5*, components of spindle checkpoint processes and segregation such as *MAD1* and *MAD2*, and components of DNA-damage–checkpoint processes such as *MEC3*, *RAD9*, and *RFX1*. Genes absent 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. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. Ama, Activator of meiotic anaphase-promoting complex; APC (or APC/C), Anaphase-Promoting Complex; Bfa, Byr-four-alike; Bm1, Biomimetic moiety glutathionesulfonic acid; Bub, Budding uninhibited by benzimidazole; Cak1, Cyclin-dependent kinase-activating kinase; cAMP, cyclic AdenosineMonoPhosphate; Cdc, Cell division cycle; Cdh, CDC20 homolog; Cdr, *Candida* drug resistance; Chk, Checkpoint kinase; Cks, Cdc28 kinase subunit; Clb, Cyclin B; Cln, Cyclin; Cyc, Cytochrome C; Dam, Duo1 and Mps1 interacting; Dbf, Dumbbell former; Ddc, DNA Damage Checkpoint; Doc, Destruction of Cyclin B; Dun, DNA-damage UNinducible; Esp1, Extra spindle pole bodies 1; Far1, Factor ARrest; FEL, faster-evolving lineage; Fob, Fork Blocking less; Fus3, cell fusion 3; Gin4, Growth inhibitory 4; Grr, Glucose repression-resistant; Hsl, Histone synthetic lethal; Irr, Irregular cell behavior; Kcc, K^+-Cl^- cotransporters; Lte, Low temperature essential; *MAD*, Mitotic Arrest-Deficient; MAPK, Mitogen-Activated Protein Kinase; Mbp, Mlul-box–binding protein; Mcd, Mitotic chromosome determinant; MCM, Mini-Chromosome Maintenance; *MEC3*, Mitosis Entry Checkpoint 3; Met30, Methionine requiring 30; Mih1, Mitotic inducer homolog; Mnd, Meiotic nuclear divisions; Mob, Mps one binder; Mps, Monopolar spindle; Mrc, Mediator of the Replication Checkpoint; Net, Nucleolar silencing establishing factor and telophase regulator; ORC, Origin Recognition Complex; Pds, Precocious Dissociation of Sisters; PHO, PHOsphate; Pom, Polarity misplaced; PP2A, Protein Phosphatase 2A; *RAD9*, RADiation sensitive; *RFX1*,; SCB, Swi4,6-dependent cell ycle box; Scc, Sister Chromatid Cohesion; SCF, S-phase kinase-associated protein, Cullin, F-box containing complex; SEL, slower-evolving lineage; Sic, Sucrose NonFermenting; Slk, Synthetic lethal karyogamy; Smc, Stability of minichromosomes; Spo, Sporulation; Swe, *Saccharomyces* Wee1; Swi, Switching deficient; Swm, Spore Wall Maturation; Tah11, Topo-A Hypersensitive; Tem, Termination of M phase; Tup, deoxythymidine monophosphate-uptake; *WHI5*, WHIskey 5; Ycg, Yeast cap G; Ycs, Yeast condensing subunit; Yhp1, Yeast Homeo-Protein 1; Yox1, Yeast homeobox 1.

https://doi.org/10.1371/journal.pbio.3000255.g003

and genes associated with spore formation such as Sporulation-specific protein 1 (*SSP1*) [60] and Glycogen 7-Interacting Protein 1 (*GIP1*) [61] (S9 Fig).

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* [62], the

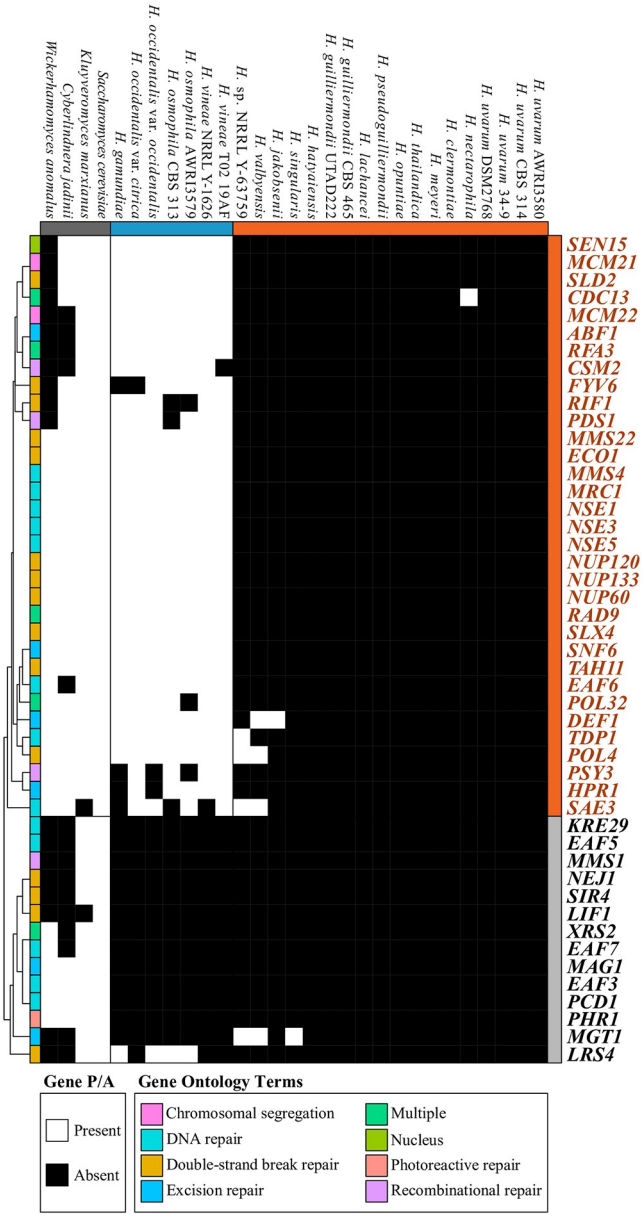

Fig 4. A panoply of genome-maintenance and DNA repair genes are absent among *Hanseniaspora*, especially in the FEL. Genes annotated as DNA repair genes according to GO (GO:0006281) and child terms were examined for presence and absence in at least two-thirds of each lineage, respectively (268 total genes). 47 genes are absent among the FEL species, and 14 genes are absent 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. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. *ABF1*, Autonomously replicating sequence-Binding Factor 1; AWRI, Australian Wine Research Institute; CBS, Centraalbureau voor Schimmelcultures; *CDC13*, Cell Division Cycle 13; *CSM2*, Chromosome Segregation in Meiosis 2; *DEF1*, RNA polymerase II Degradation Factor 1; DSM2768, Dutch State Mines 2768; *EAF6*, Essential something about silencing 2-related acetyltransferase 1-Associated Factor 6; *ECO1*, Establishment of Cohesion 1; FEL, faster-evolving lineage; *FYV6*, Function required for Yeast Viability 6; GO, gene ontology; *HPR1*, HyPerRecombination 1; *KRE29*, Killer toxin Resistant 29; *LIF1*, Ligase Interacting Factor 1; *LRS4*, Loss of RDNA Silencing 4; *MAG1*, 3-MethylAdenine DNA Glycosylase 1; *MCM21*, Mini-Chromosome Maintenance 21; *MGT1*, O-6-MethylGuanine-DNA methylTransferase 1; *MMS22*, Methyl MethaneSulfonate sensitivity 22; *MRC1*, Mediator of the Replication Checkpoint 1; *NEJ1*, Nonhomologous End-Joining defective 1; NRRL, Northern Regional Research Laboratory; *NSE1*, NonStructural maintenance of chromosomes Element 1; *NUP120*, NUclear Pore 120; *PCD1*, Peroxisomal Coenzyme A Diphosphatase 1; *PDS1*, Precocious Dissociation of Sisters 1; *PHR1*, PHotoreactivation Repair deficient 1; *POL32*, POLymerase 32; *PSY3*, Platinum SensitivitY 3; P/A, presence or absence; *RAD9*, RADiation sensitive 9; *RFA3*,

Replication Factor A 3; *RIF1*, Repressor/activator site binding protein-Interacting Factor 1; *SAE3*, Sporulation in the Absence of sporulation Eleven; SEL, slower-evolving lineage; *SEN15*, Splicing ENdonuclease 15; *SIR4*, Silent Information Regulator 4; *SLD2*, Synthetically Lethal with DNA polymerase B (II)-1 2; *SLX4*, Synthetical Lethal of unknown (X) function 4; *SNF6*, Sucrose NonFermenting 6; *TAH11*, Topo-A Hypersensitive 11; *TDP1*, Tyrosyl-DNA Phosphodiesterase 1;UTAD222, University of Trás-os-Montes and Alto Douro 222; *XRS2*, X-Ray Sensitive 2.

https://doi.org/10.1371/journal.pbio.3000255.g004

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 Tyrosyl-DNA Phosphodiesterase 1 (*TDP1*), which repairs damage caused by topoisomerase activity [63]; the DNA polymerase gene *POL32*, which participates in base-excision and nucleotide-excision repair and whose null mutants have increased genomic deletions [20]; and the *CDC13* gene, which encodes a telomere-capping protein [64].

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 and 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 ($H_{FEL–SEL\ branch}$; Fig 5B), revealed that 678 genes (68.55% of examined genes) significantly rejected the null hypothesis (Table 1; $\alpha = 0.01$; likelihood ratio test [LRT]; median FEL stem branch $\omega = 0.57$, median SEL stem branch $\omega = 0.29$, and median background $\omega = 0.060$). 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_{FEL}; Fig 5C) revealed 743

Table 1. Rate of sequence evolution hypotheses and results.

Hypotheses for Interlineage Comparisons	Parameters	Fraction of Genes Significantly Different from H_O	Median ω Values		
			$\omega_{background}$	ω_1	ω_2
H_O: Uniform rate for all branches Fig 5A	Single ω value	N/A	N/A	N/A	N/A
$H_{FEL–SEL\ branch}$: Unique rates for FEL and SEL stem Fig 5B	$\omega_{background} \neq \omega_1 \neq \omega_2$ ω_1 = FEL stem branch ω_2 = SEL stem branch	678 genes (68.55% of examined genes)	0.060	0.566	0.293
H_{FEL}: Unique rates for FEL stem and FEL crown Fig 5C	$\omega_{background} \neq \omega_1 \neq \omega_2$ ω_1 = FEL stem branch ω_2 = FEL crown branches	743 genes (75.13% of examined genes)	0.063	0.711	0.061
H_{SEL}: Unique rates for SEL stem and SEL crown Fig 5D	$\omega_{background} \neq \omega_1 \neq \omega_2$ ω_1 = SEL stem b≠anch ω_2 = SEL crown branches	528 genes (53.7% of examined genes)	0.059	0.267	0.074
$H_{FEL–SEL\ crown}$: Unique rates for FEL crown and SEL crown Fig 5E	$\omega_{background} \neq \omega_1 \neq \omega_2$ ω_1 = FEL crown branches ω_2 = SEL crown branches	717 genes (72.5% of examined genes)	0.010	0.062	0.074

Abbreviations: FEL, faster-evolving lineage; SEL, slower-evolving lineage.

https://doi.org/10.1371/journal.pbio.3000255.t001

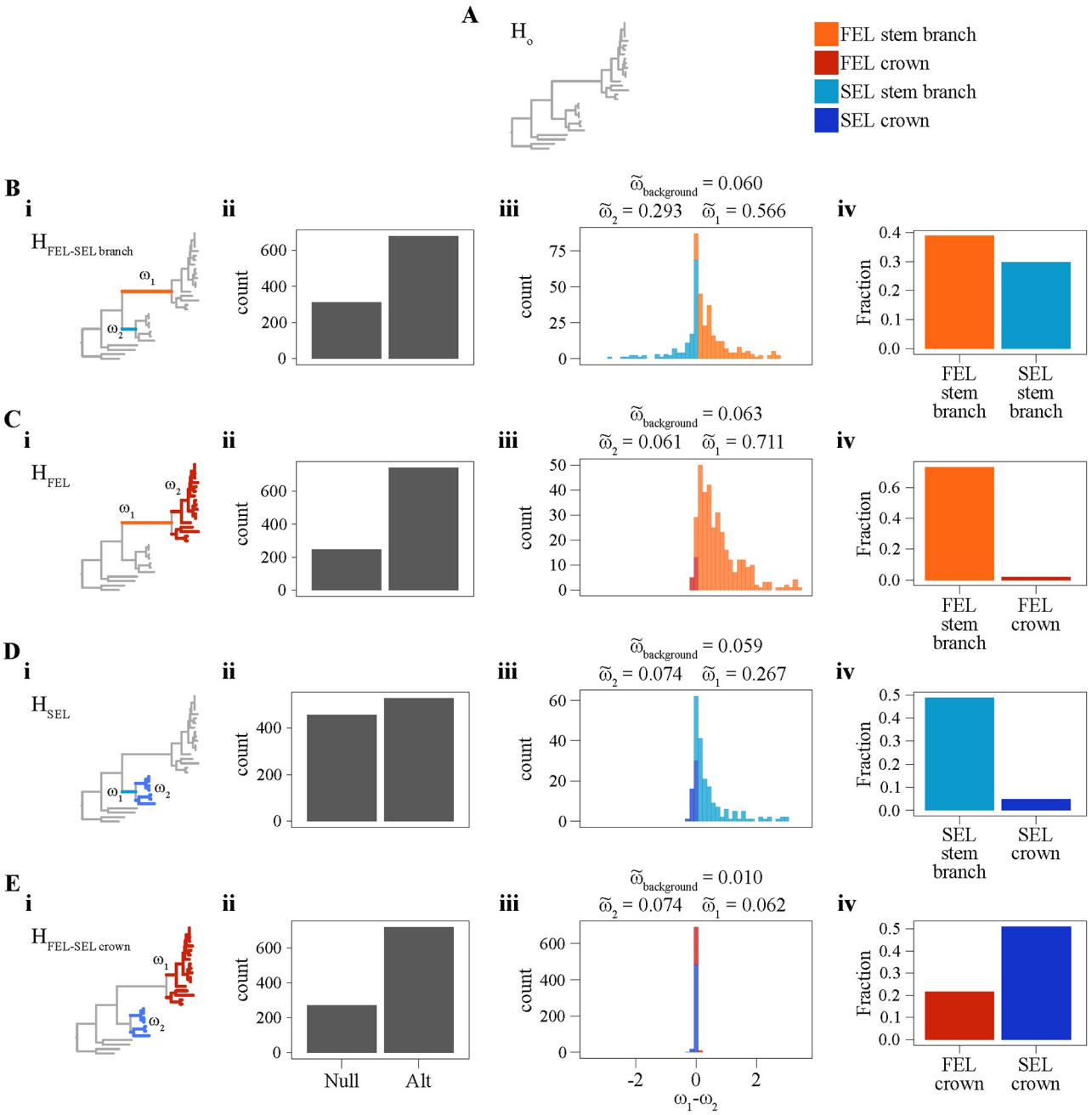

Fig 5. dN/dS (ω) analyses support a historical burst of accelerated evolution in the FEL. (A) The null hypothesis (H_O) that all branches in the phylogeny have the same ω value. Alternative hypotheses (B–E) evaluate ω along three sets of branches. (Bi) The alternative hypothesis ($H_{FEL-SEL\ branch}$) examined ω values along the FEL and SEL stem branches. (Bii) 311 (31.45%) genes supported H_O, and 678 (68.55%) genes supported $H_{FEL-SEL\ branch}$. (Biii) Among the genes that supported $H_{FEL-SEL\ branch}$, we examined the distribution of the difference between ω_1 and ω_2 as specified in part Bi. Here, a range of $\omega_1-\omega_2$ 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) 384 (38.83%) genes significantly rejected H_O and were faster in the FEL than the SEL, while 237 (23.96%) significantly rejected H_O and were faster in the SEL than the FEL. (Ci) The alternative hypothesis (H_{FEL}) examined ω values along the FEL stem branch (ω_1) and crown branches (ω_2). (Cii) 246 (24.87%) genes supported H_O, and 743 (75.13%) genes supported H_{FEL}. (Ciii) Among the genes that supported H_{FEL}, 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) 725 (73.31%) genes significantly rejected H_O and had higher ω_1 values than ω_2 values, while 18 (1.82%) genes significantly rejected H_O and had higher ω_2 than ω_1 values. (Di) The alternative hypothesis (H_{SEL}) examined ω values along the SEL stem branch (ω_1) and crown branches (ω_2). (Dii) 455 (46.29%) genes supported H_O, and 528 (53.71%) genes supported H_{SEL}. (Diii) Among the genes that supported H_{SEL}, 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) 481 (48.93%) genes significantly rejected H_O and had higher ω_1 than ω_2 values, while 47 (4.78%) genes significantly rejected H_O and had higher ω_2 than ω_1 values. (Ei) The alternative hypothesis

($H_{FEL–SEL\ crown}$) examined ω values in the FEL crown branches (ω_1) and SEL crown branches (ω_2). (Eii) 272 (27.50%) genes supported H_O, and 717 (72.50%) genes supported $H_{FEL–SEL\ crown}$. (Eiii) Among the genes that supported $H_{FEL–SEL\ crown}$, 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) 481 (21.54%) genes significantly rejected H_O and had higher ω_1 than ω_2 values, while 504 (50.96%) genes had higher ω_2 than ω_1 values. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. dN, rate of nonsynonymous substitutions; dS, rate of synonymous subsitutions; FEL, faster-evolving lineage; SEL, slower-evolving lineage.

https://doi.org/10.1371/journal.pbio.3000255.g005

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_{SEL}; 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 ($H_{FEL–SEL\ crown}$; 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 ($F(1) = 196.88$, $p < 0.001$; multifactor 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).

We next investigated differences in the direction of substitutions. Specifically, we examined if substitutions were biased in the AT direction (i.e., G|C → A|T) or GC direction (i.e., A|T → G|C) as well as whether there are differences among substitutions in these directions between the FEL and the SEL. We observed significant differences among substitutions in the AT and GC directions between the FEL and the SEL ($F(1) = 447.1$, $p < 0.001$; multifactor ANOVA), as well as between overall AT and GC bias across both lineages among G|C (n = 232,546) and A|T (n = 385,157) sites ($F(1) = 914.5$, $p < 0.001$; multifactor ANOVA) (Fig 6B). There were 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).

We next examined patterns of transition/transversion ratios and observed 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, a significant AT bias in base substitutions across all *Hanseniaspora*, and a low transition/transversion ratio across the FEL and SEL.

Examination of indels revealed that the total number of insertions or deletions was significantly greater in the FEL ($mean_{insertions}$ = 7,521.11 ± 405.34; $mean_{deletions}$ = 3,894.11 ± 208.16) compared to the SEL ($mean_{insertions}$ = 6,049.571 ± 155.85; $mean_{deletions}$ = 2,346.71 ± 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

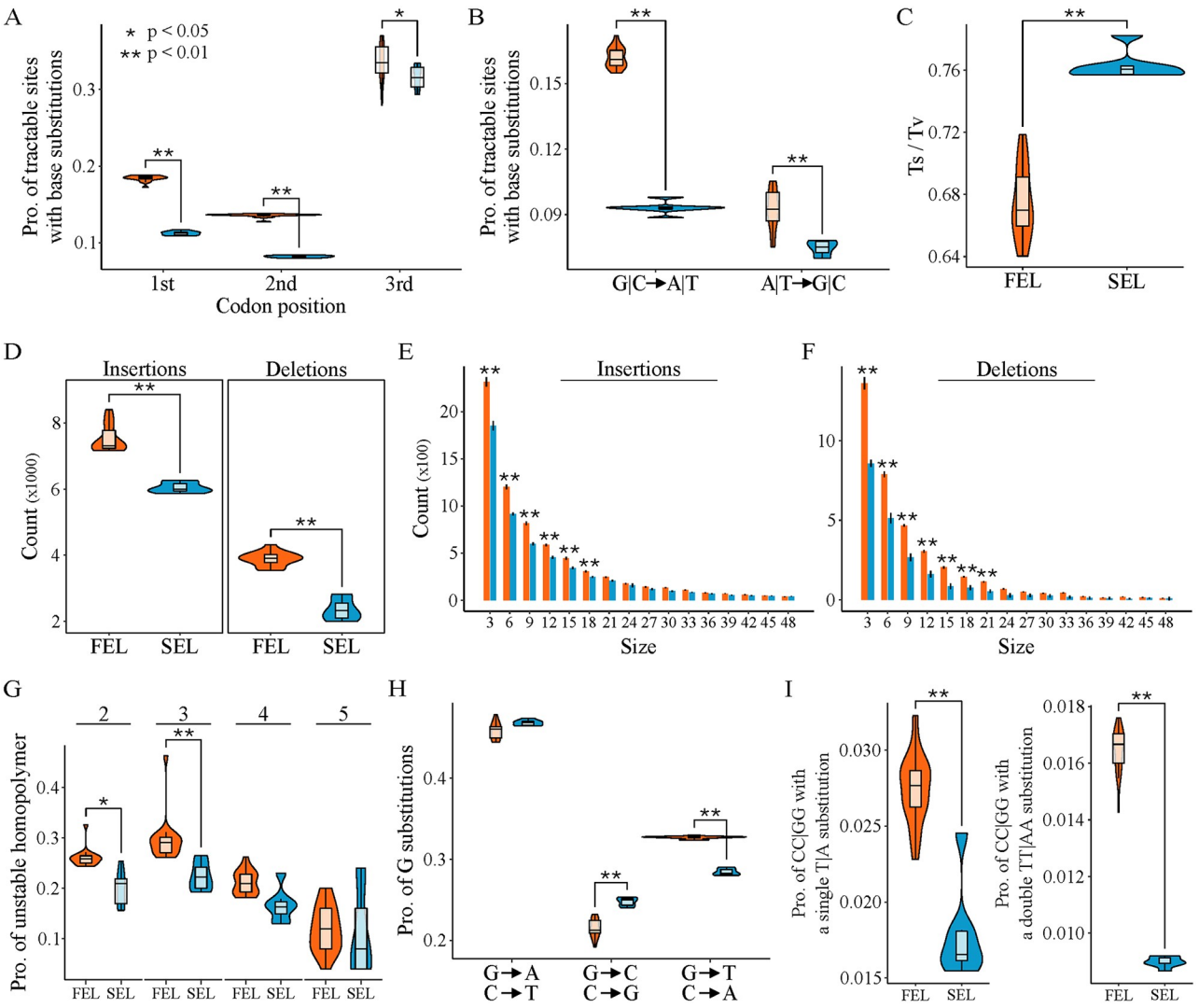

Fig 6. Analyses of base substitutions and indels reveal a higher mutational load in the FEL compared to the SEL. (A) Analyses of substitution patterns among codon-based alignments of 1,034 OGs revealed a higher number of base substitutions in the FEL compared to the SEL ($F(1) = 196.88$, $p < 0.001$; multifactor ANOVA) and an asymmetric distribution of base substitutions at codon sites ($F(2) = 1{,}691.60$, $p < 0.001$; multifactor ANOVA). A Tukey honest significance differences post hoc test revealed a higher proportion of substitutions in the FEL compared to the SEL 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) revealed significant differences between the FEL and SEL ($F(1) = 447.1$, $p < 0.001$; multifactor ANOVA) as well as differences in the directionality of base substitutions ($F(1) = 914.5$, $p < 0.001$; multifactor 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 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 sites with base substitutions in the FEL compared to the SEL ($p < 0.001$). Specifically, a higher number of base substitutions was 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 ($\bar{X}_{insertions} = 7,521.11 \pm 405.34$; $\bar{X}_{deletions} = 3,894.11 \pm 208.16$) compared to the SEL ($\bar{X}_{insertions} = 6,049.571 \pm 155.85$; $\bar{X}_{deletions} = 2,346.71 \pm 326.22$). (E and F) When adding the factor of size per insertion or deletion, significant differences were still observed between the lineages ($F(1) = 2{,}102.87$, $p < 0.001$; multifactor 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 2 ($n = 17{,}391$), 3 ($n = 1{,}062$), 4 ($n = 104$), and 5 ($n = 5$) 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 ($F(1) = 27.68$, $p < 0.001$; multifactor 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 S10 Fig. (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 ($F(2) = 5{,}682$, $p < 0.001$; multifactor 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) Signatures of UV-damage–associated single and double substitutions (i.e., C → T at CC sites and CC → TT) double substitutions are greater in the FEL compared to the SEL ($p < 0.001$ for both tests; Wilcoxon rank–sum test). figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. FEL, faster-evolving lineage; OG, orthologous gene; Pro., Proportion; SEL, slower-evolving lineage.

https://doi.org/10.1371/journal.pbio.3000255.g006

($F(1) = 2{,}102.87$, $p < 0.001$; multifactor 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., stretches of the same base) in codon-based alignments of the 1,034 OGs (i.e., [substituted bases + deleted bases + inserted bases]/total homopolymer bases) revealed significant differences between the FEL and SEL (Fig 6G; $F(1) = 27.68$, $p < 0.001$; multifactor 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 test). To gain more insight into the stability of different homopolymer runs (i.e., A|T or C|G) and the types of sequence changes that occur among 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) (S10 Fig). In addition to recapitulating differences between the types of mutations that occur at homopolymers ($F(2) = 1{,}686.70$, $p < 0.001$; multifactor 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). E.g., among A|T and C|G homopolymers of length 2 and C|G homopolymers of length 3, 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 5 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 because of 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 ($F(2) = 5{,}682$, $p < 0.001$; multifactor 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* FEL yeasts have a greater genomic signature of UV damage.** UV damage can result in C → T substitutions at CC sites and CC → TT double substitutions [20,65]. Although both the FEL and SEL have lost *PHR1*, a gene encoding a DNA photolyase that repairs pyrimidine dimers, the FEL has lost additional genes in other pathways that can repair UV damage (e.g., *POL32* in the excision repair pathways). We hypothesized the FEL would have a greater signature of UV damage due to these gene losses. We found significantly greater number of single and double substitutions in CC sites indicative of UV damage in the FEL compared to the SEL (Fig 6I; $p < 0.001$ for both tests; Wilcoxon rank–sum test).

Lastly, we examined whether 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 [66] and Epstein's coefficient of difference [67]. For both measures, we observed significantly more radical amino-acid substitutions in the FEL compared to the SEL (S11 Fig; $p < 0.001$; 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

Species in the genus *Hanseniaspora* exhibit the longest branches among budding yeasts, and their genomes have some of the lowest numbers of genes, lowest GC contents, and smallest assembly sizes in the subphylum (Fig 1, S4 Fig) [34–36]. 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 the FEL, is a lineage of long-term, hypermutator species that have undergone extensive gene loss (Figs 1–4 as well as S2, S5, S7 and S8 Figs).

Evolution by gene loss is gaining increasing attention as a major mode of genome evolution [35,68] and is mainly possible because of the dispensability of the majority of genes. E.g., 90% of *E. coli* [69], 80% of *S. cerevisiae* [70], and 73% of *Candida albicans* [71] genes are dispensable in laboratory conditions. The loss of dispensable genes can be selected for [72] and is common in lineages of obligate parasites or symbionts such as in the microsporidia, intracellular fungi that have lost key metabolic pathways such as amino-acid biosynthesis pathways [73,74], and myxozoa, a group of cnidarian obligate parasites that infect vertebrates and invertebrates [75]. Similar losses are also increasingly appreciated in free-living organisms, such as the budding yeasts [this study; 34,35,76–78] and animals [68]. E.g., the loss of *SUC2*, a gene known to enable sucrose utilization [47], in the FEL reflects the inability of species in the FEL to grow on sucrose, while its presence in the SEL reflects its species' ability to grow on sucrose (Fig 2).

However, *Hanseniaspora* species have experienced not just the typically observed losses of metabolic genes (Fig 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 [79]. 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.

Lost 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 [12,51], 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 for rapid growth in the wine fermentation environment [41].

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) [45,80]. These three lineages are distantly related to one another on the budding yeast phylogeny [35], 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* [80], 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* [80], reveals that *CDC26*, Pho85 CycLin 1 (*PCL1*), Precocious Dissociation of Sisters 1 (*PDS1*), *RFX1*, Substrate/Subunit Inhibitor of Cyclin-dependent protein kinase 1 (*SIC1*), SPOrulation 12 (*SPO12*), and *WHI5* are absent (S6 File), 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 because of 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* leads 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. E.g., *POL4* knock-out strains of *S. cerevisiae* can be rescued by the deletion of Yeast KU protein 70 (*YKU70*) [81], 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 [82]; and the 4-gene MIND complex, which is required for kinetochore biorientation and accurate chromosome segregation [83], were all lost in the FEL.

Lastly, the telomere-capping protein *CDC13* was lost in the FEL but is essential not only in *S. cerevisiae* but also in mammalian cells. However, additional losses in DNA-damage–response genes (i.e., Slow Growth Suppressor 1 [*SGS1*], EXOnuclease 1 [*EXO1*], and *RAD9*) can allow yeast cells to survive in the absence of *CDC13* [84]. In addition to *CDC13*, the FEL has also lost the checkpoint protein *RAD9* and other genes in the DNA-damage–checkpoint

pathway, including Mediator of the Replication Checkpoint 1 (*MRC1*) and *MEC3*. We hypothesize that the loss of *CDC13* was compensated by losses in the DNA-damage–response pathway, as has been observed in *S. cerevisiae* [84].

Long-term hypermutation and the subsequent slowing of sequence evolution

Estimates of the substitution rate ratio 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 [33]. Although we do not know the catalyst for this burst of sequence evolution, hypermutators may be favored in maladapted populations or in conditions in which environmental parameters frequently change [27,33]. While 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 [40,85,86]. Interestingly, grapes appear to have originated [87] 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. E.g., whereas the transition/transversion ratios of eukaryotic genomes are typically within the 1.7 and 4 range [88–91], *Hanseniaspora* ratios are approximately 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* [88,92], 0.92 in *E. coli* [93], 0.98 in *Drosophila melanogaster* [94], and 1.70 in humans [95]). 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 [96], *Drosophila* fruit flies [94], *S. cerevisiae* [88], and humans [95].

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. E.g., 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 has lost 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 [97]. 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 [98] are higher in the FEL compared to the SEL (Fig 5I) due to the loss of PHR1 and other alternative pathways that repair UV damage [20,65].

Our analyses provide the first, to our knowledge, major effort to characterize the genome function and evolution of the enigmatic genus *Hanseniaspora*. Our analyses focus on genomic differences between two lineages 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 (S12 Fig). These discoveries set the stage for further examination of intralineage or intraspecies variation in genomic features and content. More interestingly, our analyses lay the foundation for 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 [35]. The gDNA was sonicated and ligated to Illumina sequencing adaptors as previously described [99], and the libraries were submitted for paired-end sequencing (2 × 250) on an Illumina HiSeq 2500 instrument (Illumina, San Diego, CA, USA).

Phenotyping

We qualitatively measured growth of species on five carbon sources (maltose, raffinose, sucrose, melezitose, and galactose) as previously described in [35]. 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 *H. gamundiae* were obtained from Čadež and colleagues [42].

Genome assembly and annotation

To generate de novo genome assemblies, we used paired-end DNA sequence reads as input to iWGS, version 1.1 [100], a pipeline that 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) [101] as described in [35]. More specifically, sequenced reads were first quality trimmed, and adapter sequences were removed using Trimmomatic, version 0.33 [102] and Lighter, version 1.1.1 [103]. Subsequently, KmerGenie, version 1.6982 [104] 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 [105]; DISCOVAR, release 51885 [106]; MASURCA, version 2.3.2 [107]; SGA, version 0.10.13 [108]; SOAPdenovo2, version 2.04 [109]; and SPADES, version 3.7.0 [110])

were used to generate genome assemblies from the processed reads. Using QUAST, version 4.4 [111], 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., *Cyberlindnera jadinii*, *Kluyveromyces marxianus*, *S. cerevisiae*, and *Wickerhamomyces anomalus*) were generated in a recent comparative genomic study of the budding yeast subphylum [35]. The other 11 *Hanseniaspora* genomes examined here were annotated by following the same protocol as in [35].

In brief, the genomes were annotated using the MAKER pipeline, version 2.31.8 [112]. 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 [113], 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 [114]; SNAP, version 2013-11-29 [115]; and AUGUSTUS, version 3.2.2 [116], 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/; S1 File). These species and strains include *H. guilliermondii* UTAD222 [85], *H. opuntiae* AWRI3578, *H. osmophila* AWRI3579, *H. uvarum* AWRI3580 [117], *H. uvarum* 34–9, *H. vineae* T02 19AF [118], *H. valbyensis* NRRL Y-1626 [34], and *H. gamundiae* [42]. We also included *S. cerevisiae* S288C, *K. marxianus* DMKU3-1042, *W. anomalus* NRRL Y-366-8, and *C. jadinii* NRRL Y-1542, four representative budding yeast species that are all outside the genus *Hanseniaspora* [35], 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 [119], and *Hanseniaspora taiwanica*, which likely belongs to the SEL based on neighbor-joining analyses of the LSU rRNA gene sequence [56].

Assembly assessment and identification of orthologs

To determine genome assembly completeness, we calculated contig N50 [101] and assessed gene-content completeness using multiple databases of curated orthologs from BUSCO, version 3 [120]. 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 (S1 File). 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 1 kb

scaffold filter), and the average genome assembly size was 10.13 ± 1.38 Mb (9.93 ± 1.35 Mb when imposing a 1 kb scaffold filter). Notably, the genome assemblies and gene annotations created in the present project were comparable to publicly available ones. E.g., the genome size of publicly available *H. vineae* T02 19AF is 11.38 Mb with 4,661 genes, while our assembly of *H. 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. E.g., *H. 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 [117], 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 OGs among all protein coding sequences for all 29 taxa, we used ORTHOMCL, version 1.4 [121]. ORTHOMCL clusters genes into OGs using a Markov clustering algorithm [122; https://micans.org/mcl/] from gene similarity information acquired from a blastp "all-vs-all" using NCBI's BLAST+, version 2.3.0 (S2 Fig; [123]) and the proteomes of species of interest as input. The key parameters used in blastp "all-vs-all" were e-value = 1×10^{-10}, percent identity cutoff = 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 [124]. To do so, we first aligned and trimmed sequences in 1,143 OGs out of 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 [125], with the BLOSUM62 matrix of substitutions [126], a gap penalty of 1.0, 1,000 maximum iterations, the "genafpair" parameter, and TRIMAL, version 1.4 [127], 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 [128], 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 [128] were collapsed using PHYLOTREEPRUNER, version 1.0 [124], 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 [129] with concatenation [130,131] and coalescence [132] 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. [133], in which 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 [134] 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) [135]. Branch support for the concatenation and coalescence phylogenies was determined using 100 rapid bootstrap replicates [136] and local posterior support [132], respectively.

Several previous phylogenomic studies have shown that the internal branches preceding the *Hanseniaspora* FEL and SEL are long [34,36]. 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 [137] to remove the 88 single-gene trees in which 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 PAML, version 4.9 [138] and the concatenated 1,034-gene matrix. The input tree was derived from the concatenation-based ML analysis under a single LG + G4 [134] model (Fig 1A). The in-group root (i.e., the split between the FEL and the SEL) age was set between 0.756 and 1.177 time units (1 time unit = 100 mya), which was adopted from a recent study [35].

To infer the *Hanseniaspora* time tree, we first estimated branch lengths under a single LG + G4 [134] model with codeml in the PAML, version 4.9 [138] package and obtained a rough mean of the overall mutation rate. Next, we applied the approximate likelihood method [139,140] to estimate the gradient vector and Hessian matrix with Taylor expansion (option usedata = 3). Last, we assigned (i) the gamma-Dirichlet prior for the overall substitution rate (option rgene_gamma) as G(1, 1.55), with a mean of 0.64; (ii) the gamma-Dirichlet prior for the rate-drift parameter (option sigma2 gamma) as G(1, 10); and (iii) the parameters for the birth–death sampling process with birth and death rates $\lambda = \mu = 1$ and sampling fraction $\rho = 0$. 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 = 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 HMMs for each gene present in *S. 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* [141] (downloaded October 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 [142] blastp function and an e-value cutoff of 1×10^{-3}, as recommended for homology searches [143]. We used the top 100 hits for the gene of interest and aligned them using MAFFT, version 7.294b [125], 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 [144]. 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 an 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.

To determine whether any functional categories were over- or under-represented among genes present or absent among *Hanseniaspora* species, we conducted GO [145] enrichment analyses using GOATOOLS, version 0.7.9 [146]. We used a background of all *S. cerevisiae* genes and a *p*-value cutoff of 0.05 after multiple-test correction using the Holm method [147]. Plotting gene presence and absence among pathways was done by examining depicted pathways available through the KEGG project [148] and the *Saccharomyces* Genome Database [141].

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* [149]. 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. E.g., 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 (S4 File). Similarly, *S. cerevisiae* essential genes are significantly under-represented ($p < 0.001$; Fischer's exact test for both lineages) among lost genes, with 134 and 23 *S. cerevisiae* essential genes having been lost from the FEL and SEL genomes, respectively (lists of essential *S. cerevisiae* genes absent among *Hanseniaspora* genomes are available through figshare 10.6084/m9.figshare.7670756).

Ploidy estimation

To determine ploidy, we leveraged base frequency distributions at variable sites by mapping each genome's reads to its assembly. This approach is widely employed to determine ploidy from next-generation sequencing data and has been implemented in several pieces of software [150–152] and studies [153,154]. In short, examination of base frequency distributions between a frequency of 20 and 80 can provide insight into ploidy status. More specifically, haploid genomes lack biallelic sites, so their base frequency distributions will peak at high and low base frequencies and be depleted in positions with base frequencies near 50 (or a "smiley pattern"); diploid genomes typically have two alleles for a locus and are expected to exhibit a unimodal distribution centered around a base frequency of 50; finally, triploid genomes typically have one allele on one chromosome and the other allele in the other two chromosomes and are expected to exhibit a bimodal distribution centered around base frequencies of 33 and 66. Note that this approach assumes that there is a sufficient amount of heterozygosity in the genome and that ploidy changes may be go undetected in genomes lacking heterozygosity. To ensure high-quality read mapping, we first quality-trimmed reads suing Trimmomatic, version 0.36 [102], 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 [155], with the "sensitive" parameter, and we converted the resulting file to a sorted bam format using SAMtools, version 1.3.1 [156]. We next used nQuire [151], 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 [151].

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 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 [157] and calculated ω values under the different hypotheses using the CODEML module in PAML, version 4.9 [138]. 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 $H_{FEL–SEL\ branch}$ 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_{FEL} 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_{SEL} 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 $H_{FEL–SEL\ crown}$ 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 whether each of the alternative hypotheses was significantly different from the null hypothesis, we used the LRT ($\alpha = 0.01$). A few genes could not be analyzed because of fatal interruptions or errors during use in PAML, version 4.9 [138], which have been reported by other users [158]; these genes were removed from the analysis. Thus, this analysis was conducted for 989 genes for three tests ($H_{FEL–SEL\ branch}$, H_{FEL}, and H_{SEL} hypotheses) and 983 genes for one test ($H_{FEL–SEL\ crown}$ 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 whether 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 in which a nucleotide was present in all outgroup taxa but a gap was present in *Hanseniaspora* as deletions and positions in which 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 [159] and NUMPY, version 1.13.1 [160] modules.

We discovered that all *Hanseniaspora* species lack the *PHR1* gene, which is associated with the repair of UV radiation damage, but the FEL has lost additional genes that participate in other pathways that can repair UV damage such as the base-excision and nucleotide-excision repair pathway [20,65]. UV radiation induces high levels of C → T substitutions at CC sites and, more rarely, double substitutions of CC → TT [98,161]. To examine signatures of UV radiation damage across *Hanseniaspora*, we examined the number of C → T substitutions at CC sites (or G → A substitutions at GG sites) as well as the less frequent CC → TT (or GG → AA) double substitutions.

Supporting information

S1 Fig. Phylogenomics method pipeline. Using 25 *Hanseniaspora* proteomes and the proteomes of 4 outgroup taxa, we identified 11,877 orthologous groups of genes. For 1,143 orthologous groups, ≥90% of taxa were represented by a single sequence, while the others had two sequences (i.e., putative paralogs). The sequences of the 1,143 orthologous groups 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, trimmed, and had their evolutionary history inferred. Orthologous groups where the outgroup taxa were not recovered as the sister clade to the genus *Hanseniaspora* were removed, reducing the set to 1,034 orthologous groups. Among these 1,034 orthologous groups 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 orthologous group single-gene phylogenies. OG, orthologous gene.
(TIF)

S2 Fig. Concatenation and coalescence produce nearly identical and well-supported phylogenies that support two distinct *Hanseniaspora* lineages. (Left) Concatenation provides support for a lineage with a long stem branch, which we term the FEL, and another lineage with a much shorter stem branch, which we term the SEL. (Right) Coalescence supports monophyly of the FEL and SEL. Minor discrepancies are observed between the topologies. Only the values for bipartitions without full support are shown. Support for concatenation and coalescence was determined using 100 rapid bootstrap replicates and local posterior support, respectively. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. FEL, faster-evolving lineage; SEL, slower-evolving lineage.
(TIF)

S3 Fig. Internode key to accompany divergence time estimate file per internode. Internode identifiers for time tree analysis are shown in Fig 1B. Associated mean divergence time and credible intervals can be found in the S2 File.
(TIF)

S4 Fig. *Hanseniaspora* have among the smallest genome sizes, lowest number of genes, and lowest GC content percentages in the budding yeast subphylum Saccharomycotina. (A) The genus *Hanseniaspora* (family Saccharomycodaceae) includes the smallest budding yeast genome. Average genome sizes in the FEL, SEL, and the Saccharomycotina are 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. Average number of genes per genome in the FEL, SEL, and Saccharomycotina are 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. GC content values in the FEL, SEL, and Saccharomycotina are 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 gray represents all of Saccharomycotina. Blue represents the SEL, and orange represents the FEL. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. FEL, faster-evolving lineage; GC, Guanine–Cytosine; max, maximum; min, minimum; SEL, slower-evolving lineage.
(TIF)

S5 Fig. BUSCO analyses reveal extensive gene absence across various taxonomic ranks. BUSCO [120] 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 that very large numbers of BUSCO genes are absent from *Hanseniaspora* genomes and from FEL genomes in particular. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. BUSCO, Benchmarking Universal Single-Copy Orthologs; FEL, faster-evolving lineage.
(TIF)

S6 Fig. A liberal targeted gene-searching pipeline and the number of genes absent 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 *S. cerevisiae* proteome, was 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, an 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 cutoff 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 ORFs in the *S. cerevisiae* proteome. FEL, faster-evolving lineage; HMM, Hidden Markov Model; ORF, open reading frame; refseq, reference sequence; SEL, slower-evolving lineage.
(TIF)

S7 Fig. 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 the absence of key 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 absent from all *Hanseniaspora* species. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. *FBP1*, Fructose-1,6-BisPhosphatase 1; *PCK1*, Phosphoenolpyruvate CarboxyKinase 1.
(TIF)

S8 Fig. Base frequency plots reveal diversity in ploidy of *Hanseniaspora* species. (A) Plots with peaks at high and low base frequencies (or a "smiley pattern") reflect a lack of biallelic sites, which is suggestive of a haploid genome. The smiley-pattern distributions observed for the genomes of *H. occidentalis* var. *occidentalis*, *H. uvarum* CBS 314, and *H. guilliermondii* CBS 465 suggest these species have haploid genomes. (B) A unimodal distribution centered around a base frequency of 50 is consistent with the presence of two alleles at a given locus and

suggestive of a diploid genome. The unimodal distributions centered around a base frequency of 50 suggest *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) Bimodal distributions centered around base frequencies of 33 and 66 reflect one allele on one chromosome and another allele on the other two chromosomes, which is suggestive of a triploid genome. Bimodal distributions centered around 33 and 66 suggest *H. lachancei* and *H. jakobsenii* are triploid. (D) Analyses of *H. vineae* CBS 2171, *H. valbyensis*, *H.* sp. NRRL Y-63759, 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 only in *H. occidentalis* var. *citrica*. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. CBS, Centraalbureau voor Schimmelcultures; FEL, faster-evolving lineage; NRRL, Northern Regional Research Laboratory; SEL, slower-evolving lineage.
(TIF)

S9 Fig. Gene presence and absence related to yeast meiosis. Genes absent in both lineages and the FEL are colored purple and orange, respectively. Dotted lines with arrows indicate indirect links or unknown reactions. Lines with arrows indicate molecular interactions or relations. Circles indicate chemical compounds such as glucose or cAMP. figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. cAMP, cyclic AdenosineMonoPhosphate; FEL, faster-evolving lineage.
(TIF)

S10 Fig. Analyses of homopolymers by sequence length, base pair, and type of mutation. Significant differences among the proportion of mutated bases among homopolymers of various lengths were observed (Fig 6). 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 Fig 6, we observed significant differences in the numbers of mutations between the FEL and SEL ($F = 27.06$, $p < 0.001$; multifactor ANOVA) as well as in the type of mutations ($F = 1686.70$, $p < 0.001$; multifactor 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 ($p < 0.001$). figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. FEL, faster-evolving lineage; SEL, slower-evolving lineage.
(TIF)

S11 Fig. 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 radical in the FEL compared to the SEL ($p < 0.001$; Wilcoxon rank–sum test for both metrics). figshare: https://doi.org/10.6084/m9.figshare.7670756.v2. FEL, faster-evolving lineage; SEL, slower-evolving lineage.
(TIF)

S12 Fig. Mean protein similarity reveals very high evolutionary genomic diversity in *Hanseniaspora*. Mean protein similarity (measured by amino acid substitutions/site) between species in the FEL (using *H. uvarum* as the reference), species in the SEL (using *H. vineae* as a

reference), Saccharomycetaceae (*S. cerevisiae*), animals (human), and plants (thale cress *Arabidopsis thaliana*). For each lineage, mean protein similarity was estimated using a reciprocal best blast hit approach. The mean protein similarity observed in these lineages is roughly on par with genus-level differences within the family Saccharomycetaceae, humans to zebrafish, and thale cress to Japanese rice. Silhouettes were obtained under the Public Domain or Creative Commons license from phylopic.org (human, mouse, zebra finch, frog, zebrafish, and thale cress), from openclipart.org (field mustard, white spruce, wild tomato, Japanese rice; the colors of the field mustard and wild tomato original images were changed to black), or drawn by hand by Jacob L. Steenwyk (all yeasts). FEL, faster-evolving lineage; SEL, slower-evolving lineage.
(TIF)

S1 File. Summary table of genomes under study. Information including strain, genome characteristics (e.g., genome size, N50, GC content, number of genes, and other metrics), and source of strains is provided. GC, Guanine–Cytosine.
(XLSX)

S2 File. Divergence times for every internode in the *Hanseniaspora* phylogeny. Mean divergences as well as upper and lower confidence intervals are provided for every internode (internal branch length) in the *Hanseniaspora* phylogeny. Internode labels correspond to the labels shown in S4 Fig.
(XLSX)

S3 File. Summary of genes absent and present in the FEL and SEL. FEL, faster-evolving lineage; SEL, slower-evolving lineage.
(XLSX)

S4 File. GO enrichment analysis of genes absent in *Hanseniaspora*. A summary table that details over- and underenrichment analysis results of genes absent in the *Hanseniaspora* FEL and SEL. GO enrichment results are reported for genes absent in all *Hanseniaspora*, FEL, SEL, and genes uniquely absent in either lineage. FEL, faster-evolving lineage; GO, gene ontology; SEL, slower-evolving lineage.
(XLSX)

S5 File. Growth phenotypes across *Hanseniaspora* species and outgroup taxa. Growth phenotypes across eight substrates are depicted here. Among strains examined for a particular growth phenotype, ability to grow was characterized as (1) able to grow on the substrate, (2) unable to grow, or (3) display weak and delayed growth.
(XLSX)

S6 File. Summary of cell-cycle–gene presence and absence in *S. ludwigii*. A summary table of cell-cycle genes that are either present or absent in the closely related bipolar budding yeast *S. ludwigii*.
(XLSX)

Acknowledgments

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

Author Contributions

Conceptualization: Jacob L. Steenwyk, Chris Todd Hittinger, Antonis Rokas.

Data curation: Jacob L. Steenwyk, Dana A. Opulente, Jacek Kominek, Xing-Xing Shen, Xiaofan Zhou, Neža Čadež, Jeremy DeVirgilio, Amanda Beth Hulfachor.

Formal analysis: Jacob L. Steenwyk, Dana A. Opulente, Jacek Kominek, Xing-Xing Shen, Xiaofan Zhou, Abigail L. Labella, Noah P. Bradley.

Funding acquisition: Jeremy DeVirgilio, Cletus P. Kurtzman, Chris Todd Hittinger, Antonis Rokas.

Investigation: Jacob L. Steenwyk, Dana A. Opulente, Jacek Kominek, Xing-Xing Shen, Noah P. Bradley, Brandt F. Eichman, Neža Čadež, Diego Libkind.

Methodology: Jacob L. Steenwyk, Dana A. Opulente, Jacek Kominek, Xing-Xing Shen, Xiaofan Zhou, Abigail L. Labella, Brandt F. Eichman, Neža Čadež, Diego Libkind, Jeremy DeVirgilio, Amanda Beth Hulfachor.

Project administration: Amanda Beth Hulfachor, Cletus P. Kurtzman, Chris Todd Hittinger, Antonis Rokas.

Resources: Dana A. Opulente, Xiaofan Zhou, Abigail L. Labella, Noah P. Bradley, Brandt F. Eichman, Neža Čadež, Diego Libkind, Jeremy DeVirgilio, Amanda Beth Hulfachor, Cletus P. Kurtzman, Chris Todd Hittinger, Antonis Rokas.

Software: Xiaofan Zhou.

Supervision: Brandt F. Eichman, Cletus P. Kurtzman, Chris Todd Hittinger, Antonis Rokas.

Validation: Jacob L. Steenwyk, Dana A. Opulente.

Visualization: Jacob L. Steenwyk.

Writing – original draft: Jacob L. Steenwyk, Chris Todd Hittinger, Antonis Rokas.

Writing – review & editing: Jacob L. Steenwyk, Dana A. Opulente, Jacek Kominek, Xiaofan Zhou, Chris Todd Hittinger, Antonis Rokas.

References

1. Lindahl T. Quality Control by DNA Repair. Science (80-). 1999; 286: 1897–1905. https://doi.org/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. https://doi.org/10.1038/emboj.2008.15 PMID: 18285820
3. Broustas CG, Lieberman HB. DNA Damage Response Genes and the Development of Cancer Metastasis. Radiat Res. 2014; 181: 111–130. https://doi.org/10.1667/RR13515.1 PMID: 24397478
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. https://doi.org/10.1038/nature06350 PMID: 18059461
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. https://doi.org/10.1016/S0092-8674(01)00227-6 PMID: 11239397
6. Billmyre RB, Clancey SA, Heitman J. Natural mismatch repair mutations mediate phenotypic diversity and drug resistance in Cryptococcus deuterogattii. Elife. 2017; 6: e28802. https://doi.org/10.7554/eLife.28802 PMID: 28948913
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: e00595–17. https://doi.org/10.1128/mBio.00595-17 PMID: 28559486
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. https://doi.org/10.1534/g3.116.037499 PMID: 28188180

9. Barnum KJO'Connell MJ. Cell Cycle Regulation by Checkpoints. Methods Mol. Biol. 2014; 1170:29–40. https://doi.org/10.1007/978-1-4939-0888-2_2 PMID: 24906307
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. https://doi.org/10.1098/rstb.2011.0078 PMID: 22084380
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: e09492. https://doi.org/10.7554/eLife.09492 PMID: 27162172
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. https://doi.org/10.1016/j.cell.2004.05.024 PMID: 15210111
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. https://doi.org/10.1038/sj.onc.1207973 PMID: 15678131
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. https://doi.org/10.1002/embr.201338114 PMID: 24477934
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. https://doi.org/10.1371/journal.pgen.1006778 PMID: 28505153
16. Zhou B-BS, Elledge SJ. The DNA damage response: putting checkpoints in perspective. Nature. 2000; 408: 433–439. https://doi.org/10.1038/35044005 PMID: 11100718
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. https://doi.org/10.1101/gad.8.6.652 PMID: 7926756
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. https://doi.org/10.1073/pnas.1314423111 PMID: 24449905
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. https://doi.org/10.1016/j.cell.2017.09.048 PMID: 29056344
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. https://doi.org/10.1007/s002940000149 PMID: 11126776
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; 487(3–4): 137–147. https://doi.org/10.1016/S0921-8777(01)00113-6 PMID: 11738940
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. https://doi.org/10.1073/pnas.88.24.11251 PMID: 1763039
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. PMID: 2117700
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. https://doi.org/10.1093/nar/gkh868 PMID: 15475388
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; 275(42): 32925–32930. https://doi.org/10.1074/jbc.M005015200 PMID: 10922370
26. De Bont R. Endogenous DNA damage in humans: a review of quantitative data. Mutagenesis. 2004; 19: 169–185. https://doi.org/10.1093/mutage/geh025 PMID: 15123782
27. Ram Y, Hadany L. The evolution of stress-induced hypermutation in asexual populations. Evolution (N Y). 2012; 66: 2315–2328. https://doi.org/10.1111/j.1558-5646.2012.01576.x PMID: 22759304
28. Oliver A. High Frequency of Hypermutable Pseudomonas aeruginosa in Cystic Fibrosis Lung Infection. Science (80-). 2000; 288: 1251–1253. https://doi.org/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. https://doi.org/10.1126/science.1056421 PMID: 11283373

30. Healey KR, Zhao Y, Perez WB, Lockhart SR, Sobel JD, Farmakiotis D, et al. Prevalent mutator genotype identified in fungal pathogen Candida glabrata promotes multi-drug resistance. Nat Commun. 2016; 7: 11128. https://doi.org/10.1038/ncomms11128 PMID: 27020939

31. 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. https://doi.org/10.1038/42696 PMID: 9192893

32. Sniegowski PD, Gerrish PJ, Lenski RE. Evolution of high mutation rates in experimental populations of E. coli. Nature. 1997; 387: 703–705. https://doi.org/10.1038/42701 PMID: 9192894

33. 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. https://doi.org/10.1016/j.cub.2012.04.056 PMID: 22727704

34. 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. https://doi.org/10.1073/pnas.1603941113 PMID: 27535936

35. 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; 175(6): 1533–1545.e20. https://doi.org/10.1016/j.cell.2018.10.023 PMID: 30415838

36. 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. https://doi.org/10.1534/g3.116.034744 PMID: 27672114

37. 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: 1569. https://doi.org/10.3389/fmicb.2015.01569 PMID: 26834719

38. 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. https://doi.org/10.3390/beverages1040292

39. Montero CM, Dodero MCR, Sanchez DAG, Barroso CG. Analysis of low molecular weight carbohydrates in food and beverages: A review. Chromatographia. 2004; 59(1–2): 15–30.

40. 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.

41. 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 (22): e01580–17. https://doi.org/10.1128/AEM.01580-17 PMID: 28887422

42. 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. https://doi.org/10.1371/journal.pone.0210792 PMID: 30699175

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. https://doi.org/10.1093/nar/gks1116 PMID: 23180791

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. https://doi.org/10.1186/s12915-018-0498-3 PMID: 29499717

45. Kurtzman CP, Fell JW. The Yeasts—A Taxonomic Study. 4th ed. Kurtzman CP, Fell JW, editors. Amsterdam: 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. PMID: 2548922

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. https://doi.org/10.1371/journal.pbio.1001122 PMID: 21857801

48. Steenwyk J, Rokas A. Extensive Copy Number Variation in Fermentation-Related Genes Among Saccharomyces cerevisiae Wine Strains. G3 Genes, Genomes, Genet. 2017; 7: 1475–1485. Available from: http://www.g3journal.org/content/7/5/1475#ref-25

49. Gallone B, Steensels J, Prahl T, Soriaga L, Saels V, Herrera-Malaver B, et al. Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. Cell. Elsevier; 2016; 166: 1397–1410.e16. https://doi.org/10.1016/j.cell.2016.08.020 PMID: 27610566

50. Steenwyk JL, Rokas A. Copy Number Variation in Fungi and Its Implications for Wine Yeast Genetic Diversity and Adaptation. Front Microbiol. 2018; 9: 288. https://doi.org/10.3389/fmicb.2018.00288 PMID: 29520259

51. Jorgensen P. Systematic Identification of Pathways That Couple Cell Growth and Division in Yeast. Science (80-). 2002; 297: 395–400. https://doi.org/10.1126/science.1070850 PMID: 12089449

52. 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. https://doi.org/10.1016/S0092-8674(01)00596-7 PMID: 11747810

53. 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. https://doi.org/10.1128/MCB.25.23.10665-10673.2005 PMID: 16287876

54. Galgoczy DJ, Toczyski DP. Checkpoint Adaptation Precedes Spontaneous and Damage-Induced Genomic Instability in Yeast. Mol Cell Biol. 2001; 21: 1710–1718. https://doi.org/10.1128/MCB.21.5.1710-1718.2001 PMID: 11238908

55. 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. https://doi.org/10.3393/ac.2017.33.4.139 PMID: 28932723

56. 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. https://doi.org/10.1099/ijs.0.034231-0 PMID: 21841004

57. Diawara B, Kando C, Anyogu A, Ouoba LII, Nielsen DS, Sutherland JP, et al. Hanseniaspora jakobsenii sp. nov., a yeast isolated from Bandji, a traditional palm wine of Borassus akeassii. Int J Syst Evol Microbiol. 2015; 65: 3576–3579. https://doi.org/10.1099/ijsem.0.000461 PMID: 26297247

58. Jindamorakot S, Ninomiya S, Limtong S, Yongmanitchai W, Tuntirungkij M, Potacharoen W, et al. Three new species of bipolar budding yeasts of the genus Hanseniaspora and its anamorph Kloeckera isolated in Thailand. FEMS Yeast Res. 2009; 9: 1327–1337. https://doi.org/10.1111/j.1567-1364.2009.00568.x PMID: 19788563

59. Kassir Y, Granot D, Simchen G. IME1, a positive regulator gene of meiosis in S. cerevisiae. Cell. 1988; 52: 853–62. PMID: 3280136

60. Nag DK, Koonce MP, Axelrod J. SSP1, a gene necessary for proper completion of meiotic divisions and spore formation in Saccharomyces cerevisiae. Mol Cell Biol. 1997; 17: 7029–39. PMID: 9372934

61. Tachikawa H, Bloecher A, Tatchell K, Neiman AM. A Gip1p-Glc7p phosphatase complex regulates septin organization and spore wall formation. J Cell Biol. 2001; 155: 797–808. https://doi.org/10.1083/jcb.200107008 PMID: 11724821

62. Xiao W, Chow BL. Synergism between yeast nucleotide and base excision repair pathways in the protection against DNA methylation damage. Curr Genet. 1998; 33(2): 92–99. https://doi.org/10.1007/s002940050313 PMID: 9506896

63. 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. https://doi.org/10.1073/pnas.0603455103 PMID: 16751265

64. Lustig AJ. Cdc13 subcomplexes regulate multiple telomere functions. Nat Struct Biol. 2001; 8: 297–9. https://doi.org/10.1038/86157 PMID: 11276244

65. Budden T, Bowden N. The Role of Altered Nucleotide Excision Repair and UVB-Induced DNA Damage in Melanomagenesis. Int J Mol Sci. 2013; 14: 1132–1151. https://doi.org/10.3390/ijms14011132 PMID: 23303275

66. Sneath PH. Relations between chemical structure and biological activity in peptides. J Theor Biol. 1966; 12: 157–95. PMID: 4291386

67. Epstein CJ. Non-randomness of Ammo-acid Changes in the Evolution of Homologous Proteins. Nature. 1967; 215: 355–359. https://doi.org/10.1038/215355a0 PMID: 4964553

68. Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet. 2016; 17: 379–391. https://doi.org/10.1038/nrg.2016.39 PMID: 27087500

69. 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: 2006.0008. https://doi.org/10.1038/msb4100050 PMID: 16738554

70. 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. https://doi.org/10.1038/nature00935 PMID: 12140549

71. 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: e02048–18. https://doi.org/10.1128/mBio.02048-18 PMID: 30377286

72. Koskiniemi S, Sun S, Berg OG, Andersson DI. Selection-Driven Gene Loss in Bacteria. Casadesús J, editor. PLoS Genet. 2012; 8: e1002787. https://doi.org/10.1371/journal.pgen.1002787 PMID: 22761588

73. Keeling PJ, Slamovits CH. Simplicity and complexity of microsporidian genomes. Eukaryot Cell. 2004; 3: 1363–9. https://doi.org/10.1128/EC.3.6.1363-1369.2004 PMID: 15590811

74. 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. https://doi.org/10.1038/35106579 PMID: 11719806

75. 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. https://doi.org/10.1073/pnas.1511468112 PMID: 26627241

76. 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. https://doi.org/10.1073/pnas.0404319101 PMID: 15381776

77. 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. https://doi.org/10.1073/pnas.0914418107 PMID: 20479238

78. 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. https://doi.org/10.1093/femsyr/fov035 PMID: 26066552

79. Hartwell L. Defects in a cell cycle checkpoint may be responsible for the genomic instability of cancer cells. Cell. 1992; 71(4): 543–546. https://doi.org/10.1016/0092-8674(92)90586-2 PMID: 1423612

80. 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(18): e01195–18. https://doi.org/10.1128/MRA.01195-18 PMID: 30533777

81. Sterling CH. DNA Polymerase 4 of Saccharomyces cerevisiae Is Important for Accurate Repair of Methyl-Methanesulfonate-Induced DNA Damage. Genetics. 2005; 172: 89–98. https://doi.org/10.1534/genetics.105.049254 PMID: 16219787

82. 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. https://doi.org/10.1126/science.aar6436 PMID: 29724956

83. 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. https://doi.org/10.1016/j.cell.2016.10.011 PMID: 27881300

84. Ngo H-P, Lydall D. Survival and Growth of Yeast without Telomere Capping by Cdc13 in the Absence of Sgs1, Exo1, and Rad9. Copenhaver GP, editor. PLoS Genet. 2010; 6: e1001072. https://doi.org/10.1371/journal.pgen.1001072 PMID: 20808892

85. 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. https://doi.org/10.1128/genomeA.01515-16 PMID: 28153887

86. 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. https://doi.org/10.1016/j.fm.2009.05.005 PMID: 19835764

87. Wikstrom N, Savolainen V, Chase MW. Evolution of the angiosperms: calibrating the family tree. Proc R Soc B Biol Sci. 2001; 268: 2211–2220. https://doi.org/10.1098/rspb.2001.1782 PMID: 11674868

88. 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. https://doi.org/10.1073/pnas.1323011111 PMID: 24847077

89. 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. https://doi.org/10.1186/gb-2011-12-9-r84 PMID: 21917140

90. 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. https://doi.org/10.1093/bioinformatics/btu668 PMID: 25297068

91. 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. https://doi.org/10.1186/1297-9686-34-3-275 PMID: 12081799

92. 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; 105(27): 9272–9277. https://doi.org/10.1073/pnas.0803466105 PMID: 18583475

93. Lynch M. The Origins of Genome Architecture. Sunderland, MA: Sinauer Associates; 2007. https://doi.org/10.1093/jhered/esm073

94. 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. https://doi.org/10.1101/gr.091231.109 PMID: 19439516

95. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci. 2010; 107: 961–968. https://doi.org/10.1073/pnas.0912629107 PMID: 20080596

96. Hershberg R, Petrov DA. Evidence That Mutation Is Universally Biased towards AT in Bacteria. Nachman MW, editor. PLoS Genet. 2010; 6: e1001115. https://doi.org/10.1371/journal.pgen.1001115 PMID: 20838599

97. Seeberg E, Eide L, Bjørås M. The base excision repair pathway. Trends Biochem Sci. 1995; 20: 391–397. https://doi.org/10.1016/S0968-0004(00)89086-6 PMID: 8533150

98. 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. https://doi.org/10.1101/gr.220038.116 PMID: 28739859

99. 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. https://doi.org/10.1038/nature08791 PMID: 20164837

100. Zhou X, Peris D, Kominek J, Kurtzman CP, Hittinger CT, Rokas A. In Silico Whole Genome Sequencer and Analyzer (iWGS): A Computational Pipeline to Guide the Design and Analysis of de novo Genome Sequencing Studies. G3 Genes|Genomes|Genetics. 2016; 6(11): 3655–3662. https://doi.org/10.1534/g3.116.034249 PMID: 27638685

101. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012; 13: 329–42. https://doi.org/10.1038/nrg3174 PMID: 22510764

102. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

103. Song L, Florea L, Langmead B. Lighter: fast and memory-efficient sequencing error correction without counting. Genome Biol. 2014; 15: 509. https://doi.org/10.1186/s13059-014-0509-9 PMID: 25398208

104. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014; 30: 31–37. https://doi.org/10.1093/bioinformatics/btt310 PMID: 23732276

105. 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. https://doi.org/10.1101/gr.089532.108 PMID: 19251739

106. 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. https://doi.org/10.1038/ng.3121 PMID: 25326702

107. Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013; 29: 2669–2677. https://doi.org/10.1093/bioinformatics/btt476 PMID: 23990416

108. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012; 22: 549–556. https://doi.org/10.1101/gr.126953.111 PMID: 22156294

109. 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. https://doi.org/10.1186/2047-217X-1-18 PMID: 23587118

110. 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. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

111. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 29: 1072–1075. https://doi.org/10.1093/bioinformatics/btt086 PMID: 23422339

112. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011; 12: 491. https://doi.org/10.1186/1471-2105-12-491 PMID: 22192575

113. 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. https://doi.org/10.1093/nar/gkt1183 PMID: 24297253

114. 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. https://doi.org/10.1101/gr.081612.108 PMID: 18757608

115. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004; 5: 59. https://doi.org/10.1186/1471-2105-5-59 PMID: 15144565

116. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003; 19: ii215–ii225. https://doi.org/10.1093/bioinformatics/btg1080 PMID: 14534192

117. 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. https://doi.org/10.1128/genomeA.01287-16 PMID: 27856586

118. 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. https://doi.org/10.1128/genomeA.00530-14 PMID: 24874663

119. Cadez N, Raspor P, Smith MT. 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. https://doi.org/10.1099/ijs.0.64052-0 PMID: 16627671

120. 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. https://doi.org/10.1093/molbev/msx319 PMID: 29220515

121. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13: 2178–2189. https://doi.org/10.1101/gr.1224503 PMID: 12952885

122. van Dongen S. Graph clustering by flow simulation. Graph Stimul by flow Clust [Ph.D. thesis]. Utrecht (the Netherlands): University of Utrecht; 2000.

123. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10: 421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

124. 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. https://doi.org/10.4137/EBO.S12813 PMID: 24250218

125. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013; 30: 772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

126. Mount DW. Using BLOSUM in Sequence Alignments. Cold Spring Harb Protoc. 2008; 2008: pdb.top39. https://doi.org/10.1101/pdb.top39 PMID: 21356855

127. 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. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945

128. Price MN, Dehal PS, Arkin AP. FastTree 2—Approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010; 5: e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

129. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol. 1981; 17: 368–376. https://doi.org/10.1007/BF01734359 PMID: 7288891

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

131. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 2003; 425: 798–804. https://doi.org/10.1038/nature02053 PMID: 14574403

132. Edwards SV. Is a new and general theory of molecular systematics emerging? Evolution (N Y). 2009; 63: 1–19. https://doi.org/10.1111/j.1558-5646.2008.00549.x PMID: 19146594

133. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

134. Le SQ, Gascuel O. An Improved General Amino Acid Replacement Matrix. Mol Biol Evol. 2008; 25: 1307–1320. https://doi.org/10.1093/molbev/msn067 PMID: 18367465

135. 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. https://doi.org/10.1093/bioinformatics/btv234 PMID: 26072508

136. Stamatakis A, Hoover P, Rougemont J. A Rapid Bootstrap Algorithm for the RAxML Web Servers. Renner S, editor. Syst Biol. 2008; 57: 758–771. https://doi.org/10.1080/10635150802429642 PMID: 18853362

137. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010; 26: 1669–1670. https://doi.org/10.1093/bioinformatics/btq243 PMID: 20472542

138. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007; 24: 1586–1591. https://doi.org/10.1093/molbev/msm088 PMID: 17483113

139. 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. https://doi.org/10.1038/nrg.2015.8 PMID: 26688196

140. Reis M d., Yang Z. Approximate Likelihood Calculation on a Phylogeny for Bayesian Estimation of Divergence Times. Mol Biol Evol. 2011; 28: 2161–2172. https://doi.org/10.1093/molbev/msr045 PMID: 21310946

141. 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. https://doi.org/10.1093/nar/gkr1029 PMID: 22110037

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

143. Pearson WR. An Introduction to Sequence Similarity ("Homology") Searching. Curr Protoc Bioinforma. 2013; 42: 3.1.1–3.1.8. https://doi.org/10.1002/0471250953.bi0301s42 PMID: 23749753

144. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Comput Biol. 2011; 7: e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361

145. GeneOntologyConsortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004; 32: 258D–261. https://doi.org/10.1093/nar/gkh036 PMID: 14681407

146. 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. https://doi.org/10.1038/s41598-018-28948-z PMID: 30022098

147. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979; 6: 65–70.

148. 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. https://doi.org/10.1093/nar/gkv1070 PMID: 26476454

149. 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. https://doi.org/10.1126/science.285.5429.901

150. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012; 28: 423–425. https://doi.org/10.1093/bioinformatics/btr670 PMID: 22155870

151. 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. https://doi.org/10.1186/s12859-018-2128-z PMID: 29618319

152. Augusto Corrêa dos Santos R, Goldman GH, Riaño-Pachón DM. ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. Berger B, editor. Bioinformatics. 2017; 33: 2575–2576. https://doi.org/10.1093/bioinformatics/btx204 PMID: 28383704

153. Zhu YO, Sherlock G, Petrov DA. Whole Genome Analysis of 132 Clinical Saccharomyces cerevisiae Strains Reveals Extensive Ploidy Variation. G3 Genes|Genomes|Genetics. 2016; 6: 2421–2434. https://doi.org/10.1534/g3.116.029397 PMID: 27317778

154. Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, et al. The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. Elife. 2013; 2: e00731. https://doi.org/10.7554/eLife.00731 PMID: 23741619

155. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9: 357–359. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

156. 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. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

157. Suyama M, Torrents D, Bork P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006; 34: W609–12. https://doi.org/10.1093/nar/gkl315 PMID: 16845082

158. 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. https://doi.org/10.1073/pnas.1616744114 PMID: 28808022

159. 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. https://doi.org/10.1093/bioinformatics/btp163 PMID: 19304878

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

161. Ikehata H, Ono T. The Mechanisms of UV Mutagenesis. J Radiat Res. 2011; 52: 115–125. https://doi.org/10.1269/jrr.10175 PMID: 21436607