nature genetics

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[10,17], 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[13], Kirsten McLay[13], Nizar Drou[13], William Brown[14], Jacqueline Hayles[6], Rafael E Carazo Salas[8,9], Markus Ralser[7,15,16], Nikolas Maniatis[1], David J Balding[1,2,17], Francois Balloux[1,2], Richard Durbin[10] & Jürg Bähler[1,2]

© 2015 Nature America, Inc. All rights reserved.

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 usefulness of this model, we surveyed the genomic and phenotypic variation in 161 natural isolates. We sequenced the genomes of all strains, finding moderate genetic diversity ($\pi = 3 \times 10^{-3}$ substitutions/site) and weak global population structure. We estimate that dispersal of *S. pombe* began during human antiquity (~340 BCE), and ancestors of these strains reached the Americas at ~1623 CE. We quantified 74 traits, finding 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 the phenotypic variance on average, with indels having larger effects than SNPs. This analysis represents a rich resource to examine genotype-phenotype relationships in a tractable model.

Although 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 (ref. 4). Natural isolates have also been collected from vineyards in Sicily and cachaça (a sugarcane spirit) in Brazil and have been found to contribute to the microbial ecology of kombucha (fermented tea)[1,5,6]. The diverse origins of these natural isolates (**Fig. 1a** and **Supplementary Table 1**) suggest that this yeast species 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 centers 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 of known origin had been associated with human activities, providing little information about the natural environment of the species.

RESULTS

Variation and population structure

We sequenced the genome of all strains to at least 18-fold coverage, with a median of 76-fold coverage. To facilitate the detection of genetic variants, we mapped reads to the reference genome[7]. Mapping was comprehensive and accurate owing to the small, non-repetitive

[1]Department of Genetics, Evolution and Environment, University College London, London, UK. [2]University College London Genetics Institute, University College London, London, UK. [3]Department of Cell Biology, Faculty of Science, 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]Gurdon Institute, University of Cambridge, Cambridge, UK. [10]Wellcome Trust Sanger Institute, Cambridge, UK. [11]Associated Regional and University Pathologists, Inc. University of Utah, Salt Lake City, Utah, USA. [12]Commonwealth Scientific and Industrial Research Organisation (CSIRO) Mathematics, Informatics and Statistics, North Ryde, New South Wales, Australia. [13]Genome Analysis Centre, Norwich, UK. [14]Centre for Genetics and Genomics, University of Nottingham, Nottingham, UK. [15]Cambridge Systems Biology Centre, University of Cambridge, Cambridge, UK. [16]Division of Physiology and Metabolism, Medical Research Council (MRC) National Institute for Medical Research, London, UK. [17]Present addresses: Ontario Institute for Cancer Research, Toronto, Ontario, Canada (J.T.S.) and School of Biosciences and School of Mathematics and Statistics, University of Melbourne, Melbourne, Queensland, Australia (D.J.B.). Correspondence should be addressed to D.C.J. (d.jeffares@ucl.ac.uk) or J.B. (j.bahler@ucl.ac.uk).

Received 8 July 2014; accepted 14 January 2015; published online 9 February 2015; doi:10.1038/ng.3215

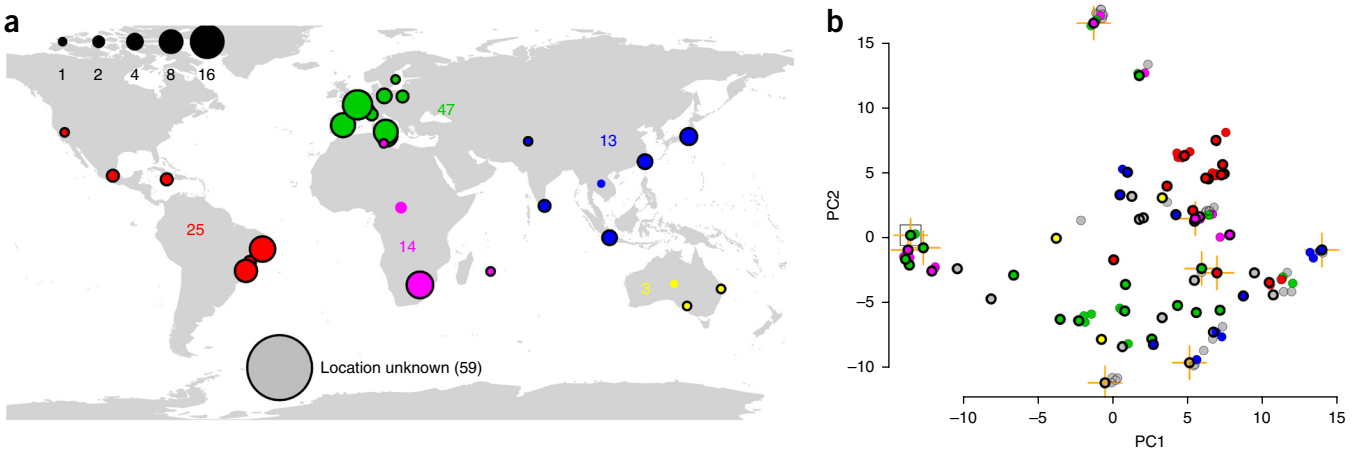

Figure 1 An overview of the strain collection. (**a**) Geographical origins of all 161 strains analyzed. Colored circles indicate the original sources of the strains used in this study, with circle sizes indicating the number of strains obtained from each site (the scale is illustrated with black circles). Strains for which only an approximate source is known (for example, Africa) lack a black border. (**b**) Principal-component projection of the 'drift distance' between strains determined using the 752 unlinked SNPs (Online 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 group of 57 non-redundant strains have a black border; strains known to contain large structural inversions[2] are indicated with an orange cross. PC, principal component.

nature of the 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 SNPs, 14,508 small indels and 1,048 long terminal repeat (LTR) insertions (**Table 1**).

Initial analysis identified 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 differed by ≥1,900 SNPs, which included 99.6% of the SNPs identified for all strains. The average pairwise diversity (π) within these 57 strains was 3.0×10^{-3} substitutions/site (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 (containing both mating types), and all 57 strains were prototrophic (able to grow on the same minimal medium as the 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 were close to linkage equilibrium (pairwise $r^2 < 0.5$) and were distributed relatively evenly across the genome (**Supplementary Fig. 1b**), the use of which better suits population genetic models that assume no linkage between variants. Principal-component analysis with 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 distances being weakly but significantly correlated ($P = 9.9 \times 10^{-5}$; **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 the observed genetic isolation had resulted in any reproductive isolation, we measured spore viability for 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's $r = 0.52$, $P = 6.5 \times 10^{-4}$; **Supplementary Fig. 1d**). This result suggests that these strains have accumulated sufficient genetic differences for reproductive barriers to emerge. Chromosomal rearrangements 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 the unsupervised genetic clustering methods Admixture[14] and fineSTRUCTURE[15], which do not take into account the geographical origin of the strains, to uncover any genetically differentiated populations. Both clustering methods identified between two and five populations that were consistent with those identified by principal-component analysis (**Supplementary Fig. 2a–c**). These results and further phylogenetic analysis showed that these groups were interbreeding populations, rather than clonally isolated lineages (**Supplementary Fig. 2d**). The F_{ST} values (representing the proportion of between-population genetic variance) for

Table 1 Genetic variation discovered in *S. pombe* strains

Annotation	Bases[a]	Genome (%)[a]	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	–
Nonsynonymous, frameshift	–	–	31,441	453	–
Pseudogene	38,896	0.3	254	19	0
Stop gain or loss	–	–	230	–	–
Start gain or loss	–	–	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 noncoding RNA	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 noncoding RNA	60,235	0.5	291	26	4

Variant counts that are enriched (above what is expected for the percentage of the genome) are shown in bold.
[a]The number of bases and percentage of nucleotides annotated refers to the reference genome.

© 2015 Nature America, Inc. All rights reserved.

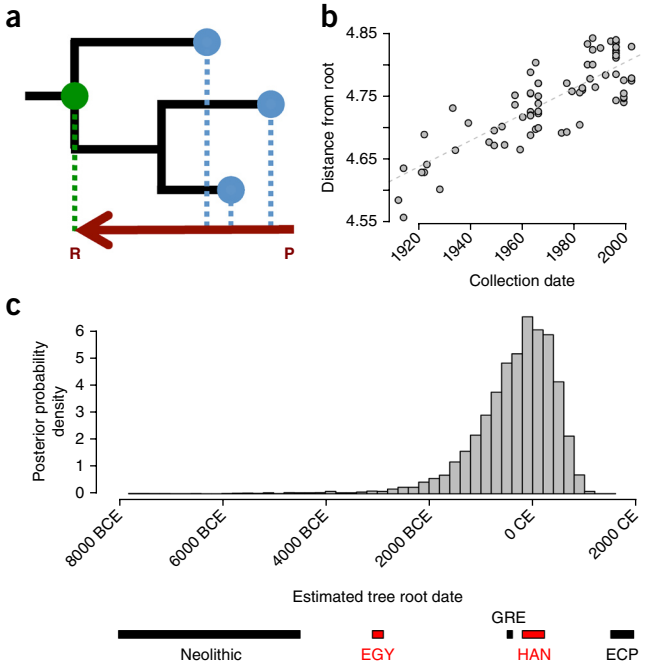

Figure 2 Recent dispersal of *S. pombe*. (**a**) Calibration of tree nodes using dated tips. With a collection of sequences sampled over various times (blue dots) up until the present day (P), we can jointly estimate the phylogenetic tree topology (black branches), the rate of evolution and the age of any node in the tree, including the root representing the most recent common ancestor of all strains (R; green dot). (**b**) Root-to-tip distances (mutations/site $\times 10^{-3}$) correlate with collection date ($P < 1 \times 10^{-16}$), indicating that the data have reasonable predictive power. Distances were estimated using BEAST[35] from mitochondrial data for the 81 strains where collection dates were available; statistical details are provided in the Online Methods. The gray line shows the linear model. (**c**) Historical context of dispersal. The posterior probability distribution for time to the most recent common ancestor 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 (500 BCE–400 BCE); EGY, First Dynasty of ancient Egypt (3190 BCE–2800 BCE); Neolithic, Neolithic era (10,00 BCE–4,500 BCE).

the five-population clustering ranged between 0.22 and 0.59 (mean of 0.40) for different pairwise comparisons, indicating considerable genetic differences among these five connected clusters.

Dating the global dispersal of *S. pombe*

Although *S. pombe* now appears to be globally distributed, we have no ecological or historical 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 (the most recent common ancestor of all strains) (**Fig. 2a**). Modeling of the evolutionary rate showed that our data had predictive power (**Fig. 2b**), and we estimated the ancestor of all strains to have lived ~2,300 years ago (~340 BCE; **Fig. 2c**). We deduced a similar timeline from the nuclear genome sequences (**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 8,000–10,000 years ago, consistent with a Neolithic expansion[17]. Furthermore, our analysis provided 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 there, which began in 1492 CE. Notably, isolates from the Americas also showed the highest genetic similarity (**Fig. 1b** and **Supplementary Note**). Taken 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 *U* test, $P = 1.5 \times 10^{-21}$ 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% versus 0.2% in the whole genome), have 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 (measuring nucleotide diversity) for the following annotation classes (**Fig. 3b**): protein-coding exons, introns, canonical RNAs (rRNAs, tRNAs, small nucleolar and small nuclear RNAs), long noncoding RNAs (lncRNAs), UTRs (of protein-coding transcripts) and the 15% of the genome not annotated as any of the preceding. Within exons, we calculated θ for one-fold-degenerate sites (where all changes to the DNA sequence lead to changes in the protein sequence) and four-fold-degenerate sites (where all changes to the DNA sequence result in the same protein sequence). Although polymorphisms at four-fold-degenerate 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 four-fold-degenerate sites, suggesting substantial evolutionary selection at post-transcriptional levels of gene regulation. Analysis of SNP and indel median minor allele frequencies (MAFs) within windows showed consistent results (**Supplementary Fig. 4a**,**b**). Although lncRNAs appeared to be subject to little or no purifying selection overall, further analyses showed that the 20% of lncRNAs that were most highly expressed were subject to detectable purifying selection (**Supplementary Fig. 4c–e**). These findings indicate that purifying selection is dominated by protein-coding genes, including their UTRs. As a consequence, we would expect fewer genetic variants to remain in gene-dense regions. Consistent with this idea, 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 has several hundred solo LTR fragments that indicate the sites of previous insertions. These elements are transcribed at low levels[20] and thus may be actively propagating. To examine this possibility, we

© 2015 Nature America, Inc. All rights reserved.

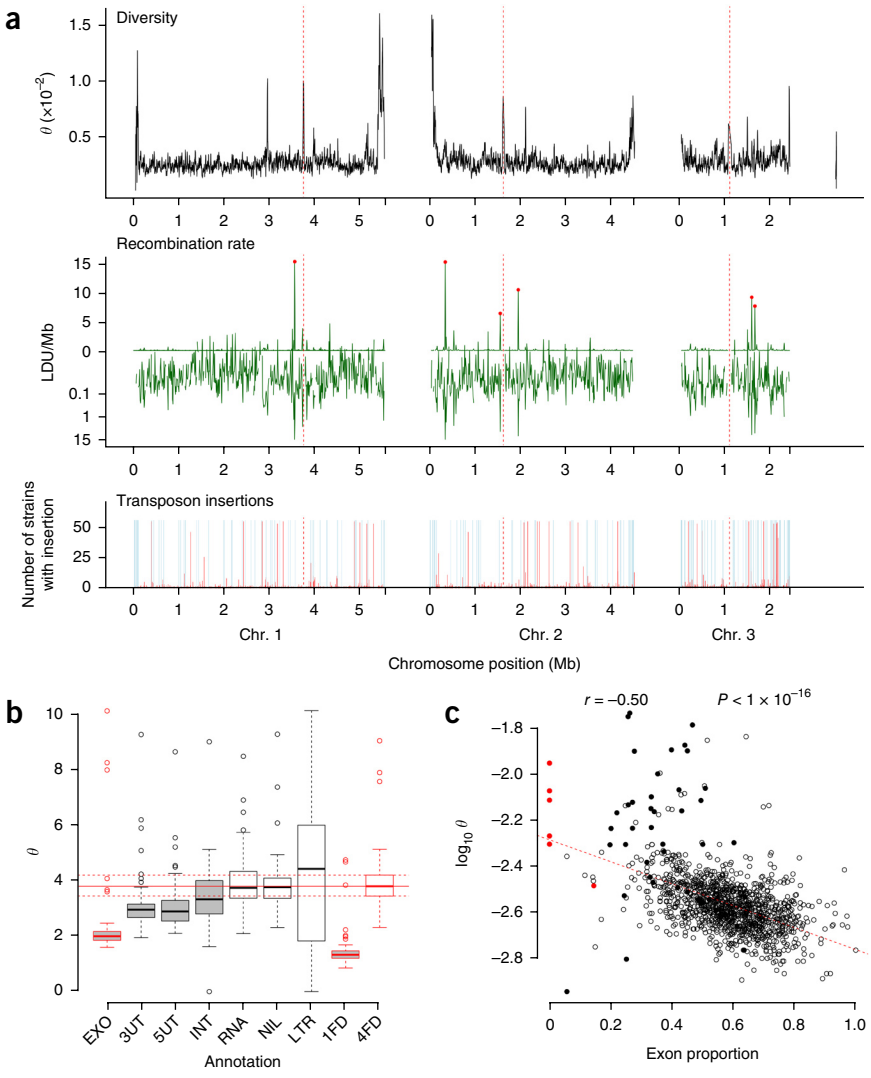

Figure 3 Relationships between genetic diversity and genome function. (**a**) Main features of diversity in the genome, with chromosome scale on the *x* axis and the mitochondrial genome on the right edge. Top, diversity (Watterson's θ) calculated using SNPs. Middle, recombination rate (scale, LDU/Mb × 10^{-3} above the *x* axis and log(1 + LDU/Mb) below the *x* axis). The six major recombination hotspots are indicated with red dots. Bottom, sites of Tf family LTR insertion (insertions present in all strains are shown in light blue) in the group of 57 non-clonal strains. (**b**) Diversity described by genome annotation. Distribution of Watterson's θ values for each centile of the genome, using only sites annotated as exons (EXO), 5′ and 3′ UTRs (5UT and 3UT), introns (INT), lncRNAs (RNA), unannotated regions (NIL), LTRs of Tf2 family transposons (LTR), and onefold-degenerate (1FD) and fourfold-degenerate (4FD) sites in exons. Protein-coding categories have red borders. The horizontal red lines correspond to the median and interquartile range for fourfold-degenerate sites; annotation classes with diversity significantly lower than the diversity for this proxy for neutral sites are shaded gray. One-sided paired Mann-Whitney *U* test *P* values in comparison to the fourfold-degenerate sites were as follows: exons, UTRs and onefold-degenerate sites, $P < 2 \times 10^{-16}$; introns, $P = 1 \times 10^{-6}$; lncRNAs, unannotated regions and LTRs, $P > 0.05$ (whiskers define the most extreme data points up to 1.5 times the interquartile range). (**c**) Diversity is negatively correlated with exon density. Diversity (θ) is plotted against the proportion of each window annotated as protein-coding exons, determined for 10-kb genomic windows. The Spearman's rank correlation and significance are shown above. Filled red circles, centromeric regions; filled black circles, telomeric regions (terminal 100 kb).

searched for new insertions of Tf elements in the 56 non-clonal strains and determined which reference LTRs were present in the other non-clonal strains. We located 1,048 LTR insertions, of which 78% were not present in the reference genome. Consistent with previous studies showing that Tf element insertions are targeted to RNA polymerase II (Pol II)-bound 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 analyzed Tf element insertions within intergenic regions containing one promoter and one terminator, as these insertions allowed 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 the level of gene expression is a main determinant for Tf element insertion. Insertions were also enriched upstream of genes without introns, 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 predictions of protein-coding genes for each strain from the assembly and attempted to locate similar genes in the reference strain genome. The strains were highly similar in their gene content; for example, 95% of the predicted encoded peptides from the divergent strain JB758 could be aligned to a reference protein with >95% identity. Curation identified only 17 putative new proteins, including 9 with strong supporting evidence (**Supplementary Table 3**). The majority of these new proteins were most similar to the products of genes from *Ascomycete* fungi, including 12 for which we could identify orthologs in related *Schizosaccharomyces* species by BLASTP (*e* value < 1×10^{-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.

© 2015 Nature America, Inc. All rights reserved.

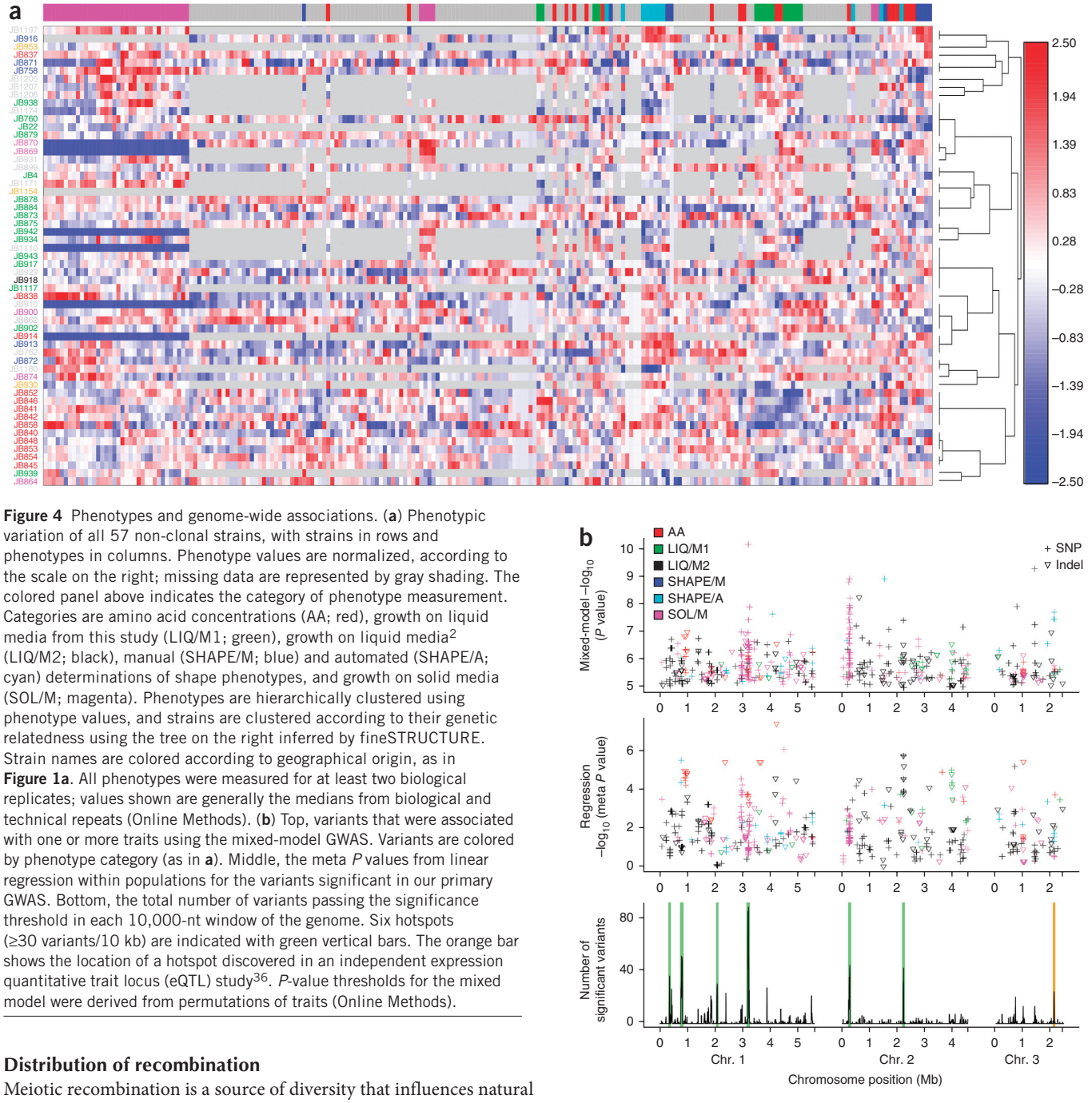

Figure 4 Phenotypes and genome-wide associations. (**a**) Phenotypic variation of all 57 non-clonal strains, with strains in rows and phenotypes in columns. Phenotype values are normalized, according to the scale on the right; missing data are represented by gray shading. The colored panel above indicates 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[2] (LIQ/M2; black), manual (SHAPE/M; blue) and automated (SHAPE/A; cyan) determinations of shape phenotypes, and growth on solid media (SOL/M; magenta). Phenotypes are hierarchically clustered using phenotype values, and strains are clustered according to their genetic relatedness using the tree on the right inferred by fineSTRUCTURE. Strain names are colored according to geographical origin, as in **Figure 1a**. All phenotypes were measured for at least two biological replicates; values shown are generally the medians from biological and technical repeats (Online Methods). (**b**) Top, variants that were associated with one or more traits using the mixed-model GWAS. Variants are colored by phenotype category (as in **a**). Middle, the meta *P* values from linear regression within populations for the variants significant in our primary GWAS. Bottom, the total number of variants passing the significance threshold in each 10,000-nt window of the 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 expression quantitative trait locus (eQTL) study[36]. *P*-value thresholds for the mixed model were derived from permutations of traits (Online Methods).

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 historical recombination rate by constructing genetic maps with distances in linkage disequilibrium units (LDUs)[29]. The estimated recombination rate was zero for genomic regions spanned by 87% of the SNPs and was distributed log normally within the 13% of sites showing recombination (**Supplementary Fig. 6a**). Six regions with very high historical recombination rates were evident (rates above the 99.99th percentile; **Fig. 3a**). These hotspots showed a weak relationship with regions of high DSB activity (Spearman's rank $\rho = 0.25$, $P = 5.2 \times 10^{-16}$), 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's $\rho = 0.43$, $P = 3.2 \times 10^{-57}$) and was primarily located away from genes (**Supplementary Fig. 6b,c**). For example, exons cover 57% of the genome, but only 26% of the 1,000 sites with the highest recombination were in exons. The result of the low-recombination regions is that, on average, LD (r^2) declines to 50% within 21 kb (**Supplementary Fig. 6d**). Hence, *S. pombe* has long haplotypes in comparison to eukaryotes of similar genome size and gene density; for example, LD in the budding yeasts *S. cerevisiae* and *Saccharomyces paradoxus* declines to 50% within 3–11 kb and 9 kb, respectively[8,9].

Phenotypic variation and genome-wide association studies

Model organisms have been used extensively to describe the complex genetics of quantitative traits[30,31], a task that is far more difficult in

© 2015 Nature America, Inc. All rights reserved.

less tractable species such as humans. It was clear that our collection contained variation in quantitative traits, both from previous studies[1,2] and our observation that some strains showed differences in cell shape and size (**Supplementary Fig. 7**). To extend these data, we measured 74 quantitative traits using 5 methods selected to sample a large variety of different phenotypes: (i) manual and (ii) automated measurements of cell shape and size, (iii) multiple growth parameters in minimal and rich liquid media, (iv) colony sizes on solid media under 42 different nutrient, drug and environmental conditions and (v) mass spectrometry measurements of intracellular amino acid concentrations. In combination with previous data[2], we analyzed 9,383 measurements for 223 phenotypes (an average of 164 values per strain) (**Fig. 4a** and **Supplementary Table 4**).

To assess the feasibility of using these data for GWAS, we estimated the heritability of each of these phenotypes using 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 for phenotypes for which data were gathered using all methods ($P < 0.05/220$; **Supplementary Fig. 8a** and **Supplementary Table 5**). Amino acid concentrations were among 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 trait measurements from biological and technical replicates also showed that the availability of data from replicates 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 according to 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 value among the five populations defined by Admixture. Only 19 of the 223 quantitative traits were significantly differentiated ($P < 0.05/220$), showing that traits are usually not stratified by population (**Supplementary Fig. 9a**).

Because our traits were highly heritable and infrequently stratified by population, we applied GWAS to search for the genetic variants associated with each of 223 quantitative traits. We used a mixed model[34], including 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 significance thresholds with a 5% family-wise error rate for each trait, we discovered 1,419 variants that were significantly associated with at least one phenotype (1,239 SNPs and 180 indels; **Fig. 4b** and **Supplementary Table 6**). Genomic inflation factors (the median of the observed test statistic divided by the median expected test statistic) indicated that the mixed model was accounting well for unequal strain relatedness (**Supplementary Fig. 9a**,**b**). As an additional critical test of these associations, we divided the 57 non-clonal strains into 3 subpopulations (with 12, 26 and 17 members, defined by Admixture[14]) and examined each of the 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 least one additional subpopulation ($P < 0.05$; **Fig. 4b** and **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 the UTRs of coding transcripts (**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 to be the most likely candidates for causal variants. These 89 variants (72 SNPs and 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 would have power to detect only variants of large effect. As for any GWAS, although estimates are globally unbiased, the largest estimates are likely to reflect a combination of genetic and stochastic effects and so tend to overestimate the true genetic variance explained, a bias known as winner's curse. In this study, the stochastic component of traits was well controlled by the use of replicate measurements (**Supplementary Fig. 8b**), which mitigates such bias.

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

The distribution of variants passing the significance threshold included six hotspots that harbored multiple variants associated with several different phenotypes (**Fig. 4b**). The most prominent of these hotspots contained 89 variants associated with 6 traits (**Supplementary Fig. 10a**), including the 3 most significant variants (3 SNPs, all in perfect LD with each other, all with $P = 7 \times 10^{-11}$). These polymorphisms were associated with growth in $MgCl_2$ and fell 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 F_1 progeny in the presence and absence of $MgCl_2$. Sequencing of this pool showed a bias for 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 toward identifying the gene(s) affected by these SNPs, we compared the growth of the standard laboratory strain to that of 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*.

DISCUSSION

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.

URLs. Gene List Analyzer, http://128.40.79.33/cgi-bin/GLA/GLA_input; FlowJo, http://www.flowjo.com/.

METHODS

Methods and any associated references are available in the online version of the paper.

© 2015 Nature America, Inc. All rights reserved.

Accession codes. Sequence data are archived in the European Nucleotide Archive under study accessions PRJEB2733 and PRJEB6284. All SNPs and indels were submitted to NCBI dbSNP under accessions 974514578–974688138 (SNPs) and 974702618–974688139 (indels).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank L. Clissold, H. Musk, D. Baker and R. Davey for their contributions to sequencing, H. Levin for discussions about transposons, and J. Mata and S. Marguerat for comments on the manuscript. This work was supported by a Wellcome Trust Senior Investigator Award to J.B. (grant 095598/Z/11/Z), by the Wellcome Trust to S.B., T.K., J.T.S. and R.D., 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 A.R. and F.B., by UK Medical Research Council grant G0901388 to D.S. and D.J.B., by a Cancer Research UK Postdoctoral Fellowship to T.M.K.C., by an ERC Starting Grant (SYSGRO) to R.E.C.S., a Wellcome Trust PhD studentship to J.L.D.L. and BBSRC grant BB/K006320/1 to R.E.C.S. and A.C., by a Wellcome Trust grant (RG 093735/Z/10/Z) and ERC Starting Grant 260809 to M.R. (M.R. is a Wellcome Trust Research Career Development and Wellcome-Beit Prize Fellow), by Czech Science Foundation grant P305/12/P040 and Charles University grant UNCE 204013 to M.P. and by Cancer Research UK to L.J. and J.H.

AUTHOR CONTRIBUTIONS

D.C.J. coordinated all analyses, isolated DNA for sequencing, analyzed and filtered SNP calls, conducted diversity analysis and GWAS and drafted the manuscript. C.R. produced phenotype data for growth on various solid media and growth rates in liquid media. A.R. conducted analysis of dating using mitochondrial data. D.S. conducted GWAS. M.P. analyzed all phenotype data. T.M. identified LTR transposon insertions and analyzed transposon insertion data. F.X.M. conducted crosses for the analysis of spore viability. Z.I. produced indel calls with Cortex. W.L. conducted analysis of recombination rate, LD decay and principal-component analysis for distance between strains. T.M.K.C. assisted with phenotype and population analysis. R.P. analyzed Cortex and GATK indel calls. M.M. conducted amino acid profiling. J.L.D.L. and A.C. produced automated measures of cell morphology. S.B. aligned reads and produced GATK SNP calls. G.H. analyzed population structure using fineSTRUCTURE. B.O'F. estimated the time to the most recent common ancestor from the nuclear genome using ACG. T.K. identified LTR transposon insertions. J.T.S. produced *de novo* assemblies. L.B. developed the custom Workspace workflow Spotsizer. B.T. assisted with sequence analysis. D.A.B. assisted with analysis of new genes. T.S. assisted with strain verification. S.C. produced images of wild strains and assisted with strain verification. J.E.E.U.H. assisted with SNP validation. L.v.T. and M.T. assisted with LTR validation. L.J. and J.-J.L. assisted with manual measures of cell morphology and FACS. S.A. produced gene expression data. M.F., K.M. and N.D. assisted with sequencing. W.B. initiated and assisted with strain collection. J.H. coordinated manual measures of cell morphology and FACS. R.E.C.S. coordinated automated measures of cell morphology. M.R. coordinated amino acid profiling. N.M. conducted analysis of recombination and LD and advised on aspects of diversity and GWAS. D.J.B. advised on GWAS. F.B. advised on population structure and supervised A.R. R.D. facilitated sequencing. J.B. contributed to the initiation and development of the project and financed the Bähler laboratory.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Gomes, F.C.O. *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.* **48**, 399–406 (2002).
2. 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).
3. Fawcett, J.A. *et al.* Population genomics of the fission yeast *Schizosaccharomyces pombe*. *PLoS ONE* **9**, e104241 (2014).
4. Osterwalder, A. *Schizosaccharomyces liquefaciens* n.sp., eine gegen freie schweflige Säure widerstandsfähige Gärhefe. *Mitt. Geb. Lebensmittelunters. Hyg.* **15**, 5–28 (1924).
5. Florenzano, G., Balloni, W. & Materassi, R. Contributo alla ecologia dei lieviti *Schizosaccharomyces* sulle uve. *Vitis* **16**, 38–44 (1977).
6. Teoh, A.L., Heard, G. & Cox, J. Yeast ecology of Kombucha fermentation. *Int. J. Food Microbiol.* **95**, 119–126 (2004).
7. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880 (2002).
8. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341 (2009).
9. Schacherer, J., Shapiro, J.A., Ruderfer, D.M. & Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458**, 342–345 (2009).
10. Avelar, A.T., Perfeito, L., Gordo, I. & Godinho Ferreira, M. Genome architecture is a selectable trait that can be maintained by antagonistic pleiotropy. *Nat. Commun.* **4**, 2235 (2013).
11. Seich Al Basatena, N.-K., Hoggart, C.J., Coin, L.J. & O'Reilly, P.F. The effect of genomic inversions on estimation of population genetic parameters from SNP data. *Genetics* **193**, 243–253 (2013).
12. Zanders, S.E. *et al.* Genome rearrangements and pervasive meiotic drive cause hybrid infertility in fission yeast. *eLife* **3**, e02630 (2014).
13. Cromie, G.A. *et al.* Genomic sequence diversity and population structure of *Saccharomyces cerevisiae* assessed by RAD-seq. *G3* **3**, 2163–2171 (2013).
14. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
15. Lawson, D.J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
16. Hornsey, I.S. *A History of Beer and Brewing* (The Royal Society of Chemistry, 2003).
17. Fay, J.C. & Benavides, J.A. Evidence for domesticated and wild populations of *Sacchoromyces cerevisiae*. *PLoS Genet.* **1**, 66–71 (2005).
18. Zhou, T., Gu, W. & Wilke, C.O. Detecting positive and purifying selection at synonymous sites in yeast and worm. *Mol. Biol. Evol.* **27**, 1912–1922 (2010).
19. Bowen, N.J., Jordan, I.K., Epstein, J.A., Wood, V. & Levin, H.L. 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 (2003).
20. Mourier, T. & Willerslev, E. Large-scale transcriptome data reveals transcriptional activity of fission yeast LTR retrotransposons. *BMC Genomics* **11**, 167 (2010).
21. Kwon, E.-J.G. *et al.* Deciphering the transcriptional-regulatory network of flocculation in *Schizosaccharomyces pombe*. *PLoS Genet.* **8**, e1003104 (2012).
22. Guo, Y. & Levin, H.L. High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in *Schizosaccharomyces pombe*. *Genome Res.* **20**, 239–248 (2010).
23. Guo, Y. *et al.* Integration profiling of gene function with dense maps of transposon integration. *Genetics* **195**, 599–609 (2013).
24. Feng, G., Leem, Y.-E. & Levin, H.L. Transposon integration enhances expression of stress response genes. *Nucleic Acids Res.* **41**, 775–789 (2013).
25. Jeffares, D.C., Penkett, C.J. & Bähler, J. Rapidly regulated genes are intron poor. *Trends Genet.* **24**, 375–378 (2008).
26. Chen, D. *et al.* Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell* **14**, 214–229 (2003).
27. Cromie, G.A. *et al.* A discrete class of intergenic DNA dictates meiotic DNA break hotspots in fission yeast. *PLoS Genet.* **3**, e141 (2007).
28. Fowler, K.R., Gutiérrez-Velasco, S., Martín-Castellanos, C. & Smith, G.R. Protein determinants of meiotic DNA break hot spots. *Mol. Cell* **49**, 983–996 (2013).
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* **99**, 2228–2233 (2002).
30. Liti, G. & Louis, E.J. Advances in quantitative trait analysis in yeast. *PLoS Genet.* **8**, e1002912 (2012).
31. Mackay, T.F.C. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.* **15**, 22–33 (2014).
32. 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).
33. Warringer, J. *et al.* Trait variation in yeast is defined by population history. *PLoS Genet.* **7**, e1002111 (2011).
34. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525–526 (2012).
35. Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
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.* **10**, 764 (2014).

© 2015 Nature America, Inc. All rights reserved.

ONLINE METHODS

Sequencing and quality control. All strains are described in **Supplementary Table 1**. Strains were sequenced with either 54-nt or 100-nt paired-end Illumina reads. A summary of read data is provided in **Supplementary Table 7**. To verify that strain identity was correct at various stages in the project, we genotyped 30 SNPs (that would distinguish all 57 non-clonal strains with at least 2 allelic differences) in the 161 extracts used for sequencing, replicate extracts of the 57 non-clonal strains, extracts from stocks obtained directly from stock centers and extracts made from cultures picked from the ROTOR phenotyping plate. Only 2 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 extracts from cultures grown on ROTOR plates were as expected.

Read mapping and SNP and indel calling. Reads were mapped to the *S. 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 the Genome Analysis Toolkit (GATK) IndelRealigner.

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

SNP and indel validation. To estimate the false discovery rate and sensitivity of SNP calling, we sequenced ~20 paired-end shotgun clones from each of 4 strains with increasing genetic distance from the reference strain using 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 Integrated Genomics Viewer (IGV) tool[41]. These windows included 47,619 nt of mappable regions and 182 known SNPs. We found that all of the SNPs were valid, whereas 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 the Illumina read alignments at 100 indels called in the same 4 strains, choosing indels that were dispersed across all chromosomes. Only four of these calls were false positives (4% false discovery rate for calling indels). A total of seven indels corresponded to 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 the soft-clipped, unaligned parts of sequence reads covering the insertion sites to distinguish between independent insertions at closely situated genomic sites, collating 1,474 predicted insertions into 820 insertion events. We assessed the target site duplication sizes from the soft-clipped reads. We used PCR to verify 90 of the RetroSeq predictions: 56 of these produced a product in both the reference and alternate strain, and 80% (45/56) of these confirmed the insertion with high confidence, whereas 93% (52/56) confirmed the insertion with at least medium-level confidence.

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 were identified by only considering LTR insertions between genes arranged in tandem (neighboring genes in the same orientation). Gene features were analyzed by the Gene List Analyzer (F. Schubert, S. Khadayate and J.B., unpublished data). The presence/absence and positions of all LTR insertions are provided in **Supplementary Table 8**. Detailed information about the PCR validation of LTR insertions is provided in **Supplementary Table 9**.

Diversity analysis. Diversity estimates were calculated using Variscan[44]. For analysis with 10-kb windows, we excluded windows with fewer than 1,000 nt of reliably called sites. To compare different genome annotation classes, we used regions that were annotated exclusively as exon, intron, noncoding RNA, etc. Median MAF was calculated for all variants passing the significance threshold in 100 windows (126 kb long) for SNPs and in 50 windows (252 kb) 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 appeared to be a homogenous population in principal-component 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 replicates of the time point at 5 h (14 probes in all), divided by the median signal for probes in the 7-probe window for the time point at 0 h. 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 values were calculated for all pairs of SNPs with MAF > 0.05 up to a distance of 250 kb, using LDMAP (for D')[45] and PLINK (for 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} value, 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 57 non-clonal strains on all SNPs using the recombination rate estimate. When using ChromoPainter, we first ran ten expectation-maximization iterations to infer the 'global mutation' and 'switch rate' parameters, averaged the inferred values for each across chromosomes with weighting by the number of SNPs and then performed a final ChromoPainter run using these weight-averaged values. Isolation by distance was calculated using geoDist values from the SoDA packages in R. See the **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, 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 (ref. 35) on both the five schemes obtained with PartitionFinder and the whole molecule. In the first case, substitution and clock models were unlinked and tree topology was assumed to be the same for the five 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 and stepping-stone sampling methods[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 four 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 4 independent a posteriori combined chains in which samples were drawn every 2,500 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.

Estimation of time to the most recent common ancestor with nuclear DNA. To obtain estimates of the time to the most recent common ancestor 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 length. To ensure 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 the mitochondrial genome. To ease the 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] and 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

© 2015 Nature America, Inc. All rights reserved.

doi:10.1038/ng.3215

except recombination rate, for which we employed an exponential prior with a mean of 100.0 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 (ref. 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 65 bp 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 new genes. To identify protein-coding genes that were present in wild strain(s) but not in the reference strain, 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+ (ref. 55), BLASTP, TBLASTN and BLASTN. Predictions of ≥100 amino acids in length that scored <80% identity in all BLAST searches were chosen as potential new 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 sequence using Emboss cons and used this consensus sequence as a query for BLASTP searches against the *S. pombe* reference protein data set and the NCBI nr protein data set. We excluded potential new genes whose best BLAST hit in the nr data set was from *S. pombe* or the phage Φx174 (likely contamination). We retained the 17 potential new genes where the ratio of the nr BLASTP bit score to the *S. pombe* bit score was >1. To examine the conservation of the 17 potential new genes in other *Schizosaccharomyces* yeasts, we used each predicted protein (from each *S. pombe* strain) for the 17 putative most promising new genes to query the predicted proteins of *S. cryophilus*, *Schizosaccharomyces japonicus* and *Schizosaccharomyces octosporus* using BLASTP, accepting BLAST hits with *e* value < 1×10^{-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. These are the phenotypes with the prefix "aaconc" in **Supplementary Tables 4** and **5**.

Triplicate cultures (1.6 ml) of each strain were grown for 8 h, cell extracts were prepared with 80 °C boiling ethanol, extracts were cleared of insoluble material by centrifugation and the supernatant was collected for liquid chromatography and tandem mass spectrometry analysis. Samples were analyzed on a liquid chromatography (Agilent 1290 Infinity) and tandem mass spectrometry (Agilent 6460) system. Amino acids were separated by hydrophilic interaction chromatography with 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]. The average of the amino acid values from the triplicates was used for further analysis. For quality control, all values with a coefficient of variation greater than two times the overall coefficient of variation (median) were eliminated. For the 19 amino acids analyzed, the median coefficients of variation were between 0.07 and 0.21 (mean of 0.13).

Growth and stresses on solid media. These are the phenotypes with the prefix "smgrowth" in **Supplementary Tables 4** and **5**.

Strains were arrayed by a ROTOR robot (Singer Instruments) onto solid YES and EMM2 media at a 1,536-spot density, with each strain represented by 4 spots. Edges of plates and various interspersed positions were inoculated with the standard laboratory strain, as well as strains with known sensitivity (*atf1Δ* and *sty1Δ*) or resistance (*pka1Δ*).

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 size was performed using the custom Workspace package with the Spotsizer custom workflow (L.B., M.P., C.R., D.C.J. and Y. Arzhaev *et al.*, unpublished data). Colonies with microbial contaminations and misidentified colonies were discarded. Median colony size for each strain was calculated for each plate and replicate. Conditions or plates showing poor reproducibility were removed from further analysis. Colony size data for strains under each condition were normalized to growth on YES medium and then to the growth of the 972 h^- reference strain under the given condition. Two or more replicate plates were analyzed for 25 of the 43 conditions, and 1 plate was analyzed for all others. Plate values were the median colony size from the four colonies analyzed per strain. The median between-plate Pearson's correlation was 0.95.

Cell growth parameters and kinetics in liquid media. These are the phenotypes with the prefix "lmgrowth" in **Supplementary Tables 4** and **5**.

All 57 non-clonal strains were cultured in a Biolector microfermenter (m2p labs) in 1.5 ml of YES or EMM2 medium (Formedium) using flowerplates from m2p labs for 24 h at 32 °C, measuring light scattering every 10 min. Growth of each strain was repeated at least in duplicate. For each replicate of the optical density data points, we used the R grofit package[59] to determine all growth parameters. Two biological repeats of Biolector cultures were grown per strain. Correlations between biological repeats were typically >0.9, and all were above 0.884. All coefficients of variation (within a strain) were above 0.075 (median for all traits = 0.034).

Manual cell morphology characterization. These are the phenotypes with the prefix "shape1" in **Supplementary Tables 4** and **5**.

Strains were grown on plates with YES medium at 32 °C and allowed to form small colonies. Cells around the edge of at least five colonies were examined using a Zeiss Axioskop microscope with both 20× LD ACROPLAN 0.4 and 50× CF plan 0.55 objectives, and the cell phenotype was described. Using the 50× CF plan 0.55 objective with 2.5× Optivar, a representative colony was photographed with the Sony NEX 5N camera. For cultures in liquid media, strains were grown to mid-log phase and examined using a Zeiss Axioskop 40 microscope with a 63× Plan APOCHROMAT 1.4 oil immersion objective. Cell length and width were 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 DNA content was estimated using FlowJo. Replicate length and width measurements were the median values for at least 34 cells (median of 53), with median coefficients of variation of 0.07 in both cases.

Automated cell morphology analysis. These are the phenotypes with the prefix "shape2" in **Supplementary Tables 4** and **5**.

Cells were grown to mid-log phase in YES medium and imaged using the OperaLX (PerkinElmer) high-throughput microscope at 60× magnification. Images were then automatically preprocessed, segmented and analyzed 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 ten independent repeats. Here only the 405-nm channel and six repeats were needed.

The symmetrized Kullback-Leibler divergence between each strain and the reference strain was used as an additional quantitative trait (shape2.KL.Predicted.* in **Supplementary Table 4**), along with the length, width and ratio of the width of both sides of the cell ('cell asymmetry'). Up to six populations of cells were analyzed per strain. Because 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), with an average of 0.87.

Heritability and genome-wide association studies. We used LDAK[32] to estimate the heritability of all traits. We report values based on quantile-normalized phenotypes (see below), but we also repeat estimates using raw values. Heritabilities estimated with raw values were strongly correlated with those from normally transformed values ($r = 0.69$, $P \leq 2.2 \times 10^{-16}$).

We performed mixed-model association analysis using Fast-LMM[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 fewer than five calls for the minor allele (MAF < 3.1%) and variants that had >5% missing calls. We estimated a trait-specific *P*-value threshold for each trait by permuting trait values for individuals 1,000 times, recording the lowest *P* value from Fast-LMM

© 2015 Nature America, Inc. All rights reserved.

analysis and using the 5% quantile (50th lowest value) as the threshold. Variants passing this threshold therefore had a 5% family-wise error rate. We also performed conditional analysis; for each of the 89 traits with at least one variant significantly associated in the primary mixed-model GWAS, we repeated the analysis, including as a covariate the genotypes for the most significant variant.

The genomic inflation factor (GIF) was calculated as GIF = $\text{median}\left(\chi^2_{\text{observed}}(P)\right)/\text{median}\left(\chi^2_{\text{expected}}(P)\right)$, and the adjusted GIF was calculated as $\text{GIF} = \text{median}\left(\chi^2_{\text{observed}}(P)\right)/\text{median}\left(\chi^2_{\text{permuted}}(P)\right)$, where $\chi^2_{\text{observed}}(P)$ is the statistic corresponding to the observed P value and $\chi^2_{\text{expected}}(P)$ is the statistic expected assuming that P values are distributed uniformly within [0, 1]. Here permuted P values were obtained by permuting trait values once for each of the 223 traits used for the GWAS. The expected median, assuming P values are uniformly distributed, is 0.454.

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

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

Statistics. All statistics were generated with R (ref. 63).

37. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
38. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
39. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
40. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
41. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
42. Keane, T.M., Wong, K. & Adams, D.J. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**, 389–390 (2013).
43. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
44. Hutter, S., Vilella, A.J. & Rozas, J. Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics* **7**, 409 (2006).
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* **23**, 517–519 (2007).
46. 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).
47. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
48. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
49. Lanfear, R., Calcott, B., Ho, S.Y.W. & Guindon, S. Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29**, 1695–1701 (2012).
50. Baele, G., Li, W.L.S., Drummond, A.J., Suchard, M.A. & Lemey, P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Mol. Biol. Evol.* **30**, 239–243 (2013).
51. O'Fallon, B.D. ACG: rapid inference of population history from recombining nucleotide sequences. *BMC Bioinformatics* **14**, 40 (2013).
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.* **10**, 512–526 (1993).
53. Simpson, J.T. & Durbin, R. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Res.* **22**, 549–556 (2012).
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* **7**, 62 (2006).
55. Camacho, C., Coulouris, G. & Avagyan, V. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
56. van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol. Biol.* **804**, 281–295 (2012).
57. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
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.* **78**, 4281–4290 (2006).
59. Kahm, M., Hasenbrink, G., Lichtenberg-Frate, H., Ludwig, J. & Kschischo, M. Grofit: fitting biological growth curves with R. *J. Stat. Softw.* **33**, 1–21 (2010).
60. Sazer, S. & Sherwood, S.W. Mitochondrial growth and DNA synthesis occur in the absence of nuclear DNA replication in fission yeast. *J. Cell Sci.* **97**, 509–516 (1990).
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* **31**, 227–239 (2014).
62. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
63. The R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2013).

© 2015 Nature America, Inc. All rights reserved.