Supplementary Note for:

The Genomic and Phenotypic Diversity of *Schizosaccharomyces pombe*

Daniel C. Jeffares[1*], Charalampos Rallis[1], Adrien Rieux[1], Doug Speed[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[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[2], Francois Balloux[2], Richard Durbin[10], Jürg Bähler[1*]

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.
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.
12. 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.
†Current Address: Ontario Institute for Cancer Research, Toronto, Canada.
*Correspondence to: d.jeffares@ucl.ac.uk and j.bahler@ucl.ac.uk

Contains:
Supplementary Note
References

Supplementary Note

1. Population structure.

To determine whether there were discrete populations in our strain collection, we employed the *Admixture* program[1,2] with the 752 unlinked SNPs, and the non-redundant 57 strains, using K values (the predetermined number of populations) from 2 to 20, running each k value in triplicate. The cross-validation error was lowest with values 2-5, suggesting at most 5 populations (data not shown). However, these groups did not coincide well with the geographic groups (Supplementary Fig. 2a). The five *Admixture*-defined populations were similar to the groupings defined by principal component analysis (Supplementary Fig. 2b).

We next employed *ChromoPainter* and *fineSTRUCTURE*[3], which model the sharing of haplotypes. 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. This analysis also indicated that genetically-defined population groups do not coincide well with geographic groups (Supplementary Fig. 2c).

To confirm this finding with a simple metric, we used all SNPs to estimate F_{ST} for all pairwise combinations of populations, including Europe, Asia, Africa and the Americas. In support of the analysis from *Admixture*, PCA and *fineSTRUCTURE*, values are relatively low, as below.

F_{ST}	Europe	Asia	America
Africa	0.000	0.099	0.178
Europe	-	0.260	0.258
Asia	-	-	0.175

However, defining groups of strains according to their SNP variants with *Admixture* (using the optimal value of k = 5) produced much higher F_{ST} values, as below (mean pairwise F_{ST} = 0.40). These results indicate that highly differentiated populations are present, and that some strains have been displaced sufficiently in recent times to allow little gene flow. For example, in the projection of first two principal components considering their genetic profile (Fig. 1b) African strains (pink) cluster with either European (green) or South American strains (red).

F_{ST}	Group 2	Group 3	Group 4	Group 5
Group 1	0.43	0.22	0.40	0.52
Group 2	-	0.23	0.27	0.59
Group 3	-	-	0.29	0.47
Group 4	-	-	-	0.57
Group 5	-	-	-	-

Finally, the principal component projection and *fineSTRUCTURE* analysis suggested that haplotypes (alleles) had been shared between all populations by recombination. This contrasts with budding yeast, where it is reported that some lineages (‘clean lines’) share very few haplotypes[4]. To test explicitly whether this was the case for the *S. pombe* collection, we examined whether phylogenetic trees were similar across 100 different regions of the genome for the non-redundant set of 57 strains. This showed that no clades of the tree were well supported by a large percentage of regions, so there were no ‘clean lines’ (Supplementary Fig. 2d).

2. Estimates of the TMRCA

There are two main caveats to our mitochondrial estimate of the time to the most recent common ancestor (TMRCA). First, we used only the mitochondrial genome, that could have been subject to a recent selective sweep, so would have a more recent coalescence time than the remainder of the genome. Second, only 81 of the 161 strains had a reliable collection date, so it is possible that some strains exist with an older TMRCA.

The first possibility appears unlikely, because the mitochondrial genome is not significantly lower in diversity from the remainder of the genome (median for all 1 kb windows $\pi_{mito} = 2.7x10^{-3}$, $\pi_{nuc} = 2.6x10^{-3}$, Mann-Whitney test P = 0.74). To test this possibility more rigorously, we used the ACG software[5] which can estimate the TMRCA for recombining genome data. Because the TMRCA in regions of the genome is determined using segregating sites, it is not independent from the degree of background selection. The major factor influencing the extent of purifying selection is the density of protein-coding genes (see main text). Therefore, to select regions that would experience similar background selection to the mitochondrial genome, we chose 160 mitochondria-sized (20 kb) regions of the nuclear genome that were between 50% and 60% exon density, close to the 57% exon-density of the *S. pombe* mitochondria. These regions were evenly distributed throughout chromosomes 1-3. We estimated the TMRCA (in substitutions per site) for each of these regions with AGC, and produced an estimate from the mitochondrial genome using the same method. The mitochondrial estimate (0.0030 subs/site) was close to the mean of nuclear regions (0.0033).

To investigate the second possibility (that some undated strains exist with an older TMRCA), we ran *BEAST* with the alignment of all mitochondrial genomes, using the same parameters and number of iterations as the initial TMRCA estimate, except that: a) we assume a strict clock, and b) we estimate the dates of the undated strains by sampling the age in a uniform distribution. The 95% *Highest Posterior Density* (*HPD*) intervals of the age of the TMRCA of all strains overlapped with the TMRCA of the 81 dated strains (data not shown). We consider this tree to be robust because the deepest three nodes (and most others) have a posterior probability >0.95.

3. Analysis of diversity in long non-coding RNAs.

The analysis of SNP diversity (θ_w) showed that exons contained the lowest diversity, followed by 5'- and 3'-UTRs and introns (Fig. 3b). All these groups showed significantly lower diversity than four-fold degenerate sites (4FD sites). Non-coding RNAs (ncRNAs), un-annotated ('intergenic') regions and LTRs showed higher diversity not significantly different from 4FD sites. Watterson's θ was calculated using only sites that fell exclusively within each annotation class for each 100^{th} of the genome (each 126 kb window). θ could be estimated because our SNP-calling methods identified all callable sites, polymorphic or not. Hence we calculated θ per callable site.

A limitation of θ to detect the effects of purifying selection is that our power to detect segregating sites will not be the same in all regions of the genome. Low complexity regions, such as introns for example, contain fewer 'callable' sites than exons, and so we may record fewer segregating sites. Allele frequencies will be less subject to this issue because the complexity will be the same (or very similar) for all strains, so frequencies should be correctly determined.

The expectation for sites under stronger purifying selection than a neutral standard is an excess of rare allele frequency variants. Due to the linkage of variants, comparing raw rare allele frequencies could inflate P-values. Therefore, to assess relative levels of purifying selection in annotated regions of the genome, we used the same 100 windows of the genome (each 126 kb), and calculated the median allele frequency per window as a summary statistic

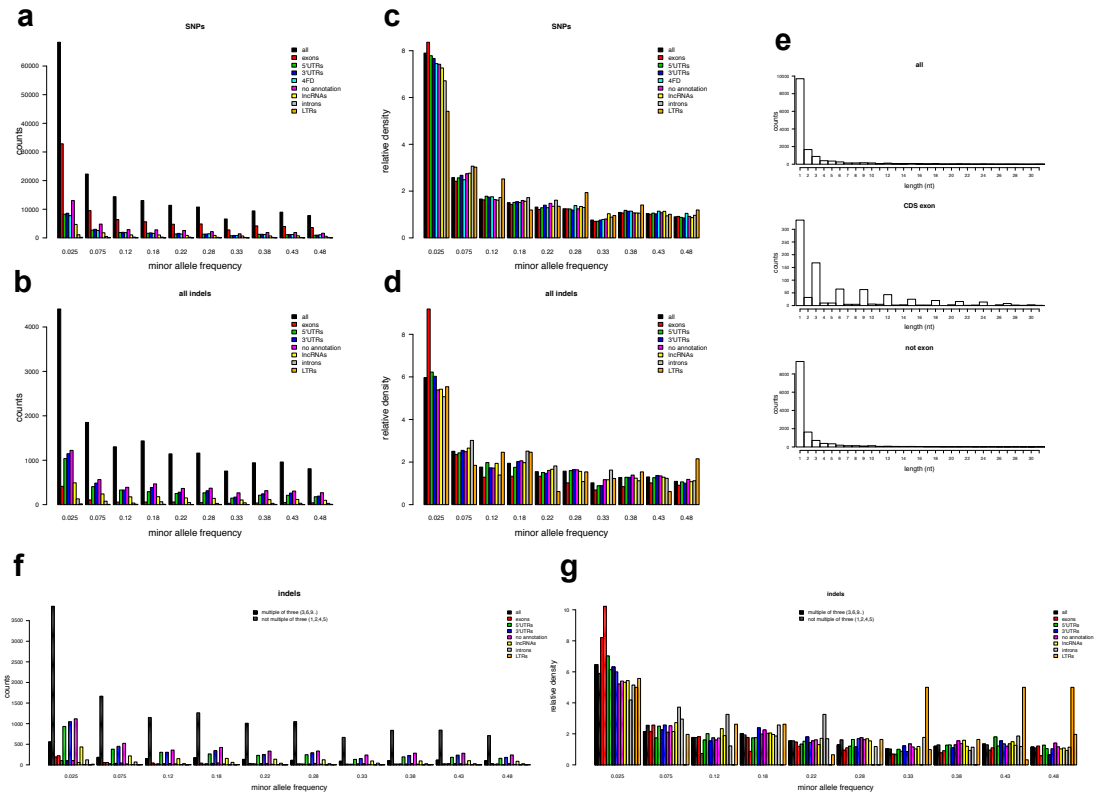
(for variants with exactly one annotation). Using SNP data, both 5'- and 3'-UTRs showed significantly lower median allele frequency than 4FD sites, but again lncRNAs are not lower than 4FD sites (Supplementary Fig. 4a). We obtained similar result with indels, using un-annotated regions as our neutral proxy, except that 3'-UTRs are not significant (Supplementary Fig. 4b).

It is possible that a subset of the non-canonical lncRNAs is subject to purifying selection. To investigate this possibility, we used the expression levels of lncRNAs that have been quantified in copies per cell (CPC) in two physiological states to define subsets of lncRNAs. Our expectation is that more highly expressed lncRNAs may be subject to stronger purifying selection. We divided lncRNAs into five categories based on the maximum CPC obtained in either state (first 5^{th} of expression, second 5^{th} and so on). It is also possible that a subset of 4FD sites would be a more appropriate neutral proxy (since some sites may be subject to weak selection). To test this we divided protein-coding RNAs into 10 expression-fractions using the same data, and for each of 100 windows of the genome, calculated SNP diversity (θ_w) from these 4FD site and lncRNA fractions.

There was little difference between the 4FD site fractions, validating our use of all 4FD sites as a neutral proxy ((ig. 4c). However, we observed that only the most highly-expressed 5^{th} of the lncRNAs are significantly less diverse than the neutral proxy (Supplementary Fig. 4c), suggesting that these lncRNAs are subject to purifying selection.

To confirm this, we calculated the median minor allele frequency using SNPs and indels, using for each of the 100 genomic windows the same 4FD and lncRNA fractions. Again, there was little difference between high and low expression 4FD sites (Supplementary Fig. 4d). Comparing the lncRNA statistics against a neutral proxy again showed that only the most highly expressed 5^{th} of lncRNAs were subject to purifying selection (Supplementary Fig. 4d,e). These lncRNAs are estimated to be expressed at 0.41-1300 CPC [6].

In summary, analysis of SNP θ_w, and SNP and indel median allele frequencies suggests that exons, 5'- and 3'-UTRs are subject to stronger purifying selection than the 4FD or un-annotated region neutral proxies. There was no evidence that lncRNAs as a class were subject to purifying selection. However, a subset of the 20% most highly expressed lncRNAs showed consistent signals of purifying selection using all these three parameters. The majority of these conclusions are also supported by the excess of rare alleles in the raw minor allele frequency spectrum (below). In this figure we show **a, b,** Raw counts of MAF for SNPs show that exons contain a large proportion of the SNPs, but a much smaller proportion of the indels. Indels are much more frequent in UTRs, un-annotated regions and lncRNAs. **c, d,** Relative frequencies of SNP and indel MAF. After all the distribution of all SNPs (black) categories are sorted from according to their preference for rare alleles in SNPs. Exons and UTRs show a stronger bias to rare SNPs than four-fold degenerate sites, whereas un-annotated regions, introns, and lncRNAs do not. This supports the conclusion from θ and median MAF that purifying selection is dominated by exons and UTRs. Indels show an even stronger bias to rare alleles in exons, consistent with many being strongly deleterious. The indel length distribution shows that the majority of indels are short (the 95^{th} percentile is 12nt), that indels in exons (middle panel) are strongly biased to multiple-of-three lengths, which includes many more long indels than in other locations of the genome (the 95^{th} percentile is 30nt). No multiple-of-three bias is observable in non-exonic regions (lower panel). **f,g,** Because of this length bias, raw indel counts are mainly described by short indels UTRs, un-annotated regions and lncRNAs. The strongest bias to rare alleles is in non-multiple-of-three indels in exons.



4. Recombination and DSB hotspots.

We obtained genome-wide DSB rates (data set S1) from Cromie et al.[7]. We processed this data by calculating the median signal for all 14 probes from a 7 probe window, pooling experiment 1 and experiment 2 for the 5 h time point, and the median signal for all 7 probes for the 0 h time point. We then used the ratio median 5 h/median 0 h. The average rate (per 1 kb window) was correlated with the average historic recombination rate (Spearman rank ρ = 0.25, P = $7x10^{-17}$). If we define DSB and historic recombination hotspots as the 1% of 1 kb windows with the highest rates, then 62 of the 118 recombination hotspots are in DSB hotspots. There is a weak, but significant correlation between the average recombination rate (LDU/Mb) and the count of recombination events in 54 segregants of the cross between JB50 and JB759 (both calculated over 100 kb) (Pearson test r = 0.20, P = 0.023) (unpublished data, Mathieu Clement-Ziza and Andreas Beyer).

5. GWAS

The hotspot from a previous study.
The vertical orange bar in Fig. 4b (lower panel) indicates the position of the *swc5* gene (SPCC576.13), which is implicated as causal for the major hotspot in a previous study that has pleiotropic effects on gene expression[8]. Our analysis does not call the frame-shifting indel in *swc5*, but we do observe 23 variants that are significantly associated with traits in the 10kb around this gene (see table below). This includes three traits, consistent with the *swc5* variant being pleiotropic.

chr	position	mixed model P-value	trait	type

III	2097636	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2097898	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2098270	1.83E-06	smgrowth.MgCl2.0.1.M	snp
III	2098378	3.27E-06	smgrowth.MgCl2.0.1.M	snp
III	2098907	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2098940	1.83E-06	smgrowth.MgCl2.0.1.M	snp
III	2099316	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2099368	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2099511	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2099526	3.53E-06	shape2.KL.Predicted.Tshape	snp
III	2099714	3.53E-06	shape2.KL.Predicted.Tshape	snp
III	2100595	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2100959	3.07E-06	wb.NiCl20.75mM.Efficiency	snp
III	2101069	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2101074	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2101157	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2102584	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2103803	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2103938	2.10E-08	shape2.KL.Predicted.Tshape	snp
III	2104066	2.10E-08	shape2.KL.Predicted.Tshape	snp
III	2105767	3.76E-08	shape2.KL.Predicted.Tshape	snp
III	2106879	3.27E-06	smgrowth.MgCl2.0.1.M	snp
III	2107041	3.76E-08	shape2.KL.Predicted.Tshape	snp

Associated variants that were rare or not present in the set of 57 non-clonal strains.
The variants for GWAS were filtered to have minor allele count ≥ 5 in the entire collection of 161 strains (108,453 SNPs and 8417 indels). This selected 8740 SNPs that have a minor allele count of < 2 in the non-clonal 57 strains (8% of the SNPs used the GWAS). Only two of these produced significant associations from the mixed model GWAS. Only 17 of the 1239 associated SNPs from the GWAS (1.3%) had a minor allele count in the non-clonal 57 set that was less than 5. Two of these variants private to strains other than the non-clonal 57 set (not segregating in the non-clonal 57 set). Since neither of these could be validated by regression using the 3 *Admixture*-defined populations, we would not regard these as very strong candidates.

The mixed model accounts for unequal strain relatedness
If the mixed model were not accounting well for population structure, then any excess of associated variants should be most severe traits that are stratified according to the genetic structure of the strains. In such traits, variants that tag populations will co-associate with high/low trait values. We had used the *Admixture* program to cluster strains using the 752 unlinked SNPs as input (which also defined relatedness in Fast-LMM). This approach identified 5 'populations' (Supplementary Fig. 2a), and we used these 5 populations to examine this possibility. For each quantitative trait, we tested for significant differences between the 5 populations applying a Kolmogorov-Smirnov test. Only 19 of the 220 traits were significantly differentiated after Bonferroni correction, showing that traits are usually not stratified by populations (Supplementary Figure 9a). There was no correlation between the number of passing variants and the KS test P-value (Supplementary Figure 9a), consistent

with the mixed model controlling well for population structure. Only 6 traits that were stratified by population contained variants that passed our P-value threshold.

Additionally, we would expect inflation of many P-values above the expected distribution. To examine this possibility, we used genomic inflation factors (GIFs, Supplementary Figure 9b), calculated as median(observed many P-value)/(median expected P-value). With a very large sample size and low LD, the median expected P-value = 0.5. However, with a small sample size GIF varies, as expected (Supplementary Fig. 9b). To examine this variation under a null model, we calculated the median P-value from permutated data (one permutation per trait). Adjusted GIFs calculates as median(observed P-value)/(median permuted P-value) are centred around 1.0. We note that some inflation of genomic-control lambda may be due to multiple causal variants and high LD.

References

1. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).
2. Alexander, D. H., Novembre, J. & Lange, K. Admixture 1.22 Software Manual. (2012).
3. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet* **8,** e1002453 (2012).
4. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458,** 337–341 (2009).
5. O'Fallon, B. D. ACG: rapid inference of population history from recombining nucleotide sequences. *BMC Bioinformatics* **14,** 40 (2013).
6. Marguerat, S. *et al.* Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell* **151,** 671–683 (2012).
7. Cromie, G. A. *et al.* A discrete class of intergenic DNA dictates meiotic DNA break hotspots in fission yeast. *PLoS Genet* **3,** e141 (2007).
8. Clément-Ziza, M. *et al.* Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast. *Mol Syst Biol* **10,** 764 (2014).
9. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
10. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved Heritability Estimation from Genome-wide SNPs. *Am. J. Hum. Genet.* **91,** 1011–1021 (2012).
11. Brown, W. R. A. *et al.* A geographically diverse collection of Schizosaccharomyces pombe isolates shows limited phenotypic variation but extensive karyotypic diversity. *G3* **1,** 615–626 (2011).
12. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21,** 456–463 (2005).