Published in final edited form as:
Nat Genet. 2015 March ; 47(3): 235–241. doi:10.1038/ng.3215.

The Genomic and Phenotypic Diversity of *Schizosaccharomyces pombe*

Daniel C. Jeffares[1,*], **Charalampos Rallis**[1], **Adrien Rieux**[1,2], **Doug Speed**[1,2], **Martin Převorovský**[3], **Tobias Mourier**[4], **Francesc X. Marsellach**[1], **Zamin Iqbal**[5], **Winston Lau**[1], **Tammy M.K. Cheng**[6], **Rodrigo Pracana**[1], **Michael Mülleder**[7], **Jonathan L.D. Lawson**[8,9], **Anatole Chessel**[7], **Sendu Bala**[10], **Garrett Hellenthal**[1,2], **Brendan O'Fallon**[11], **Thomas Keane**[10], **Jared T. Simpson**[10,†], **Leanne Bischof**[12], **Bartlomiej Tomiczek**[1], **Danny A. Bitton**[1], **Theodora Sideri**[1], **Sandra Codlin**[1], **Josephine E.E.U. Hellberg**[1], **Laurent van Trigt**[1], **Linda Jeffery**[6], **Juan-Juan Li**[6], **Sophie Atkinson**[1], **Malte Thodberg**[4], **Melanie Febrer**[12], **Kirsten McLay**[12], **Nizar Drou**[12], **William Brown**[13], **Jacqueline Hayles**[6], **Rafael E. Carazo Salas**[8,9], **Markus Ralser**[7,14,15], **Nikolas Maniatis**[1], **David J. Balding**[1,2], **Francois Balloux**[1,2], **Richard Durbin**[10], and **Jürg Bähler**[1,2,*]

[1]Department of Genetics, Evolution & Environment, University College London, London, UK

[2]UCL Genetics Institute, University College London, London, UK

[3]Department of Cell Biology, Charles University in Prague, Prague, Czech Republic

[4]Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

[5]Wellcome Trust Centre for Human Genetics, Oxford, UK

[6]Cell Cycle Laboratory, Cancer Research UK London Research Institute, London, UK

[7]Department of Biochemistry, University of Cambridge, Cambridge, UK

[8]Department of Genetics, University of Cambridge, Cambridge, UK

[*]Correspondence to: d.jeffares@ucl.ac.uk and j.bahler@ucl.ac.uk.
[†]Current Address: Ontario Institute for Cancer Research, Toronto, Canada.

Author contributions

DCJ coordinated all analyses, isolated DNA for sequencing, analysed and filtered SNP calls, conducted diversity analysis and GWAS and drafted the manuscript. CR produced phenotype data for growth on various solid media and growth rates in liquid media. AR conducted analysis of dating using mitochondrial data. DS conducted GWAS. MP analysed all phenotype data. TM identified LTR transposon insertions and analysed transposon insertion data. FXM conducted crosses for analysis of spore viability ZI produced indel calls with Cortex. WL conducted analysis of recombination rate, linkage disequilibrium decay and PCA for distance between strains. TMKC assisted with phenotype and population analysis. RP analysed Cortex and GATK indel calls. MM conducted amino acid profiling. JLDL and AC produced automated measures of cell morphology. SB aligned reads and produced GATK SNP calls. GH analysed population structure using *fineSTRUCTURE*. BO'F estimated the TMRCA from the nuclear genome using ACG. TK identified LTR transposon insertions JTS produced *de novo* assemblies. LB developed the custom Workspace workflow *Spotsizer*. BT assisted with sequence analysis. DAB assisted with analysis of novel genes. TS assisted with strain verification. SC produced images of wild strains and assisted with strain verification. JEEUH assisted with SNP validation. LvT and MT assisted with LTR validation. LJ and JL assisted with manual measures of cell morphology and FACS. SA produced gene expression data. MF, KM and ND assisted with sequencing. WB initiated and assisted with strain collection. JH coordinated manual measures of cell morphology and FACS. RECS coordinated automated measures of cell morphology. MR coordinated amino acid profiling. NM conducted analysis of recombination, linkage disequilibrium and advised on aspects of diversity and GWAS. DJB advised on GWAS. RD facilitated sequencing. JB contributed to the initiation and development of the project and financed the JB laboratory.

Accessions

Sequence data are archived in the European Nucleotide Archive (www.ebi.ac.uk/ena/), Study Accessions PRJEB2733 and PRJEB6284 (Supplementary Table 7). All SNPs and indels were submitted to NCBI dbSNP (www.ncbi.nlm.nih.gov/SNP/). Accessions are 974514578-974688138 (SNPs) and 974702618-974688139 (indels).

Europe PMC Funders Author Manuscripts

[9]The Gurdon Institute, University of Cambridge, Cambridge, UK

[10]Wellcome Trust Sanger Institute, Cambridge, UK

[11]ARUP Labs, University of Utah, Salt Lake City, USA

[12]CSIRO Mathematics, Informatics and Statistics, North Ryde, Australia; The Genome Analysis Centre, Norwich, UK

[13]Centre for Genetics and Genomics, The University of Nottingham, Nottingham, UK

[14]Cambridge Systems Biology Centre, University of Cambridge, Cambridge, UK

[15]Division of Physiology and Metabolism, MRC National Institute for Medical Research, London, UK

Abstract

Natural variation within species reveals aspects of genome evolution and function. The fission yeast *Schizosaccharomyces pombe* is an important model for eukaryotic biology, but researchers typically use one standard laboratory strain. To extend the utility of this model, we surveyed the genomic and phenotypic variation in 161 natural isolates. We sequenced the genomes of all strains, revealing moderate genetic diversity ($\pi = 3 \times 10^{-3}$) and weak global population structure. We estimate that dispersal of *S. pombe* began within human antiquity (~340 BCE), and ancestors of these strains reached the Americas at ~1623 CE. We quantified 74 traits, revealing substantial heritable phenotypic diversity. We conducted 223 genome-wide association studies, with 89 traits showing at least one association. The most significant variant for each trait explained 22% of variance on average, with indels having higher effects than SNPs. This analysis presents a rich resource to examine genotype-phenotype relationships in a tractable model.

Introduction

While the standard laboratory strain of *S. pombe* has been extensively studied, genetic variation and phenotypic diversity have been analyzed only in preliminary ways[1-3]. Remarkably little is known about the evolutionary history or ecology of this model organism. It was first described in East African millet beer in 1893, and the standard laboratory strain was isolated from French wine in 1924[4]. Natural isolates have also been collected from vineyards in Sicily, Cachaça (sugarcane spirit) in Brazil, and found to contribute to the microbial ecology of Kombucha (fermented tea)[1,5,6]. The diverse origins of these natural isolates (Fig. 1a; Supplementary Table 1) suggest that this yeast is now widely distributed.

To further describe *S. pombe*, we analyzed the genetic and phenotypic variation in natural isolates. Because the natural environment is not known, we collected all isolates available from the major stock centres and those given to us by microbial ecologists (Supplementary Table 1). These 161 strains had been collected over the last 100 years, in over 20 countries across the globe, primarily from cultivated fruit or various fermentations. Notably, the strains with known origin had been associated with human activities, providing little information about the natural environment of the species.

Results and Discussion

Variation and population structure

We sequenced the genome of all strains to at least 18-fold coverage, with a median 76-fold coverage. To facilitate detection of genetic variants, we mapped reads to the reference genome[7]. Mapping was comprehensive and accurate owing to the small, non-repetitive genome, allowing us to query 93% of the genome with high confidence (11.8 Mb of 12.6 Mb). We identified 172,935 high-quality single-nucleotide polymorphisms (SNPs), 14,508 small insertion and deletions (indels), and 1,048 long terminal repeat (LTR) insertions (Table 1).

Initial analysis revealed 25 clusters of near-identical strains that differed by <150 SNPs (Supplementary Fig. 1a). As most clusters were isolated from a single location, they probably derive from isolated, mitotically reproducing populations or from repeat depositions of the same strain to stock centers. By excluding such 'clonal' strains, we identified a set of 57 strains that each differ by ≥1,900 SNPs, which includes 99.6% of the SNPs present in all strains. The average pairwise diversity (π) within these 57 strains was 3.0×10^{-3} (3 SNPs/kb), slightly lower than the diversity within the budding yeast *Saccharomyces cerevisiae* ($\pi = 5.7 \times 10^{-3}$)[8,9]. Flow cytometry indicated that all but one (JB1207/NBRC10570) of these strains were haploid. Also, 34 of 39 strains were homothallic (i.e. contained both mating types), and all 57 strains were prototrophic (i.e. able to grow on same minimal medium as reference strain).

To describe the relatedness among these 57 strains, we analyzed SNPs in the nuclear genome. Some strains carry large inversions and translocations[2,10], which bias estimates of population structure when large regions of chromosomes are inherited without recombination[11]. Therefore, we selected a set of 752 SNPs that are close to linkage equilibrium (pairwise r^2 <0.5) and are distributed relatively evenly across the genome (Supplementary Fig. 1b), which better suits population genetic models that assume no linkage between variants. Principal component analysis of these SNPs showed weak clustering of strains by geography (Fig. 1b). Moreover, a pattern of genetic isolation by distance was evident, with genetic and physical distance being weakly, but significantly correlated (Supplementary Fig. 1c). This result suggests that there is some global population structure, which has been obscured by recent dispersal and intermixing of some strains. To examine whether this genetic isolation has resulted in any reproductive isolation, we measured spore viability between 43 crosses that spanned the range of genetic distances, avoiding crosses that involved known structural variants[2]. We found a significant correlation between genetic distance and spore viability (Pearson r = 0.52, $P = 6.5 \times 10^{-4}$, Supplementary Fig. 1d). This result suggests that these strains have accumulated sufficient genetic differences for reproductive barriers to emerge. Chromosomal rearrangements will also contribute to reproductive isolation[10,12].

The budding yeast *S. cerevisiae* shows strong clustering of strains, determined both by geography and cultural uses[8,13]. To assess the situation for *S. pombe*, we applied unsupervised genetic clustering methods, *Admixture*[14] and *fineSTRUCTURE*[15], which are oblivious to the geographic origin of the strains, to uncover any genetically differentiated

populations. Both clustering methods identified between two and five populations that were consistent with the principal component analysis (Supplementary Figure 2a-c). These results and further phylogenetic analysis showed that these groups were interbreeding populations, rather than clonally-isolated lineages (Supplementary Figure 2d). The F_{ST} values (proportion of between population genetic variance) for the five-population clustering ranged between 0.22 and 0.59 (mean 0.40) for different pairwise comparisons, indicating considerable genetic differences between these five connected clusters.

Dating the global dispersal of *S. pombe*

While *S. pombe* now appears globally distributed, we have no ecological or historic context to this dispersal, except that most strains were isolated from brewed beverages. The available strains were collected between 1912 and 2002, which allowed us to estimate the age of every node in the phylogenetic tree from the mitochondrial genomes, including the root (most recent common ancestor of all strains) (Fig. 2a). Modelling of the evolutionary rate showed that our data had predictive power (Fig. 2b), and we estimate the ancestor of all strains to have lived ~2,300 years ago (~340 BCE, Fig. 2c). A similar timeline could be deduced from the nuclear genome (Supplementary Note). This estimate points to an evolutionarily recent worldwide dispersal, perhaps associated with the spreading of technologies for brewing or other fermentations[16]. In comparison, it has been estimated that domesticated strains of *S. cerevisiae* dispersed 8-10,000 years ago, consistent with a Neolithic expansion[17]. Furthermore, our analysis provides a mean estimate of 1623 CE for the arrival of *S. pombe* in the Americas (95% confidence interval 1422-1752 CE), coincident with European colonialism of this continent, which began in 1492 CE. Notably, isolates from the Americas also showed the highest genetic similarity (Supplementary Note, Fig. 1b). Together, these findings suggest a recent European origin for *S. pombe* in the Americas.

Genetic diversity and genome function

Genetic variation data also contain signals of selection, which can be used to describe genome function. For example, both background selection and adaptive evolution reduce diversity most strongly in genetic elements that contribute to cell function. A consistent reduction in diversity is therefore a signature of functional elements, as reflected in the biased distribution of SNPs and indels (Table 1). Variation was significantly higher in the terminal 100 kb of all chromosomes and in centromeric regions (Mann-Whitney tests, $P=1.5\times10^{-21}$ and 3.2×10^{-7}, respectively) (Fig. 3a). These regions are unusual in that they contain no essential genes, have an excess of pseudogenes (19% *vs* 0.2% in genome), an excess of LTR insertions, and show low gene expression during vegetative growth, stationary phase and meiotic differentiation (Supplementary Fig. 3).

To systematically explore the relationship between genetic diversity and genome function, we calculated Watterson's θ (which measures nucleotide diversity) for the following annotation classes (Fig. 3b): protein-coding exons, introns, canonical RNAs (rRNAs, tRNAs, snoRNAs, snRNAs), long non-coding RNAs (lncRNAs), UTRs (untranslated regions of protein-coding transcripts), and the 15% of the genome not annotated as any of the above. Within exons, we calculated θ for one-fold degenerate sites (where all changes to DNA sequence lead to changes in protein sequence) and four-fold degenerate sites (4FD,

where all changes to DNA sequence result in same protein sequence). While polymorphisms in 4FD sites are not truly neutral, they are subject to much weaker selection[18]. As expected, protein-coding exons were the least diverse regions of the genome (Fig. 3b). Additionally, 5'- and 3'-UTRs and introns were all significantly less diverse than 4FD sites, suggesting substantial evolutionary selection at post-transcriptional levels of gene regulation. Analysis of SNP and indel median minor allele frequencies within windows showed consistent results (Supplementary Fig. 4a,b). While lncRNAs appeared to be subject to little or no purifying selection overall, further analyses revealed that the 20% most highly expressed lncRNAs were subject to detectable purifying selection (Supplementary Fig. 4c-e). These findings indicate that purifying selection is dominated by protein-coding transcripts, including their UTRs. As a consequence, we would expect fewer genetic variants to remain in gene-dense regions. Consistently, θ was strongly negatively correlated with protein-coding exon density, with outliers mainly derived from telomeric regions that lack essential genes (Fig. 3c).

Variation in transposon insertions and gene content

Transposons create another source of genomic variation, which may contain signatures of evolutionary processes. *S. pombe* has only one family of mobile elements, the *Tf*-type LTR retrotransposons[19]. The reference genome contains only 13 full-length *Tf* elements, but also several hundred solo LTR fragments that indicate the sites of previous insertions. These elements are transcribed at low levels[20], so may be actively propagating. To examine this possibility, we searched for novel insertions of *Tf*-elements in the non-clonal strains and determined which reference LTRs were present in the other 56 non-clonal strains. We located 1048 LTR insertions, of which 78% were not present in the reference. Consistent with previous studies showing that *Tf*-element insertions are targeted to RNA polymerase II (Pol II) promoters[21,22], we observed a sharp peak of insertions upstream of transcription start sites (Supplementary Fig. 5), and few insertions in exons (Table 1). The majority of the insertions (593 loci, 57%) were present only in a single strain, suggesting recent transposon integration and loss.

Transposon integration has been proposed to occur during cellular stress[23,24]. To examine this model, we analysed *Tf*-element insertions within intergenic regions containing one promoter and one terminator, as these insertions allow us to determine which promoter had been targeted by the insertion. Analysis of this set of 998 insertion sites upstream of 354 genes showed that insertions were more abundant upstream of genes with high Pol II occupancy, suggesting that gene expression level is a main determinant for *Tf*-element insertion. Insertions were also enriched upstream of intronless genes, which tend to be rapidly regulated[25], and of *sty1*-activated stress-response genes[26] (Supplementary Table 2). These observations corroborate the experimental finding that stress-response genes are targeted by *Tf*-insertions[22], and support the model that transposon integration occurs during stress, but also preferentially occurs in highly expressed genes.

To gauge how much our collection differed in gene content, we used *de novo* assemblies of the 57 non-redundant strains to identify genes that were present in at least one strain, but not present in the well-annotated reference strain. We created protein-coding gene predictions for each strain from the assembly and attempted to locate similar genes in the reference

strain. The strains were highly similar in their gene content; for example, 95% of the predicted peptides from the divergent strain JB758 could be aligned to a reference protein with >95% identity. Curation produced only 17 putative novel proteins, including nine with strong supporting evidence (Supplementary Table 3). The majority of these novel proteins were most similar to genes from *Ascomycete* fungi, including 12 for which we could identify orthologs in related *Schizosaccharomyces* species by blastp (e-value < 10^{-20}), suggesting ancient ancestry and subsequent gene loss in the reference strain. A notable exception was a protein most similar to the OsmC family from the plant pathogenic enterobacterium *Brenneria salicis*, with highly-conserved OsmC sequences being present in 29 of the 57 strains. This finding may reflect horizontal gene transfer, raising the possibility of an ecological association between *S. pombe* and plants.

Distribution of recombination

Meiotic recombination is a source of diversity that influences natural selection and also reflects population history. Recombination events are initiated via double-stranded breaks (DSBs) that occur preferentially at hotspots in the *S. pombe* genome[27,28]. To examine the distribution of recombination, we estimated the historic recombination rate by constructing genetic maps with distances in Linkage Disequilibrium Units (LDU)[29]. The rate estimate was zero for genomic regions spanned by 87% of the SNPs, and was log-normally distributed within the 13% of sites showing recombination (Supplementary Fig. 6a). Six regions with very high historic recombination rates were evident (rates >99.99th percentile; Fig. 3a). These hotspots showed a weak relationship with regions of high DSB activity (Spearman rank ρ=0.25, P=5.2×10^{-16}), but only 52% of the most recombinogenic SNPs were in DSB hotspots (Supplementary Note). As in other species, recombination positively correlated with genetic diversity (Spearman ρ=0.43, P=3.2×10^{-57}) and was primarily located away from genes (Supplementary Fig. 6b,c). For example, exons cover 57% of the genome, but only 26% of the 1000 highest recombination sites were in exons. The result of the low recombination regions is that on average linkage disequilibrium (r^2) declines to 50% within 21 kb (Supplementary Fig. 6d). Hence *S. pombe* shows long haplotypes compared to eukaryotes of similar genome size and gene density; for example, linkage disequilibrium in the budding yeasts *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* decline to 50% within 3-11 kb and 9 kb, respectively[8,9].

Phenotypic variation and genome-wide association studies

Model organisms have been utilized extensively to describe the complex genetics of quantitative traits[30,31], a task which is far more difficult in less tractable species such as humans. It was clear that our collection contained quantitative trait variation, both from previous studies[1,2] and from our observation that some strains showed differences in cell shape and size (Supplementary Fig. 7). To extend this data, we measured 74 quantitative traits using five methods selected to sample a large variety of different phenotypes: 1) manual and 2) automated measurements of cell shape and size, 3) multiple growth parameters in minimal and rich liquid media, 4) colony sizes on solid media under 42 different nutrient, drug and environmental conditions, and 5) mass-spectrometry measurements of intracellular amino-acid concentrations. Combined with previous data[2], we

analyzed 9,383 measurements for 223 phenotypes (an average of 164 values per strain) (Fig. 4a; Supplementary Table 4).

To assess the feasibility of using these data for genome-wide association studies (GWAS), we estimated the heritability of each of these phenotypes using the LDAK software[32], which considers additive genetic contributions without accounting for genetic interactions. These narrow-sense heritability estimates were significantly greater than zero for 130 of the 223 phenotypes, including phenotypes gathered using all methods (Supplementary Fig. 8a; Supplementary Table 5). Amino-acid concentrations were amongst the most heritable phenotypes, indicating a high metabolic diversity with little contribution from genetic interactions (which are not measured by narrow-sense heritability). Analysis of biological and technical repeat trait measurements also showed that the availability of repeats substantially increased the power of GWAS by reducing the non-genetic component of variance (Supplementary Fig. 8b).

GWAS would also be challenging if quantitative traits were clustered along with the population structure of the strains, as they are in budding yeast[33]. To examine this possibility, we tested each trait for significant differences in values between the 5 populations defined by *Admixture*. Only 19 of the quantitative 223 traits were significantly differentiated after Bonferroni correction, showing that traits are usually not stratified by populations (Supplementary Fig. 9a).

Since our traits were highly heritable and infrequently stratified by populations, we applied GWAS to search for genetic variants associated with each of 223 quantitative traits. We used a mixed model[34], utilizing all SNP and indel variants with minor allele counts ≥5 (108,105 SNPs and 8,543 indels). Mixed model linear regression accounts for unequal relatedness between individuals. Using trait-specific thresholds with a 5% family-wise error rate per trait, we discovered 1,419 variants that were significantly associated with at least one phenotype (1239 SNPs and 180 indels; Fig. 4b, Supplementary Table 6). Genomic inflation factors (median of observed test statistic divided by expected median) indicated that the mixed model was accounting for unequal strain relatedness well (Supplementary Fig. 9a,b). As an additional critical test of these associations, we divided the 57 non-clonal strains into three sub-populations (with 12, 26 and 17 members, defined by *Admixture*[14]), and examined each of these 1,419 variants for significant association using linear regression. Despite the small sample sizes, 67 of these variants were nominally associated with the trait and replicated in at one more sub-populations ($P < 0.05$; Fig. 4b, Supplementary Note).

Overall, we found that 1% of SNPs and 2% of indels were significantly associated with one or more traits (χ^2 test $P = 3.0 \times 10^{-15}$). Associated indels also explained higher proportions of trait variance (Supplementary Fig. 9c), consistent with indels being more destructive variants. Many of the indels used in the GWAS were in untranslated regions of coding transcripts (UTRs, Supplementary Fig. 9d), which we showed are subject to selective constraint, suggesting that indels contribute to phenotypic change by altering gene regulation.

For 89 of the 223 traits examined, at least one variant passed the significance threshold. We considered the most significant variants as the most likely candidates for causal variants. These 89 variants (72 SNPs, 18 indels) showed no bias for any genomic regions (Supplementary Fig. 9d) and explained 12-60% of trait variance, consistent with the expectation that the small sample size will have power to detect only variants of large effect. As for any GWAS, while estimates are globally unbiased, the largest estimates are likely to reflect a combination of genetic and stochastic effects and so tend to over-estimate the true genetic variance explained, a bias known as the winner's curse. In this study, the stochastic component of traits was well controlled by repeat measurements (Supplementary Fig. 8b), which will mitigate such bias.

Because of the extensive linkage disequilibrium (LD) in this collection, many variants will be significant because they are in LD to a causal variant. To locate further variants that are independently associated with traits, we re-applied the mixed model for each of these 89 traits, conditioning on the most significant variant. This approach uncovered 18 further variants (10 SNPs, 8 indels, Supplementary Table 6). These conditional hits explain 12-50% of the remaining trait variance.

The distribution of passing variants included six hotspots that harboured multiple variants associated with several different phenotypes (Fig. 4b). The most prominent of these hotspots contained 89 variants associated with six traits (Supplementary Fig. 10a), including the most significant three variants (all SNPs, all with $P = 7\times10^{-11}$, all of which have pairwise $r^2 = 1$). These polymorphisms are associated with growth in $MgCl_2$, and fall in the intergenic region between *nsk1* (encoding a microtubule-binding protein) and *sod2* (encoding a predicted manganese superoxide dismutase).

To experimentally validate this association, we crossed two strains that showed clear differences for this trait and contained the alternative haplotypes. We grew the pool of F1 progeny in the presence and absence of $MgCl_2$. Sequencing of this pool showed a bias to the expected allele, supporting a role for this variant in these two genetic backgrounds (Supplementary Fig. 10b,c). These results provide experimental support for a causal role for this variant or the tightly linked SNPs. As a first step towards identifying the gene(s) affected by these SNPs, we compared the growth of the standard laboratory strain to strains with either *nsk1* or *sod2* deleted. Both deletion strains were sensitive to $MgCl_2$ (Supplementary Fig. 10d), consistent with the haplotype affecting a bidirectional promoter between *nsk1* and *sod2*.

In conclusion, this study contributes to the understanding of *S. pombe* in several areas. Our analysis is limited by the available strains collected from human-associated samples that share a relatively recent common ancestor. However, we show that GWAS are feasible with this strain collection, and uncover a large number of potential causal variants. The effectiveness of GWAS, despite the low number of strains, was probably enabled by the relatively small genome and the quantitative phenotyping under tightly controlled conditions, which is obviously not possible with humans. We expect that the rich natural genetic and phenotypic variation presented here will provide a valuable resource to

understand the complexities and subtleties of genetic architecture and genome function in this model species.

Online Methods

Sequencing and quality control

All strains are described in Supplementary Table 1. Strains were sequenced with either 54 or 100 nt paired-end Illumina reads. To verify that strain identity was correct at various stages in the project we genotyped 30 SNPs (that would distinguish all the 57 non-clonal strains with at least two allelic differences) from the 161 extracts used for sequencing, repeat extracts of the 57 non-clonal strains, extracts from stocks obtained directly from stock centres extracts made from cultures picked from the ROTOR phenotyping plate. Only two of the 232 sets of genotypes were not as expected, and neither of these were members of the 57 non-clonal strains. All of the ROTOR plate extracts were as expected.

Read mapping, SNP and indel calling

Reads were mapped to the *Schizosaccharomyces pombe* 972 h^- reference genome (May 2011 Version)[7] with Stampy (v1.0.17)[18,37]. After detection of possible indel sites alignments were realigned with GATK IndelRealigner.

SNPs were called with the GATK UnifiedGenotyper and filtered using custom parameters (available on request). Indels were identified using the Genome Analysis Toolkit (GATK) HaplotypeCaller[38] and Cortex[39] both filtered using custom parameters. Cortex and HaplotypeCaller call sets were by merging any two indels from each set that were positioned within 3 nucleotides of each other, within a 30% length range and differing by a maximum of 1 minor allele count.

SNP and indel validation

To estimate false discovery rate and sensitivity of SNP calling, we sequenced ~20 paired end shotgun clones from each of four strains with increasing genetic distance from the reference with an ABI capillary machine. Reads were then mapped to the reference genome using BWA *mem*[40]. We then manually examined 85 windows of the genome using the IGV tool[41]. This included 47,619 nt of mappable regions, and 182 known SNPs. We found that all of these were valid, while 17 were discovered in alignments that were not called by our SNP calling pipeline (8.5% false negative rate).

To estimate the false discovery rate of indel calling, we manually inspected Illumina read alignments at 100 indels called in the same four strains, choosing indels that were dispersed across all chromosomes. Only 4 of these calls were false positives for an indel occurring at the site (4% false discovery rate for calling an indel). A total of 7 indels contained at least one strain with an incorrect allele call.

Locating Tf retrotransposons

We used RetroSeq[42] to locate insertions in the 57 strains that were not present in the reference strain. As LTR insertions are highly targeted in *S. pombe*[22], we used soft-clipped,

Europe PMC Funders Author Manuscripts

unaligned parts of a sequence reads covering the insertions sites to distinguish between independent insertions at closely situated genomic sites, collating 1474 predicted insertions into 820 insertion events (Supplementary Table 8). We assed the target site duplication (TSD) sizes from from the soft-clipped reads. We used PCR to verify 90 of the RetroSeq predictions. 56 of these produced a product in both reference and alternate strain and 80% (45/56) of these confirmed the insertion with high confidence, while 93% (52/56) confirmed the insertion with at least medium confidence (Supplementary Table 9).

To determine which reference LTR elements were present in each wild strain, we used *delly* (Version 0.0.6)[43] to locate deletions in the same position as a reference LTR sequence. Genes targeted by LTR insertions only considered LTR insertions between genes arranged in tandem (i.e. neighboring genes in the same orientation). Gene features were analysed by the GeneListAnalyser (http://128.40.79.33/cgi-bin/GLA/GLA_input).

Diversity analysis

Diversity estimates were calculated using Variscan[44]. For 10kb window analysis we excluded windows with less than 1000nt of reliably called sites. To compare annotations of the genome, we used regions that were annotated exclusively as exon/intron/ncRNA, *etc*. Median minor allele frequency was calculated from all passing SNPs or indels, in the 100 (126 kb long) windows of the genome for SNPs, and in 50 (252 kb) windows for indels.

Recombination rate, hotspots and linkage disequilibrium maps

We used LDMAP[45] to construct LDU maps from the SNPs segregating in 46 unrelated strains that looked to be a homogenous population from principal components analysis, excluding SNPs with MAF <0.05. We calculated the DSB rate (per microarray probe) from the data of Cromie et al.[27], as the median signal for all probes in a 7-probe window, using both repeats of the 5h time point (14 probes in all), divided by the median signal for probes in the 7-probe window for the 0h time point. For both recombination rate and the DSB rate, we then calculated the mean signal over non-overlapping 1 kb windows of the genome. Pairwise D′ and r^2 were calculated between all pairs of SNPs with a minor allele frequency >0.05 up to 250 kb distance, using LDMAP (for D′)[45] and PLINK (r^2)[46]. Mean values were calculated from ≥500,000 pairwise comparisons for each 1 kb window.

Population structure

For analysis tools that assume variants are independent, we used 752 SNPs that were unlinked (pairwise r^2 <0.5) ('unlinked SNPs'). We used *vcftools*[47] to estimate the Weir and Cockerham weighted F_{ST}, using all SNPs, for all pairwise combinations of populations. *Admixture* (Version 1.22)[14] was run with k=1 to k=20. *ChromoPainter* and *fineSTRUCTURE*[15] were run using only the non-clonal 57 strains, using all SNPs, utilizing the recombination rate estimate. When using *ChromoPainter*, we first ran 10 Expectation-Maximisation (E-M) iterations to infer the "global mutation" and "switch rate" parameters, then averaged the inferred values for each across chromosomes, weighting by the number of SNPs, and performed a final *ChromoPainter* run using these weight-averaged values. Isolation by distance was calculated using the using geoDist from SoDA packages in R. See Supplementary Note for more details.

Dating strain divergence with mitochondrial data

This analysis used only the 84 strains with recorded sampling dates, which contained 204 SNPs. The *Schizosaccharomyces cryophilus* mitochondrial genome (Genbank accession ACQJ00000000.2, Supercontig_3.27), was used as the outgroup, aligned to the *S. pombe* strains using Muscle[48].

We used PartitionFinder[49] to choose the optimal partitioning scheme (K= 5) and nucleotide substitution model. Phylogenetic analyses were performed with *BEAST* 1.7.4[35] on both the 5 schemes obtained with PartitionFinder and the whole molecule. In the first case, substitution and clock models were unlinked while tree topology was assumed to be the same between the 5 schemes. Log-normal relaxed clocks were compared to strict clocks through the evaluation of Bayes factors. To do so, marginal likelihood was computed using both path (PS) and stepping-stone (SS) sampling method[50]. To minimize demographic assumptions, we adopted a Bayesian skyline plot approach to integrate over different coalescent histories. Rate variation among sites was modeled with a discrete gamma distribution with 4 rate categories. Posterior distributions of parameters, including divergence times and substitution rates, were estimated by Markov chain Monte Carlo (MCMC) sampling in *BEAST*. For each analysis, we ran four independent a posteriori combined chains in which samples were drawn every 2500 MCMC steps from a total of 25,000,000 steps, after a discarded burn-in of 2,500,000 steps. Convergence to the stationary distribution was assessed by inspection of posterior samples.

TMRCA estimate with nuclear DNA

To obtain TMRCA estimates for the nuclear genome, we produced independent runs of ACG[51] for the full mitochondrial genome and for 160 regions of the nuclear genome, each 20 kb in size. So that background selection between the mitochondrial and nuclear genome fractions would be approximately similar, we selected nuclear regions to have an exon density of 50-60%, similar to that of the mitochondria. To ease computational burden and aid convergence of the chains, we randomly chose 15 of the samples for inclusion. For each region an ACG run of 5×10^7 steps was conducted using a Metropolis-coupled MCMC scheme with 8 chains. The first 25% of steps were discarded as burn-in. We estimated posterior distributions of the parameters of the substitution matrix assuming the TN93 model[52], the ancestral recombination graph (ARG), recombination rate, substitution rate, and locations of recombination breakpoints from the data. Flat (uniform) priors were assumed for all parameters except the recombination rate, for which we employed an exponential prior with mean 100.0 in units of recombinations per unit of branch length. Convergence of chains was assessed by visual examination of the likelihood of the data conditional on the ARG.

De novo assembly

De novo assemblies were performed using SGA version 0.9.35[53]. Error correction used 41-mer frequencies to identify and correct sequencing errors. For the contig-assembly step, the minimum overlap length was set to 65bp for the strains with 100 nt reads. For strains with 54 nt reads, a minimum overlap of 45 bp was required instead. Evidence from a minimum of five read pairs was required to build contigs into a scaffold.

Locating novel genes

To identify protein-coding genes that were present in a wild strain(s) but not in the reference, we produced gene predictions from each *de novo* assembly with Augustus[54] using default parameters. We then compared each predicted protein to the *S. pombe* reference using BLAST+[55] *blastp*, *tblastn* and *blastn*. Predictions ≥ 100 amino acids that scored < 80% identity from all of blast searches were chosen as potential novel genes (800 predicted peptides). We used Markov clustering[56] to group these peptides into 32 clusters of similar peptides and 5 singletons. We then aligned each cluster with Clustal Omega[57], produced a consensus using Emboss *cons*, and used this consensus as a query for *blastp* searches against the *S. pombe* reference protein data set, and the NCBI nr protein data set. We excluded potential novel genes whose best nr blast hit was from *S. pombe*, or from the phage Φ×174 (likely contamination). We retained the 17 potential novel genes where the ratio of (nr blastp bit score)/(*S. pombe* bit score) was >1. To examine the conservation of the 17 potential novel genes in other Schizosaccharomyces yeasts, we used each predicted protein (from each *S. pombe* strain) from the 17 putative most promising novel genes to query the predicted proteins of *S. cryophilus*, *S. japonicus* and *S. octosporus* using blastp, accepting blast hits with an e-value < 10^{-20} in one or more species.

Phenotyping

A summary of all phenotype measurements is provided in Supplementary Table 4, and the specific approaches are described below.

Amino acid quantification

Phenotypes with prefix "aaconc" in Supplementary Tables 4-5—Triplicate cultures (1.6 ml) of each strain were cultured for 8 hours, cells extracted with 80°C boiling ethanol, extracts were cleared from insoluble material by centrifugation and the supernatant collected for LC-MS/MS analysis. Samples were analysed on a LC (Agilent 1290 Infinity) - MS/MS (Agilent 6460) system. Amino acids were separated by hydrophilic interaction chromatography by gradient elution using an ACQUITY UPLC BEH amide column.

Amino acid concentrations were determined by external calibration. Dilution was corrected by probabilistic quotient normalization[58]. Repeats: The average of the amino acid values from the triplicates was used for further analysis. For quality control, all values with a CV greater than two times the overall CV (median) were eliminated. For the 19 amino acids, median coefficients of variation were between 0.07 and 0.21 (mean of 0.13).

Growth and stresses on solid media

Phenotypes with prefix "smgrowth" in Supplementary Tables 4-5—Strains were arrayed by a RoToR robot (Singer Instruments) onto solid YES and EMM2 media at 1536-spot density, with each strain represented by 4 spots. Edges of plates and various interspersed positions were inoculated with the standard strain, as were strains with known sensitivity (*atf1*Δ and *sty1*Δ) or resistance (*pka1*Δ).

Plates were incubated at 32°C and high-resolution images of the plates were acquired using a UVP Multi-DocIt transillumination system. Two biological replicates were performed.

Quantification of colony sizes was then performed using the custom Workspace package with the *Spotsizer* custom workflow (manuscript in prep.). Colonies with microbial contaminations and misidentified colonies were discarded. Median strain colony size was then calculated for each plate and replicate. Conditions or plates showing poor reproducibility were removed from further analysis. Strain colony size data per condition were normalized to the growth on YES, and then to the growth of the 972 h^- reference strain under the given condition. Repeats: Two or more replicate plates were analysed for 25 of the 43 conditions, and one plate for all others. Plate values were the median colony size from the four colonies per strains. The median between-plate Pearson correlation was 0.95.

Cell growth parameters/kinetics in liquid media

Phenotypes with prefix "lmgrowth" in Supplementary Tables 4-5—All 57 non-clonal strains were cultured in a Biolector micro-fermenter (m2p labs) in 1.5 ml of YES/EMM2 media (Formedium) using m2p labs flowerplates for 24 hours at 32°C, measuring light scattering every 10 minutes. Each strain was repeated in at least in duplicate. For each replicate of optical density data points we used the R grofit package[59] to determine all growth parameters. Repeats: Two biological repeats Biolector cultures were grown per strain. Correlations between biological repeats were typically >0.9, and all above 0.884. All coefficients of variation (within a strain) were above 0.075 (median for all traits = 0.034).

Manual cell morphology characterization

Phenotypes with prefix "shape1" in Supplementary Tables 4-5—Strains were grown on YES plates at 32°C and allowed to form small colonies. Cells around the edge of at least 5 colonies were examined using a Zeiss Axioskop microscope using both a X20 LD ACROPLAN 0.4 and a X50 CF plan 0.55 objective and the cell phenotype described. Using X50 CF plan 0.55 objective with 2.5× Optivar, a representative colony was photographed using Sony NEX 5N camera. For liquid media, strains were grown mid log and examined using a Zeiss Axioskop 40 with a X63 Plan APOCHROMAT 1.4 oil immersion objective. Cell length and width was measured for a minimum of 30 septated cells using ImageJ. FACS analysis was carried out as described[60]. The percentage of cells with 1C, 2C, 2-4C and >4C was estimated using FlowJo, http://www.flowjo.com. Repeats: Length and width were the median of at least 34 cells (median of 53), with the median coefficients of variation of 0.07 in both cases.

Automated cell morphology

Phenotypes with prefix "shape2" in Supplementary Tables 4-5—Cells were grown to mid log phase in YES medium and imaged using the OperaLX (PerkinElmer, USA) high-throughput microscope at 60×. Images were then automatically pre-processed, segmented and analysed to give 54 independent measurements of phenotypic features for all strains.

The occurrence of stereotypical *S. pombe* cell shape phenotypes (wild-type, long, stubby, curved, branched, round, skittle and kinked;) was assessed for each strain using SVM classifiers. This method is described fully in Graml et al.[61] where cells were imaged using

405 nm and 488 nm exposure channels with 10 independent repeats. Here, only the 405 nm channel and 6 repeats were needed.

The symmetrized Kullback–Leibler divergence between each strain and the reference was used as an additional quantitative trait (the 'shape2.KL.Predicted.*' in Supplementary Table 4), along with the length, width, and the ratio of width of both sides of the cell (i.e. 'cell asymmetry'). Repeats: Up to 6 populations of cells per strain. Since measurements were generally non-Gaussian, variation within populations was assessed using the median of absolute deviation (MAD) divided by the median. MAD values ranged from 0.04 (length) to 1.42 (ks.predicted.long), average 0.87.

Heritability and Genome-wide association studies

We used LDAK[32] to estimate heritability of all traits. We report values based on quantile normalized phenotypes (see below) but we also repeated estimates using raw values. Heritability estimated with raw values were strongly correlated normally transformed values ($r = 0.69$, $P = < 2.2\times10^{-16}$).

We performed mixed model association analysis using FastLMM[34], version 2.07. The mixed model adds to the standard linear regression model a polygenic term, designed to "soak up" the effects attributable to relatedness and population structure[62]. We first normalized each phenotype by replacing observed values with the corresponding quantile from a standard normal distribution. We excluded variants with less than 5 calls for the minor allele (MAF < 3.1%), and variants that had >5% of missing calls. We estimated a trait-specific P-value threshold for each trait by permuting trait values between individuals 1000 times, recording the lowest P-value from Fast-LMM analysis and using the 5% quantile (50^{th} lowest value) as the threshold. Passing variants therefore have a 5% family-wise error rate. We also performed conditional analysis; for each of the 89 traits with at least one variance significant from the primary mixed model GWA, we repeated the analysis, including as a covariate the genotypes from the most significant variant.

Genomic inflation factors (GIFs) were calculated as (GIFs) were calculated as: GIF = median $(\chi^2_{observed}(P))/(\text{median } \chi^2_{expected}(P))$, and adjusted GIFs as: GIF = median $(\chi^2_{observed}(P))/(\text{median } \chi^2_{permuted}(P))$. Where $\chi^2_{observed}(P)$ are the chi-squared statistics corresponding to the observed P-values and $\chi^2_{expected}(P)$ are those expected assuming P-values are distributed uniformly within [0,1]. Permuted P-values were contained by permuting trait values, once for each of the 223 traits used for the GWAS. The median permuted GIF from all traits was 0.454.

To validate the results from the association analyses, we split the 57 non-clonal strains into 3 datasets (3 populations defined by *Admixture*, on the 752 independent SNPs). Each dataset was therefore a homogeneous group of relatively unrelated members. The three datasets had 12, 26 and 17 members but 2 out of the 57 strains were excluded because they were not members of any of the 3 populations. The association analysis was based on a linear regression of every trait on each of the 1,567 markers that passed the GWAS threshold from the initial analysis using the pooled data and the mixed model. We then meta-analysed only those that were replicated (showed nominal statistical evidence of association in at least 2

out of the 3 *k* datasets). The P-values from the linear regression from each dataset for the same trait and marker was combined using Fisher's combined probability test:

$$\chi^2 = -2\sum_{i=1}^{k} \ln(P),$$

The meta P-value was obtained for 6 degrees of freedom (2*k*).

A summary of all the validated signals using linear regression together with their meta P-values and the P-values from the pooled data using the mixed model are presented in Supplementary Table 6.

Statistics

All statistics were produced with R[63].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Leah Clissold, Heather Musk, David Baker and Robert Davey for their contribution to sequencing, Henry Levin for discussions about transposons, and Juan Mata and Samuel Marguerat for comments on the manuscript. This work was supported by a Wellcome Trust Senior Investigator Award to JB (grant # 095598/Z/11/Z), by the Wellcome Trust to SB, TK, JTS and RD, by grant 260801-BIG-IDEA from the European Research Council (ERC) and grant BB/H005854/1 from the Biotechnology and Biological Sciences Research Council (BBSRC) to AR and FB, by Medical Research Council grant G0901388 to DS and DJB, by a Cancer Research UK Postdoctoral Fellowship to TMKC, by a ERC Starting Grant (SYSGRO) to REC-S, a Wellcome Trust PhD studentship to JLDL, and a BBSRC grant BB/K006320/1 to REC-S and AC, by a Wellcome Trust grant (RG 093735/Z/10/Z) and ERC Starting Grant 260809 to MM and MR, MR is a Wellcome Trust Research Career Development and Wellcome-Beit Prize Fellow, by the Czech Science Foundation grant P305/12/P040 and Charles University grant UNCE 204013 to MP, and by Cancer Research UK to LJ and JH.

References

1. Gomes FCO, et al. Physiological diversity and trehalose accumulation in Schizosaccharomyces pombe strains isolated from spontaneous fermentations during the production of the artisanal Brazilian cachaça. Can J Microbiol. 2002; 48:399–406. [PubMed: 12109879]
2. Brown WRA, et al. A geographically diverse collection of Schizosaccharomyces pombe isolates shows limited phenotypic variation but extensive karyotypic diversity. G3. 2011; 1:615–626. [PubMed: 22384373]
3. Fawcett JA, et al. Population Genomics of the Fission Yeast Schizosaccharomyces pombe. PLoS ONE. 2014; 9:e104241. [PubMed: 25111393]
4. Osterwalder A. Schizosaccharomyces liquefaciens n.sp., eine gegen freie schweflige Säure widerstandsfähige Gärhefe. Mitt Geb Lebensmittelunters Hyg. 1924; 15:5–28.
5. Florenzano G, Balloni W, Materassi R. Contributo alla ecologia dei lieviti Schizosaccharomyces sulle uve. Vitis. 1977; 16:38–44.
6. Teoh AL, Heard G, Cox J. Yeast ecology of Kombucha fermentation. Int J Food Microbiol. 2004; 95:119–126. [PubMed: 15282124]
7. Wood V, et al. The genome sequence of Schizosaccharomyces pombe. Nature. 2002; 415:871–880. [PubMed: 11859360]

8. Liti G, et al. Population genomics of domestic and wild yeasts. Nature. 2009; 458:337–341. [PubMed: 19212322]
9. Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L. Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. Nature. 2009; 458:342–345. [PubMed: 19212320]
10. Teresa Avelar A, Perfeito L, Gordo I, Godinho Ferreira M. Genome architecture is a selectable trait that can be maintained by antagonistic pleiotropy. Nat Commun. 2013; 4:2235. [PubMed: 23974178]
11. Seich Al Basatena N-K, Hoggart CJ, Coin LJ, O'Reilly PF. The effect of genomic inversions on estimation of population genetic parameters from SNP data. Genetics. 2013; 193:243–253. [PubMed: 23150602]
12. Zanders SE, et al. Genome rearrangements and pervasive meiotic drive cause hybrid infertility in fission yeast. eLife. 2014; 3:e02630. [PubMed: 24963140]
13. Cromie GA, et al. Genomic Sequence Diversity and Population Structure of Saccharomyces cerevisiae Assessed by RAD-seq. G3. 2013 doi:10.1534/g3.113.007492.
14. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19:1655–1664. [PubMed: 19648217]
15. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012; 8:e1002453. [PubMed: 22291602]
16. Hornsey, IS. A History of Beer and Brewing. The Royal Society of Chemistry; 2003. p. X001-X004.doi:10.1039/9781847550026
17. Fay JC, Benavides JA. Evidence for domesticated and wild populations of Sacchoromyces cerevisiae. PLoS Genet. 2005; 1:66–71. [PubMed: 16103919]
18. Zhou T, Gu W, Wilke CO. Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol. 2010; 27:1912–1922. [PubMed: 20231333]
19. Bowen NJ, Jordan IK, Epstein JA, Wood V, Levin HL. 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. 2003; 13:1984–1997. [PubMed: 12952871]
20. Mourier T, Willerslev E. Large-scale transcriptome data reveals transcriptional activity of fission yeast LTR retrotransposons. BMC Genomics. 2010; 11:167. [PubMed: 20226011]
21. Kwon E-JG, et al. Deciphering the transcriptional-regulatory network of flocculation in Schizosaccharomyces pombe. PLoS Genet. 2012; 8:e1003104. [PubMed: 23236291]
22. Guo Y, Levin HL. High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in Schizosaccharomyces pombe. Genome Res. 2010; 20:239–248. [PubMed: 20040583]
23. Guo Y, et al. Integration profiling of gene function with dense maps of transposon integration. Genetics. 2013; 195:599–609. [PubMed: 23893486]
24. Feng G, Leem Y-E, Levin HL. Transposon integration enhances expression of stress response genes. Nucleic Acids Res. 2013; 41:775–789. [PubMed: 23193295]
25. Jeffares DC, Penkett CJ, Bähler J. Rapidly regulated genes are intron poor. Trends Genet. 2008; 24:375–378. [PubMed: 18586348]
26. Chen DR, et al. Global transcriptional responses of fission yeast to environmental stress. Mol. Biol. Cell. 2003; 14:214–229. [PubMed: 12529438]
27. Cromie GA, et al. A discrete class of intergenic DNA dictates meiotic DNA break hotspots in fission yeast. PLoS Genet. 2007; 3:e141. [PubMed: 17722984]
28. Fowler KR, Gutiérrez-Velasco S, Martín-Castellanos C, Smith GR. Protein Determinants of Meiotic DNA Break Hot Spots. Molecular Cell. 2013 doi:10.1016/j.molcel.2013.01.008.
29. Maniatis N, et al. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci USA. 2002; 99:2228–2233. [PubMed: 11842208]
30. Liti G, Louis EJ. Advances in quantitative trait analysis in yeast. PLoS Genet. 2012; 8:e1002912. [PubMed: 22916041]

31. Mackay TFC. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet. 2014; 15:22–33. [PubMed: 24296533]
32. Speed D, Hemani G, Johnson MR, Balding DJ. Improved Heritability Estimation from Genome-wide SNPs. Am. J. Hum. Genet. 2012; 91:1011–1021. [PubMed: 23217325]
33. Warringer J, et al. Trait variation in yeast is defined by population history. PLoS Genet. 2011; 7:e1002111. [PubMed: 21698134]
34. Listgarten J, et al. Improved linear mixed models for genome-wide association studies. Nature Methods. 2012; 9:525–526. [PubMed: 22669648]
35. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012; 29:1969–1973. [PubMed: 22367748]
36. Clément-Ziza M, et al. Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast. Mol Syst Biol. 2014; 10:764. [PubMed: 25432776]

Methods only references

37. Lunter G, Goodson M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2010 doi:10.1101/gr.111120.110.
38. Depristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011; 43:491–498. [PubMed: 21478889]
39. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature Genetics. 2012; 44:226–232. [PubMed: 22231483]
40. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26:589–595. [PubMed: 20080505]
41. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013; 14:178–192. [PubMed: 22517427]
42. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics. 2013; 29:389–390. [PubMed: 23233656]
43. Rausch T, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012; 28:i333–i339. [PubMed: 22962449]
44. Hutter S, Vilella AJ, Rozas J. Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics. 2006; 7
45. Lau W, Kuo T-Y, Tapper W, Cox S, Collins A. Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. Bioinformatics. 2007; 23:517–519. [PubMed: 17142813]
46. Purcell S, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007; 81:559–575. [PubMed: 17701901]
47. Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27:2156–2158. [PubMed: 21653522]
48. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]
49. Lanfear R, Calcott B, Ho SYW, Guindon S. Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol. 2012; 29:1695–1701. [PubMed: 22319168]
50. Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics. Mol Biol Evol. 2013; 30:239–243. [PubMed: 23090976]
51. O'Fallon BD. ACG: rapid inference of population history from recombining nucleotide sequences. BMC Bioinformatics. 2013; 14:40. [PubMed: 23379678]
52. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993; 10:512–526. [PubMed: 8336541]

53. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012; 22:549–556. [PubMed: 22156294]
54. Stanke M, Schoffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006; 7
55. Camacho C, Coulouris G, Avagyan V. BLAST+: architecture and applications. BMC Evol Biol. 2009:421–430.
56. Van Dongen S, Abreu-Goodger C. Using MCL to extract clusters from networks. Methods Mol Biol. 2012; 804:281–295. [PubMed: 22144159]
57. Sievers F, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011; 7:539. [PubMed: 21988835]
58. Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Anal. Chem. 2006; 78:4281–4290. [PubMed: 16808434]
59. Kahm M, Hasenbrink G. grofit: Fitting biological growth curves with R. Journal of Statistical Software. 2010; 33:1–21. [PubMed: 20808728]
60. Sazer S, Sherwood SW. Mitochondrial growth and DNA synthesis occur in the absence of nuclear DNA replication in fission yeast. J. Cell. Sci. 1990; 97(Pt 3):509–516. [PubMed: 2074269]
61. Graml V, et al. A Genomic Multiprocess Survey of Machineries that Control and Link Cell Shape, Microtubule Organization, and Cell-Cycle Progression. Dev. Cell. 2014; 31:227–239. [PubMed: 25373780]
62. Yu JM, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics. 2006; 38:203–208. [PubMed: 16380716]
63. R Core, T. R: A Language and Environment for Statistical Computing. 2013. R-project.org at <http://www.R-project.org>

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

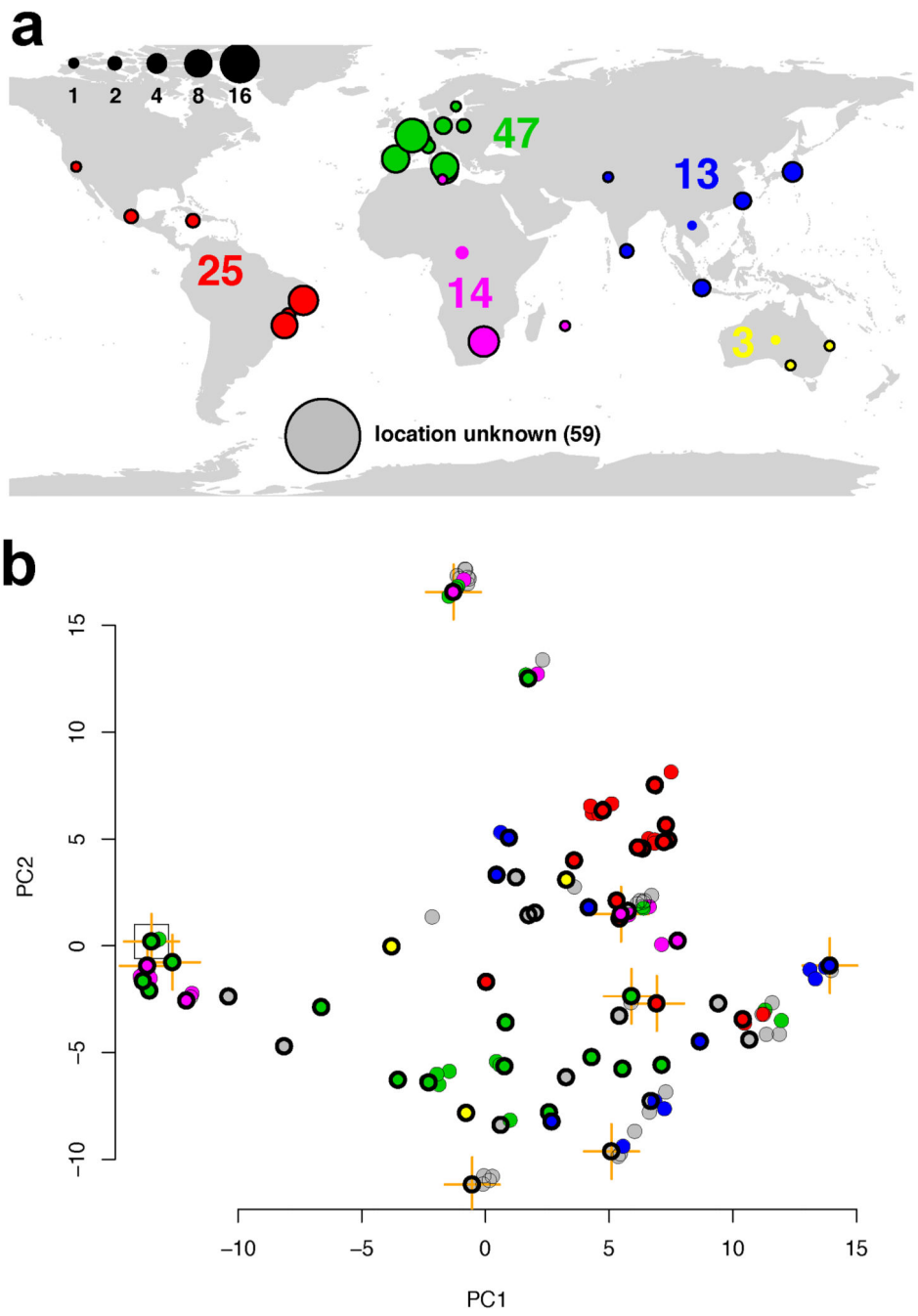

Figure 1. An overview of the strain collection

a, Geographic origins of all 161 strains analyzed. Colored circles indicate the original sources of strains used in this study, with circle sizes indicating the number of strains obtained from each site (as in scale of black circles, top left). Strains for which only an approximate source is known (e.g. Africa) lack the black border. **b,** principal components projection of 'drift distance' between strains determined using the 752 unlinked SNPs (see Methods). The color scheme is as in (a). Leupold's 972 reference strain is indicated with an open black square; strains that are members of the non-redundant group of 57 strains have a

black border; strains known to contain large structural inversions[2] are indicated with an orange cross.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

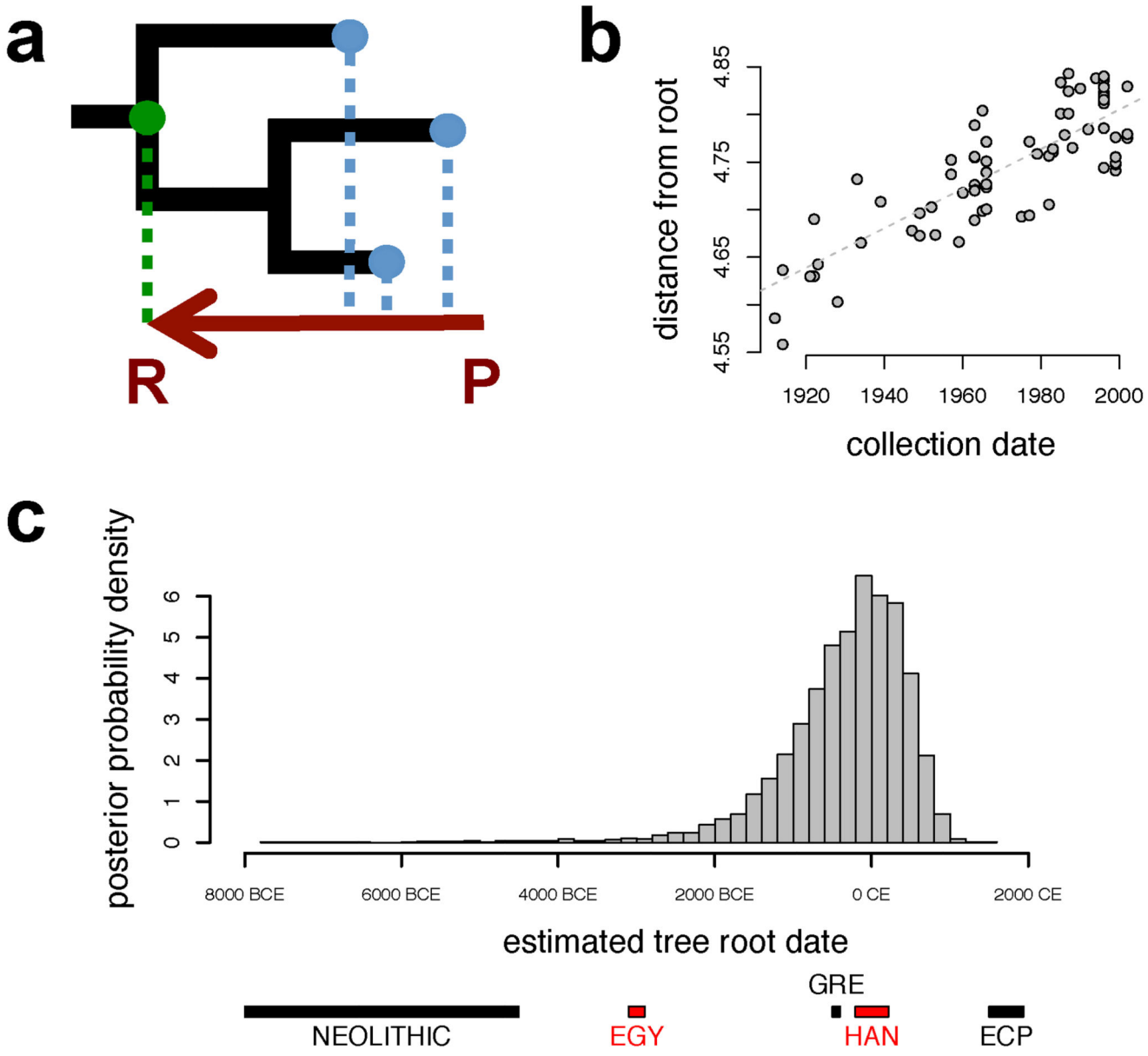

Figure 2. Recent dispersal of *S. pombe*

a, Calibration of tree nodes using dated tips. With a collection of sequences sampled over various times (blue dots) until the present day (P), we can jointly estimate the phylogenetic tree topology (in black), the rate of evolution and the age of any node in the tree, including the root, the most recent common ancestor of all strains (R, green dot). **b,** Root to tip distances (mutations/site × 10^{-3}) correlate with collection date ($P < 10^{-16}$), showing the data has reasonable predictive power. Distances were estimated using *BEAST*[35] from mitochondrial data of the 81 strains where collection dates were available, statistical details are provided in Methods. The grey line shows the linear model. **c,** Historic context of dispersal. The posterior probability distribution for time to most recent common ancestor (TMRCA) of the 81 collection-dated strains estimated using *BEAST*. The mean estimate was

340 BCE (95% confidence interval: 1875 BCE-1088 CE). Approximate historical periods are shown for context: ECP, European Colonial Period (~1500-1940 CE), HAN, Han Dynasty in China (206 BCE-220 CE), GRE, Classical Greece (400 BCE-500 BCE), EGY, First Dynasty of ancient Egypt (2890 BCE-3100 BCE), NEOLITHIC, Neolithic Era (4,500 BCE-10,000 BCE).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

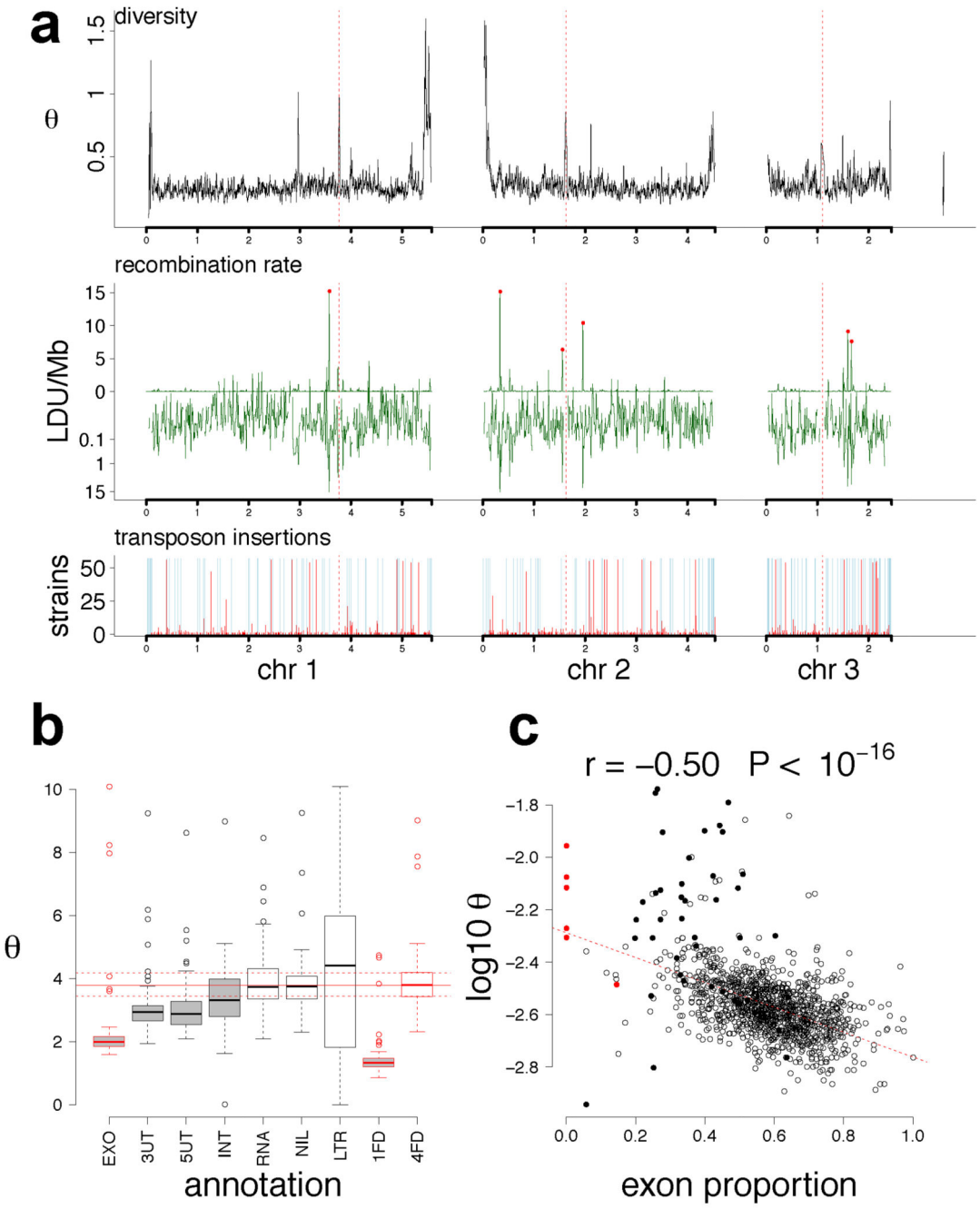

Figure 3. Relationships between genetic diversity and genome function

a, Main features of diversity in the genome, with chromosome scale in Mb on *x*-axis, and mitochondrial genome on right edge. Top panel, diversity (Watterson's θ) calculated using SNPs (scale: $\theta \times 10^{-2}$). Middle panel, recombination rate (scale: LDU/Mb $\times 10^{-3}$ above *x*-axis and log(1+LDU/Mb) below *x*-axis). The six major recombination hotspots are indicated with red dots. Bottom panel, sites of *Tf*-family LTR insertions (scale: number of strains containing each insertion, with insertions present in all strains shown in light blue) in the group of 57 strains. **b,** Diversity described by genome annotation. Distribution of

Europe PMC Funders Author Manuscripts

Watterson's θ values for each 100th of genome, using only annotated sites annotated as: exons (EXO), 5'- and 3'-UTRs (5UT, 3UT), introns (INT), long non-coding RNAs (RNA), un-annotated regions (NIL), LTRs of *Tf2*-family transposons (LTR), one-fold (1FD) and four-fold (4FD) degenerate sites of exons. Protein-coding categories have red borders. The horizontal red lines indicate the median and interquartile range for 4FD sites, annotation classes significantly lower than this neutral proxy shaded grey. One-sided paired Mann-Whitney test P-values vs the FFD site neutral proxy were; exons, UTRs and one-fold degenerate sites all $P < 2\times10^{-16}$, introns $P = 1\times10^{-6}$, lncRNAs, un-annotated regions and LTRs $P > 0.05$. **c**, Diversity is negatively correlated with exon density. Diversity (θ) and proportion of each window annotated to protein-coding exons determined for 10 kb genomic windows. The Spearman rank correlation and significance are shown on top. Filled red circles: centromeric regions; filled black circles: telomeric regions (terminal 100 kb).

Europe PMC Funders Author Manuscripts

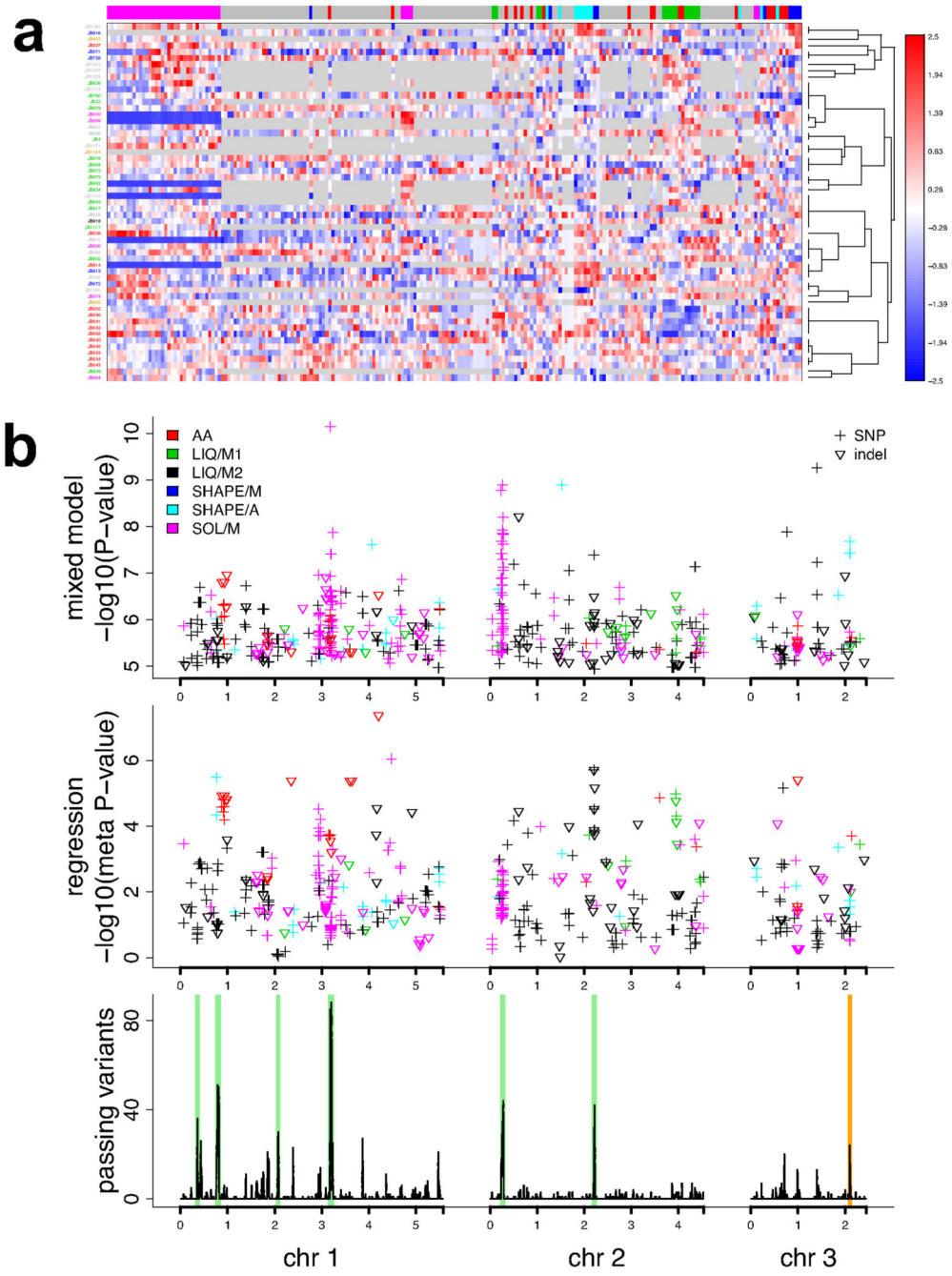

Figure 4. Phenotypes and genome-wide associations

a, Phenotypic variation of all 57 non-clonal strains, with strains in rows and phenotypes in columns. Phenotype values are normalized, according to the scale at right, missing data are colored grey. The colored panel above each row indicates the category of phenotype measurement. Categories are amino-acid concentrations (AA, red), growth on liquid media from this study (LIQ/M1, green), growth on liquid media (LIQ/M2, black)[2], manual (SHAPE/M, blue) and automated (SHAPE/A, cyan) shape phenotypes, growth on solid media (SOL/M, magenta). Phenotypes are hierarchically clustered using phenotype values,

and strains are clustered according to their genetic relatedness using tree at right inferred by *fineSTRUCTURE*. Strain names are colored according to their geographic origin, as in Fig. 1a. All phenotypes were measured for at least two biological replicates, values shown are generally medians from biological and technical repeats (see Methods). **b**, Top panel shows variants that were associated with one or more traits using the mixed model GWAS. Variants are shown as crosses (SNPs) or triangles (indels), colored by phenotype category (as above). The horizontal scale shows the physical distance in Mb. The middle panel shows, for variants significant in our primary GWAS, the meta-P-values from linear regression within populations. The lower panel shows the total number of passing variants in each 10,000 nt window of genome. Six hotspots (≥30 variants/10 kb) are indicated with green vertical bars. The orange bar shows the location of a hotspot discovered in an independent eQTL study[36]. P-values thresholds for the mixed model are derived from permutations of traits (Methods).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Genetic variation discovered in *S. pombe* strains. Variant counts that are enriched (above what is expected for percentage of genome) are in bold text, with the most enriched annotation shown in bold. The number of bases and percentage of nucleotides annotated refers to the reference genome.

Annotation	Bases	% Genome*	SNPs	Indels	LTRs
Genome	12,591,251	100	172,935	14,508	1,048
Exon	7,204,824	57.2	78,567	882	41
synonymous/frame conserving	-	-	46,624	882	-
non-synonymous/frame shift	-	-	31,441	453	-
pseudogenes	38,896	0.3	254	19	0
stop gained/lost	-	-	230	-	-
start gained/lost	-	-	18	-	-
5′ or 3′ UTR	3,270,717	26	**48,839**	**6,947**	**298**
No annotation	1,851,692	14.7	**35,306**	**4,464**	**598**
Non-canonical ncRNA	1,722,785	13.7	**27,866**	**2,851**	**223**
Intron	213,282	1.7	**3,709**	**570**	4
Transposon LTR	76,038	0.6	806	66	-
Canonical ncRNA	60,235	0.5	291	26	4