Accurate Tracking of the Mutational Landscape of Diploid Hybrid Genomes

Lorenzo Tattini,*[,1] Nicolò Tellini,[1] Simone Mozzachiodi,[1] Melania D'Angiolo,[1] Sophie Loeillet,[2] Alain Nicolas,[2] and Gianni Liti*[,1]

[1]CNRS UMR7284, INSERM, IRCAN, Université Côte d'Azur, Nice, France

[2]CNRS UMR3244, Institut Curie, PSL Research University, Paris, France

***Corresponding authors:** E-mails: lorenzo.tattini@univ-cotedazur.fr; gianni.liti@unice.fr.

Associate Editor: Banu Ozkan

Abstract

Mutations, recombinations, and genome duplications may promote genetic diversity and trigger evolutionary processes. However, quantifying these events in diploid hybrid genomes is challenging. Here, we present an integrated experimental and computational workflow to accurately track the mutational landscape of yeast diploid hybrids (MuLoYDH) in terms of single-nucleotide variants, small insertions/deletions, copy-number variants, aneuploidies, and loss-of-heterozygosity. Pairs of haploid *Saccharomyces* parents were combined to generate ancestor hybrids with phased genomes and varying levels of heterozygosity. These diploids were evolved under different laboratory protocols, in particular mutation accumulation experiments. Variant simulations enabled the efficient integration of competitive and standard mapping of short reads, depending on local levels of heterozygosity. Experimental validations proved the high accuracy and resolution of our computational approach. Finally, applying MuLoYDH to four different diploids revealed striking genetic background effects. Homozygous *Saccharomyces cerevisiae* showed a ∼4-fold higher mutation rate compared with its closely related species *S. paradoxus*. Intraspecies hybrids unveiled that a substantial fraction of the genome (∼250 bp per generation) was shaped by loss-of-heterozygosity, a process strongly inhibited in interspecies hybrids by high levels of sequence divergence between homologous chromosomes. In contrast, interspecies hybrids exhibited higher single-nucleotide mutation rates compared with intraspecies hybrids. MuLoYDH provided an unprecedented quantitative insight into the evolutionary processes that mold diploid yeast genomes and can be generalized to other genetic systems.

Key words: **genome evolution, mutation rate, hybrid genomes, heterozygosity, loss-of-heterozygosity, *Saccharomyces paradoxus*.**

Introduction

High-throughput sequencing (HTS) technologies, both short- and long-read, have had a massive impact on genome research, enabling previously unimaginable and detailed dissection of the genomic landscape with outstanding speed and low costs (Stephens et al. 2015; Goodwin et al. 2016; Magi et al. 2016). The occurrence of variation in sequence, structure, and size of a genome in time is triggered by several factors including DNA mutations, recombination, and whole-genome duplication. These factors contribute to diversity within a population, translate into quantitative phenotypic variation, and may eventually result in speciation. Integrated bioinformatic pipelines, along with high-quality reference assemblies, are fundamental to successfully depict the mutational landscape of genomes (Magi et al. 2015; Wong et al. 2018). However, de novo whole-genome assembly and phasing is still highly challenging and results in incomplete sequences (Garg et al. 2018; Sedlazeck et al. 2018). Thus, the mutational landscape of diploid or polyploid organisms has been characterized through resequencing studies which are based on mapping short reads against a single consensus reference, although the latter inherently misses what defines the genetic identity of one individual (Church 2018). For example, the human genome was assembled using the DNA of ∼50 individuals with just one of them accounting for ∼70% of the sequence, whereas the yeast reference genome was produced from a single laboratory strain (namely S288C) and its derivatives (Cherry et al. 2012; Engel et al. 2014). Recently, high-quality panels of reference sequences (Maretty et al. 2017; Yue et al. 2017; Ameur et al. 2018) and novel standards for genome assembly (Editorial 2018) have been reported, whereas graph-based models have been suggested to overcome the limits imposed by reference bias (Eggertsson et al. 2017; Paten et al. 2017; Garrison et al. 2018). Nevertheless, using a single reference sequence is a convenient simplification (Church 2018) and current technologies are boosting genome quality (Pennisi 2017). Resequencing studies have been proven successful whenever the level of heterozygosity is sufficiently low (e.g., the percentage of polymorphic loci in humans is <0.16% [1000 Genomes Project Consortium et al. 2015]) or for homozygous genomes. Yet, mapping short reads against a reference genome prevents probing variation in genomic regions missing in the reference as well as calling variants with direct phasing, that

Article

Mol. Biol. Evol. 36(12):2861–2877 doi:10.1093/molbev/msz177 Advance Access publication August 9, 2019

is, assigning a heterozygous variant to one of the two homologous chromosomes. Whole-genome resequencing methods do not provide phasing information by default (Browning and Browning 2011; Zhang et al. 2017) although haplotype phasing impacts several aspects of population and medical genetics, including characterizing the relationship between genetic variation and disease susceptibility, inferring human demographic history, detecting points of recombination, recurrent mutation, signatures of selection, and modeling *cis*-regulation of gene expression. For example, missense, synonymous, and *cis*-regulatory mutations have been shown to collectively provide phenotypic diversity in haploid segregants (She and Jarosz 2018). Variant phasing has been the focus of several computational studies (based on a consensus reference) but solving exactly the problem is highly demanding since it is NP-hard (Sedlazeck et al. 2018). Thus, current methods for assembling haplotype-resolved sequences rely on computational and experimental techniques that require trio data (Koren et al. 2018) or population-based statistical phasing and long reads to maximize the performance (Choi et al. 2018).

Natural diploid genomes harbor varying levels of heterozygosity (Peter et al. 2018). Analyzing diploid hybrid genomes, characterized by high heterozygosity, against a reference poses the problem of spurious read mapping, which in turn may lead to false positive (FP) calls of both single-nucleotide variants and small insertions/deletions (SNVs and indels, respectively). High levels of heterozygosity allow for mapping short-read data against hybrid genome assemblies obtained by concatenating the two parental subgenomes ("competitive mapping") (Smukowski Heil et al. 2017; Langdon et al. 2018). This strategy provides direct variant phasing but, as we discuss in the following sections, it is risky whenever the number of genetic markers, namely preexisting variants that differentiate subgenomes, is low. In fact, this will result in genomic regions characterized by reads with non-unique mapping, preventing the assessment of de novo small variants (namely SNVs and indels).

The study of the role of hybridization in species fitness is an active field of research in evolutionary biology (Taylor and Larson 2019). Unfortunately, notwithstanding the importance of experimental and computational validation of the methods based on HTS data (Escalona et al. 2016; Stephens et al. 2016; Semeraro et al. 2018), none of the approaches tailored to the analysis of hybrid genomes has been automatized nor tested through simulations (Laureau et al. 2016; Dutta et al. 2017; Smukowski Heil et al. 2017). In this context *Saccharomyces cerevisiae*, along with its closely related species, is a leading-edge eukaryotic model system that has long been exploited in genetics, cell biology, and systems biology (Magi et al. 2012; Duina et al. 2014; Marsit et al. 2017; She and Jarosz 2018; Coelho et al. 2019). The *S. cerevisiae* genome was the first fully sequenced eukaryotic genome (Goffeau et al. 1996), and more recently it has also played a crucial role in understanding key principles in evolutionary genomics (Dujon 2010; Hittinger 2013; Marcet-Houben and Gabaldón 2015). Species from the *Saccharomyces* genus have been shown to be prone to intra- and inter-species hybridization (Liti et al. 2006; Dujon 2010; Gallone et al. 2016). Hybridization occurs ubiquitously with natural hybrids associated with multiple fermenting environments (Lopandic 2018; Mixão and Gabaldón 2018; Monerawela and Bond 2018; Peris et al. 2018). Outbreeding has also shaped the *S. cerevisiae* population structure with several groups of strains showing mosaic genomes that result from ancient admixtures of extant lineages (Liti et al. 2009).

The precise laboratory control of the sexual and asexual phases is a major strength of yeast genetics and enables to combine different haploid species and isolates into designed hybrid ancestors. These diploids can be evolved under various laboratory protocols such as return-to-growth (RTG) (Laureau et al. 2016), adaptive evolution (Barrick and Lenski 2013; Long et al. 2015), and mutation accumulation (Lynch et al. 2008; Nishant et al. 2010; Serero et al. 2014; Zhu et al. 2014; Sharp et al. 2018). RTG experiments generate genome-wide recombinant hybrids characterized by loss-of-heterozygosity (LOH) events. LOHs allow the expression of recessive alleles (Gerstein et al. 2014; Vázquez-García et al. 2017; Sharp et al. 2018) as well as the formation of new combinations of haplotypes and provide an alternative approach for the analysis of complex traits (Laureau et al. 2016; Sadhu et al. 2016). Adaptive evolution experiments quantify the preferential accumulation of preexisting and de novo genetic variants that are selected in a controlled environment due to their contribution to organismal fitness. On the contrary, in mutation accumulation experiments, a bottleneck of one or few individuals is imposed on a population, allowing for nonlethal mutations to accumulate with slight or no filtering by natural selection. Forcing population bottlenecks provides a means to evaluate mutational rates and signatures. Compared with fluctuation assay (Lang and Murray 2008), mutation accumulation experiments yield unbiased genome-wide estimations of the rates but, so far, they have been mostly restricted to laboratory strains, mutator backgrounds, and haploids or homozygous diploids, although estimates using other model organisms have been reported (Fry et al. 1999; Sharp and Agrawal 2012; Yang et al. 2015). Thus, a global picture of the mutational landscape, including genetic background effects and a quantitative measure of the impact of LOH, is still missing. In this study, we present MuLoYDH, a general framework for the comprehensive characterization of the Mutational Landscape of Yeast Diploid Hybrids in terms of SNVs, indels, copy-number variants (CNVs), aneuploidies, and LOHs. The genetic cross of haploid parents with fully assembled genomes enables to reconstruct a phased diploid genome that serves as ancestor. The latter is otherwise impossible to obtain from direct sequencing of hybrid diploids. After extensive benchmarking against both simulated and experimentally designed diploid *Saccharomyces* hybrids, we use MuLoYDH to accurately characterize intra- and inter-species mutation accumulation lines (MALs) obtained by crossing domesticated and natural strains. Our strategy reveals striking genetic background effects and quantifies 1) the rates of the different events contributing to the evolution of hybrid genomes and 2) the corresponding fraction of the genome affected.

Results and Discussion

Overview of the MuLoYDH Strategy

The MuLoYDH workflow begins with experimentally generating ancestor hybrids by combining two haploid founder strains with fully assembled and annotated genomes. This allows the investigation of the fully phased genome of the derived hybrids and to assess the impact of heterozygosity on genome evolution. *Saccharomyces cerevisiae* is an ideal genetic system for this approach since haploid strains can be crossed to produce diploids with a broad range of heterozygosity (fig. 1*A* and *B*). Designed *Saccharomyces* diploids can range from complete homozygous (0% heterozygosity, when a single strain is used), low (0.1%, in intraspecies hybrids derived from strains of the same subpopulation), moderate-high (0.5–4%, crossing strains from diverged subpopulations), and extremely high heterozygosity (8–35%, in interspecies hybrids) (Smukowski Heil et al. 2017; Yue et al. 2017; Peter et al. 2018). Following mitotic hybrid evolution under different defined laboratory conditions (fig. 1*C*), the corresponding short-read data are processed using a novel pipeline that combines competitive and standard mapping (fig. 1*D*–*F*).

The former consists in mapping short-read data against the union of the two parental assemblies, whereas the latter refers to mapping against a single parental assembly. The computational strategy implemented in MuLoYDH for tracking the mutational events relies on the presence of single-nucleotide markers (in short "markers" in the following sections) between the two parental subgenomes (supplementary figs. S1–S4, Supplementary Material online). The genomic density and distribution of markers are fundamental for our purposes since they are probes for LOH detection and, as detailed in the following section, they allow direct phasing of de novo small variants (namely variants arisen during the experiment) from competitive mapping (in short "variants" in the following sections). In addition, markers positions are determined from the assemblies and can be used to set up a rational quality threshold for LOH detection as well as for filtering de novo small variants (Materials and Methods).

As expected, the number of markers detected aligning *S. paradoxus*/*S. cerevisiae* assemblies was ∼15-fold higher compared with *S. cerevisiae*/*S. cerevisiae* assemblies (table 1). Markers were classified as lying in collinear regions or lying within structural rearrangements, namely inversions or translocations, and the corresponding fractions were calculated (f_c and f_r, respectively). Using these values, we were able to further differentiate the backgrounds beyond the typical heterozygosity measures based on sequence divergence. All the hybrids derived from the strains UWOPS03-461.4 (MA in short in the following sections, from the Malaysian *S. cerevisiae* clade) and UFRJ50816 (from the South American *S. paradoxus* clade) showed the largest fraction of markers within structurally rearranged regions (fig. 1*A*), consistently with multiple genomic rearrangements that occurred within these lineages (Yue et al. 2017).

The genomic distribution of markers represents a key feature of the hybrid genome and highly impacts the accuracy of de novo small variants detection. Therefore, we calculated the low-marker-density-regions (LMDRs) fraction, that is, the fraction of genomic regions characterized by <1 marker in 300 bp, namely twice the read length of the sequencing experiments reported in this study (table 1). Pairs of genomes characterized by a small number of markers showed higher values of the LMDRs fraction. In order to minimize the LMDRs fraction, MuLoYDH can be run with two different settings (collinear/rearranged) exploiting a priori knowledge of parental genomes reciprocal structure (supplementary fig. S5, Supplementary Material online).

The fully phased hybrid genome assembly can be exploited to perform a competitive mapping of the reads obtained from evolved hybrids. This approach, compared with a standard mapping against a single assembly or an unphased reference genome, is expected to provide a larger number of mapped reads in regions unique to one parental assembly. Indeed, competitive mapping in *S. cerevisiae* hybrids reduced the number of unmapped reads by ∼8% on average (supplementary table S7, Supplementary Material online). However, it represents a challenge regarding reads mapping to identical regions within the two parental assemblies. In fact, the fraction of reads showing a mapping quality (MAPQ) value equal to zero (Li et al. 2008), thus reflecting nonunique mapping, is also expected to increase as the level of heterozygosity decreases. As expected, the number of reads showing MAPQ = 0 increased in the competitive mapping (∼43%) with respect to the standard mappings (∼15%) in intraspecies hybrids (supplementary table S8, Supplementary Material online). In summary, the crossing phase (fig. 1*B*) did provide the unique opportunity to generate diploid hybrids with phased genome to benchmark the computational approaches for studying their evolution.

Benchmarking MuLoYDH against Simulated Data Sets

Highly similar DNA sequences may occur on different genomic scales, from short stretches (such as homopolymers), to complex events (e.g., segmental duplications), up to chromosome level (i.e., homologous chromosomes) (Weisenfeld et al. 2017). These repetitive sequences are characterized by nearly 100% sequence identity and represent a major challenge of HTS data analysis (Treangen and Salzberg 2012). The number and the distribution of markers affect the performance of small variants calling from competitive mapping. As the number of markers and the level of heterozygosity decrease, the mapping algorithm produces a progressively increasing number of reads characterized by MAPQ = 0. In turn, this affects the small variants calling algorithms, since reads characterized by nonunique mapping (i.e., MAPQ = 0) are filtered out. Thus, we investigated the impact of the level of heterozygosity on the performance of MuLoYDH in calling small variants from competitive mapping in simulated genomes (fig. 2*A* and *B*). The number of simulated markers was chosen to mimic experimental data (table 1). As expected, the F_1 score, namely the harmonic mean of precision and recall (the Simulated Data section reports the mathematical definition), sharply decreased with the number of markers. Nevertheless, when the percentage of markers was ∼0.5, the score tended to a

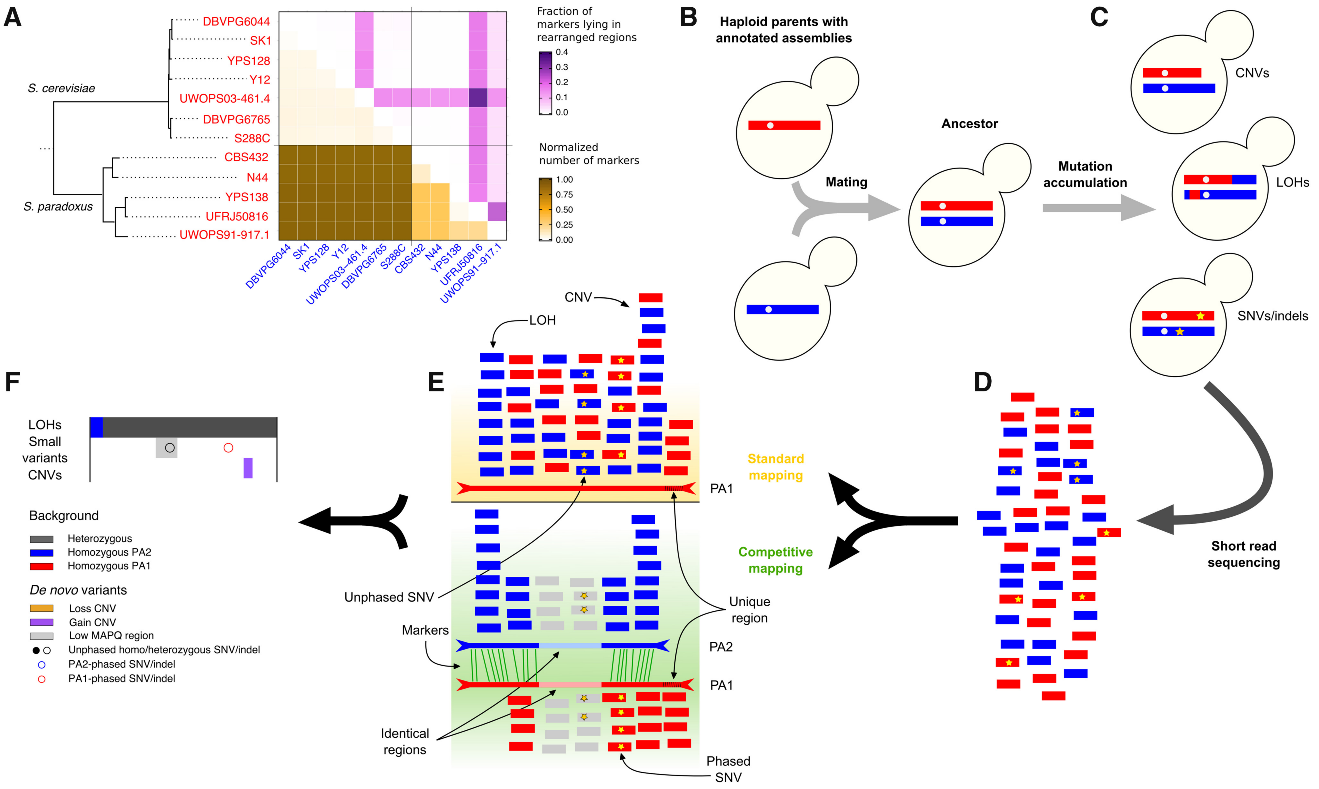

Fig. 1. MuLoYDH overview. MuLoYDH comprises (*A*) founder parent strains characterization, (*B*) hybrid generation and (*C*) evolution, (*D*) resequencing of evolved strain, and (*E* and *F*) tracking their mutational events. (*A*) Complete genome assemblies and annotations guide a rational selection of founder strains with desired genomic distances (bottom left heatmap) and interchromosomal rearrangements (top right heatmap). The phylogenetic tree is reproduced from previous work (Yue et al. 2017). (*B*) Selected parental strains are combined into the diploid ancestor. The genetic crossing eludes the problems in assembling and phasing diploid genomes. (*C*) Hybrids are evolved and accumulate different types of de novo variants. (*D*) Genomes from evolved hybrids are sequenced at high coverage by short-read sequencing. (*E*) Short reads are mapped against the assemblies of the two parental genomes separately (standard mappings) and against the concatenation of the two assemblies (competitive mapping). For simplicity, the standard mapping against only one parental assembly (PA) is reported. In the competitive mapping, reads from parent 1 (red) are expected to map to the assembly of parent 1 (PA1) on the basis of the presence of markers (green lines). Conversely, reads from parent 2 (blue) are expected to map to the assembly of parent 2 (PA2). Regions bearing no marker due to high sequence identity are characterized by reads with MAPQ = 0 (light gray reads). These regions are probed for small variants from standard mapping, without direct phasing, against one parental assembly (light red region in PA1), whereas the other one is masked (light blue segment in PA2). (*F*) Small variants obtained from competitive and standard mappings are combined into a single set of calls. CNVs and LOHs are called from standard mappings.

Table 1. Statistics of Markers for the Constructed Hybrids.

Species	Background (short ID)	Number of Markers	Markers %	f_c	f_r	LMDRs Fraction	
						Assembly 1	Assembly 2
S.c./S.c.	SK1/S288C	75,547	0.62	0.98	0.02	0.32	0.32
S.c./S.c.	UWOPS03-461.4/YPS128 (MA/NA)	63,926	0.54	0.81	0.19	0.30	0.31
S.c./S.c.	YPS128/DBVPG6765 (NA/WE)	78,064	0.66	0.99	0.01	0.24	0.24
S.p./S.c.	N17/DBVPG6765 (N17/WE)	1,095,399	9.19	0.96	0.04	0.11	0.11

Note.—*S.c.* and *S.p.* refer to *S. cerevisiae* and *S. paradoxus*, respectively. For each cross, we report the number of markers, the genome-wide percentage of markers, the fraction of markers lying in collinear regions (f_c), the fraction of markers lying in rearranged regions (f_r), and LMDRs fractions. The latter is the fraction of core genomic regions characterized by <1 marker in 300 bp, calculated from pairwise alignment of different pairs of assemblies.

value close to 1 (supplementary fig. S6, Supplementary Material online).

As the competitive approach has never been systematically benchmarked on a data set of simulated variants and inconsistencies among small variants callers have been reported (Yu and Sun 2013; Ghoneim et al. 2014; Li 2014; Pabinger et al. 2014; Sandmann et al. 2017; Rajaby and Sung 2018), we compared the performance of calling small variants, with both SAMtools and FreeBayes, from competitive and standard mappings as a function of the coverage. As expected, using standard mapping for complete homozygous diploids, the F_1 score increased with coverage showing saturation at 50× (fig. 2C). On the contrary, calling small variants from heterozygous diploid data mapped

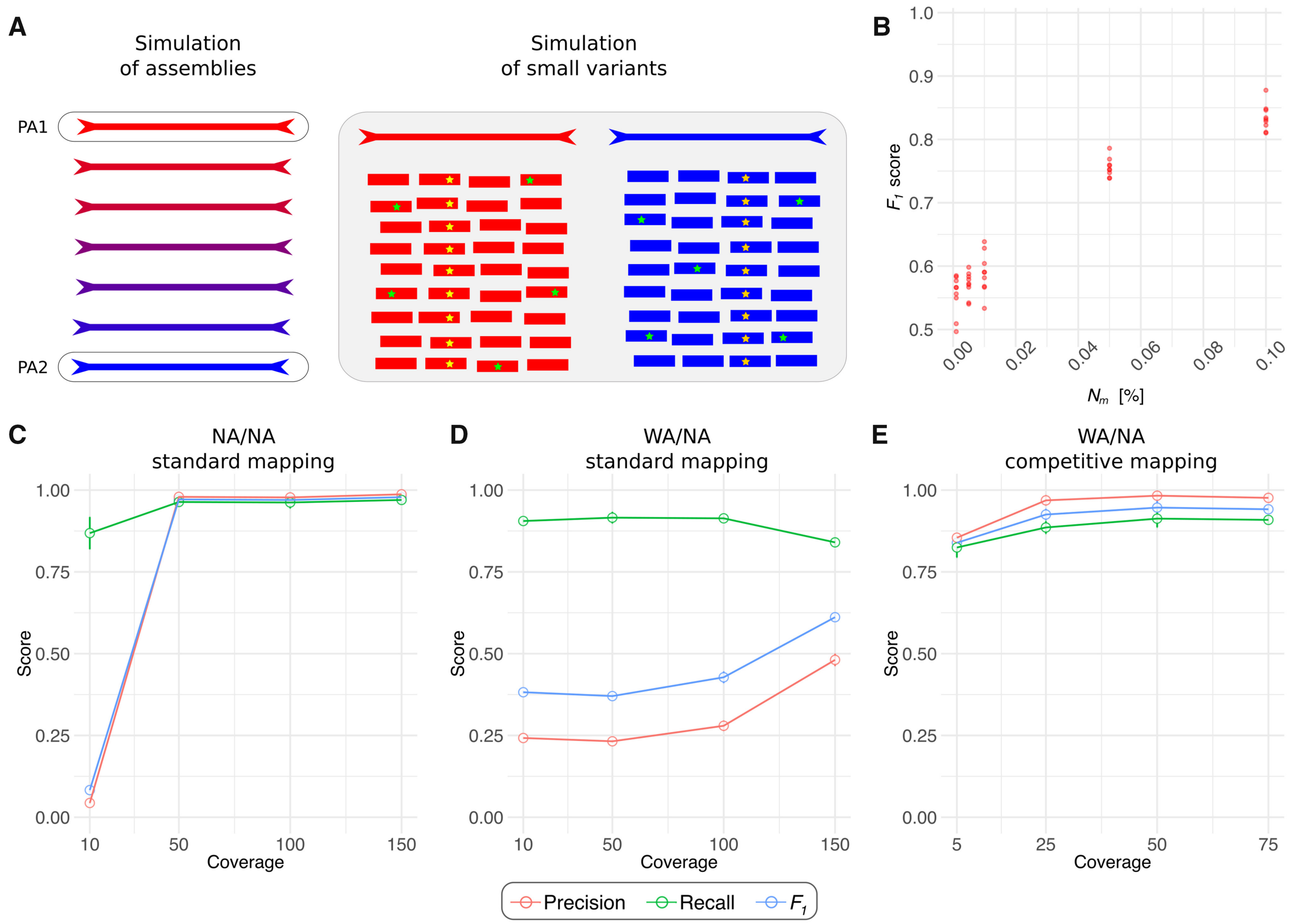

Fig. 2. High-accuracy variants detection. (*A*) Given two assemblies, various levels of heterozygosity are simulated by progressively replacing the markers in parental assembly 2 (PA2) with the corresponding alleles detected in parental assembly 1 (PA1). Small variants (yellow stars) are simulated in known genomic positions to assess the performance of variant calling algorithms from competitive and standard mapping separately, with Illumina reads bearing sequencing errors (green stars). (*B*) The overall performance (F_1 score, defined in the Simulated Data section) of small variants detection with competitive mapping decreases with the number of markers (N_m). For each value, three hybrid genomes were simulated from the WE/NA hybrid with three different replicates of short reads, carrying different variants. The lowest coverage showing F_1 score saturation (25×, panel *E*) was chosen to generate the short-read data set. DBVPG6044 (a West African *Saccharomyces cerevisiae* strain, WA in short) and NA assemblies are exploited to compare the performance of variant calling from competitive and standard mappings separately. Precision (red), recall (green), and F_1 score (blue) are reported as a function of coverage for (*C*) NA complete homozygous diploid data with standard mapping, (*D*) WA/NA heterozygous diploid data with standard mappings against both parental assemblies, and (*E*) WA/NA heterozygous diploid data with competitive mapping. Since we compared the same set of reads in the three approaches, competitive mappings show half of the coverage with respect to standard mappings. Solid lines serve as an eye guide.

with the standard approach provided low F_1 score with limited benefits by increasing coverage (fig. 2*D*). This effect is explained by spurious mapping of reads from parent 2 against the assembly of parent 1 (and vice versa) which leads to FPs. In fact, the low F_1 score can be ascribed to low precision. Instead, the competitive mapping of heterozygous diploid data (fig. 2*E*) yielded a large number of true positives (TPs) and high F_1 score, with a trend similar to the results (fig. 2*C*) obtained from the standard mapping of the complete homozygous diploid. Therefore, competitive mapping can be exploited to call small variants with direct phasing, although the overall performance is constrained by the number of false negatives (FNs) (recall in fig. 2*E*). Thus, we included in MuLoYDH a module that automatically calculates the boundaries of regions characterized by reads with low mapping quality (i.e., MAPQ $\leq$ 5). These regions are investigated through standard mapping. Although this prevents direct variant phasing, it allows for testing the presence of small variants in the whole accessible regions of the genome.

We also investigated the efficiency of the small variants detection strategy in genomic regions which are present only in one parent and which are not reported in the *S. cerevisiae* S288C reference genome. Using DBVPG6765/YPS128 (WE/NA in short, from the Wine/European and North American oak clades, respectively) hybrid data and the corresponding annotated assemblies, we calculated the F_1 score considering only the variants lying within a unique region of the WE strain on chromosome XV, derived from a 65-kb horizontal gene transfer (HGT) from *Torulaspora microellipsoides*, previously described (Marsit et al. 2015; Yue et al. 2017). We obtained

$F_1 = 0.96$ (TP = 14, FN = 1, and FP = 0) on the basis of 15 variants (14 SNV and a 1-bp insertion) combining all the simulated short-read data (20 experiments). Hence, MuLoYDH allows for calling small variants in regions which are not reported in the S288C reference genome.

Another aspect of the small variants calling procedure is whether MuLoYDH can correctly genotype de novo variants within LOH regions. LOH arises from double-strand break (DSB) repair processes thorough different homologous recombination pathways resulting in the loss of the DSB affected allele (Carr and Gottschling 2008). Depending on segregation patterns, recombination may produce interstitial/terminal LOHs. Interstitial events may result from single/double crossing over and gene conversion, whereas terminal LOHs may arise from single crossing over and break-induced replication (Laureau et al. 2016). Hence, LOH regions may carry homozygous de novo small variants (occurred before DSB repair) and heterozygous de novo small variants (occurred after DSB repair). Thus, we compared the genotypes of simulated variants with those reported by MuLoYDH, in WE/NA hybrids. MuLoYDH correctly called and genotyped 1,840 variants in the simulated LOH regions (691 homozygous and 1,149 heterozygous variants), producing 62 FPs and 207 FNs ($F_1 = 0.93 \pm 0.01$) with, as expected, a larger number of missed events in heterozygous state (121 heterozygous vs. 86 homozygous) (supplementary fig. S7, Supplementary Material online).

Overall, these results demonstrate that both competitive and standard mapping are required to maximize small variants calling performance. Competitive mapping provides direct variant phasing although it can be used only in regions characterized by a sufficient number of markers, whereas standard mapping is locally necessary if no marker exists in regions larger than twice the read length.

Applying the MuLoYDH Workflow to a Mutator Strain

We applied MuLoYDH to a SK1/BY hybrid with a mutator background (*tsa1*Δ/*tsa1*Δ, see data set 1 in Materials and Methods) evolved for 25 consecutive single-cell bottlenecks (Huang et al. 2003; Serero et al. 2014). This hybrid evolved drastically from its ancestor, thus providing a challenging test for our workflow and a benchmark to compare competitive and standard approaches for calling de novo small variants. It accumulated both de novo small variants and a series of complex LOH events (fig. 3A and B and supplementary tables S9–S11, Supplementary Material online). MuLoYDH provided a robust genotyping approach of markers positions determined by aligning the parental assemblies (fig. 3C). Exploiting high-quality markers, genotyped against both parental assemblies (Materials and Methods), 43 LOHs ranging from 97 to 591 kb were detected.

We further characterized the mutational landscape which included 34 SNVs, 1 indel, and 2 CNVs (fig. 3D–F). Twenty-four out of 34 SNVs were phased (12 to SK1, 12 to BY background) as well as the indel that occurred on the BY chromosome IV. The remaining 10 SNVs were called without phasing. Six of them were detected in BY LOH regions (four heterozygous, two homozygous), one in SK1 LOH regions (homozygous), while three variants were called from standard mapping. For validation, 11 variants (3 phased and 8 unphased) were tested through polymerase chain reaction and Sanger sequencing. All of them were validated as TPs. Six out of 8 unphased variants were heterozygous, whereas 2 SNVs, lying in LOH regions, were genotyped as homozygous and thus further supported that they resulted from a mitotic recombination event. We detected a short LOH segment (SK1 allele, 473 bp), supported by four markers, bearing the tRNA-Ser (AGA) gene. The SK1 LOH region lay within a large (>450 kb) BY LOH region. The latter carried a validated homozygous missense variant (C → T, YDR484W, fig. 3B, yellow star) that likely occurred before the large event. Remarkably, one validated intergenic heterozygous variant (fig. 3B, green star) lays within the aforementioned LOH region (BY chromosome IV). This mutational status suggests that a recombination event led to a short LOH followed by 1) the occurrence of a SNV (fig. 3B, yellow star), 2) a larger LOH, and finally 3) one heterozygous SNV (fig. 3B, right-most green star). Annotated electropherograms with validated variants and LOH markers are reported in supplementary figures S8 and S9, Supplementary Material online. Altogether, these results demonstrate that the markers quality-filtering approach for LOH detection provided accurate results also for events supported by few markers (supplementary fig. S10, Supplementary Material online). In addition, validations of the de novo small variants showed that the combination of two callers (both applied to competitive and standard mappings) and the implemented filtering strategy yielded reliable tracking of genomic variants (fig. 3D).

Evolution through Complex CNVs

Changes in copy number, from single gene to whole-chromosome events, have been observed in both natural and laboratory evolved strains (Dunham et al. 2002; Gresham et al. 2008; Dunn et al. 2013; Lauer et al. 2018). MuLoYDH produces CNV calls through Control-FREEC (Boeva et al. 2011) normalizing the read count (RC) signal for GC-content and mappability (Tattini et al. 2015). We tracked genome evolution in one intraspecies MA/NA *S. cerevisiae* hybrid evolved via the RTG protocol (data set 2 in Materials and Methods) (Laureau et al. 2016). Remarkably, a large fraction of the ancestor hybrid genome is noncollinear, due to a massive genome instability occurred in the Malaysian lineage (Yue et al. 2017). In particular, the MA chromosome VIII consists of a 350-kb collinear region that spans the centromere and a 390-kb translocation derived from chromosome VII, whereas the MA chromosome VII is a complex mosaic harboring two distinct regions from chromosome VIII (fig. 4A). Recombination between noncollinear homologous chromosome potentially results in complex CNVs.

The combination of CNV profiles and B-allele frequencies (BAFs) of markers, both calculated from standard mapping, shed light on complex events. The MA/NA hybrid showed multiple LOHs with two events occurring on both arms of chromosome VIII. Two DSBs likely occurred in the MA

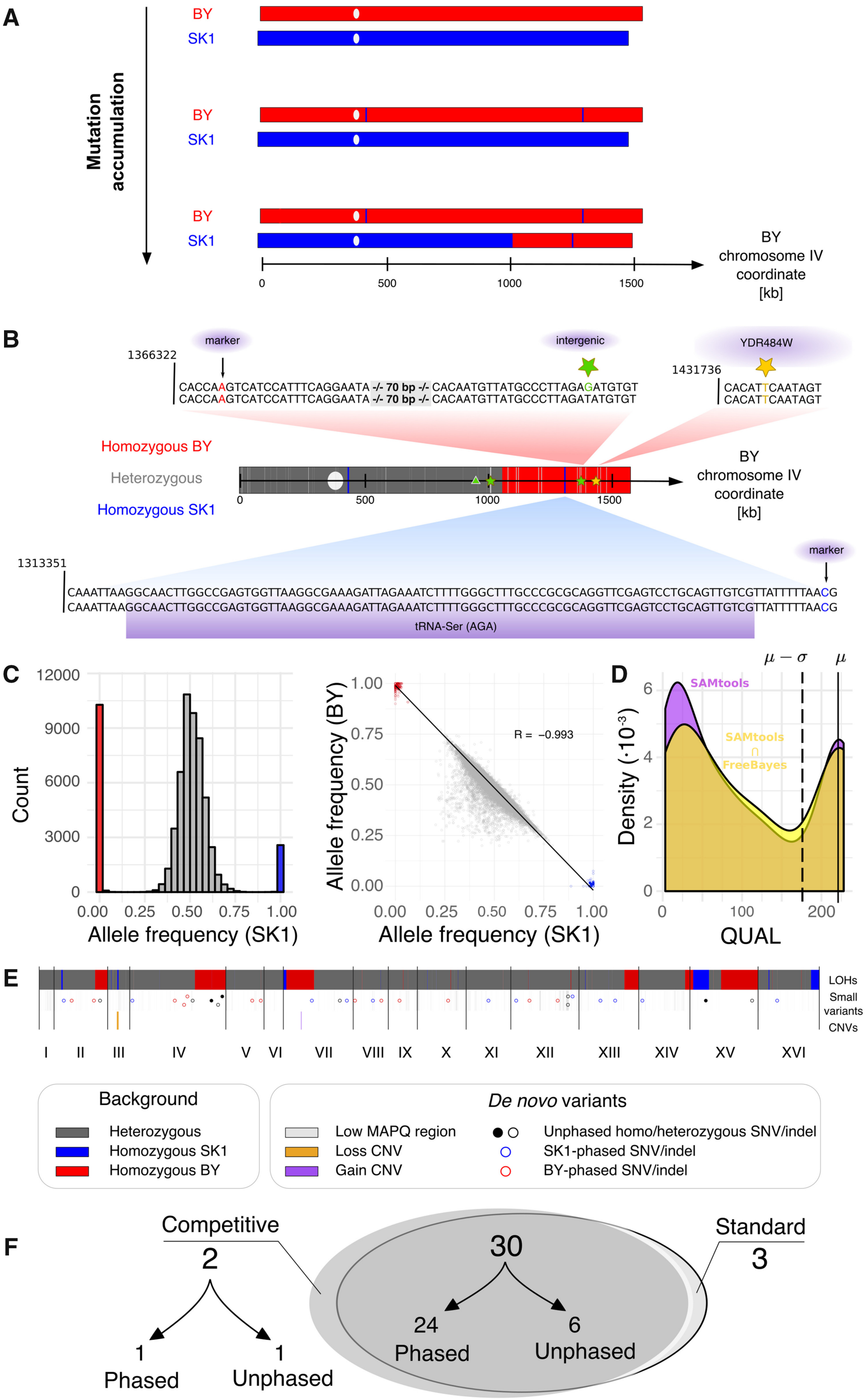

Fig. 3. MuLoYDH provides accurate tracking of the mutational landscape in a SK1/BY *tsa1*Δ/*tsa1*Δ MAL. (*A*) Hybrid evolution leads to LOH and (*B*) to small variants. One homozygous (yellow star) and one heterozygous (green star) SNVs were detected within LOH regions on chromosomes IV (red, BY; blue, SK1; dark gray, heterozygous segments; white oval, centromere). The presence of markers (black arrows) allows direct variant phasing through competitive mapping. A 1-bp deletion was detected in a heterozygous segment and phased to the BY chromosome (green triangle). One heterozygous SNV (green star) was detected from standard mapping (light gray segment). (*C*) The strategy implemented in MuLoYDH for the detection of LOHs allows noise mitigation (see also supplementary fig. S10, Supplementary Material online) as shown by the clear separation of genotypes with different allele frequencies and by the high negative correlation of allele frequencies. R is the Pearson correlation coefficient. Red (blue) dots/columns refer to homozygous BY (SK1) markers, whereas gray dots/columns refer to heterozygous markers. Markers are filtered on the basis of their quality values. (*D*) The same strategy is applied to de novo small variants. Variants with a quality value (QUAL)

chromosome VIII (fig. 4A, purple and yellow stars) and were repaired using the homologous NA chromosome VIII region. Chromosome VIII-L repair occurred within the collinear region and resulted in a simple LOH event without an associated CNV. The same holds for the collinear region in chromosome VIII-R spanning from the DSB to the breakpoint of the chromosome VIII/VII translocation. These regions showed RC $\simeq$ 2 and BAF $\simeq$ 0 (fig. 4B). In contrast, the rearranged chromosome VIII region embedded in the LOH was subjected to CNV, as shown by RC $\simeq$ 3 and BAF $\simeq$ 0.3. The latter supported the presence of two copies of NA alleles and one copy of MA allele. This complex genomic configuration was further confirmed by the RC signal and the BAF data from chromosome VII, given the loss of the MA chromosome VII translocated regions (fig. 4C and supplementary fig. S11, Supplementary Material online). Thus, combining the knowledge of the exact chromosomal configurations of the ancestor enabled to dissect complex CNVs.

Mutational Rates across Different Genetic Backgrounds

Mutational rates and signatures are key hallmarks of genome evolution but how these vary across different genetic backgrounds has remained largely unexplored. We applied the MuLoYDH workflow to investigate the effect of the genetic background on mutational rates. We constructed four yeast diploids that enabled multiple comparisons (data set 3 in Materials and Methods). We used a single *S. cerevisiae* background (WE) and crossed it to itself, to a different subpopulation of the same species (*S. cerevisiae* NA) and to a different species (*S. paradoxus* N17). This resulted in three diploids (with 0%, ∼0.5%, and ∼10% heterozygosity) that enabled the investigation of the effect of heterozygosity in laboratory evolution experiments. We also generated a complete homozygous *S. paradoxus* N17 diploid to compare *S. cerevisiae* to its closely related species and performed mutation accumulation experiments using 8 replicated lines for each of the 4 diploids subjected to 120 consecutive single-cell bottlenecks. The corresponding number of generations was estimated measuring the colony cell population size, observing minimal differences between the four diploids (supplementary table S2, Supplementary Material online). Taking advantage of two parental assemblies, we were able to accurately calculate the mutation rates for SNVs, indels, CNVs, aneuploidies, and LOHs (supplementary table S2, Supplementary Material online). Remarkably, although indirect estimates of LOH rates were derived exploiting de novo small variants in homozygous diploids (Sharp et al. 2018), we performed the first direct measurement of LOH rate and genome-wide patterns. Moreover, we assessed if selection occurred during the propagation of MALs and confirmed a scenario close to neutral evolution with no signature of selection (supplementary fig. S15 and supplementary tables S3 and S4, Supplementary Material online).

The SNVs (supplementary tables S12–S20, Supplementary Material online) and indels (supplementary tables S21–S24, Supplementary Material online) mutation rate per base-pair per generation ($R_{\mathrm{bp,grt}}$) in *S. cerevisiae* (supplementary table S2, Supplementary Material online) was consistent with previous estimates using diploid laboratory strains, indicating no major differences with the WE background used in this study (Sharp et al. 2018). However, we surprisingly observed a 4-fold lower (P value < 0.0005, Welch's t-test) SNVs mutation rate in *S. paradoxus* ($R_{\mathrm{bp,grt}}$ [95% CI t-distribution]: $7.27\,[3.39, 11.1] \times 10^{-11}$) compared with *S. cerevisiae* ($R_{\mathrm{bp,grt}}$ [95% CI t-distribution]: $2.82\,[1.98, 3.67] \times 10^{-10}$). The same trend, characterized by the slower evolution in *S. paradoxus* genomes, was observed considering the number of base-pairs shaped by indels per generation (fig. 5A and B) as well as for aneuploidies (fig. 5C and supplementary tables S25–S32, Supplementary Material online), whereas no CNV was detected in both the homozygous backgrounds. Thus, *S. paradoxus* diploids showed an overall higher genome stability compared with *S. cerevisiae*. The SNVs mutation rates of the two heterozygous hybrids were also different (P value $= 0.008$, one-sided Welch's t-test) with the highly heterozygous interspecies hybrid N17/WE showing higher rate ($R_{\mathrm{bp,grt}}$ [95% CI t-distribution]: $2.24\,[1.73, 2.74] \times 10^{-10}$) compared with the intraspecies hybrid NA/WE ($R_{\mathrm{bp,grt}}$ [95% CI t-distribution]: $1.49\,[1.09, 1.89] \times 10^{-10}$). The SNVs mutation rate of N17/WE hybrids was higher than the rate calculated for N17/N17 diploids (P value < 0.00001, Welch's t-test), whereas no significant difference was observed comparing N17/WE hybrids against WE/WE diploids (P value $= 0.18$, Welch's t-test). Although both background genotype and whole-genome heterozygosity may have an impact, we speculate that the WE haplotype is dominant with respect to mutation rate.

Mitotic recombination occurs rarely and usually requires a selectable marker to detect events at specific genomic loci (Lee et al. 2009). Nevertheless, given the large number of generations performed in our study, we were able to observe a considerable number of LOH events in the evolved hybrids. The two heterozygous diploids (fig. 5D) showed a substantial difference in terms of LOH rates per genome per generation ($R_{\mathrm{gnm,grt}}$ [95% CI t-distribution]: $6.25\,[4.79, 7.71] \times 10^{-3}$ and $2.98\,[2.29, 3.68] \times 10^{-3}$, respectively for NA/WE and N17/WE), with intraspecies hybrids showing higher rate with respect to the interspecies diploids (P value $= 0.00037$, Welch's t-test). Intraspecies hybrids showed a larger number of LOHs

FIG. 3. Continued
$< \mu - \sigma$ are masked. μ is the mean quality and σ is the standard deviation calculated from markers. Kernel density estimations are calculated from quality values of SAMtools calls (purple) and the intersection of SAMtools and FreeBayes calls (yellow). (*E*) The genome-wide mutational landscape includes CNVs: one gain event in chromosome VII (three BY copies) and one loss event of the BY allele in chromosome III. (*F*) Small variants detected from competitive and standard mapping are reported in the Venn diagram. Variants from competitive mapping are classified as phased and unphased. The latter were all detected within LOH regions.

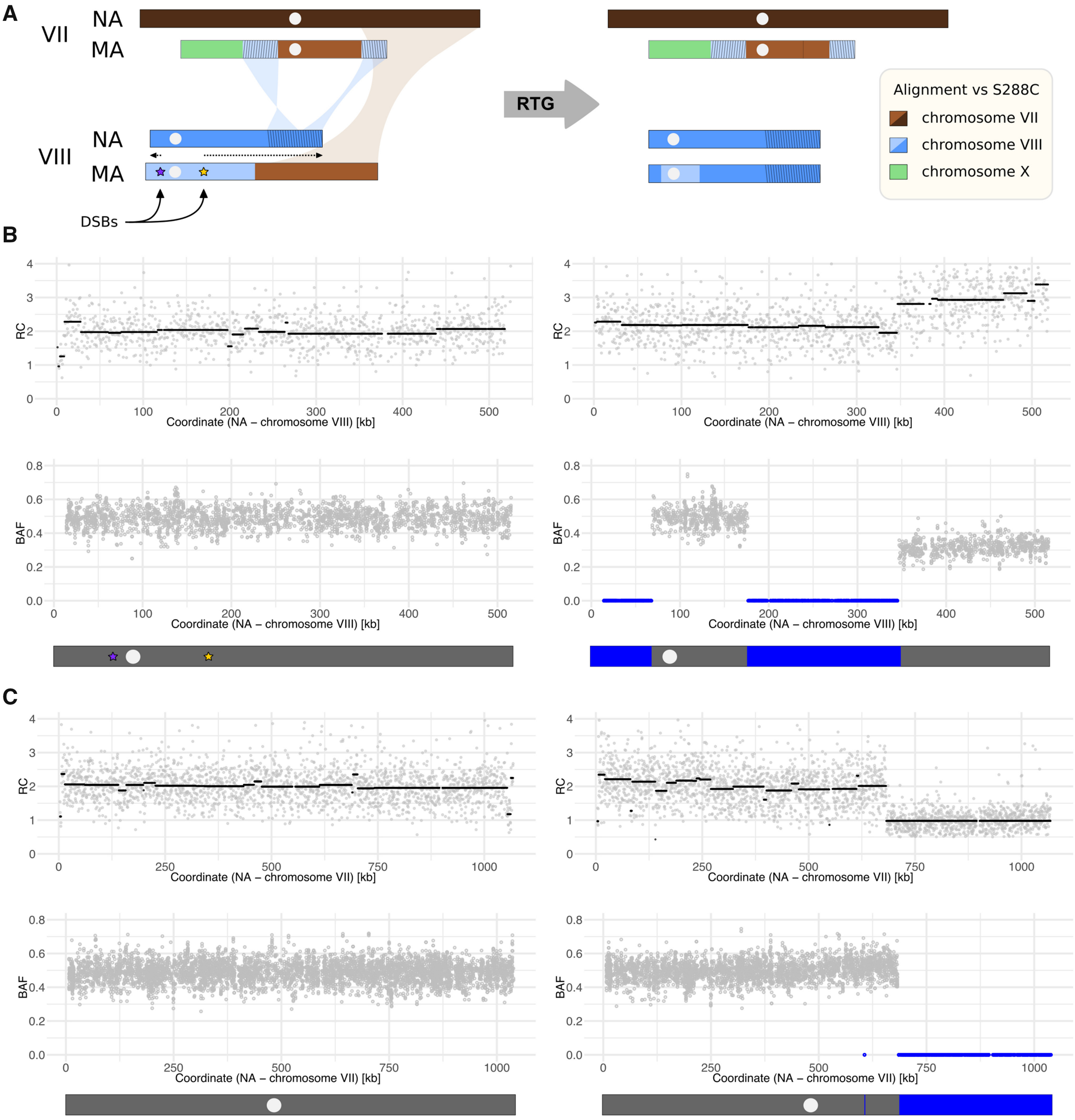

Fig. 4. Resolving complex CNVs. (*A*) Ancestor and RTG-evolved karyotype in a MA/NA hybrid. Light and dark colors encode for MA and NA alleles respectively. Chromosomal identities were assigned by homologous centromeres. DSBs in the MA chromosome VIII (purple and yellow stars) were repaired (dotted arrows) using the homologous chromosome from NA. These chromosomes bear collinear regions and large interchromosomal rearrangements, including a large inversion. Repairing a DSB (purple star) in a collinear region with the homologous chromosome leads to a terminal LOH. Repairing a DSB (yellow star) in a chromosome arm bearing collinear and rearranged regions leads to an interstitial LOH and to a complex CNV. (*B*) Chromosome VIII RC data support the presence of three copies of a large segment on the right arm. BAF data from markers using mapping against the NA assembly support the presence of a three-copy region (two copies of NA and one copy of MA). The two stretches of BAF values at zero refer to the terminal and interstitial LOH regions (both NA alleles). The former results from DSB repair (purple star, panel *A*) in chromosome VIII-L, whereas the latter is the outcome of DSB repair (yellow star, panel *A*) in chromosome VIII-R. (*C*) RC signal and BAF data support the deletion of a large segment of MA chromosome VII-R.

(111) compared with the interspecies hybrids (53) along with (fig. 5*E*) a 5-fold higher fraction of the genome in LOH (0.02 ± 0.01 and 0.004 ± 0.007, corresponding to ∼250 and ∼50 bp per generation undergoing LOH, respectively), and (fig. 5*F*) a higher number of large events (>25 kb), namely 3 and 0.4 on average per sample (supplementary tables S33–S37, Supplementary Material online). However, both hybrids were characterized by a large fraction of small interstitial

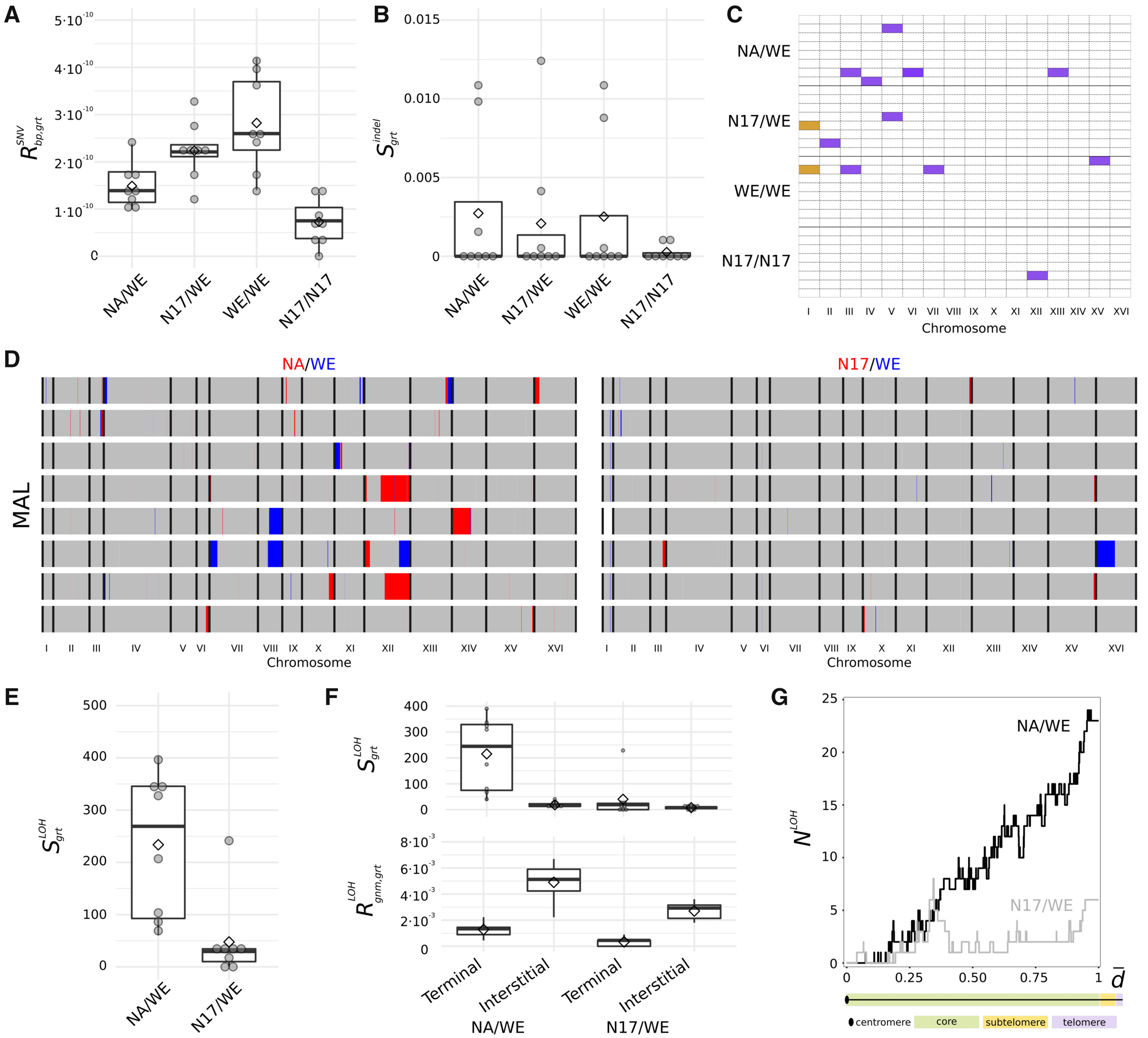

Fig. 5. Mutation signatures in different diploid backgrounds. Box plots of (*A*) SNVs mutation rate per base-pair per generation ($R^{\mathrm{SNV}}_{\mathrm{bp,grt}}$) and (*B*) the number of base-pairs shaped by indels per generation ($S^{\mathrm{indel}}_{\mathrm{grt}}$), for different homozygous and heterozygous diploid backgrounds. Diamonds represent mean values. (*C*) Aneuploidies patterns (orange and violet for loss and gain, respectively). (*D*) LOH events detected in NA/WE and N17/WE hybrids. LOHs are depicted in red (NA and N17 homozygous regions) and blue (WE homozygous regions), whereas heterozygous regions are represented in gray. White regions were masked due to aneuploidies resulting in whole-chromosome loss. (*E*) Box plots of the number of base-pairs shaped by LOHs per generation ($S^{\mathrm{LOH}}_{\mathrm{grt}}$) for intraspecies (NA/WE) and interspecies (N17/WE) hybrids. (*F*) Box plots of the number of base-pairs shaped by terminal and interstitial LOHs per generation ($S^{\mathrm{LOH}}_{\mathrm{grt}}$) for intraspecies (NA/WE) and interspecies (N17/WE) hybrids (top) and box plots of the rate of terminal and interstitial LOHs per genome, per generation ($R^{\mathrm{LOH}}_{\mathrm{gnm,grt}}$) (bottom). (*G*) Number of LOHs (N^{LOH}) detected in intraspecies (NA/WE) and interspecies (N17/WE) hybrids as a function of the relative distance ($\bar{d}$) from the centromere (normalized for chromosome-arm length).

events and few large terminal LOHs (fig. 5*F*), with the number of detected LOHs nonuniformly distributed along the genome (supplementary fig. S12 and supplementary file 8, Supplementary Material online) and increasing with the distance from the centromeres (fig. 5*G*). We also detected an LOH upstream the large HGT region unique to the WE strain (chromosome XV-R). The LOH occurred in a syntenic region producing an N17 homozygous region as well as the loss of the nonshared HGT, demonstrating how LOH may lead to CNVs (supplementary fig. S17, Supplementary Material online).

No bias toward one of the parental haplotypes was detected for LOHs, as in the case of SNVs (supplementary tables S35 and S16, Supplementary Material online, respectively). Overall, we provided a detailed overview of the mutational events that occurred in the four backgrounds. These results provided an unprecedented quantitative genome-wide measure of the importance of LOHs in shaping polymorphisms patterns in diploid hybrid genomes and how this process was strongly inhibited by very high heterozygosity.

Conclusions

Although HTS has proven revolutionary for genomic sciences, resequencing studies show intrinsic limits, particularly in the context of hybrid genomes. Variation graphs will contribute overcoming this deficiency but they will require extensive efforts to exhaustively support the shift to the new graph-based paradigm (Church 2018; Garrison et al. 2018). Long-read sequencing is a valuable approach to provide novel reference genomes by means of de novo assembly. The availability of novel reference genomes opens new perspectives on resequencing approaches, allowing for investigations of the genomic mutational landscape with unprecedented resolution via short-read experiments. Still, current methods for assembling and phasing diploid genomes are costly and yield to limited contiguity (Sedlazeck et al. 2018). Recently, long-read sequencing has been exploited to evaluate the performance of small variants callers through a synthetic-diploid benchmark (Li et al. 2018). Here, we extended the benefits of long-read sequencing beyond synthetic diploids to evolved *Saccharomyces* hybrids. MuLoYDH provides a framework for studying genome dynamics by tracking the mutational landscape of designed yeast diploid hybrids. Moreover, as the sequencing technologies enhance both read length and accuracy, they will soon allow to produce fully assembled and phased natural hybrid diploid genomes. At this stage, the initial crossing phase implemented in MuLoYDH to generate designed hybrids will be possibly bypassed while our computational strategy will be readily appropriate for the application to natural genomes.

MuLoYDH was developed to take advantage of the fully phased diploid genome assembly of the ancestor and Illumina short reads from the evolved hybrids. Haploid parents were combined into diploid hybrids with varying levels of heterozygosity which were evolved in different laboratory settings. Remarkably, the hybridization process can be easily automated (Hallin et al. 2016) and as the number of available annotated assemblies increases, the number of potential hybrids grows quadratically.

The presence of markers provided a reliable quality threshold for filtering de novo small variants, thus bypassing the need of multiple hard filters. Moreover, it enabled the phasing of de novo SNVs, indels and CNVs as well as the precise characterization of LOHs. Our method was designed to yield a quantitative measure of the fraction of the genome which cannot be probed for direct phasing through competitive mapping and to perform variant calling in these regions by means of a standard approach based on a single reference. It was devised to resolve the drawbacks of using a single consensus reference for the analysis of diploid hybrid genomes, namely 1) spurious read mapping, which may lead to FP calls of both SNV and indels, 2) the impossibility of probing variation in genomic regions which are not reported in the S288C reference, as well as 3) the impracticable direct phasing of de novo variants. Experimental validation of variants detected by MuLoYDH in a mutator MAL showed that it can be used to trace the time course of mutation occurrence with direct phasing information. Moreover, the analysis of the MA/NA hybrid proved that MuLoYDH can be used to dissect complex CNVs.

Although several studies focused on the mutation rates in haploid and complete homozygous diploid *S. cerevisiae* laboratory strains (Lynch et al. 2008; Nishant et al. 2010; Zhu et al. 2014; Sharp et al. 2018), we used MuLoYDH to track the mutational landscape of four diploid genetic backgrounds with varying levels of heterozygosity, providing the first measurement of mutation rates in *S. paradoxus*. Surprisingly, we observed low mutation rates in the natural *S. paradoxus* N17 homozygous diploids compared with *S. cerevisiae* WE. These results suggest that the lower mutation rate of *S. paradoxus* might have contributed to the slower evolution of this species compared with *S. cerevisiae*, as shown by the overall branch length differences observed between the two species since the split from their last common ancestor (Kellis et al. 2003; Yue et al. 2017).

Rates in intra- and inter-species hybrids revealed that SNVs and LOHs in particular, are major sources of genomic variability that play a key role in genome evolution. Although the mutation rate of SNVs was higher in interspecies hybrids compared with intraspecies diploids, the opposite was observed for LOHs. These results suggest that higher heterozygosity promotes genome evolution through SNVs, whereas, on the other hand, it inhibits recombination. However, similar patterns of recombination were observed, with both the hybrids showing a relatively large number of small interstitial events compared with the number of terminal LOHs detected. Centromere-proximal regions showed refractory to LOHs, recapitulating historical patterns of recombination occurred during species evolution (Peter et al. 2018). Remarkably, taking advantage of both parental assemblies, we were able to highlight LOHs resulting in unbalanced events, such as complex CNVs, in genomic regions that are present only in one parental subgenome. Given the high LOH rate observed, this might be a crucial mechanism for promoting genome evolution.

In view of the modular implementation of our computational approach, novel features can be added to extend the pipeline to nonclonal data from adaptive evolution experiments (Li et al. 2019) in order to monitor allele frequency shifts, to map interspecies introgressions (Almeida et al. 2014; Peter et al. 2018), to characterize industrial hybrids with highly complex genomes, or to study recombination in gametes obtained from these hybrids. Our approach can be readily applied to investigate the mechanisms and signatures underlying de novo mutations in mutator backgrounds or to explore how environmental factors impact mutation rates and spectra (Liu and Zhang 2019). Finally, extending MuLoYDH to other *Saccharomyces* hybrids that encompass the whole spectrum of heterozygosity will enable to disentangle heterozygosity effects from those background-related in shaping genome evolution.

Materials and Methods

Simulated Data

Simulations of Hybrid Genomes with Varying Levels of Heterozygosity

Diploid genomes with varying levels of heterozygosity were simulated by custom R scripts, modifying the number of markers between the two parental subgenomes. Given two input assemblies (WE and NA), marker positions were determined by MUMmer (NUCmer) (Kurtz et al. 2004). Decreasing values of markers percentage were obtained by progressively replacing the allele of assembly 1 with the corresponding allele, as determined by NUCmer, of assembly 2 in known positions. Starting from 0.66%, the substitution step was repeated in order to provide different levels of markers (0.5%, 0.1%, 0.05%, 0.01%, 0.005%, and 0.001%). For each value, three replicates were simulated.

Simulations of Short Reads for Heterozygous Hybrids

Simulated paired-end short reads were generated using the DWGSIM package (Escalona et al. 2016). In order to produce simulated short-read data from genome assemblies, two input reference assemblies were concatenated to produce a single multi-FASTA, which was sampled to build simulated paired-end (150 bp, insert size 500 bp) Illumina experiments with different coverage levels (10×, 50×, 100×, and 150×). The mutation rate was set to 10^{-5} with the purpose of balancing a relevant number of small variants (~240 per genome) with the storage and the computational resources required for data processing. All the simulations were performed using the following parameters: 0.01 error rate for both forward and reverse read (according to estimations from Illumina data), and 0.1 indel/SNV ratio. Base-quality parameters were set according to the experimental data reported in this study. Each simulation was performed in five replicates. The command line is reported in supplementary material, Supplementary Material online.

Simulations of Short Reads for Hybrids Bearing LOH Regions

Short-read data of WE/NA hybrids bearing LOHs (with WE alleles) were obtained using DWGSIM with heterozygous genomes with the exception of chromosome I for which two copies of the FASTA sequence of WE were used as input. In order to have a robust statistic, the mutation rate was set to 10^{-3} for chromosome I and to 10^{-5} for all the other chromosomes. The average coverage was set to 50× on the basis of the short-read simulations (fig. 2E). Ten replicates were produced. All the other parameters were set as described above. Overall, we simulated 2,304 variants in heterozygous regions of the genome and 2,081 in LOH regions (787 homozygous and 1,294 heterozygous).

Simulations of Short Reads from Simulated Hybrid Genomes

Short reads from simulated hybrid genomes with different levels of heterozygosity (as described above) were obtained using DWGSIM with the parameters reported above. The average coverage was set at 50× on the basis of the short-read simulations (fig. 2E). Each simulation was performed in three replicates.

Performance of Small Variants Calling

Given a set of relevant elements (i.e., the simulated variants) and a set of selected elements (i.e., the called variants), we classified each element (namely each variant) as TP, FP, or FN. We calculated precision as $P = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and recall as $R = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. The performance of the small variants calling was quantified in terms of the F_1 score which was calculated as the harmonic mean of (P) and (R) according to $F_1 = 2(P \times R)/(P + R)$. All the calculations were performed after filtering out marker positions (both single-nucleotide markers and indels) determined by NUCmer as described below. In order to fairly compare competitive and standard mapping, the latter approach was run using a control sample for variant subtraction. This allowed for filtering out marker positions (both single-nucleotide markers and indels) which could not be detected by NUCmer.

Experimental Data

Samples from data sets 1 and 2 were chosen from large sequencing data sets to benchmark MuLoYDH against genomes harboring a large number of both small variants and LOHs (data set 1), and complex CNVs (data set 2). Sample IDs are reported in supplementary material, Supplementary Material online.

Data set 1 comprises the sequencing data from a mutation accumulation experiment using a mutator SK1/BY hybrid evolved for 25 single-cell bottlenecks (*MATa/MATα*; *ARG4/arg4-nsp*, *bgl*; *his3Δ1/HIS3*; *leu2Δ0/leu2*; *met15Δ0/MET15*; *ura3Δ0/ura3*; *tsa1::KanMX/tsa1::KanMX*) generated as previously described (Serero et al. 2014), and the corresponding ancestor. This data set was analyzed using the SK1 and S288C assemblies included in MuLoYDH.

Data set 2 consists of the short reads from the UWOPS03-461.4/YPS128 (MA/NA in short) hybrid (low sequence divergence; noncollinear genomes with chromosomal rearrangements) evolved under the RTG protocol (adapted from Laureau et al. [2016]) and the corresponding ancestor. The hybrid (*MATa/MATα*, *ho::HygMX/ho::HygMX*, *ura3::KanMX/ura3::KanMX*, *leu2Δ0/LEU2*, *met15Δ0/MET15*, *LYS2/lys2::URA3*) was patched from the −80 °C glycerol stock on YPD solid media (1% yeast extract, 2% peptone, 2% dextrose, and 2% agar) and incubated overnight at 30 °C. The following day the strain was streaked to minimal solid media not supplemented with uracil and the plate is incubated at 30 °C for 48 h. Different single colonies of the hybrid strain were taken and inoculated separately in 10 ml of presporulation media YPEG (1% yeast extract, 2% peptone, 3% ethanol, and 3% glycerol) for 15 h at 30 °C with shaking at 220 rpm. Each presporulation culture was washed twice with sterile water and resuspended in 2% potassium acetate ($\mathrm{OD}_{600} = 0.5$) using a 250-ml flask that was incubated at 23 °C with shaking at 220 rpm. One milliliter was collected from the culture at the beginning of sporulation and another 1-ml sample after 6 h of incubation. The two samples were washed twice with

1-ml YPD and incubated in 1-ml YPD for 18 h at 30 °C without shaking. The following day the YPD liquid cultures were vortexed and 20 µl of each culture were plated on minimal media containing 1 mg/ml 5-fluoroorotic acid (5-FOA) and spread with glass beads (Vázquez-García et al. 2017). The plate was then incubated at 30 °C for 48 h.

Data set 3 is composed by the sequencing data of MALs (and the corresponding ancestors) from four prototroph diploid backgrounds: N17/DBVPG6765 (N17/WE in short), YPS128/DBVPG6765 (NA/WE in short), N17/N17, and DBVPG6765/DBVPG6765 (WE/WE in short). All the lines were isogenic in respect of the markers used for construction. Each mutation accumulation experiment consisted of eight independently propagated lines. *Saccharomyces cerevisiae* WE homozygous diploids (*MATa/MATα*, *ho::HygMX/ho::HygMX*, *ura3::KanMX/ura3::KanMX*, *LYS2/lys2::URA3*) were derived from the Wine/European subpopulation. *Saccharomyces paradoxus* N17 homozygous diploids (*MATa/MATα*, *ho::HygMX/ho::HygMX*, *ura3::KanMX/ura3::KanMX*, *LYS2/lys2::URA3*) were derived from the European subpopulation. NA/WE *S. cerevisiae* intraspecies hybrids (*MATa/MATα*, *ho::HygMX/ho::HygMX*, *ura3::KanMX/ura3::KanMX*, *LYS2/lys2::URA3*) were obtained by mating of North American (NA) and Wine/European (WE) haploid strains. N17/WE interspecies hybrids (*MATa/MATα*, *ho::HygMX/ho::HygMX*, *ura3::KanMX/ura3::KanMX*, *LYS2/lys2::URA3*) were obtained by mating a *S. paradoxus* haploid strain from the European subpopulation (N17) and a *S. cerevisiae* haploid strain from the Wine/European subpopulation (WE). MALs were propagated from each parental background on YPD solid medium and passed through a single-cell bottleneck every ~48 h (~20 generations) at 30 °C, for a total of 120 bottlenecks (~2,400 generations). At each single-cell bottleneck, a random colony was streaked to isolate the next single colony. To avoid any involuntary selection, at each streak, the closest colony to the center of the plate was picked, independently of its size. To determine the number of generations passed after 48 h, three colonies for each parental background were independently resuspended in 100 µl of sterile water and serially diluted. Twenty microliters of each dilution were plated on solid YPD medium and grown for ~48 h at 30 °C. The number of colonies was manually counted in the plate with suitable dilution and the number of generations (*G*) was estimated according to $G = \log_2(n \times d)$, where *n* is the number of cells counted on the plate and *d* is the corresponding dilution factor. After 120 single-cell bottlenecks, cells were inoculated in 5-ml liquid YPD cultures and grown overnight at 30 °C in a shaking incubator. DNA was extracted using "Yeast Masterpure" kit (Epicentre, USA) following the manufacturer's instructions.

Sequencing

Illumina paired-end libraries (2×150 bp) were prepared according to manufacturer's standard protocols and sequenced with an HiSeq 2500 instrument, at the NGS platform of Institut Curie. Coverage statistics are reported in supplementary table S38, Supplementary Material online.

Experimental Validation of Variants

Seven markers supporting two LOHs and nine SNVs variants from the SK1/BY hybrid were validated by Sanger sequencing. SNVs were randomly selected to avoid any bias. A pair of primers (upstream and downstream) was designed for each SNV using Unipro UGENE (Okonechnikov et al. 2012). Polymerase chain reaction products were sequenced by Eurofins Genomics. The presence and the genotype of the variants were checked by visual inspection of the electropherograms.

Data Analysis

Parental Assemblies

All the parental assemblies reported in this work were downloaded from the "Population-level Yeast Reference Genomes" website (https://yjx1217.github.io/Yeast_PacBio_2016/welcome/; last accessed August 07, 2019). The assembly of *S. paradoxus* strain N17 was obtained correcting the genome sequence of its close relative CBS432, for which a complete assembly is available (Liti et al. 2009; Yue et al. 2017). The correction was performed using Pilon (Walker et al. 2014) with short-read data from Illumina sequencing of a diploid homozygous N17 sample. The command line is reported in supplementary material, Supplementary Material online.

MuLoYDH General Description

The MuLoYDH pipeline requires as input 1) a data set of short-read sequencing experiments from yeast diploid hybrids and 2) the two parental genomes which were used to produce the hybrids in FASTA format as well as the corresponding annotations in the "general feature format" (GFF) (supplementary fig. S18, Supplementary Material online). Reads from hybrid data are mapped against the assemblies of the two parental genomes separately (standard mappings) and against the union of the two aforementioned assemblies (namely a multi-FASTA obtained concatenating the two original assemblies) to produce the competitive mappings (fig. 1*E*). In the latter case, reads from parent 1 are expected to map to the assembly of parent 1 on the basis of the presence of single-nucleotide markers. Conversely, reads from parent 2 are expected to map to the assembly of parent 2. Standard mappings are used to determine the presence of CNVs. The latter are also exploited to discriminate LOHs due to recombination from those resulting by deletion of one parental allele. The markers between the parental assemblies are determined by the NUCmer algorithm and are exploited to map LOH segments. Markers are genotyped from standard mappings. De novo small variants are determined from both competitive and standard mappings. Competitive mapping allows for direct variant phasing in heterozygous regions. Variant calling from competitive mapping is performed setting ploidy = 1 in heterozygous regions and ploidy = 2 in LOH blocks. Regions characterized by reads with low mapping quality (MAPQ ≤ 5 in the competitive mapping) are assessed from standard mapping using arbitrarily the assembly from parent 1. All the scripts described in the following sections are embedded in MuLoYDH.

Quality Check, Mapping, Mapping Refinement, and Coverage Calculation

Data quality is assessed by FastQC version 0.11.4. Competitive and standard mappings of Illumina reads are performed with BWA version 0.7.12-r1039 using the MEM algorithm (Li and Durbin 2009). Assemblies can be downloaded from the "Population-level Yeast Reference Genomes" website (https://yjx1217.github.io/Yeast_PacBio_2016/welcome/; last accessed August 07, 2019). Duplicates are removed by SAMtools 1.3.1 (using HTSlib 1.3.1). Depth of coverage is calculated with SAMtools (depth) and awk scripts (supplementary table S38, Supplementary Material online).

Determination of Single-Nucleotide Marker Positions

Single-nucleotide marker positions are determined through the NUCmer algorithm (MUMmer version 3) [with show-snps -ClrT] (Kurtz et al. 2004). In order to obtain reliable marker positions and take advantage of the "seed and extend" strategy of the algorithm, markers are calculated in both direct (assembly 1 vs. assembly 2) and reverse (assembly 2 vs. assembly 1) ways. The intersection of the two sets is retained for LOH detection and to calculate statistics. In the collinear mode, markers are determined chromosome-by-chromosome, aligning a chromosome of parent 1 against the corresponding homologous from parent 2, whereas with the rearranged option, they are calculated through a single whole-genome alignment of parental assemblies.

Classification of Single-Nucleotide Markers

Markers are classified as lying in collinear or rearranged regions as determined by MUMmer and custom R scripts. The fraction of markers within collinear regions (f_c) is calculated as $f_c = 1 - f_r$, where f_r is the fraction of markers lying within rearranged regions, namely inter- and intra-chromosome inversions and translocations.

Markers Genotyping, Small Variants Calling, Annotation, and Filtering

Markers calling and genotyping is performed using SAMtools (mpileup) [-u -min-MQ5 –skip-indels -E] and BCFtools (call) [-c -Oz] from standard mappings. Markers are quality-filtered removing those with quality $< (\mu - \sigma)$, where μ is the sample marker mean quality value and σ is the corresponding standard deviation.

The strategy implemented in MuLoYDH for calling de novo small variants relies on a stringent procedure to limit the number of FPs and keep the number of FNs as low as possible. Thus, in order to balance performance (in terms of F_1 score) and both the required computational resources and running time, two general-purpose small variants callers are implemented in MuLoYDH. De novo SNVs and indels are called with 1) SAMtools (mpileup) [-u -min-MQ5 -E] and BCFtools (call) [-c -Oz] and 2) FreeBayes (Li 2011a; Garrison and Marth 2012). Only variants called by both are retained. Both callers are exploited using competitive and standard mappings as described above. Regions characterized by reads with MAPQ $\leq$ 5 in competitive mappings are determined by custom R scripts, bash scripts and bedtools (Quinlan and Hall 2010). Parental and control hybrid variation is subtracted from hybrids data using custom bash scripts, VCFtools (Danecek et al. 2011) and tabix (Li 2011b). The resulting variants are quality-filtered masking those characterized by quality $< (\mu - \sigma)$, where μ is the sample markers mean quality value and σ is the corresponding standard deviation. Variants bearing marker alleles are filtered out, whereas those lying within (sub)telomeric regions are masked. Small variants are annotated by means of SnpEff (Cingolani et al. 2012). SnpEff database is built exploiting the annotations from the "Population-level Yeast Reference Genomes" website.

CNVs Calling and Annotation

CNVs are estimated by means of Control-FREEC with no matched normal samples, using standard mappings against both parental genomes (Boeva et al. 2011). RC data are normalized by GC-content and mappability. Mappability is calculated with GEM-mappability (Derrien et al. 2012). Results are annotated with *P* values calculated with both Kolmogorov–Smirnov and Wilcoxon Rank-Sum tests.

BAF Calculation

BAF values are calculated from standard mapping as $N_a/(N_r + N_a)$, where N_a is the number of read bearing the most abundant alternative (nonreference) allele and N_r is the number of reads bearing the reference allele at each marker position.

LOH Detection and Annotation

LOH regions are determined and annotated using custom R scripts. Considering standard mappings of each hybrid against both parental assemblies, marker positions characterized by nonmatching genotype or alternate allele are filtered out, as well as multiallelic sites, whereas those lying in subtelomeric and telomeric regions are masked. Moreover, being our approach based on both parental assemblies, MuLoYDH calls LOHs without filtering out small events using an arbitrary threshold based on the number of supporting markers (Laureau et al. 2016; Dutta et al. 2017). This aspect is crucial since we aim at comparing LOH rates in *S. cerevisiae*/*S. cerevisiae* and *S. paradoxus*/*S. cerevisiae* crosses. Markers involved in large deletions, as predicted by Control-FREEC, are masked. Finally, stretches of consecutive marker positions are grouped in LOH regions. Genomic coordinates of each LOH event are determined using both the "first/last" coordinates and the "start/end" coordinates. First/last coordinates are determined using the coordinates of the first and the last markers of the event. Start/end coordinates are calculated using the average coordinate of the first (last) marker and the last (first) marker of the adjacent event. LOH regions are annotated as terminal/interstitial as well as with genomic features embedded and those potentially involved in breakpoints. Annotation is performed based on the genomic features downloaded from the "Population-level Yeast Reference Genomes" website. Interstitial LOHs are defined as homozygous segments that are flanked on both sides by heterozygous markers. Terminal

LOHs are defined as homozygous regions extended to the end of the chromosomal arm.

Calculation of LMDRs

Regions characterized by <1 marker in 300 bp are calculated using custom R scripts which are embedded in the MuLoYDH pipeline.

Platform

MuLoYDH was developed, tested and optimized using a Linux environment (OS openSUSE 13.2 x86_64), equipped with 64 Intel Xeon CPUs (E7-4820, 2.0GHz).

Variants Filtering in MALs and Calculation of Mutation Rates

Small variants in WE and N17 homozygous backgrounds were quality-filtered on the basis of the values calculated from the markers of N17/WE hybrids as described above. All the small variants called by MuLoYDH were checked by visual inspection using IGV (Robinson et al. 2011). We also refined the lists of called CNVs by visual inspection in order to 1) avoid *FP*s due to small events which were not called in the control sample and 2) merge large events (e.g., aneuploidies) which were called as multiple shorter events. For each sample, we calculated the mutation rates dividing the number of variants detected and verified by visual inspection for number of generations calculated and for the length of the corresponding genome. Subtelomeric and telomeric regions were excluded from the calculation of small variants to avoid errors due to repeated regions.

Analysis of Homozygous Diploids

In order to analyze data from homozygous diploids, we set up a dedicated pipeline (mirroring MuLoYDH) which is described in the following section. Reads from homozygous diploids were mapped against the proper assembly with BWA version 0.7.12-r1039 (MEM algorithm). Assemblies were downloaded from the "Population-level Yeast Reference Genomes" website. Duplicates were removed by means of SAMtools 1.3.1 (using HTSlib 1.3.1). Depth of coverage was calculated with SAMtools (depth) and awk scripts. Following duplicates removal, small variants were called with SAMtools and FreeBayes. The intersection of their outputs was retained and variants reported in control samples were removed. Small variants were annotated by means of SnpEff. SnpEff database was built exploiting the annotation data downloaded from the "Population-level Yeast Reference Genomes" website. The presence of CNVs was assessed by means of Control-FREEC with no matched normal samples. RC data were normalized by GC-content and mappability, whereas the latter was calculated by means of GEM-mappability. Results were annotated with *P* values calculated with both Kolmogorov–Smirnov and Wilcoxon Rank-Sum tests. The pipeline is available at https://bitbucket.org/lt11/muloydhom; last accessed August 07, 2019.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by 1) Agence Nationale de la Recherche (ANR-11-LABX-0028-01, ANR-13-BSV6-0006-01, ANR-16-CE12-0019, ANR-16-JPEC-0005, ANR-18-CE12-0004, and ANR-10-EQPX-03 Investissement d'Avenir to the IC NGS platform) and Fondation ARC pour la Recherche sur le Cancer (PJA20151203273) and 2) French government, through the UCAJEDI Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01. S.M. is funded by the convention CIFRE 2016/0582 between Meiogenix and ANRT. We thank Matteo De Chiara for discussions, Olivier Croce for technical support, Agnès Llored for experimental help, Gilles Fischer for critical reading of the manuscript, and the NGS platform of the Institut Curie for NGS sequencing.

Author Contributions

L.T. designed and implemented computational methods, performed simulations, analyzed data, and wrote the manuscript. N.T. analyzed data and performed simulations. S.M. tested computational methods and performed experimental validations. M.D.A. performed mutation accumulation experiments. S.L. conducted experiments. A.N. revised the manuscript. G.L. coordinated and designed the study and wrote the manuscript. All authors discussed, critically revised, and approved the final version of the manuscript.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* 526(7571):68–74.

Almeida P, Gonçalves C, Teixeira S, Libkind D, Bontrager M, Masneuf-Pomarède I, Albertin W, Durrens P, Sherman DJ, Marullo P, et al. 2014. A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. *Nat Commun.* 5(1):4044.

Ameur A, Che H, Martin M, Bunikis I, Dahlberg J, Höijer I, Häggqvist S, Vezzi F, Nordlund J, Olason P, et al. 2018. De novo assembly of two Swedish genomes reveals missing segments from the human grch38 reference and improves variant calling of population-scale sequencing data. *Genes (Basel)* 9(10):486.

Barrick JE, Lenski RE. 2013. Genome dynamics during experimental evolution. *Nat Rev Genet.* 14(12):827–839.

Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, Barillot E. 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* 27(2):268–269.

Browning SR, Browning BL. 2011. Haplotype phasing: existing methods and new developments. *Nat Rev Genet.* 12(10):703–714.

Carr LL, Gottschling DE. 2008. Does age influence loss of heterozygosity? *Exp Gerontol.* 43(3):123–129.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. *Saccharomyces* genome database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40(D1):D700–D705.

Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ. 2018. Comparison of phasing strategies for whole human genomes. *PLoS Genet.* 14(4):e1007308.

Church DM. 2018. Genomes for all. *Nat Biotechnol*. 36(9):815–816.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2):80–92.

Coelho MC, Pinto RM, Murray AW. 2019. Heterozygous mutations cause genetic instability in a yeast model of cancer evolution. *Nature* 566(7743):275–278.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and vcftools. *Bioinformatics* 27(15):2156–2158.

Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast computation and applications of genome mappability. *PLoS One* 7(1):e30377.

Duina AA, Miller ME, Keeney JB. 2014. Budding yeast for budding geneticists: a primer on the *Saccharomyces cerevisiae* model system. *Genetics* 197(1):33–48.

Dujon B. 2010. Yeast evolutionary genomics. *Nat Rev Genet*. 11(7):512–524.

Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D. 2002. Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A*. 99(25):16144–16149.

Dunn B, Paulish T, Stanbery A, Piotrowski J, Koniges G, Kroll E, Louis EJ, Liti G, Sherlock G, Rosenzweig F. 2013. Recurrent rearrangement during adaptive evolution in an interspecific yeast hybrid suggests a model for rapid introgression. *PLoS Genet*. 9(3):e1003366.

Dutta A, Lin G, Pankajam AV, Chakraborty P, Bhat N, Steinmetz LM, Nishant KT. 2017. Genome dynamics of hybrid *Saccharomyces cerevisiae* during vegetative and meiotic divisions. *G3 (Bethesda)* 7(11):3669–3679.

Editorial. 2018. A reference standard for genome biology. *Nat Biotechnol*. 36:1121.

Eggertsson HP, Jonsson H, Kristmundsdottir S, Hjartarson E, Kehr B, Masson G, Zink F, Hjorleifsson KE, Jonasdottir A, Jonasdottir A, et al. 2017. Graphtyper enables population-scale genotyping using pangenome graphs. *Nat Genet*. 49(11):1654–1660.

Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, et al. 2014. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)* 4(3):389–398.

Escalona M, Rocha S, Posada D. 2016. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat Rev Genet*. 17(8):459–469.

Fry JD, Keightley PD, Heinsohn SL, Nuzhdin SV. 1999. New estimates of the rates and effects of mildly deleterious mutation in drosophila melanogaster. *Proc Natl Acad Sci U S A*. 96(2):574–579.

Gallone B, Steensels J, Prahl T, Soriaga L, Saels V, Herrera-Malaver B, Merlevede A, Roncoroni M, Voordeckers K, Miraglia L, et al. 2016. Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* 166(6):1397–1410.e16.

Garg S, Rautiainen M, Novak AM, Garrison E, Durbin R, Marschall T. 2018. A graph-based approach to diploid genome assembly. *Bioinformatics* 34(13):i105–i114.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.

Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, et al. 2018. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*. 36(9):875–879.

Gerstein AC, Kuzmin A, Otto SP. 2014. Loss-of-heterozygosity facilitates passage through Haldane's sieve for *Saccharomyces cerevisiae* undergoing adaptation. *Nat Commun*. 5(1):3819.

Ghoneim DH, Myers JR, Tuttle E, Paciorkowski AR. 2014. Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. *BMC Res Notes* 7(1):864.

Goffeau A, Barrell B, Bussey H, Davis R, Dujon B, Feldmann H, Galibert F, Hoheisel J, Jacq C, Johnston M, et al. 1996. Life with 6000 genes. *Science* 274(5287):546–567.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 17(6):333–351.

Gresham D, Desai MM, Tucker CM, Jenq HT, Pai DA, Ward A, DeSevo CG, Botstein D, Dunham MJ. 2008. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet*. 4(12):e1000303.

Hallin J, Märtens K, Young AI, Zackrisson M, Salinas F, Parts L, Warringer J, Liti G. 2016. Powerful decomposition of complex traits in a diploid model. *Nat Commun*. 7(1):13311.

Hittinger CT. 2013. *Saccharomyces* diversity and evolution: a budding model genus. *Trends Genet*. 29(5):309–317.

Huang M-E, Rio A-G, Nicolas A, Kolodner RD. 2003. A genomewide screen in *Saccharomyces cerevisiae* for genes that suppress the accumulation of mutations. *Proc Natl Acad Sci U S A*. 100(20):11529–11534.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937):241–254.

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol*. 36(12):1174.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol*. 5(2):R12.

Lang GI, Murray AW. 2008. Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. *Genetics* 178(1):67–82.

Langdon QK, Peris D, Kyle B, Hittinger CT. 2018. sppider: a species identification tool to investigate hybrid genomes with high-throughput sequencing. *Mol Biol Evol*. 35(11):2835–2849.

Lauer S, Avecilla G, Spealman P, Sethia G, Brandt N, Levy SF, Gresham D. 2018. Single-cell copy number variant detection reveals the dynamics and diversity of adaptation. *PLoS Biol*. 16(12):e3000069.

Laureau R, Loeillet S, Salinas F, Bergström A, Legoix-Né P, Liti G, Nicolas A. 2016. Extensive recombination of a yeast diploid hybrid through meiotic reversion. *PLoS Genet*. 12(2):e1005781.

Lee PS, Greenwell PW, Dominska M, Gawel M, Hamilton M, Petes TD. 2009. A fine-structure map of spontaneous mitotic crossovers in the yeast *Saccharomyces cerevisiae*. *PLoS Genet*. 5(3):e1000410.

Li H. 2011a. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21):2987–2993.

Li H. 2011b. Tabix: fast retrieval of sequence features from generic tab-delimited files. *Bioinformatics* 27(5):718–719.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30(20):2843–2851.

Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D. 2018. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods*. 15(8):595–597.

Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25(14):1754–1760.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 18(11):1851–1858.

Li J, Vázquez-García I, Persson K, González A, Yue J-X, Barré B, Hall MN, Long A, Warringer J, Mustonen V, et al. 2019. Shared molecular targets confer resistance over short and long evolutionary timescales. *Mol Biol Evol*. 36(4):691–708.

Liti G, Barton DBH, Louis EJ. 2006. Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. *Genetics* 174(2):839–850.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* 458(7236):337–341.

Liu H, Zhang J. 2019. Yeast spontaneous mutation rate and spectrum vary with environment. *Curr Biol*. 29(10):1584–1591.e3.

Long A, Liti G, Luptak A, Tenaillon O. 2015. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nat Rev Genet*. 16(10):567–582.

Lopandic K. 2018. *Saccharomyces* interspecies hybrids as model organisms for studying yeast adaptation to stressful environments. *Yeast* 35(1):21–38.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A*. 105(27):9272–9277.

Magi A, D'Aurizio R, Palombo F, Cifola I, Tattini L, Semeraro R, Pippucci T, Giusti B, Romeo G, Abbate R, et al. 2015. Characterization and identification of hidden rare variants in the human genome. *BMC Genomics*. 16:340.

Magi A, Pisanti N, Tattini L. 2016. The source of the data flood: sequencing technologies. *Ercim News* 104:25–26.

Magi A, Tattini L, Benelli M, Giusti B, Abbate R, Ruffo S. 2012. WNP: a novel algorithm for gene products annotation from weighted functional networks. *PLoS One* 7(6):e38767.

Marcet-Houben M, Gabaldón T. 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol*. 13(8):e1002220.

Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, Skov L, Belling K, Theil Have C, Izarzugaza JMG, et al. 2017. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* 548(7665):87–91.

Marsit S, Leducq J-B, Durand É, Marchant A, Filteau M, Landry CR. 2017. Evolutionary biology through the lens of budding yeast comparative genomics. *Nat Rev Genet*. 18(10):581–598.

Marsit S, Mena A, Bigey F, Sauvage F-X, Couloux A, Guy J, Legras J-L, Barrio E, Dequin S, Galeote V. 2015. Evolutionary advantage conferred by an eukaryote-to-eukaryote gene transfer event in wine yeasts. *Mol Biol Evol*. 32(7):1695–1707.

Mixão V, Gabaldón T. 2018. Hybridization and emergence of virulence in opportunistic human yeast pathogens. *Yeast* 35(1):5–20.

Monerawela C, Bond U. 2018. The hybrid genomes of *Saccharomyces pastorianus*: a current perspective. *Yeast* 35(1):39–50.

Nishant KT, Wei W, Mancera E, Argueso JL, Schlattl A, Delhomme N, Ma X, Bustamante CD, Korbel JO, Gu Z, et al. 2010. The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. *PLoS Genet*. 6(9):e1001109.

Okonechnikov K, Golosova O, Fursov M, UGENE Team. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28(8):1166–1167.

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*. 15(2):256–278.

Paten B, Novak AM, Eizenga JM, Garrison E. 2017. Genome graphs and the evolution of genome inference. *Genome Res*. 27(5):665–676.

Pennisi E. 2017. New technologies boost genome quality. *Science* 357(6346):10–11.

Peris D, Pérez-Torrado R, Hittinger CT, Barrio E, Querol A. 2018. On the origins and industrial applications of *Saccharomyces cerevisiae* x *Saccharomyces kudriavzevii* hybrids. *Yeast* 35(1):51–69.

Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A, et al. 2018. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* 556(7701):339–344.

Quinlan AR, Hall IM. 2010. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.

Rajaby R, Sung W-K. 2018. Transurveyor: an improved database-free algorithm for finding non-reference transpositions in high-throughput sequencing data. *Nucleic Acids Res*. 46(20):e122.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol*. 29(1):24–26.

Sadhu MJ, Bloom JS, Day L, Kruglyak L. 2016. Crispr-directed mitotic recombination enables genetic mapping without crosses. *Science* 352(6289):1113–1116.

Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, Dugas M. 2017. Evaluating variant calling tools for non-matched next-generation sequencing data. *Sci Rep*. 7(1):43169.

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*. 19(6):329–346.

Semeraro R, Orlandini V, Magi A. 2018. Xome-blender: a novel cancer genome simulator. *PLoS One* 13(4):e0194472.

Serero A, Jubin C, Loeillet S, Legoix-Né P, Nicolas AG. 2014. Mutational landscape of yeast mutator strains. *Proc Natl Acad Sci U S A*. 111(5):1897–1902.

Sharp NP, Agrawal AF. 2012. Evidence for elevated mutation rates in low-quality genotypes. *Proc Natl Acad Sci U S A*. 109(16):6142–6146.

Sharp NP, Sandell L, James CG, Otto SP. 2018. The genome-wide rate and spectrum of spontaneous mutations differ between haploid and diploid yeast. *Proc Natl Acad Sci U S A*. 115(22):E5046–E5055.

She R, Jarosz DF. 2018. Mapping causal variants with single-nucleotide resolution reveals biochemical drivers of phenotypic change. *Cell* 172(3):478–490.e15.

Smukowski Heil CS, DeSevo CG, Pai DA, Tucker CM, Hoang ML, Dunham MJ. 2017. Loss of heterozygosity drives adaptation in hybrid yeast. *Mol Biol Evol*. 34(7):1596–1612.

Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK. 2016. Simulating next-generation sequencing datasets from empirical mutation and sequencing models. *PLoS One* 11(11):e0167047.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big data: astronomical or genomical? *PLoS Biol*. 13(7):e1002195.

Tattini L, D'Aurizio R, Magi A. 2015. Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng Biotechnol*. 3:92.

Taylor SA, Larson EL. 2019. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nat Ecol Evol*. 3(2):170–177.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 13(1):36–46.

Vázquez-García I, Salinas F, Li J, Fischer A, Barré B, Hallin J, Bergström A, Alonso-Perez E, Warringer J, Mustonen V, et al. 2017. Clonal heterogeneity influences the fate of new adaptive mutations. *Cell Rep*. 21(3):732–744.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9(11):e112963.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res*. 27(5):757–767.

Wong KHY, Levy-Sakin M, Kwok P-Y. 2018. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat Commun*. 9(1):3040.

Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen J-Q, Hurst LD, Tian D. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* 523(7561):463–467.

Yu X, Sun S. 2013. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics* 14:274.

Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergström A, Coupland P, Warringer J, Lagomarsino MC, et al. 2017. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet*. 49(6):913–924.

Zhang F, Christiansen L, Thomas J, Pokholok D, Jackson R, Morrell N, Zhao Y, Wiley M, Welch E, Jaeger E, et al. 2017. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat Biotechnol*. 35(9):852–857.

Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci U S A*. 111(22):E2310–E2318.