Bulk Segregant Analysis Reveals the Genetic Basis of a Natural Trait Variation in Fission Yeast

Wen Hu[1,2,†], Fang Suo[2,†], and Li-Lin Du[2,*]

[1]PTN Graduate Program, School of Life Sciences, Tsinghua University, Beijing, China

[2]National Institute of Biological Sciences, Beijing, China

†These authors contributed equally to this work.

*Corresponding author: E-mail: dulilin@nibs.ac.cn.

Accepted: November 24, 2015

Abstract

Although the fission yeast *Schizosaccharomyces pombe* is a well-established model organism, studies of natural trait variations in this species remain limited. To assess the feasibility of segregant-pool-based mapping of phenotype-causing genes in natural strains of fission yeast, we investigated the cause of a maltose utilization defect (Mal$^-$) of the *S. pombe* strain CBS5557 (originally known as *Schizosaccharomyces malidevorans*). Analyzing the genome sequence of CBS5557 revealed 955 nonconservative missense substitutions, and 61 potential loss-of-function variants including 47 frameshift indels, 13 early stop codons, and 1 splice site mutation. As a side benefit, our analysis confirmed 146 sequence errors in the reference genome and improved annotations of 27 genes. We applied bulk segregant analysis to map the causal locus of the Mal$^-$ phenotype. Through sequencing the segregant pools derived from a cross between CBS5557 and the laboratory strain, we located the locus to within a 2.23-Mb chromosome I inversion found in most *S. pombe* isolates including CBS5557. To map genes within the inversion region that occupies 18% of the genome, we created a laboratory strain containing the same inversion. Analyzing segregants from a cross between CBS5557 and the inversion-containing laboratory strain narrowed down the locus to a 200-kb interval and led us to identify *agl1*, which suffers a 5-bp deletion in CBS5557, as the causal gene. Interestingly, loss of *agl1* through a 34-kb deletion underlies the Mal$^-$ phenotype of another *S. pombe* strain CGMCC2.1628. This work adapts and validates the bulk segregant analysis method for uncovering trait-gene relationship in natural fission yeast strains.

Key words: *Schizosaccharomyces pombe*, natural variation, maltose utilization.

Introduction

The fission yeast *Schizosaccharomyces pombe* is an important eukaryotic model organism that has provided key insights into fundamental cellular processes such as cell cycle control, DNA damage response, epigenetics, and cell morphogenesis (Forsburg 1999; Yanagida 2002; Goto and Nakayama 2012; Hachet et al. 2012; Rhind and Russell 2012). In recent years, studies on the population genetics and genomics of this species have demonstrated that *S. pombe* is also an excellent model for investigating natural variation and evolution (Brown et al. 2011; Rhind et al. 2011; Avelar et al. 2013; Fawcett et al. 2014; Farlow et al. 2015; Jeffares et al. 2015).

Virtually, all currently used laboratory strains of *S. pombe* derive from one natural strain, which was isolated in 1921 by A. Osterwalder from sulphited grape juice from Montpellier, France, and was deposited by him as CBS1042 at the Dutch culture collection Centraalbureau voor Schimmelcultures (CBS) under the species name *Schizosaccharomyces liquefaciens* (Osterwalder 1924; http://www-bcf.usc.edu/~forsburg/history/osterwalder.html, last accessed December 11, 2015; CBS online strain database). In 1947, Urs Leupold, the founder of fission yeast genetics, acquired this strain from CBS and chose it for his genetic studies (Leupold 1950, 1993; Munz et al. 1989; Sipiczki 1989; Barnett 2007). The Leupold strain 968 (h^{90} mating type) corresponds to CBS1042, whereas Leupold strains 972 (h^{-S} mating type) and 975 (h^{+N} mating type) are spontaneously occurring heterothallic derivatives of CBS1042 (Munz et al. 1989). The *S. pombe* reference genome is that of a Leupold 972 strain called "PN1" in the strain collection of Paul Nurse's lab (Wood et al. 2002; McDowall et al.

2015; Valerie Wood, personal communication). Thus, the reference genome of *S. pombe* is essentially that of a natural strain. This is different from the situation of the other model yeast species, *Saccharomyces cerevisiae*, whose reference strain S288C is created in the laboratory from multiple distinct progenitor strains (Engel et al. 2014).

Besides CBS1042, at least 56 other distinct natural strains of *S. pombe* have been collected from locations throughout the world (Jeffares et al. 2015). These strains exhibit variations in many different traits, including growth properties and stress resistance (Gomes et al. 2002; Brown et al. 2011; Jeffares et al. 2015), ability to invade solid substrates (Dodgson et al. 2010), cell morphology (Jeffares et al. 2015), intracellular amino acid concentrations (Jeffares et al. 2015), and transcriptome profiles (Clément-Ziza et al. 2014).

CBS5557 is a *S. pombe* strain isolated from Listan grapes grown near Jerez de la Frontera, Spain, and deposited at CBS under the species name *Schizosaccharomyces malidevorans* (Rankine and Fornachon 1964). This strain exhibits maltose utilization deficiency (Rankine and Fornachon 1964), which is a rare trait among natural strains of fission yeast (Brown et al. 2011). A previous analysis has suggested that this is a Mendelian trait and placed the causal locus on chromosome I because of linkage with the *leu2* gene (Sipiczki et al. 1982), but the causal gene remains unidentified.

Up to now, studies on the causes of natural trait variation in *S. pombe* have been mainly carried out through association mapping (Clément-Ziza et al. 2014; Jeffares et al. 2015). It would be desirable to harness additional tools and methodology for such investigations. Bulk segregant analysis (BSA) is a gene mapping method firstly developed and implemented in plants (Michelmore et al. 1991). The key design of this method is bulk genotyping of a pool of segregants sharing the same phenotype. DNA polymorphisms causing the phenotype or closely linked to the causal locus are enriched in the pool and thus stand out with a high proportion (theoretically reaching 100% for the causal recessive mutation). The advances in high-throughput genotyping technologies such as microarray and next-generation sequencing (NGS) have greatly enhanced the power of BSA (Brauer et al. 2006; Segrè et al. 2006; Schneeberger et al. 2009; Doitsidou et al. 2010). NGS-assisted BSA has been successfully applied in the budding yeast *Sa. cerevisiae* to uncover the genetic basis of Mendelian traits (Birkeland et al. 2010; Wenger et al. 2010) and multi-gene traits (Ehrenreich et al. 2010; Magwene et al. 2011; Parts et al. 2011; Swinnen et al. 2012). In this study, we applied this method to identify the causal gene of the Mal$^-$ phenotype of the fission yeast strain CBS5557. Our work demonstrates the feasibility and power of NGS-assisted BSA in studying natural variation of fission yeast and establishes tools and data sets useful for the implementation of this method.

Materials and Methods

Strains and Media

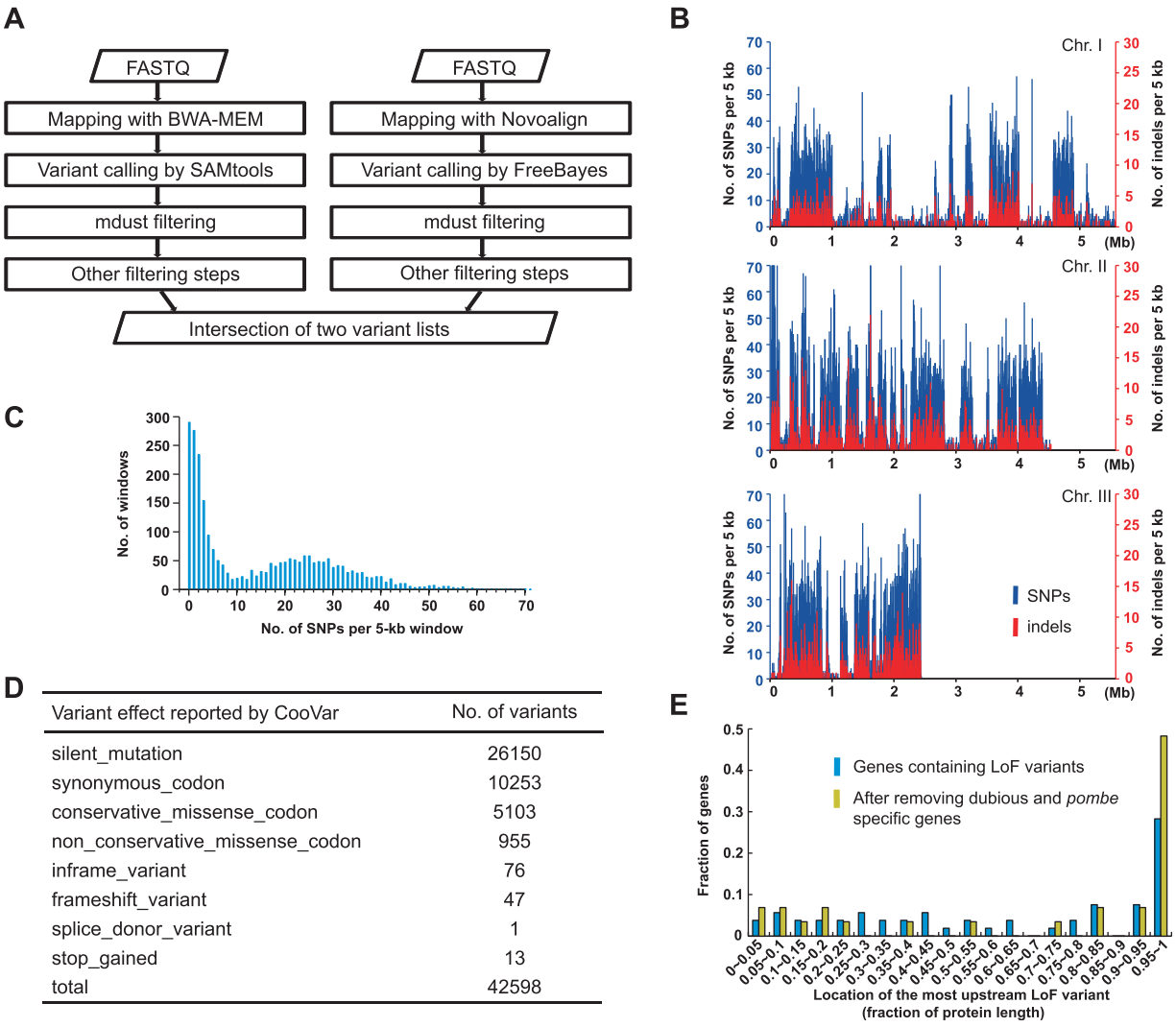
CBS5557 was obtained from the CBS Fungal Biodiversity Centre in Utrecht, the Netherlands (http://www.cbs.knaw.nl/, last accessed December 11, 2015). It is a homothallic strain. A spontaneous heterothallic derivative of CBS5557, DY5945, was obtained by selecting colonies that fail to be stained by iodine on a sporulation medium. It is of the h^+ mating type. The laboratory strain clone used as the mating partner of DY5945 in the first BSA analysis is LD775 (*h$^-$ leu1-32*). We created the artificial inversion in the laboratory strain clone LD1 (*h$^-$ ura4-D18 leu1-32*). The resulting strain was named DY8531 (*h$^-$ ura4-D18 leu1-32 inversion_junction_1::Padh1-kanMX inversion_junction_2::ura4$^+$*) and used in the second BSA analysis. The details of DY8531 construction are described below. Strains CGMCC2.1621 and CGMCC2.1628 were obtained from the China General Microbiological Culture Collection Center (CGMCC) in Beijing, China (http://www.cgmcc.net/, last accessed December 11, 2015). Standard culturing media were used (Forsburg and Rhind 2006). Glucose-containing rich medium used in this study is YES, and minimal medium used in this study is EMM2. Maltose-containing medium was made according to the recipe of YE medium, except substituting glucose with maltose.

Genome Sequencing and Sequence Variant Analysis

Genomic DNA was extracted from DY5945 and DY8531, respectively, using the MasterPure Yeast DNA Purification Kit (Epicentre). Illumina sequencing libraries were constructed using NEBNext DNA Library Prep Master Mix (NEB). Paired-end sequencing was performed using the Illumina HiSeq 2000 sequencer. Sequencing data were deposited at NCBI SRA under the accession numbers SRX1052152 (DY5945) and SRX1052153 (DY8531).

For sequencing data analysis, reference genome DNA sequence is based on the FASTA file Schizosaccharomyces_pombe.ASM294v1.18.dna.toplevel.fa.gz (last modified April 29, 2013) downloaded from ftp://ftp.ensemblgenomes.org/pub/fungi/release-18/fasta/schizosaccharomyces_pombe/dna/ (last accessed December 11, 2015). Reference genome annotations are based on the GTF file Schizosaccharomyces_pombe.ASM294v2.25.gtf.gz (last modified December 24, 2014) downloaded from ftp://ftp.ensemblgenomes.org/pub/fungi/release-25/gtf/schizosaccharomyces_pombe/ (last accessed December 11, 2015). Gene names and gene product descriptions are based on the sysID2product.tsv file (last modified January 26, 2015) downloaded from ftp://ftp.ebi.ac.uk/pub/databases/pombase/pombe/Mappings/ (last accessed December 11, 2015).

In the two variant-finding pipelines (fig. 1*A*), read mapping to the reference genome was carried out using BWA-MEM



Variant effect reported by CooVar	No. of variants
silent_mutation	26150
synonymous_codon	10253
conservative_missense_codon	5103
non_conservative_missense_codon	955
inframe_variant	76
frameshift_variant	47
splice_donor_variant	1
stop_gained	13
total	42598

Fig. 1.—Sequence variants in the CBS5557 genome. (*A*) The two data analysis pipelines used for identifying the sequence variants. (*B*) Distribution of SNPs and indels in the CBS5557 genome. The numbers of SNPs (blue, left *y* axis) and indels (red, right *y* axis) in 5-kb windows are depicted as vertical bars. (*C*) The numbers of CBS5557 SNPs in 5-kb windows show a bimodal distribution, with about half of the windows containing more than ten SNPs and the other half containing much fewer SNPs. (*D*) The functional impact of CBS5557 sequence variants assessed using CooVar. (*E*) The LoF variants are enriched in the last 5% of the CDS.

version 0.7.7 (Li 2013) and Novoalign version 3.02.04 (http://novocraft.com/, last accessed December 11, 2015), respectively. Duplicate reads were removed using SAMtools' rmdup command. Variant calling using SAMtools version 0.1.18 (Li et al. 2009) was conducted with mpileup options -B -q 10 -m 3 -F 0.2. We filtered the VCF file generated by SAMtools/BCFtools by requiring the variants to have read depth between 10 and 200 for DY5945 and between 10 and 300 for DY8531, to have a homozygous genotype, to have a quality score no smaller than 30, and to have a single allele. Variant calling using FreeBayes version 0.9.14 (Garrison and Marth 2012) was conducted on reads with mapping qualities no smaller than 10. We filtered the VCF file generated by FreeBayes by requiring the variants to have read depth between 10 and 200 for DY5945 and between 10 and 300 for DY8531, to have a homozygous genotype, to have a quality score no smaller than 300, and to have a single allele. Using mdust (downloaded from http://sourceforge.net/projects/gicl/files/other/, last modified date July 22, 2010, last accessed December 11, 2015), we identified 3,584 low-complexity regions (LCRs) in the reference nuclear genome (supplementary table S4, Supplementary Material online). Sequence variants falling within 3 bp of an LCR were filtered out (Li 2014). The excluded genomic sequences total 135,730 bp (about 1% of the reference genome). To compare and merge the variants found by the two pipelines, we used vcflib vcfallelicprimitives (https://github.com/ekg/vcflib, last accessed

December 11, 2015) to decompose complex variants into single-nucleotide polymorphisms (SNPs) and indels.

We assessed the functional impact of sequence variants using CooVar version 0.07 (Vergara et al. 2012). A single-nucleotide interval in the coding sequence (CDS) annotation of *SPBC3F6.03* caused a CooVar software error and we therefore removed the annotation of *SPBC3F6.03* from the input GTF file. Upon inspection of the CooVar output, we manually corrected the erroneous results of 12 "boundary indels" (indels locating at the start codons, stop codons, or splice sites), changing the "mutation effect" of these indels from frameshift_variant, inframe_variant, or splice_acceptor_variant, to silent_mutation (supplementary table S2, Supplementary Material online).

Bulk Segregant Analysis

CBS5557-derived h^+ strain DY5945 was crossed with an h^- laboratory strain clone (LD775 for the first BSA analysis and DY8531 for the second BSA analysis). Random spores were germinated on YES plates to form colonies. To score the maltose utilization phenotype, the colonies were inoculated into 96-well microplates containing either glucose-containing medium or maltose-containing medium. After incubating at 30 °C for 24 h, the extent of growth in each well was determined by visual inspection. The progenies that grew in both types of media were deemed Mal^+, whereas the progenies that only grew in glucose-containing medium were deemed Mal^-. The Mal^+ progenies were pooled in equal amounts to form the Mal^+ pool, and the Mal^- progenies were pooled in equal amounts to form the Mal^- pool. Only Leu^- progenies were used for constructing the pools derived from the DY5945 × LD775 cross. Genomic DNA extraction and sequencing library construction were conducted as described above. Single-read sequencing was performed using Illumina HiSeq 2000. Sequencing data were deposited at NCBI SRA under the accession numbers SRX1052154 (SRR2054742: DY5945 × LD775 Mal^+ pool; SRR2054743: DY5945 × LD775 Mal^- pool) and SRX1052164 (SRR2054745: DY5945 × DY8531 Mal^+ pool; SRR2054747: DY5945 × DY8531 Mal^- pool).

Read mapping to the reference genome was carried out using BWA-MEM version 0.7.7. After duplicate removal using SAMtools' rmdup command, we obtained the read counts of different types of bases at each SNP position using the software bam-readcount version 0.7.4 (https://github.com/genome/bam-readcount/, last accessed December 11, 2015) with the option -q 10. Reference allele frequencies were calculated using the bam-readcount output. To better visualize the trend of allele frequency differences in the scatter plots, we drew LOESS regression lines in the plots (Minevich et al. 2012) by using Matlab's smooth function with the span parameter set at values corresponding to 60 kb (0.0108 for chromosome I, 0.0132 for chromosome II, and 0.0245 for chromosome III).

Construction of the Laboratory Strain Clone Containing the Artificial Inversion

A schematic of the procedure is depicted in figure 3. Using the plasmid pBS-AS-ura4PB-kanloxNco (a gift from Kojiro Ishii) (Ishii et al. 2008) as template, we amplified fragment I shown in figure 3 through two rounds of polymerase chain reaction (PCR) with each round adding 40 bp of sequence to the homologous arms, transformed it into a laboratory strain clone LD1 (*h⁻ ura4-D18 leu1-32*), and selected for Ura^+ transformants. Through homologous recombination, a *Padh1* promoter and a *ura4*$^+$ marker flanked by two *loxP* sites were inserted at the left junction of the inversion. We verified the insertion by PCR using primers JI_5 and JI_6 and selected a correct transformant (DY8186) for the next step. DY8186 was transformed with a Cre-expressing episomal plasmid pAW8 (Watson et al. 2008) (http://www.euroscarf.de/plasmid_details.php?accno=P30545, last accessed December 11, 2015). Leu^+ transformants were selected on -Leu plates. Cre recombinase was induced by growing cells in thiamine-free liquid medium. Cells that had lost the *ura4*$^+$ marker were selected on yeast extract-based medium containing 5-fluoroorotic acid (5-FOA). The colonies were replica plated onto -Leu plates. Ura^- Leu^- colonies that had also lost the plasmid pAW8 were picked. We verified the excision of the *ura4*$^+$ marker by PCR using primers JI_2 and JI_5 and comparing the size of the PCR product to that of the amplicon derived from the parental strain DY8186. A correct clone (DY8359) was selected for the next step. Using the plasmid pBS-AS-ura4PB-kanloxNco as template, we amplified fragment II shown in figure 3 through two rounds of PCR, transformed it into DY8359, and selected for Ura^+ transformants. Through homologous recombination, a *ura4*$^+$ marker and a promoter-less *kanMX* ORF, together with a *loxP* site in between, were inserted at the right junction of the inversion. We verified the insertion by PCR using the primer pair JII_5 and JII_6 and the primer pair JII_7 and JII_8, and selected a correct transformant (DY8430) for the next step. DY8430 was transformed with pAW8 and Leu^+ transformants were selected. Cre recombinase was induced by growing cells in thiamine-free liquid medium. G418-resistant colonies resulted from the joining of the promoter-less *kanMX* ORF with the *Padh1* promoter were selected by plating cells on G418-containing rich medium plates. Colonies were replica plated onto -Leu plates and Leu^- colonies that had lost the pAW8 plasmid were picked. One G418-resistant and Leu^- clone (DY8531) was used for all further experiments. Sequences of the primers used in this procedure are listed in supplementary table S5, Supplementary Material online.

Complementation of the Mal^- Phenotype Using an Agl1-Expressing Plasmid

The *agl1* gene was amplified from laboratory strain genomic DNA and inserted into the integrating vector pHIS3K-Ptub1 (http://dna.brc.riken.jp/en/yoshidayeast_310en.html, last

accessed December 11, 2015) (Matsuyama et al. 2008). Sequences of the PCR primers used for amplifying *agl1* are listed in supplementary table S5, Supplementary Material online. The resulting plasmid, pDB1730, was cut with NotI and introduced into the Mal$^-$ fission yeast strains by transformation. Transformants containing the correctly integrated plasmid were selected by the ability to grow on G418-containing YES plates and the inability to grow on -His plates.

Results

Surveying Sequence Variants in the CBS5557 Genome

To generate a catalog of DNA polymorphisms in CBS5557, we sequenced the genome of a heterothallic derivative of CBS5557 using paired-end Illumina sequencing. As a control, we also sequenced the genome of a laboratory strain of *S. pombe*. For clarity, we use the term "laboratory strain clone" to refer to any derivative of CBS1042 first adopted by Urs Leupold for genetic analysis in 1947 and later becoming virtually the only *S. pombe* strain used for modern fission yeast research. After duplicate removal, we obtained about 70× coverage for the CBS5557 nuclear genome and about 100× coverage for the nuclear genome of the laboratory strain clone.

We used the Illumina sequencing data to identify sequence variants relative to the reference genome. To avoid software-specific errors and artifacts, we employed two variant-finding pipelines consisting of different read mapping softwares (BWA-MEM and Novoalign, respectively) and different variant calling softwares (SAMtools and FreeBayes, respectively) (fig. 1*A*). Both pipelines include a filtering step (mdust filtering) that removes variants falling into LCRs, as recommended by a recent study on the sources of variant calling errors (Li 2014). For CBS5557, the two pipelines agree with each other on > 98% of the SNPs and >88% of the indels. We conservatively selected the variants supported by both pipelines as our final sets of variants that distinguish the genomes sequenced here and the reference genome, which include 42,763 variants in CBS5557 and 202 variants in the laboratory strain clone. Intriguingly, 165 variants are shared between CBS5557 and the laboratory strain clone, representing more than 80% of the variants found in the laboratory strain clone (supplementary table S1, Supplementary Material online).

To verify the sequence variant calls using an independent approach, we performed PCR and Sanger sequencing analysis on 70 of the 165 shared variants using the genomic DNA from our laboratory strain clone as PCR template. We obtained PCR products and high-quality Sanger sequencing data for 61 variants and the results were unanimously consistent with the Illumina sequencing-based variant calls (supplementary table S1, Supplementary Material online).

Reference Sequence Errors and Gene Annotation Revision

It is unlikely that a clonal variation arising after Urs Leupold selected CBS1042 as the laboratory strain in 1947 happens to be identical to an inter-strain difference between CBS5557 and CBS1042. Thus, most of the 165 variants shared by CBS5557 and our laboratory strain clone should represent the pre-1947 state of CBS1042, whereas the reference alleles at these 165 positions probably stem from either reference sequence errors or post-1947 clone-specific mutations in the laboratory strain clone from which the reference genome sequence was derived. Nearly 50% (79/165) of these shared variants are indels, a percentage significantly higher than the level of indels among spontaneous mutations (Farlow et al. 2015), suggesting that these variants are more likely to be due to reference errors.

A great majority (146/165, 88.5%) of the variants shared between CBS5557 and our laboratory strain clone are among the 190 "sequence discrepancies" uncovered at the Broad Institute through Illumina sequencing of the genome of a laboratory strain clone yFS101 (supplementary table S1, Supplementary Material online) (Chad Nusbaum, personal communication; http://www.pombase.org/status/sequence-updates-pending, last accessed December 11, 2015). yFS101 is from the strain collection of Nicholas Rhind's lab and is a direct passage from Paul Russell lab's strain PR37, which is a "972 strain" that came from Paul Nurse's lab (Nicholas Rhind, personal communication). Thus, yFS101 is essentially identical to the clone from which the reference genome sequence was derived. Our results confirmed that these 146 "sequence discrepancies," including 75 indels and 71 SNPs, are indeed reference genome sequence errors; namely, the variant alleles at these 146 positions represent the true sequence of the genome of the laboratory strain (CBS1042), whereas the reference alleles are erroneous. Among the other 19 shared variants found by us and the remaining 44 "sequence discrepancies" found at the Broad Institute, there may be additional reference errors that are called as variants by only one study due to different variant calling criteria.

Among the 75 confirmed indel-type reference errors, 36 are located within CDSs (supplementary table S1, Supplementary Material online). We manually inspected these 36 intragenic indel errors and found that 31 of them affect the gene structure annotation of 27 genes (table 1). Using protein sequence alignment of orthologs among fission yeast species as guide, we revised the gene structure annotation of these 27 genes (table 1 and supplementary file S1, Supplementary Material online). Thirteen of these 27 genes have been previously reported to contain indel-type sequence errors (table 1) (Hayashi et al. 2006; Matsuyama et al. 2006; Yokoyama et al. 2008). Even though the reference genome sequence has not yet been updated accordingly, for 11 of these 13 genes, the current PomBase annotation apparently

Table 1

The 27 Genes Whose Gene Structure Annotations Are Revised as a Result of Reference Sequence Indel Error Correction (see supplementary file S1, Supplementary Material online, for Further Details)

Systematic ID	Gene Name	Chr.	Indel Position	Ref. Seq.	True Seq.	Annotation Change
SPAC1F8.07c[a]		I	101871	A	AG	A 2-bp intron becomes part of an exon
SPAC22F3.11c[a]	snu23	I	682993	TC	T	An intron becomes part of an exon
SPAC3A12.04c	rpp1	I	1424708	CA	C	CDS is extended at 3′-end
SPAP27G11.10c	nup184	I	1625092	T	TC	CDS is extended at 3′-end
SPAC17G8.01c	trl1	I	2343703	G	GA	An intron becomes part of an exon
SPAC823.04	rrp36	I	2588021	C	CA	A 2-bp intron becomes part of an exon
			2588066	C	CA	An intron becomes part of an exon
SPAC688.08	srb8	I	3125118	A	AT	An intron becomes part of an exon
SPAC1486.05	nup189	I	3197528	A	AG	An intron becomes part of an exon
SPAC3A11.09	sod22	I	3450130	GT	G	CDS is extended at 3′-end
SPAC3A11.06	mvp1	I	3460318	T	TC	One boundary of an intron is moved
SPAC1071.01c	pta1	I	3855790	GT	G	CDS is extended at 3′-end
SPAC29E6.03c	uso1	I	4407494	T	TG	An intron becomes part of an exon
SPAC29E6.04[b,c]	nnf1	I	4410191	CG	C	A 1-bp intron no longer exists
SPAC29A4.03c		I	5142627	A	AG	A 2-bp intron becomes part of an exon
SPAC4D7.09	tif223	I	5368262	TC	T	Three amino acids are altered
			5368273	G	GT	
SPBC16E9.16c[d]	lsd90	II	1948953	GA	G	A 1-bp intron no longer exists
			1950050	A	AG	A 2-bp intron becomes part of an exon
SPBC1E8.03c		II	1960392	A	AG	CDS is extended at 3′-end
SPBC1A4.06c[a]	tam41	II	1987101	CG	C	A 2-bp intron becomes part of an exon
			1987117	TG	T	
SPBC29A3.06[a,b]	utp18	II	2049891	AT	A	A 1-bp intron no longer exists
SPBC29A3.08[a]	pof4	II	2053516	G	GC	A 1-bp intron becomes part of an exon
SPBC23G7.06c[a]		II	2108180	T	TA	A 2-bp intron becomes part of an exon
SPBC14C8.09c[a]	dbl3	II	2219928	A	AT	A 2-bp intron becomes part of an exon
SPBC4F6.10[a]	vps901	II	2709414	G	GC	A 2-bp intron becomes part of an exon
SPBC32F12.08c[a]	duo1	II	2798040	CT	C	CDS is extended at 3′-end
SPBC13E7.01[a]	cwf22	II	3040332	C	CG	A 2-bp intron becomes part of an exon
SPBC16D10.10	tad2	II	3619003	A	AG	Both boundaries of an intron are moved
SPCC1442.04c[a]		III	1774235	T	TGATC	A 2-bp intron becomes part of an exon

[a]Frameshifts have been reported by Matsuyama et al. (2006).
[b]Amino acid sequence of the gene product is unchanged by the proposed gene structure revision.
[c]Frameshift has been reported by Hayashi et al. (2006).
[d]Frameshifts have been reported by Yokoyama et al. (2008).

seeks to minimize the impact of indel errors by the use of "frameshift introns," which are 1- or 2-bp-long artificial introns employed by the genome annotators to rectify frameshifts caused by sequence errors (table 1) (Hubbard et al. 2007). The side-by-side depiction of the current PomBase annotation and our revised annotation shows that these "frameshift introns" have by and large fulfilled their promise, with the 1-bp introns effectively offsetting 1-bp insertion errors and the 2-bp introns resulting in the loss of one codon in the cases of 1-bp deletion errors (supplementary file S1, Supplementary Material online). One exception is *SPBC29A3.08* (*pof4*), where a 1-bp deletion error should have been remedied by a 2-bp intron but is instead tackled with a 1-bp intron. Thus, even for previously documented indel errors, our analysis has revealed uncorrected and miscorrected annotations. Supplementary file S1, Supplementary Material online, provides a thorough documentation detailing the effect of these 31 indel errors on gene structure annotations. We have contacted the PomBase curators, who will incorporate the findings reported here into future releases of the reference genome (Valerie Wood, personal communication).

Distribution Pattern and Functional Impact Prediction of Sequence Variants in CBS5557

Excluding the 165 shared variants, we uncovered in the CBS5557 nuclear genome a total of 42,598 sequence variants, among which 38,783 are SNPs (3.1 SNPs/kb) and 3,815 are indels (0.3 indel/kb) (supplementary table S2, Supplementary Material online). This level of overall diversity between CBS5557 and CBS1042 is similar to the average pairwise diversity (3.0 SNPs/kb) among the 57 natural strains recently analyzed (Jeffares et al. 2015). The distribution of the variants along the chromosomes shows an uneven pattern

(fig. 1*B*). The variant-rich regions occupy about half of the nuclear genome (50.3% of the 5-kb windows have more than ten SNPs) (fig. 1*C*). Such a mosaic pattern has probably resulted from infrequent outcrossing between strains of distinct lineages, with the regions exhibiting low diversity between CBS5557 and CBS1042 coming from the same lineage. Similar mosaic patterns of variant distribution have been observed when the genomes of other *S. pombe* strains, including SPK1820 (also called YFS276 or *S. pombe var. kambucha*) and CBS2777, were compared to the reference genome (Rhind et al. 2011; Brown et al. 2014), and is a common phenomenon among *Sa. cerevisiae* strains (Liti et al. 2009). It has been postulated that the mosaic *Sa. cerevisiae* strains have arisen due to strain migration and mixing brought about by human activities (Liti et al. 2009; Bergström et al. 2014). The same argument can be applied to *S. pombe* because currently available *S. pombe* isolates seem to have all been associated with human activities (Jeffares et al. 2015). Pure lineages of truly wild *S. pombe* strains may still wait to be uncovered by future field collection efforts.

We used the software CooVar to evaluate the functional impact of sequence variants on protein-coding genes (supplementary table S2, Supplementary Material online) (Vergara et al. 2012). About 85.5% (36,403/42,598) of the variants are silent or synonymous changes, 14.2% (6,058/42,598) are missense SNPs, 0.18% (76/42,598) are inframe indels, and 0.14% (61/42,598) are potential loss-of-function (LoF) variants including frameshift indels, nonsense SNPs, and splice site mutations (fig. 1*D*). Based on the Grantham score (Grantham 1974), CooVar classifies missense SNPs into either conservative missense or nonconservative missense (Vergara et al. 2012). The former far outnumbers the latter, with a ratio of about 5:1 (5,103 vs. 955).

We paid special attention to the 61 variants that can potentially cause LoF effect (hereafter referred to as LoF variants), as these variants are more likely to have a phenotypic consequence than other variants. A total of 53 genes are inflicted by LoF variants in CBS5557 (supplementary table S3, Supplementary Material online). Among them, genes whose PomBase gene product descriptions are "dubious" (12 out of 53 genes) or "*Schizosaccharomyces pombe* specific protein" (12 out of 53 genes) are significantly enriched, with *P* values of 7.29e-12 and 1.11e-10, respectively (one-tailed Fisher's exact test). Among the 12 dubious genes, the expression status of 11 genes was recently analyzed using RNA-seq and ribosome profiling data (Duncan and Mata 2014). Four of the 11 were deemed by that study as "not expressed" (no detectable mRNA) and the other 7 were classified as "not translated" (supplementary table S3, Supplementary Material online). Thus, most if not all of the 12 dubious genes harboring LoF variants in CBS5557, and perhaps some of the 12 LoF-containing pombe-specific genes as well, are not protein-coding genes but rather nongenic sequences or noncoding genes. We analyzed the locations of the LoF variants relative to the lengths of the protein products (fig. 1*E*). For genes containing multiple LoF variants, we selected the most upstream LoF variant for this analysis. Interestingly, LoF variants locating at the extreme 3′-region (last 5%) of the CDSs are strongly enriched, especially so after removing the "dubious" and "*Schizosaccharomyces pombe* specific protein" genes (*P* values of 3.10e-8 and 2.31e-11, respectively, exact binomial test). Thus, the protein products of many genes in this class only suffer a small C-terminal alteration, which may not disrupt protein functions. Similar 3′-end enrichments of LoF variants have been observed in *Sa. cerevisiae* and humans (Liti et al. 2009; MacArthur et al. 2012; Bergström et al. 2014). If we remove "dubious" and "*Schizosaccharomyces pombe* specific protein" genes and also remove genes that retain more than 95% of their codons, only 15 of the 53 genes are left (*agl1*, *clu1*, *hri2*, *hsp3104*, *lsc1*, *mfs1*, *pet801*, *rex2*, *SPAC11D3.11c*, *SPAC2C4.08*, *SPAC57A7.13*, *SPBC1348.12*, *SPBC25H2.10c*, *SPBC337.02c*, and *SPBC460.04c*). Manual inspection of these 15 genes suggested that the gene structures of two genes, *lsc1* and *pet801*, may be misannotated, so that the indels in these two genes actually fall into an intron and the 5′-noncoding region, respectively, and are thus silent mutations instead of frameshift mutations (supplementary file S2, Supplementary Material online). The gene product of another gene, *SPAC11D3.11c*, is annotated by PomBase as "zn(2)-C6 fungal-type DNA-binding transcription factor, truncated," and the indel variant in CBS5557 renders the protein product much longer, reaching a length similar to those of its homologs in other *Schizosaccharomyces* species (supplementary file S2, Supplementary Material online). Therefore, among the 42,598 variants that we uncovered in the CBS5557 genome, truly LoF variants may affect as few as 12 functional genes. We hypothesize that a similar number of functional genes may suffer LoF variants in other fission yeast strains, including the laboratory strain (CBS1042). The indel in *SPAC11D3.11c* is an example of laboratory-strain-specific LoF variants. Another example is the DNA repair gene *apn1*, which is disrupted by a nonsense mutation in the laboratory strain (Laerdahl et al. 2011). Our manual inspection found that *apn1* is intact in CBS5557. In this case, CooVar reported the effect of the SNP variant in CBS5557 as "silent_mutation" because *apn1* is annotated as a pseudogene. Further analysis will be needed to comprehensively detect such "reverting-back-to-normal" variants and reveal the full extent of strain-specific pseudogenes in the reference genome.

Locating the Mal$^-$ Trait Locus to within the 2.23-Mb Inversion on Chromosome I

To perform the BSA analysis, we crossed a laboratory strain clone LD775, which is of the h^- mating type, to DY5945, an h^+ derivative of CBS5557, and carried out random spore analysis. Consistent with the proposition that the Mal$^-$ phenotype is a single-gene controlled trait (Sipiczki et al. 1982), the ratio of

Mal$^+$ versus Mal$^-$ progenies was approximately 1:1 (data not shown). We combined 42 Mal$^+$ progenies to form a Mal$^+$ pool and combined 42 Mal$^-$ progenies to form a Mal$^-$ pool. Genomic DNA was extracted from the two pools separately, and single-read Illumina sequencing was performed to obtain about 13× coverage for the Mal$^+$ pool and about 12× coverage for the Mal$^-$ pool after duplicate removal. Among the 38,783 CBS5557 SNPs identified by the analysis described above, we selected the ones covered by at least five confidently mapped reads in both pools. A total of 32,890 SNPs on chromosome I, II, and III met this criteria. For each of these SNPs, we calculated an allele frequency difference between the pools by subtracting the reference allele frequency of the Mal$^-$ pool from that of the Mal$^+$ pool. SNPs tightly linked to the Mal$^-$ trait locus should have allele frequency differences approaching 1, whereas SNPs not linked to the locus should have allele frequency differences around 0. We plotted the allele frequency differences in scatter plots and drew a local regression line in each plot to better visualize the trend (fig. 2*A*). Consistent with the prediction that the causal gene of the Mal$^-$ phenotype is located on chromosome I (Sipiczki et al. 1982), the only chromosome where allele frequency differences strongly deviate from 0 is chromosome I. Interestingly, we did not see a peak-like linkage pattern. Instead, all SNPs within a region more than 2 Mb long on chromosome I exhibited allele frequency differences close to 1. The boundaries of this region coincide with the breakpoints of a 2.23-Mb inversion found in most natural strains including CBS5557 (Brown et al. 2011 and our unpublished observations), indicating that the inversion region behaves like a single locus in this cross. This is consistent with the observation that meiotic recombination was strongly reduced within this inversion region when a laboratory strain clone was crossed to another natural strain harboring this inversion (Clément-Ziza et al. 2014). We concluded that the causal gene of the Mal$^-$ phenotype is situated within the inversion region. This conclusion agrees with the previously observed linkage between the Mal$^-$ trait locus and the *leu2* gene (Sipiczki et al. 1982), because *leu2* (CDS coordinates 4440096–4442372 of chromosome I) is also situated inside of the inversion region (coordinates 2683632–4911514 of chromosome I) (Brown et al. 2011).

Constructing a Laboratory Strain Clone Containing an Artificial 2.23-Mb Inversion

To determine where in the inversion region the Mal$^-$ trait locus is located, we need to remove the impediment to meiotic recombination. We hypothesized that engineering an artificial inversion in one of the two parental strains may restore normal recombination. To achieve this goal, we chose to modify the genome of the laboratory strain. This choice is based on technical convenience (availability of complete genome sequence and auxotrophic mutants for the laboratory strain), as well as the wider use of the resulting strain, which will have a chromosome I collinear with that in most of the natural strains. We employed a Cre-*loxP*-based strategy previously used for deleting a centromere in fission yeast (Ishii et al. 2008) (fig. 3). Two *loxP* sites locating at the two inversion junctions, respectively, were introduced through two rounds of PCR-based gene targeting. Cre-mediated recombination was carried out twice, first to remove a *ura4*$^+$ marker and a second time to create the inversion. Correct inversion resulted in the joining of a *Padh1* promoter with a promoter-less *kanMX* ORF. Clones containing the 2.23-Mb inversion were selected based on the *kanMX*-conferred antibiotic resistance.

The Artificial Inversion Improves the BSA Mapping of the Mal$^-$ Trait Locus

We crossed an inversion-harboring laboratory strain clone DY8531, which is of the *h*$^-$ mating type, to DY5945, an *h*$^+$ derivative of CBS5557. We combined 35 Mal$^+$ progenies to form a Mal$^+$ pool and combined 29 Mal$^-$ progenies to form a

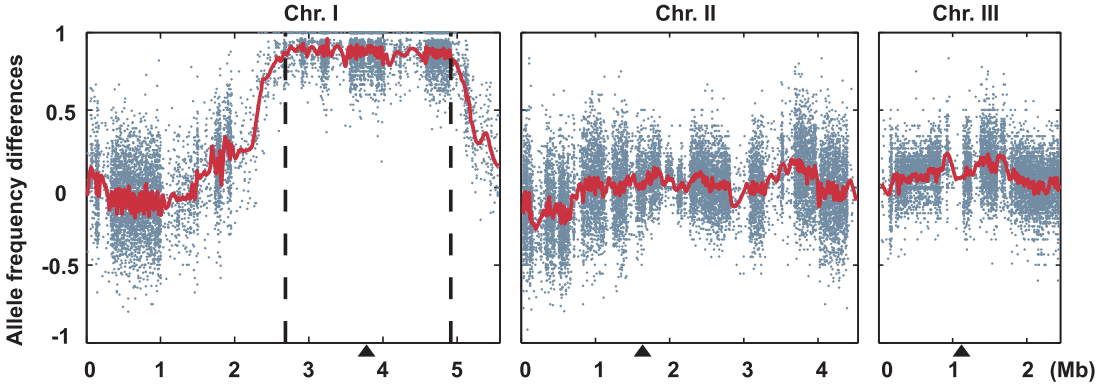


Fig. 2.—BSA on the cross between DY5945 and LD775. Scatter plots depict the differences of reference allele frequencies between the Mal$^+$ pool and the Mal$^-$ pool at SNP positions. The allele frequency differences are expected to be around 0 in most regions of the genome and reach 1 at the Mal$^-$ trait locus. Local regression lines are displayed to better visualize the trend. Two dashed vertical lines mark the boundaries of the 2.23-Mb inversion. Black triangles mark the positions of centromeres.

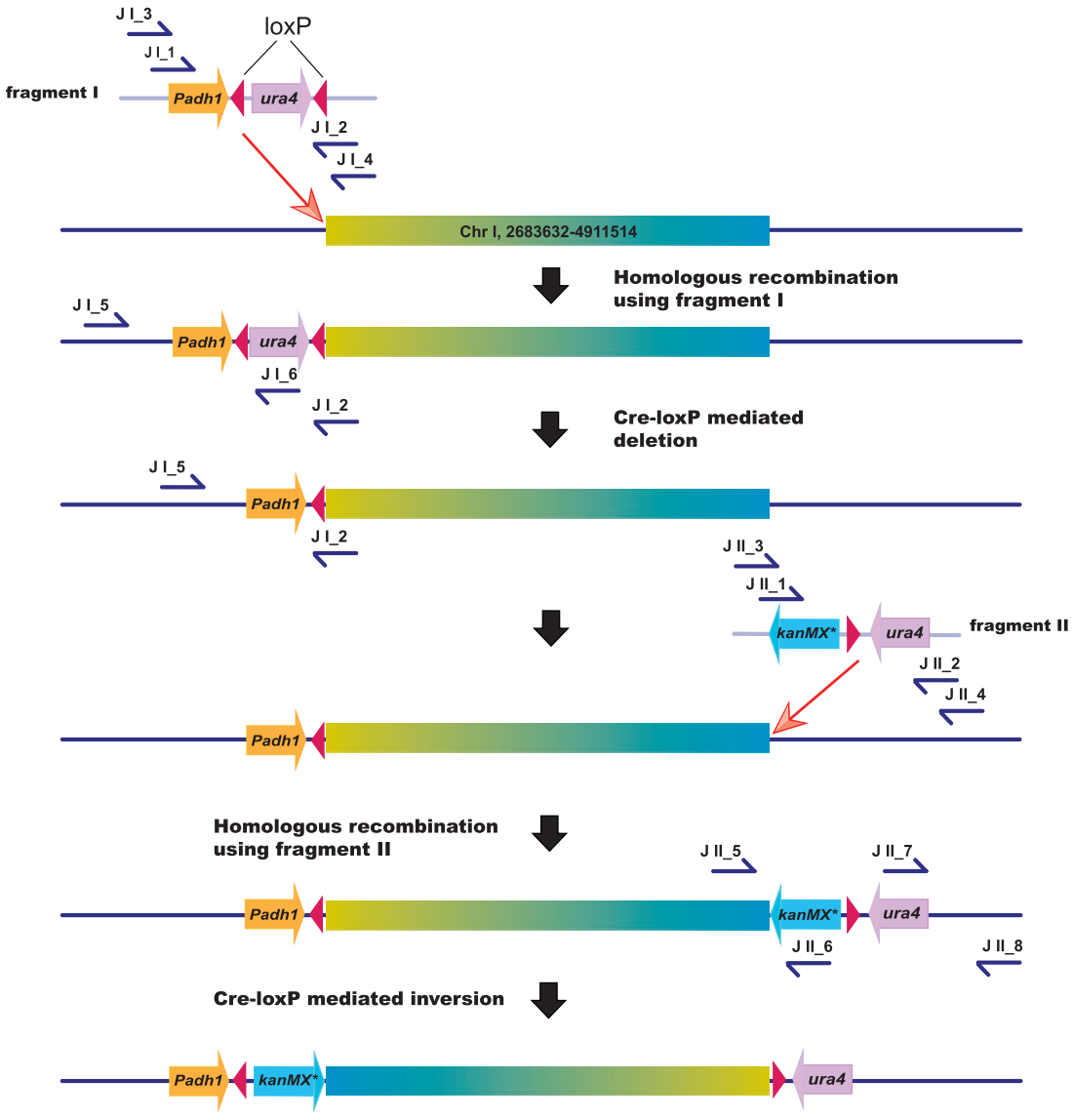


Fig. 3.—Constructing an artificial inversion on chromosome I of the laboratory strain. A schematic depicting the procedure of strain construction. See the Materials and Methods section for details. *kanMX** denotes the promoter-less *kanMX* ORF that only became expressed after being placed immediately downstream of the *Padh1* promoter by the inversion.

Mal$^-$ pool. Genomic DNA was extracted from the two pools separately and single-read Illumina sequencing was performed to obtain about 16× coverage for the Mal$^+$ pool and about 17× coverage for the Mal$^-$ pool after duplicate removal. A total of 36,041 SNPs on chromosome I, II, and III were selected for allele frequency difference calculation using the criteria described above for the first BSA cross. We plotted allele frequency differences the same way as before. Unlike the inversion-region-wide linkage pattern of the first BSA cross, we observed a peak-like pattern within the inversion region (fig. 4*A*). Thus, the inversion-harboring laboratory strain clone allowed linkage mapping within the inversion region, presumably by lifting the impediment to meiotic recombination within this region.

In this cross, all progenies in the BSA pools should harbor the inversion. Therefore, plotting using the reference genome coordinates of the SNPs, as shown in figure 4*A*, distorted the spatial relationship between SNPs situated on opposite sides of an inversion breakpoint. To more truthfully represent chromosome I of these progenies, we redrew the plot of chromosome I using adjusted coordinates that match the inversion (Fig. 4*B*). This new plot eliminated artifacts of plotting using the reference genome coordinates and clearly showed that there is a single linkage peak on chromosome I.

To define the location of the Mal$^-$ trait locus using the improved BSA mapping result, we inspected the linkage peak summit region (coordinates 2684001–3164000 of

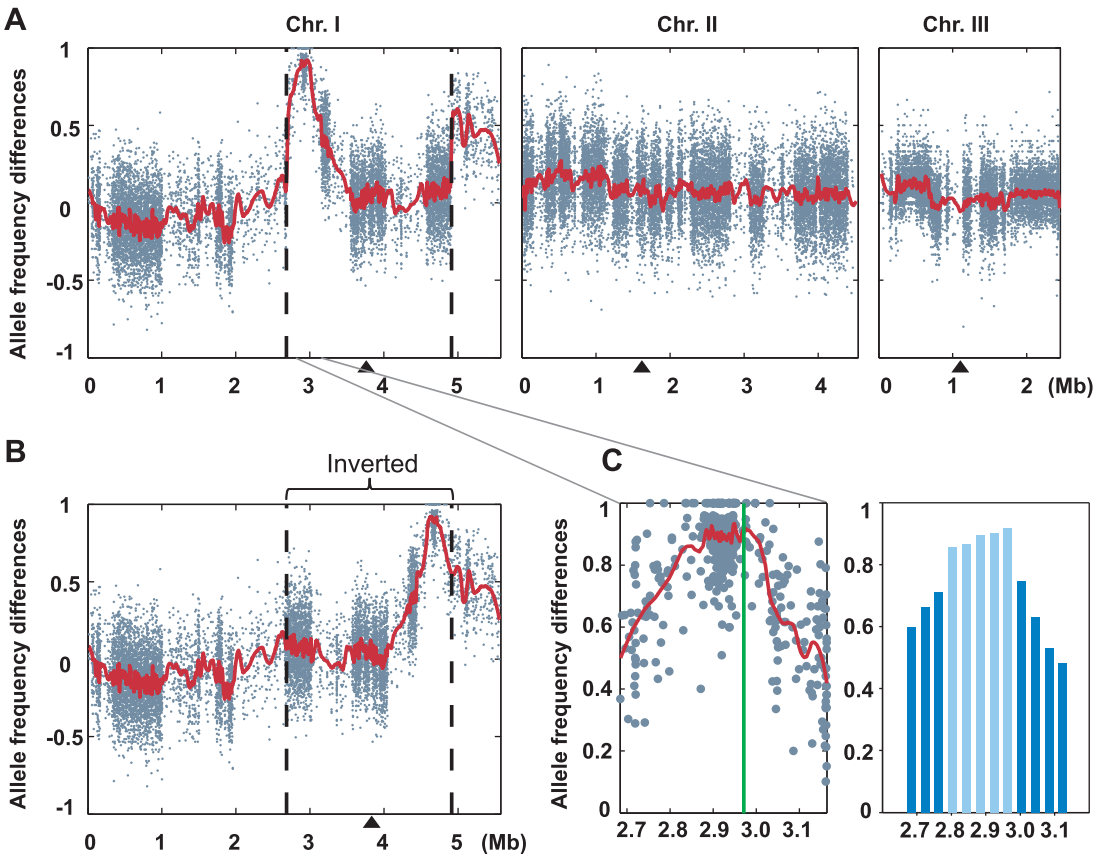


Fig. 4.—BSA on the cross between DY5945 and DY8531. (*A*) Scatter plots depicting the allele frequency differences are drawn as in figure 2. (*B*) Scatter plot of chromosome I is redrawn, so that the genomic coordinates of data points within the inversion region are adjusted to match the inversion. (*C*) Close-up views of the linkage peak summit region (coordinates 2684001–3164000 of chromosome I). The scatter plot was drawn as in (*A*). The green line marks the position of the *agl1* gene. The histogram depicts average allele frequency differences in 40-kb bins. The five bins with average allele frequency differences >0.8 are highlighted as light blue bars.

chromosome I) using both a close-up view of the scatter plot and a histogram depicting average allele frequency differences of 40-kb bins (Fig. 4*C*). Using a cutoff of 0.8 for the average allele frequency differences, we narrowed down the location of the Mal$^-$ trait locus to within a 200-kb region (coordinates 2804001–3004000 of chromosome I).

Identification of *agl1* as the Causal Gene of the Mal$^-$ Trait of CBS5557

Within the 200-kb region where the Mal$^-$ trait locus may reside, our analysis of sequence variants in the CBS5557 genome showed that only five genes are affected by mutations that may severely impact functions (supplementary table S2, Supplementary Material online). Four of them, *mdb1*, *cta4*, *ams1*, and *SPAPB2C8.01*, contain nonconservative missense mutations and one gene, *agl1*, suffers a frameshift mutation. These are the candidate genes that may underlie the Mal$^-$ trait.

We paid special attention to the gene *agl1*, not only because it is the only gene affected by a LoF mutation within the 200-kb region but also because it encodes an extracellular alpha-glucosidase involved in maltose utilization (Kato et al. 2013). In the reference genome, the protein product of this gene is 969-amino acid long. Illumina sequencing of CBS5557 genome showed that a deletion of five nucleotides, GTTAA (coordinates 2970788–2970792 of chromosome I), occurs in this gene (fig. 5*A*). We confirmed this deletion by PCR and Sanger sequencing (fig. 5*B*). This 5-bp deletion results in a frameshift after the first 65 codons. Such an early frameshift is very likely to abolish the functions of *agl1*.

To determine whether the LoF mutation in *agl1* causes the Mal$^-$ phenotype of CBS5557, we introduced into CBS5557 an integrating plasmid expressing the laboratory strain version of *agl1*. The plasmid rescued the Mal$^-$ phenotype (fig. 5*C*). Thus, we conclude that *agl1* is the causal gene of the Mal$^-$ trait in CBS5557.

Loss of *agl1* through a 34-kb Deletion Causes Mal$^-$ Phenotype of Another Natural *S. pombe* Strain

In a survey of the growth-related phenotypes of about 80 natural *S. pombe* isolates, two isolates of European origins,

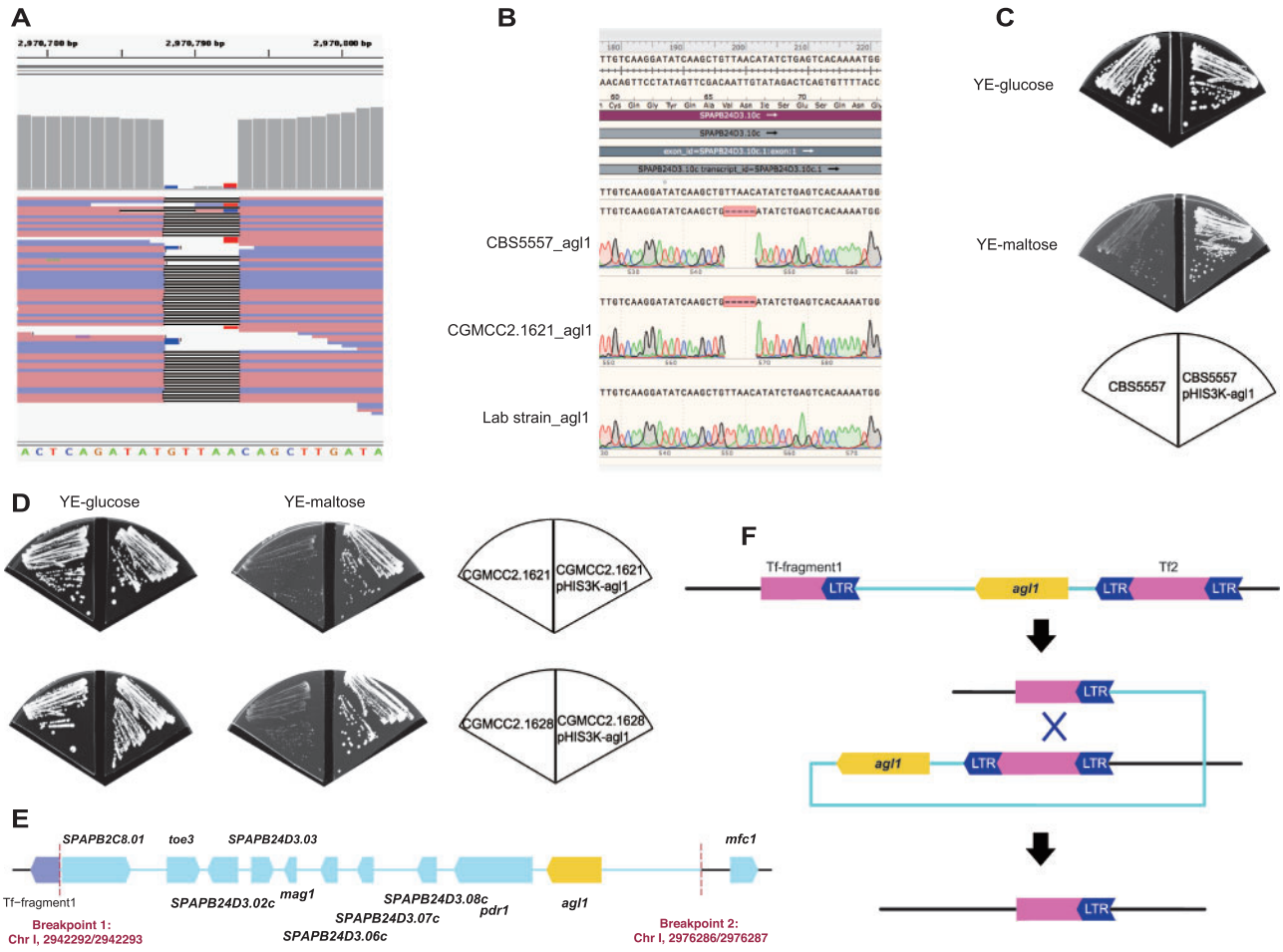


Fig. 5.—*agl1* is the causal gene of the Mal$^-$ phenotype of CBS5557 and CGMCC2.1628. (*A*) Illumina sequencing reveals a 5-bp deletion in the CDS of *agl1* in the CBS5557 genome. (*B*) PCR and Sanger sequencing confirm the 5-bp deletion in the CDS of *agl1* in the CBS5557 genome. The sequences are in a reverse complement orientation relative to that of the reference genome. (*C*) The Mal$^-$ phenotype of CBS5557 can be rescued by a plasmid expressing the laboratory strain version of *agl1*. The parental strain and transformant were streaked on agar plates containing glucose or maltose as the carbon source. To avoid the rescue of Mal$^-$ phenotype by the diffusion of extracellular glucose generated by nearby Mal$^+$ colonies, gaps were created between sectors streaked with different strains by cutting out agar slices. (*D*) Two strains deposited in the culture collection under the name *Schizosaccharomyces malidevorans*, CGMCC2.1621 and CGMCC2.1628, are Mal$^-$ *Schizosaccharomyces pombe* strains, and their Mal$^-$ phenotype can be rescued by a plasmid expressing the laboratory strain version of *agl1*. The phenotype analysis was performed as in (*C*). (*E*) A 34-kb chromosome I region containing *agl1* is deleted in CGMCC2.1628. The deletion breakpoints are denoted by vertical dashed lines. (*F*) A schematic depicting a possible scenario of how the 34-kb deletion in CGMCC2.1628 may have formed.

NCYC2387 and DBVPG4435, were found to have maltose utilization defect (Brown et al. 2011). Recent genomic analysis indicated that these two isolates are essentially identical to CBS5557 (Jeffares et al. 2015). Thus, CBS5557 stands out as the only known natural strain that has the Mal$^-$ phenotype.

CBS5557 was originally classified as *S. malidevorans* (Rankine and Fornachon 1964). Before its merging under the species *S. pombe* (Vaughan Martini 1991; Vaughan-Martini and Martini 1998), the defining taxonomic feature of *S. malidevorans* is its inability to utilize maltose (Yarrow 1984; Kocková-Kratochvílová et al. 1985; Yamada et al. 1987). Thus, we reasoned that if there are Mal$^-$ *S. pombe* strains different from CBS5557, they might have also been deposited in microbial strain collections under the species name *S. malidevorans*. We acquired from China General Microbiological Culture Collection Center (CGMCC) two strains with the species name *S. malidevorans*, CGMCC2.1621 and CGMCC2.1628. These two strains were deposited at CGMCC on May 6, 1990, but their origins are uncertain (Feng-Yan Bai, personal communication). As expected, both of these strains are defective in maltose utilization (fig. 5*D*). Introducing the plasmid expressing the laboratory version of *agl1* into these strains rescued their growth defect on maltose plates (fig. 5*D*), suggesting that like CBS5557, the Mal$^-$ phenotype of these two strains is also caused by a lack of functional *agl1*.

Our preliminary genomic analysis showed that CGMCC2.1621 is essentially identical to CBS5557, whereas CGMCC2.1628 is distinct from CBS5557 (Fang Suo, unpublished observations). PCR-sequencing analysis showed that CGMCC2.1621 has the same 5-bp *agl1* deletion as that of CBS5557 (fig. 5*B*). In contrast, in CGMCC2.1628, a 34-kb chromosome I region surrounding *agl1* (coordinates 2942293–2976286 of chromosome I) is deleted (fig. 5*E*). The left deletion breakpoint is immediately adjacent to the Tf-fragment1 element (coordinates 2939711–2942292 of chromosome I), which is a truncated version of the Tf2 retrotransposon (Bowen et al. 2003) (fig. 5*E*). Because the intact upstream LTR of Tf-fragment1 is situated immediately next to the breakpoint, a possible scenario that may have given rise to the deletion is a Tf2 insertion between coordinates 2976286 and 2976287 of chromosome I in the ancestor of CGMCC2.1628, followed by recombination between this Tf2 and Tf-fragment1 (fig. 5*F*). Besides *agl1*, nine other protein-coding genes (*SPAPB2C8.01*, *toe3*, *SPAPB24D3.02c*, *SPAPB24D3.03*, *mag1*, *SPAPB24D3.06c*, *SPAPB24D3.07c*, *SPAPB24D3.08c*, *pdr1*) are lost due to this 34-kb deletion. The loss of some of these genes may have also resulted in permanent physiological changes like that caused by the loss of *agl1*.

Discussion

In this study, we performed detailed analysis of the sequence variants in CBS5557 genome and demonstrated that NGS-assisted BSA is a powerful tool for elucidating the genetic basis of natural trait variations in fission yeast by mapping and identifying the gene underlying the Mal$^-$ phenotype of CBS5557.

Our sequence variant analysis tallied and categorized DNA polymorphisms of the nonreference strain CBS5557 and paved the way for connecting genotype and phenotype in this strain. The in-depth inspection of LoF variants in CBS5557 provided the first glimpse of the extent of strain-specific pseudogenes in fission yeast. The data sets generated here will be a useful resource for further investigation of intraspecific sequence variations in fission yeast. In addition to new knowledge gained on a nonreference strain, our investigation also led to the identification and correction of sequence and annotation errors in the reference genome. The improvement on reference genome annotation will benefit future research using the laboratory strain.

For a strong growth-related phenotype, the easiest way to perform the BSA analysis is to pool all segregants together and then select for those that can grow under a particular condition, as has been done previously with the budding yeast *Sa. cerevisiae* (Segrè et al. 2006; Ehrenreich et al. 2010; Parts et al. 2011). Unfortunately, in our study, we could not use this approach because the Mal$^-$ phenotype can be rescued by the presence of Mal$^+$ cells in the same culture (Kato et al. 2013). However, as has been shown before (Birkeland et al. 2010), and also demonstrated here, it takes only a few dozen individually phenotyped segregants to perform the BSA analysis and therefore, compared to the bulk selection method, only a small amount of extra work is required.

In our BSA analysis, we chose to cross CBS5557 to a laboratory strain clone. The selection of the laboratory strain as the mating partner allows us to take advantage of the extensive knowledge on the laboratory strain and the genetic resources available for it. However, the laboratory strain is different from most of the other natural strains due to the presence of the 2.23-Mb inversion in chromosome I (Brown et al. 2011). We found that this inversion impedes mapping of genes within this region, which contains 18.1% (931/5,132) of the protein-coding genes in the nuclear genome. We constructed a laboratory strain clone containing the 2.23-Mb inversion and showed that using this strain clone as the mating partner allows mapping within the inversion region. This strain clone (DY8531) will be useful for future application of BSA in mapping causal genes of natural traits.

We found that *agl1* is the causal gene of the Mal$^-$ phenotype of two distinct natural strains, CBS5557 and CGMCC2.1628. This is consistent with the recent report that *agl1* is required for maltose utilization by the laboratory strain (Kato et al. 2013). Agl1 protein belongs to the GH31 family of glycoside hydrolase and can hydrolyze maltose, isomaltose, and soluble starch (Okuyama et al. 2001). It is the only abundant protein in the supernatant of glucose-limited fission yeast cultures (Jansen et al. 2006) and can be expressed to g/L level, corresponding to $>$ 1% of the total cellular protein (Jansen et al. 2006; Klein et al. 2014). The transcription of *agl1* is strongly induced ($>$ 100 folds) by glucose depletion (Rhind et al. 2011; Kato et al. 2013). This induction depends on two transcription factors Atf1 and Pcr1, which are also essential for maltose utilization by the laboratory strain (Kato et al. 2013).

The requirement of an extracellular enzyme for maltose utilization by *S. pombe* cells stands in sharp contrast to the situation in the budding yeast *Sa. cerevisiae*, where maltose is first imported into the cell by a permease and then hydrolyzed by an intracellular maltase belonging to the GH13 family of glycoside hydrolase (Charron et al. 1989; Naumov et al. 1994; Gabriško 2013). Interestingly, a homolog of budding yeast maltase exists in *S. pombe* (Chi et al. 2008; Brown et al. 2010). However, this protein, Mal1, does not play an obvious role in maltose utilization (Kato et al. 2013), perhaps due to the absence of a fission yeast homolog of the budding yeast maltose permease (Brown et al. 2010).

Fission yeast Agl1 is closely related to the extracellular alpha-glucosidase AgdA of the fungal species belonging to the *Aspergillus* genus (Kimura et al. 1992; Minetoki et al. 1995; Nakamura et al. 1997; Okuyama et al. 2001; Kato et al. 2002; Yuan et al. 2008; Vongsangnak et al. 2009). Some of the *Aspergillus* species, such as *Aspergillus oryzae*, possess a set of maltose utilization genes similar to those of *Sa. cerevisiae*

and consume maltose intracellularly like *Sa. cerevisiae*, whereas other *Aspergillus* species, such as *Aspergillus niger*, lack a full set of budding yeast homologs and hydrolyze maltose extracellularly like *S. pombe* (Vongsangnak et al. 2009; Hasegawa et al. 2010; vanKuyk et al. 2012).

The expression and secretion of an extracellular digestive enzyme by a microbial cell is considered a social trait, because the freely diffusible digestion products can benefit neighboring cells, including those that produce the same enzyme ("cooperators"), as well as those that do not produce the enzyme ("cheaters"). A classic example is the secretion of invertase by *Sa. cerevisiae* (Greig and Travisano 2004). Invertase, encoded by the *SUC* genes in *Sa. cerevisiae*, hydrolyzes extracellular sucrose into glucose and fructose, which are more preferable sugars than sucrose. When analyzed separately, Suc^+ cells that produce invertase grow better on sucrose medium than Suc^- cells that do not produce invertase. However, in a mixed culture of Suc^+ and Suc^- cells, the Suc^- cells can gain a growth advantage because they do not have to shoulder the cost of invertase production but still reap the benefits of invertase digestion. The sharing of invertase has become a useful model for studying the evolution of cooperation (Greig and Travisano 2004; Craig Maclean and Brandon 2008; Gore et al. 2009; MaClean et al. 2010; Datta et al. 2013; Van Dyken et al. 2013) and the evolution of multicellularity (Koschwanez et al. 2011, 2013). Natural variation of the invertase production trait exists in *Sa. cerevisiae*. In a survey of 91 *Sa. cerevisiae* natural isolates, it was found that 80 strains are Suc^+ and 11 strains are Suc^- (Naumov et al. 1996). Such a pattern of variation has been used as evidence for the co-existence of "cooperators" and "cheaters" in natural environments (Greig and Travisano 2004; Craig Maclean and Brandon 2008; Gore et al. 2009). A more recent survey of 80 *Saccharomyces paradoxus* strains and 30 *Sa. cerevisiae* strains isolated from nonhuman-related environments failed to find any Suc^- strains, thus leading to the suggestions that "cheaters" may not be as prevalent as previously thought and Suc^- strains may result from adaptation to low-sucrose environment rather than social conflict (Bozdag and Greig 2014). Either of these two evolutionary scenarios may apply to the natural variation of the *agl1* gene in *S. pombe*.

Supplementary Material

Supplementary tables S1–S5 and files S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Hua-Lin Li for contributing to the verification of sequence variants; Kojiro Ishii for providing the plasmid pBS-AS-ura4PB-kanloxNco; Feng-Yan Bai for providing information on strains from China General Microbiological Culture Collection Center (CGMCC). We are grateful to Nicholas Rhind and Valerie Wood for critically reading the manuscript and providing strain information. This work was supported by a grant from the National Basic Research Program of China (973 Program, 2014CB849901) to L.-L.D.

Literature Cited

Avelar AT, Perfeito L, Gordo I, Ferreira MG. 2013. Genome architecture is a selectable trait that can be maintained by antagonistic pleiotropy. Nat Commun. 4:2235.

Barnett JA. 2007. A history of research on yeasts 10: foundations of yeast genetics. Yeast 24:799–845.

Bergström A, et al. 2014. A high-definition view of functional genetic variation from natural yeast genomes. Mol Biol Evol. 31:872–888.

Birkeland SR, et al. 2010. Discovery of mutations in *Saccharomyces cerevisiae* by pooled linkage analysis and whole-genome sequencing. Genetics 186:1127–1137.

Bowen NJ, Jordan IK, Epstein JA, Wood V, Levin HL. 2003. Retrotransposons and their recognition of pol II promoters: a comprehensive survey of the transposable elements from the complete genome sequence of *Schizosaccharomyces pombe*. Genome Res. 13:1984–1997.

Bozdag GO, Greig D. 2014. The genetics of a putative social trait in natural populations of yeast. Mol Ecol 23:5061–5071.

Brauer MJ, Christianson CM, Pai DA, Dunham MJ. 2006. Mapping novel traits by array-assisted bulk segregant analysis in *Saccharomyces cerevisiae*. Genetics 173:1813–1816.

Brown CA, Murray AW, Verstrepen KJ. 2010. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. Curr Biol. 20:895–903.

Brown WRA, et al. 2011. A geographically diverse collection of *Schizosaccharomyces pombe* isolates shows limited phenotypic variation but extensive karyotypic diversity. G3 (Bethesda) 1:615–626.

Brown WRA, et al. 2014. Kinetochore assembly and heterochromatin formation occur autonomously in *Schizosaccharomyces pombe*. Proc Natl Acad Sci U S A. 111:1903–1908.

Charron MJ, Read E, Haut SR, Michels CA. 1989. Molecular evolution of the telomere-associated MAL loci of *Saccharomyces*. Genetics 122:307–316.

Chi Z, Ni X, Yao S. 2008. Cloning and overexpression of a maltase gene from *Schizosaccharomyces pombe* in *Escherichia coli* and characterization of the recombinant maltase. Mycol Res. 112:983–989.

Clément-Ziza M, et al. 2014. Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast. Mol Syst Biol. 10:764.

Craig Maclean R, Brandon C. 2008. Stable public goods cooperation and dynamic social interactions in yeast. J Evol Biol. 21:1836–1843.

Datta MS, Korolev KS, Cvijovic I, Dudley C, Gore J. 2013. Range expansion promotes cooperation in an experimental microbial metapopulation. Proc Natl Acad Sci U S A. 110:7354–7359.

Dodgson J, Brown W, Rosa CA, Armstrong J. 2010. Reorganization of the growth pattern of *Schizosaccharomyces pombe* in invasive filament formation. Eukaryot Cell 9:1788–1797.

Doitsidou M, Poole RJ, Sarin S, Bigelow H, Hobert O. 2010. *C. elegans* mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. PLoS One 5:e15435.

Duncan CD, Mata J. 2014. The translational landscape of fission-yeast meiosis and sporulation. Nat Struct Mol Biol. 21:641–647.

Ehrenreich IM, et al. 2010. Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature 464:1039–1042.

Engel SR, et al. 2014. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. G3 4:389–398.

Farlow A, et al. 2015. The spontaneous mutation rate in the fission yeast *Schizosaccharomyces pombe*. Genetics 201:737–744.
Fawcett JA, et al. 2014. Population genomics of the fission yeast *Schizosaccharomyces pombe*. PLoS One 9:e104241.
Forsburg SL. 1999. The best yeast? Trends Genet 15:340–344.
Forsburg SL, Rhind N. 2006. Basic methods for fission yeast. Yeast 23: 173–183.
Gabriško M. 2013. Evolutionary history of eukaryotic α-glucosidases from the α-amylase family. J Mol Evol. 76:129–145.
Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. ArXiv e-Prints 1207:3907.
Gomes FCO, et al. 2002. Physiological diversity and trehalose accumulation in *Schizosaccharomyces pombe* strains isolated from spontaneous fermentations during the production of the artisanal Brazilian cachaça. Can J Microbiol. 48:399–406.
Gore J, Youk H, van Oudenaarden A. 2009. Snowdrift game dynamics and facultative cheating in yeast. Nature 459:253–256.
Goto DB, Nakayama J. 2012. RNA and epigenetic silencing: insight from fission yeast. Dev Growth Differ 54:129–141.
Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864.
Greig D, Travisano M. 2004. The prisoner's dilemma and polymorphism in yeast SUC genes. Proc Biol Sci. 271(Suppl 3):S25–S26.
Hachet O, Bendezú FO, Martin SG. 2012. Fission yeast: in shape to divide. Curr Opin Cell Biol. 24:858–864.
Hasegawa S, Takizawa M, Suyama H, Shintani T, Gomi K. 2010. Characterization and expression analysis of a maltose-utilizing (MAL) cluster in *Aspergillus oryzae*. Fungal Genet Biol. 47:1–9.
Hayashi A, Asakawa H, Haraguchi T, Hiraoka Y. 2006. Reconstruction of the kinetochore during meiosis in fission yeast *Schizosaccharomyces pombe*. Mol Biol Cell 17:5173–5184.
Hubbard TJP, et al. 2007. Ensembl 2007. Nucleic Acids Res. 35: D610–D617.
Ishii K, et al. 2008. Heterochromatin integrity affects chromosome reorganization after centromere dysfunction. Science 321:1088–1091.
Jansen MLA, et al. 2006. Physiological characterization and fed-batch production of an extracellular maltase of *Schizosaccharomyces pombe* CBS 356. FEMS Yeast Res. 6:888–901.
Jeffares DC, et al. 2015. The genomic and phenotypic diversity of *Schizosaccharomyces pombe*. Nat Genet. 47:235–241.
Kato H, Kira S, Kawamukai M. 2013. The transcription factors Atf1 and Pcr1 are essential for transcriptional induction of the extracellular maltase Agl1 in fission yeast. PLoS One 8:e80572.
Kato N, et al. 2002. Novel alpha-glucosidase from *Aspergillus nidulans* with strong transglycosylation activity. Appl Environ Microbiol. 68:1250–1256.
Kimura A, et al. 1992. Complete amino acid sequence of crystalline alpha-glucosidase from *Aspergillus niger*. Biosci Biotechnol Biochem. 56:1368–1370.
Klein T, et al. 2014. Overcoming the metabolic burden of protein secretion in *Schizosaccharomyces pombe*—a quantitative approach using 13C-based metabolic flux analysis. Metab Eng 21:34–45.
Kocková-Kratochvílová A, Sláviková E, Zemek J, Kadlčcíkova B, Kuniak L. 1985. Hydrolytic activity in the genus *Schizosaccharomyces* Lindner. Folia Microbiol. 30:443–451.
Koschwanez JH, Foster KR, Murray AW. 2011. Sucrose utilization in budding yeast as a model for the origin of undifferentiated multicellularity. PLoS Biol. 9:e1001122.
Koschwanez JH, Foster KR, Murray AW. 2013. Improved use of a public good selects for the evolution of undifferentiated multicellularity. Elife 2:e00367.
Laerdahl JK, et al. 2011. *Schizosaccharomyces pombe* encodes a mutated AP endonuclease 1. DNA Repair 10:296–305.
Leupold U. 1950. Die Vererbung von Homothallie und Heterothallie bei *Schizosaccharomyces pombe*. Compt Rend Lab Carlsberg 24: 381–480.
Leupold U. 1993. The origin of *Schizosaccharomyces pombe* genetics. In: Hall MN, Linder P, editors. The early days of yeast genetics. Cold Spring Harbor Laboratory Press. p. 125–128.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv e-Prints 1303:3997.
Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30:2843–2851.
Li H, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.
Liti G, et al. 2009. Population genomics of domestic and wild yeasts. Nature 458:337–341.
MacArthur DG, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828.
MaClean RC, Fuentes-Hernandez A, Greig D, Hurst LD, Gudelj I. 2010. A mixture of 'cheats' and 'co-operators' can enable maximal group benefit. PLoS Biol. 8:e1000486.
Magwene PM, Willis JH, Kelly JK. 2011. The statistics of bulk segregant analysis using next generation sequencing. PLoS Comput Biol. 7:e1002255
Matsuyama A, Shirai A, Yoshida M. 2008. A novel series of vectors for chromosomal integration in fission yeast. Biochem Biophys Res Commun. 374:315–319.
Matsuyama A, et al. 2006. ORFeome cloning and global analysis of protein localization in the fission yeast *Schizosaccharomyces pombe*. Nat Biotechnol 24:841–847.
McDowall MD, et al. 2015. PomBase 2015: updates to the fission yeast database. Nucleic Acids Res. 43:D656–D661.
Michelmore RW, Paran I, Kesseli RV. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A. 88:9828–9832.
Minetoki T, Gomi K, Kitamoto K, Kumagai C, Tamura G. 1995. Nucleotide sequence and expression of alpha-glucosidase-encoding gene (agdA) from *Aspergillus oryzae*. Biosci Biotechnol Biochem. 59:1516–1521.
Minevich G, Park DS, Blankenberg D, Poole RJ, Hobert O. 2012. CloudMap: a cloud-based pipeline for analysis of mutant genome sequences. Genetics 192:1249–1269.
Munz P, Wolf K, Kohli J, Leupold U. 1989. Genetics overview. In: Nasim A, Young PG, Johnson BF, editors. Molecular biology of the fission yeast. Academic Press, Inc. p. 1–30.
Nakamura A, et al. 1997. Cloning and sequencing of an alpha-glucosidase gene from *Aspergillus niger* and its expression in *A. nidulans*. J Biotechnol 53:75–84.
Naumov GI, Naumova ES, Michels CA. 1994. Genetic variation of the repeated MAL loci in natural populations of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*. Genetics 136:803–812.
Naumov GI, Naumova ES, Sancho ED, Korhola MP. 1996. Polymeric SUC genes in natural populations of *Saccharomyces cerevisiae*. FEMS Microbiol Lett. 135:31–35.
Okuyama M, et al. 2001. Carboxyl group of residue Asp647 as possible proton donor in catalytic reaction of alpha-glucosidase from *Schizosaccharomyces pombe*. Eur J Biochem. 268:2270–2280.
Osterwalder A. 1924. *Schizosaccharomyces liquefaciens* n. sp., eine gegen freie schweflige Säure widerstandsfähige Gärhefe. Mitt Gebiete Lebensmittelunters Hyg 15:5–28.
Parts L, et al. 2011. Revealing the genetic structure of a trait by sequencing a population under selection. Genome Res. 21:1131–1138.
Rankine BC, Fornachon JC. 1964. *Schizosaccharomyces malidevorans* sp.n., a yeast decomposing L-malic acid. Antonie Van Leeuwenhoek 30:73–75.

Rhind N, Russell P. 2012. Signaling pathways that regulate cell division. Cold Spring Harb Perspect Biol. 4:a005942.
Rhind N, et al. 2011. Comparative functional genomics of the fission yeasts. Science 332:930–936.
Schneeberger K, et al. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods 6:550–551.
Segrè AV, Murray AW, Leu J-Y. 2006. High-resolution mutation mapping reveals parallel experimental evolution in yeast. PLoS Biol. 4:e256.
Sipiczki M. 1989. Taxonomy and phylogenesis. In: Nasim A, Young PG, Johnson BF, editors. Molecular biology of the fission yeast. Academic Press, Inc. p. 431–452.
Sipiczki M, Kucsera J, Ulaszewski S, Zsolt J. 1982. Hybridization studies by crossing and protoplast fusion within the genus *Schizosaccharomyces* Lindner. J Gen Microbiol. 128:1989–2000.
Swinnen S, et al. 2012. Identification of novel causative genes determining the complex trait of high ethanol tolerance in yeast using pooled-segregant whole-genome sequence analysis. Genome Res. 22:975–984.
Van Dyken JD, Müller MJI, Mack KML, Desai MM. 2013. Spatial population expansion promotes the evolution of cooperation in an experimental Prisoner's Dilemma. Curr Biol. 23:919–923.
vanKuyk PA, Benen JAE, Wösten HAB, Visser J, de Vries RP. 2012. A broader role for AmyR in *Aspergillus niger*: regulation of the utilisation of D-glucose or D-galactose containing oligo- and polysaccharides. Appl Microbiol Biotechnol 93:285–293.
Vaughan Martini A. 1991. Evaluation of phylogenetic relationships among fission yeast by nDNA/nDNA reassociation and conventional taxonomic criteria. Yeast 7:73–78.
Vaughan-Martini A, Martini A. 1998. *Schizosaccharomyces* Lindner. In: Kurtzman CP, Fell JW, editors. The yeasts—a taxonomic study. Amsterdam: Elsevier Science Publishers. p. 391–394.
Vergara IA, Frech C, Chen N. 2012. CooVar: co-occurring variant analyzer. BMC Res Notes 5:615.
Vongsangnak W, Salazar M, Hansen K, Nielsen J. 2009. Genome-wide analysis of maltose utilization and regulation in aspergilli. Microbiology 155:3893–3902.
Watson AT, Garcia V, Bone N, Carr AM, Armstrong J. 2008. Gene tagging and gene replacement using recombinase-mediated cassette exchange in *Schizosaccharomyces pombe*. Gene 407:63–74.
Wenger JW, Schwartz K, Sherlock G. 2010. Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from *Saccharomyces cerevisiae*. PLoS Genet. 6:e1000942.
Wood V, et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. Nature 415:871–880.
Yamada Y, Aizawa K, Matsumoto A, Nakagawa Y, Banno I. 1987. An electrophoretic comparison of enzymes in strains of species in the fission yeast genera *Schizosaccharomyces*, *Octosporomyces*, and *Hasegawaea*. J Gen Appl Microbiol. 33:363–369.
Yanagida M. 2002. The model unicellular eukaryote, *Schizosaccharomyces pombe*. Genome Biol. 3:COMMENT2003.
Yarrow D. 1984. *Schizosaccharomyces* Lindner. In: Kreger-van Rij NJW, editor. The yeasts—a Taxonomic Study. Amsterdam: Elsevier Science Publishers. p. 414–422.
Yokoyama K, et al. 2008. Expression of a novel 90-kDa protein, Lsd90, involved in the metabolism of very long-chain fatty acid-containing phospholipids in a mitosis-defective fission yeast mutant. J Biochem. 143:369–375.
Yuan X-L, et al. 2008. *Aspergillus niger* genome-wide analysis reveals a large number of novel alpha-glucan acting enzymes with unexpected expression profiles. Mol Genet Genomics 279:545–561.

Associate editor: Bill Martin