Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast.

October 3, 2016

Authors

Daniel C. Jeffares[1,2], Clemency Jolly[1], Mimoza Hoti[1,], Doug Speed[2], Liam Shaw[1,2], Charalampos Rallis[1,2,3], Francois Balloux[1,2], Christophe Dessimoz[1,4,5,6*], Jürg Bähler[1,2*], Fritz J. Sedlazeck[7*]

Affiliations:

1. Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom.
2. UCL Genetics Institute, University College London, London WC1E 6BT, United Kingdom.
3. Current address: School of Health, Sport and Biosciences, University of East London, London E15 4LZ, United Kingdom.
4. Department of Computer Science, University College London, London WC1E 6BT, United Kingdom.
5. Department of Ecology and Evolution & Center for Integrative Genomics, University of Lausanne, Biophore, 1015 Lausanne, Switzerland.
6. Swiss Institute of Bioinformatics, Biophore, 1015 Lausanne, Switzerland
7. Department of Computer Science, Johns Hopkins University, 21218 Baltimore, USA

* Correspondence should be addressed to C.D. (christophe.dessimoz@unil.ch), J.B. (j.bahler@ucl.ac.uk) or F.J.S. (fritz.sedlazeck@jhu.edu)

ABSTRACT

Large structural variations (SVs) in the genome are harder to identify than smaller genetic variants but may substantially contribute to phenotypic diversity and evolution. Here we analyze the effects of SVs on gene expression, quantitative traits, and intrinsic reproductive isolation in the yeast *Schizosaccharomyces pombe*. We establish a high-quality curated catalog of SVs in the genomes of a worldwide library of *S. pombe* strains, including duplications, deletions, inversions and translocations. We show that copy number variants (CNVs) frequently segregate within closely related clonal populations, are weakly linked to single nucleotide polymorphisms (SNPs), and show other genetic signals consistent with rapid turnover. These transient CNVs produce stoichiometric effects on gene expression both within and outside the duplicated regions. CNVs make substantial contributions to quantitative traits such as cell shape, cell growth under diverse conditions, sugar utilization in winemaking, whereas rearrangements are strongly associated with reproductive isolation. Collectively, these findings have broad implications for evolution and for our understanding of quantitative traits including complex human diseases.

Keywords: structural variants, yeast, copy number variants, reproductive isolation, quantitative genetics, next generation sequencing

A variety of genetic changes can influence the biology of species, including single-nucleotide polymorphisms (SNPs), small insertion-deletion events (indels), transposon insertions and large structural variations. Structural variations (SVs), including deletions, duplications, insertions, inversions and translocations, are the most difficult to type and consequently the least well described.

Nevertheless, it is clear that SVs have strong effects on various biological processes. Copy number variants (CNVs) in particular influence quantitative traits in microbes, plants and animals, including agriculturally important traits and a variety of human diseases[1-5]. Inversions are known to influence reproductive isolation[6-13] and other evolutionary processes such as recombination[8] and hybridization between species[14], with a variety of consequences[15].

We and others have recently begun to develop the fission yeast *Schizosaccharomyces pombe* as a model for population genomics and quantitative trait analysis [6,7,16-18]. This model organism combines the advantages of a small, well-annotated haploid genome[19], abundant tools for genetic manipulation and high-throughput phenotyping[20], and considerable resources of genome-scale and gene-centric data[21-23].

Previous analyses of fission yeast have begun to describe both naturally occurring and engineered inversions and reciprocal translocations[6,7,18]. Given this evidence for SVs and their effects in this model species, we recognized that a systematic survey of SVs would progress our understanding of their biological influence. Here, we utilize the recent availability of 161 fission yeast genomes and extensive data on quantitative traits and reproductive isolation[17] to describe the nature and effects of SVs in *S. pombe*.

We show that SVs have strong effects on various aspects of biology. They contribute an average of 11% of trait variance (the much more abundant SNPs contribute 24% on average), with the largest effects coming from CNVs. We show that CNVs are transient within clonal populations, and are frequently not well tagged by SNPs. We also show that rearrangements (inversions and translocations) contribute to reproductive isolation, whereas CNVs do not.

RESULTS

Genome- and population-wide detection of structural variations

To predict an initial set of SVs, we applied four inference software packages (Delly, Lumpy, Pindel and cn.MOPs)[24-27] to existing short-read data[17], using parameters optimized on simulated data (Methods). We then filtered these initial predictions, accepting SVs detected by at least two callers, to obtain 315 variant calls (141 deletions, 112 duplications, 26 inversions, 36 translocations). We release this pipeline as an open-source tool called SURVIVOR (Methods). To ensure a high specificity, we further filtered the 315 variants by removing SV calls whose breakpoints overlapped with low complexity regions or any that corresponded to previously annotated long terminal repeats (LTRs)[17]. Finally, we manually vetted all the remaining SVs by visual inspection of read alignments in multiple strains for all remaining candidates. This meticulous approach aimed to ensure a high quality call set, to mitigate against the high uncertainty associated with SV calling[25].

This curation produced a set of 113 SVs, comprising 23 deletions, 64 duplications, 11 inversions and 15 translocations **(Figure 1a)**. Reassuringly, when

applying our variant calling methods to an engineered knockout strain, we correctly identified the known deletions and called no false positives. Attempts to validate all rearrangements by PCR and BLAST searches of *de novo* assemblies positively verified 76% of the rearrangements, leaving only a few PCR-intractable variants unverified (see Methods for details).

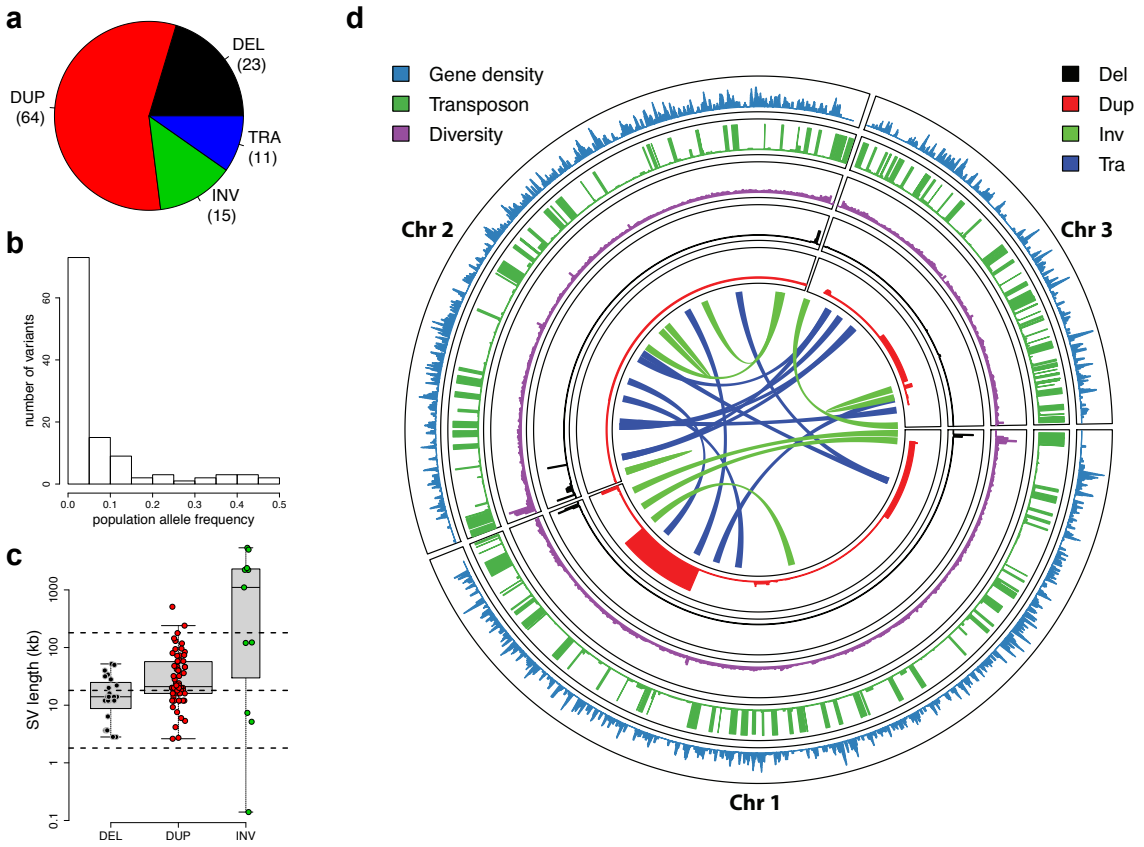


Figure 1. Characteristics of SVs in *S. pombe*. (**a**) Relative proportions of SVs identified. Duplications (DUP) were the most abundant SVs, followed by deletions (DEL), inversions (INV), and translocations (TRA). (**b**) Population allele frequency distribution of SVs, showing the frequencies of less abundant alleles in the population (minor allele frequencies). (**c**) Length distributions of SVs, $\log_{10}$ scale. Deletions were smallest (2.8–52 kb), duplications larger (2.6–510 kb), and inversions often even larger, spanning large portions of chromosomes (0.1 kb–5,374 kb, see (d)). Horizontal dotted lines show the size of chromosome regions that contain an average of 1, 10 and 100 genes in this yeast. (**d**) Locations of SVs on the three chromosomes compared to other genomic features. From outside: density of essential genes, locations of *Tf*-type retrotransposons, diversity (π, average pairwise diversity from SNPs), deletions (black), duplications (red),

and breakpoints of inversions and translocations as curved lines inside the concentric circles (green and blue, respectively). Bar heights for retrotransposons, deletions and duplications are proportional to minor allele frequencies. Diversity and retrotransposon frequencies were calculated from 57 non-clonal strains as described by Jeffares, et al. [17].

Most SVs were present at low frequencies, with 28% discovered in only one of the strains analyzed **(Figure 1b)**. The deletions were generally slightly smaller (median length 14 kb, **Figure 1c)** than duplications (median length of 21 kb), with the largest duplication extending to 510 kb and covering 200 genes (a singleton in strain JB1207/NBRC10570). The majority of CNVs were present in copy numbers varying between zero and sixteen (subsequently we refer to amplifications of two or more copies as 'duplications').

All SVs, particularly deletions and duplications, were biased towards the ends of chromosomes (**Figure 1d, Supplementary Figures 1 and 2**), which are characterized by high genetic diversity, frequent transposon insertions, and a paucity of essential genes[17], similar to *Saccharomyces cerevisiae* and *S. paradoxus*[28,29]. All SVs preferentially occurred in positions of low gene density and were strongly under-enriched in essential genes (**Supplementary Figure 2**).

To describe SVs further, we conducted gene enrichment analysis with the AnGeLi tool (**Supplementary Table 1**), which interrogates gene lists for functional enrichments using multiple qualitative and quantitative information sources[30]. The CNV-overlapping genes were enriched for caffeine/rapamycin induced genes and genes induced during meiosis (P = $4x10^{-7}$ and $1x10^{-5}$, respectively); they also showed lower relative RNA polymerase II

136 occupancy and were less likely to contain genes known to produce abnormal cell
137 phenotypes ($P = 1.8x10^{-5}$ and $3x10^{-5}$, respectively). These analyses are all broadly
138 consistent with a paucity of CNVs in genes that encode essential mitotic functions.
139 Rearrangements disrupted only a few genes and showed no significant enrichments.
140
141 **Duplications are transient within clonal populations**
142 Our previous work identified 25 clusters of near-clonal strains, which differed by <150
143 SNPs within each cluster[17]. We expect that these clusters reflect either repeat depositions
144 of strains differing only at few sites (e.g. mating-type variants of reference strains h^{90} and
145 h^{-} differ by 14 SNPs) or natural populations of strains collected from the same location.
146 Such 'clonal populations' reflect products of mitotic propagation from a very recent
147 common ancestor, without any outbreeding. We therefore expected that SVs should be
148 largely shared within these clonal populations.
149 Surprisingly, our genotype predictions indicated that most SVs present in clonal
150 populations were segregating, i.e. were not fixed within the clonal population (68/95 SVs,
151 72%). Furthermore, we observed instances of the same SVs that were present in two or
152 more different clonal populations that were not fixed within any clonal population. These
153 SVs could be either incorrect allele calls in some strains, or alternatively, recent events
154 that have emerged during mitotic propagation. To distinguish between these two
155 scenarios, we re-examined the read coverage of all 49 CNVs present within at least one
156 clonal population. Since translocations and inversions were more challenging to
157 accurately genotype, we did not re-examine these variants. This analysis verified that 40
158 of these 49 CNVs (37 duplications, 3 deletions) were clearly segregating within at least

one clonal cluster **(Supplementary Figure 3)**. For example, one clonal population of seven closely related strains, collected together in 1966 from grape must in Sicily, have an average pairwise difference of only 19 SNPs (diversity $\pi = 1.5 \times 10^{-6}$). Notably, this collection showed four non-overlapping segregating duplications (**Fig. 3c**). This striking finding suggests that CNVs can arise or disappear frequently during evolution.

To examine whether this transience is a general feature of CNVs in this population, we quantified the variation in copy number of each CNV relative to mutations in the adjacent region of the genome. If a CNV was subject only to the same processes as these adjacent regions, we would expect a strong correlation between the total mutation in these regions and the total variation in copy number of the CNV. However, the variation in copy number of CNVs across the dataset was only weakly correlated with SNP variation in nearby regions of the genome (Spearman rank correlation $\rho = 0.22$, $P = 0.041$), indicating that CNVs are subject to additional or different evolutionary processes (**Figure 2a**). Furthermore, some CNVs showed high rates of variation within closely-related clusters relative to their variation in the rest of the dataset (**Figure 2b and 2c, Supplementary Table 2**, **Supplementary Figure 4**). Finally, we found that many CNVs represented the rare allele within the cluster, consistent with events that have short half-lives (**Supplementary Figure 5**). Taken together, these results indicate that CNVs are transient and variable features of the genome, even within extremely closely related strains.

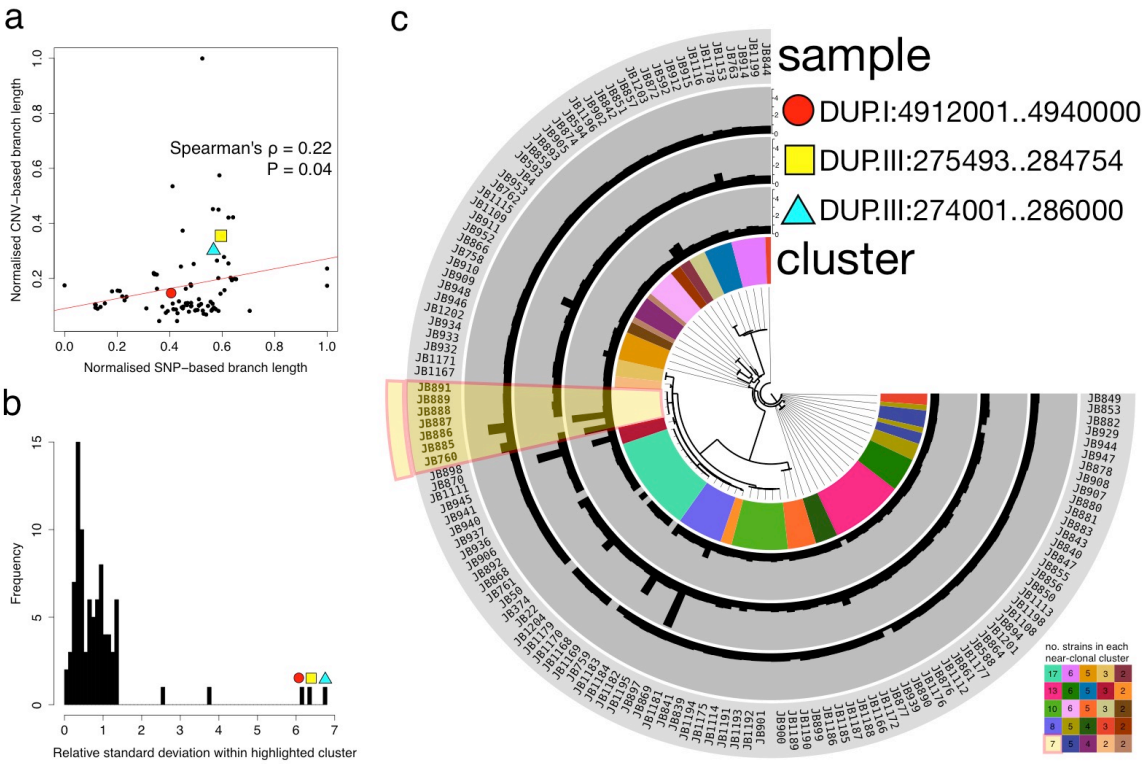
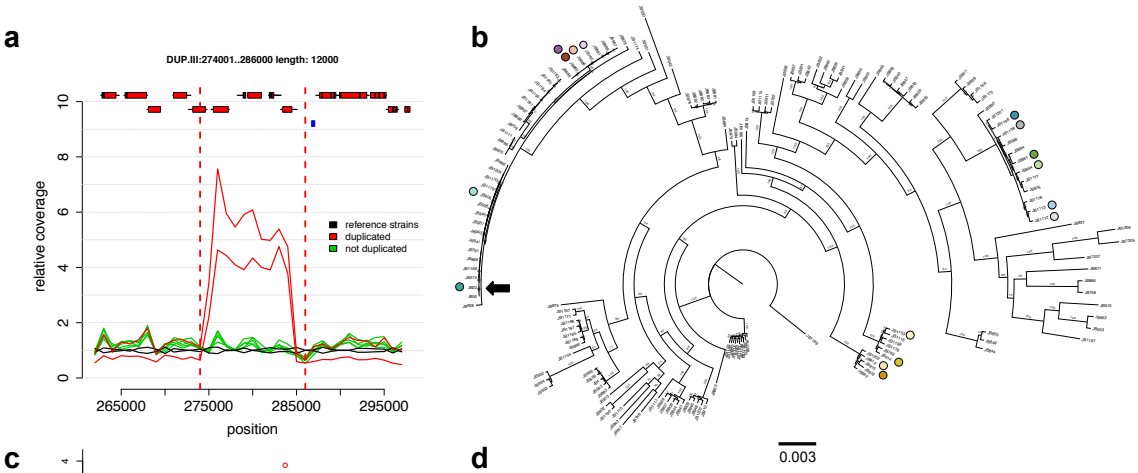


Figure 2. Copy number variants are transient within fission yeast. (a) For each of the 87 CNVs we calculated the genetic distance between strains using SNPs in the region around the CNV (20 kb up- and down-stream of the CNV, merged) as the total branch length from an approximate maximum-likelihood tree (*x*-axis, SNP-based branch length normalised to maximum value). We further calculated a CNV-based distance using the total branch length from a neighbour-joining tree constructed from Euclidean distances between strains based on their copy numbers (*y*-axis, CNV-based branch length normalised to maximum value). The weak correlation indicates that CNVs are subject to additional or different evolutionary processes. **(b)** Histogram of the standard deviation of each CNV within a near-clonal cluster (see also Figure 2a), relative to its standard deviation across strains not in the near-clonal cluster. Standard deviation is highly

193 correlated with CNV-based branch length (Spearman rank correlation $\rho = 0.90$, $P <$
194 0.001) (**Supplementary Figure 4b**). The highlighted CNVs have unusually high rates of
195 variation within this cluster compared to other clusters. **(c)** Copy number variation of
196 these highlighted CNVs plotted on a SNP-based phylogeny (20kb up- and down-stream
197 of the DUP.III:274001..286000 CNV) shows their relative transience within the cluster,
198 as well as their variation across other near-clonal clusters. SNP-based phylogenies for the
199 other two selected CNVs also do not separate the strains with different copy numbers
200 (individual plots for each CNV across clusters for its corresponding SNP-based
201 phylogeny are available as Supplementary data).
202
203
204 **Transient duplications affect gene expression**
205 Partial aneuploidies of 500-700 kb in the *S. pombe* reference strain are known to alter
206 gene expression levels within and, to some extent, outside of the duplicated region[31]. The
207 naturally occurring duplications described here are typically smaller (median length: 21
208 kb), including an average of 6.5 genes. To examine whether naturally occurring CNVs
209 have similar effects on gene expression, we examined eight pairs of closely related strains
210 (<150 SNPs among each pair) that contained at least one unshared duplication **(Figure 3,**
211 **Supplementary Table 3)**. Several of these strain pairs have been isolated from the same
212 substrate at the same time, and all pairs are estimated to have diverged approximately 50
213 to 65 years ago (**Supplementary Table 3**). We assayed transcript expression from log
214 phase cultures using DNA microarrays, each time comparing a duplicated to a non-
215 duplicated strain from within the same clonal population. In seven of the eight strain
216 pairs, the expression levels of genes within duplications were significantly induced,

217 although the degree of expression changes between genes was variable **(Figure 3c,**
218 **Supplementary Figure 6)**. The increased transcript levels correlated with the increased
219 genomic copy numbers, so that higher copy numbers produced correspondingly more
220 transcripts (Spearman rank correlation $\rho = 0.71$, $P = 0.014$, **Supplementary Figure 7**).
221 No changes in gene expression were evident immediately adjacent to the duplications
222 (**Supplementary Figure 7)**, suggesting that the local chromatin state was not strongly
223 altered by the CNVs. This result not only confirms the previous observation that CNVs
224 alter the gene expression levels, but more importantly it reveals large copy number
225 differences between two genomes that are only 19 SNPs apart.
226 Interestingly, some genes outside the duplicated regions also showed altered
227 expression levels **(Figure 3d, Supplementary Table 4)**. For example, two strain pairs
228 differ by a single 12 kb duplication. Here, five of seven genes within the duplication
229 showed induced expression, while 45 genes outside the duplicated region also showed
230 consistently altered expression levels (38 protein-coding genes, 7 non-coding RNAs)
231 (**Figure 3d**, arrays 7 and 8). As environmental growth conditions were tightly controlled,
232 these changes in gene expression could be due to either compensatory effects of the
233 initial perturbation caused by the 12kb duplication or changes that arise due to SNPs or
234 indels that segregate between the strains (**Supplementary Figure 6**). We conclude that
235 these evolutionary unstable duplications reproducibly affect the expression of distinct sets
236 of genes and thus have the potential to influence cellular function and phenotypes.



			inside DUP		outside DUP	
array	SNP/indel	DUP (kb)	up	down	up	down
1	52/469	+84 +130	29	0	54	34
2	**43/506**	**+20 +46**	**19**	**0**	**1**	**15**
3	**70/509**	**+20 +46**	**19**	**0**	**1**	**15**
4	12/37	+18	1	0	85	47
5	85/489	+96	15	0	101	44
6	37/545	+46	7	0	44	241
7	**16/49**	**+12**	**5**	**0**	**33**	**12**
8	**15/53**	**+12**	**5**	**0**	**33**	**12**

Figure 3. Transient duplications affect gene expression. (**a**) Duplications occur within near-clonal strains. Plot showing average read coverage in 1 kb windows for two clonal strains (JB760, JB886) with the duplication (red), five strains without duplication (green), and two reference strains (h^+, and h-) (black). Genes (with exons as red rectangles) and retrotransposon LTRs (blue rectangles) are shown on top (see **Supplementary Table 3** for details). (**b**) Eight pairs of closely related strains, differing by one or more large duplications, selected for expression analysis. The tree indicates the relatedness of these strain pairs (dots colored as in d). The position of the reference strain (Leupold's 972, JB22) is indicated with a black arrow. The scale bar shows the length of 0.003 insertions per site. (**c**) Gene expression increases for most genes within duplicated regions. For each tested strain pair, we show the relative gene expression (strains with duplication/strains without duplication) for all genes outside the duplication (as boxplot) and for all genes

within the duplication (red strip chart). In all but one case (array 4), the genes within the duplication tend to be more highly expressed than the genes outside of the duplication (all Wilcoxon rank sum test P-values $<1.5x10^{-3}$). (**d**) Summary of expression arrays 1-8, with strains indicated as colored dots (as in b), showing number of single-nucleotide polymorphism differences between strains (SNP), sizes of duplications in kb (DUP, where '+X +Y' indicates two duplications with lengths X and Y, respectively). We show total numbers of induced (up) and repressed (down) genes, both inside and outside the duplicated regions. Arrays 2,3 and 7,8 (in yellow shading) are replicates within the same clonal population that contain the same duplications, so we list the number of up- and down-regulated genes that are consistent between both arrays. See **Supplementary Tables 3 and 4** for details.

Copy number variants contribute to quantitative trait variance

To test whether SVs affect phenotypes, we examined the contributions of SNPs, CNVs and rearrangements to 227 quantitative traits (**Supplementary Table 5**), including 20 cell shape parameters, colony size on solid media assaying 42 stress and nutrient conditions[17], 126 growth parameters in liquid media conditions[7] and three biochemical parameters from wine fermentation[32]. For each phenotype, we used mixed model analysis to estimate the total proportion of variance explained by the additive contribution of genomic variants (the narrow-sense heritability).

When we determined heritability using only SNP data, estimates varied between 0% and 74% (median 30%). After adding CNVs and rearrangements to SNPs in a composite model, the estimated overall heritability increased for nearly all traits,

explaining up to ~40% of trait variance (**Figure 4a**). This finding indicates that the CNVs and rearrangements can explain a substantial proportion of the trait variance. Using this composite model, we quantified the individual contributions of heritability best explained by SNPs, CNVs and rearrangements (**Figure 4b**). On average, SNPs explained 24% of trait variance, CNVs 7%, and rearrangements 4% (**Supplementary Table 5**). Analysis of simulated data confirmed that the contribution of CNVs could not be explained by linkage to causal SNPs alone (**Supplementary Figure 6**).

Many trait measures gathered using the same method (e.g., growth on solid media, cell shape) are strongly correlated[17]. Thus, some groups of traits have consistently larger contributions from SVs (**Figure 4b**) than from SNPs alone. These traits include intracellular amino acid concentrations, growth under stress and several traits measured during wine fermentation (**Figure 4c**). Since many of these strains have been collected from fermentations (**Supplementary Table 6**), the substantial influence of CNVs may represent recent strong selection and adaptation to fermentation conditions that has occurred via recent CNV acquisition.

Our analysis of heritability showed that SNPs are generally able to capture most, but not all, of the genetic contribution of SVs (**Figure 4**). To examine whether trait-influencing SVs would be effectively detected from SNPs alone in this population, we examined the linkage of all 113 SVs with SNPs. We found that only 63 of these SVs (55%) are in strong linkage to SNPs (r^2 >0.6), leaving 45% of the SVs weakly linked. This lack of linkage is consistent with SVs being transient, rather than persisting within haplotypes. Such weakly linked SVs may be missed in SNP-only association studies.

To examine this possibility, and to locate specific SVs that affect these traits, we

296 performed mixed model genome-wide association studies, using all 68 SVs with minor
297 allele counts >5 (i.e. occurring in at least 5 strains) as well as 139,396 SNPs and 22,058
298 indels with minor allele counts >5. Trait-specific significance thresholds for 5%
299 familywise error rates were computed via permutation analysis, and were approximately
300 10^{-4} (SVs) and 10^{-6} (SNPs and indels). Nineteen SVs (28%) were significantly associated
301 with traits (15 duplications, 5 deletions, 1 translocation), as well as 228 SNPs (0.16%),
302 and 93 indels (0.42%) (**Supplementary Table 7**). SVs were associated with 20 different
303 traits, including amino acid concentrations, mating traits, and stress resistance in solid
304 and liquid media. Nine of these SVs were not strongly linked to SNPs ($r^2 < 0.6$). The
305 median effect size of these SVs was 14% (range 6-33%). While more detailed analyses of
306 these associations will be required to confirm any particular association, our findings are
307 consistent with the heritability analysis.
308 Collectively, these analyses indicate that even a small collection of SVs, most
309 notably CNVs, can contribute substantially to quantitative traits. Thus, GWAS analyses
310 conducted without genotyping SVs could fail to capture these important genetic factors.
311

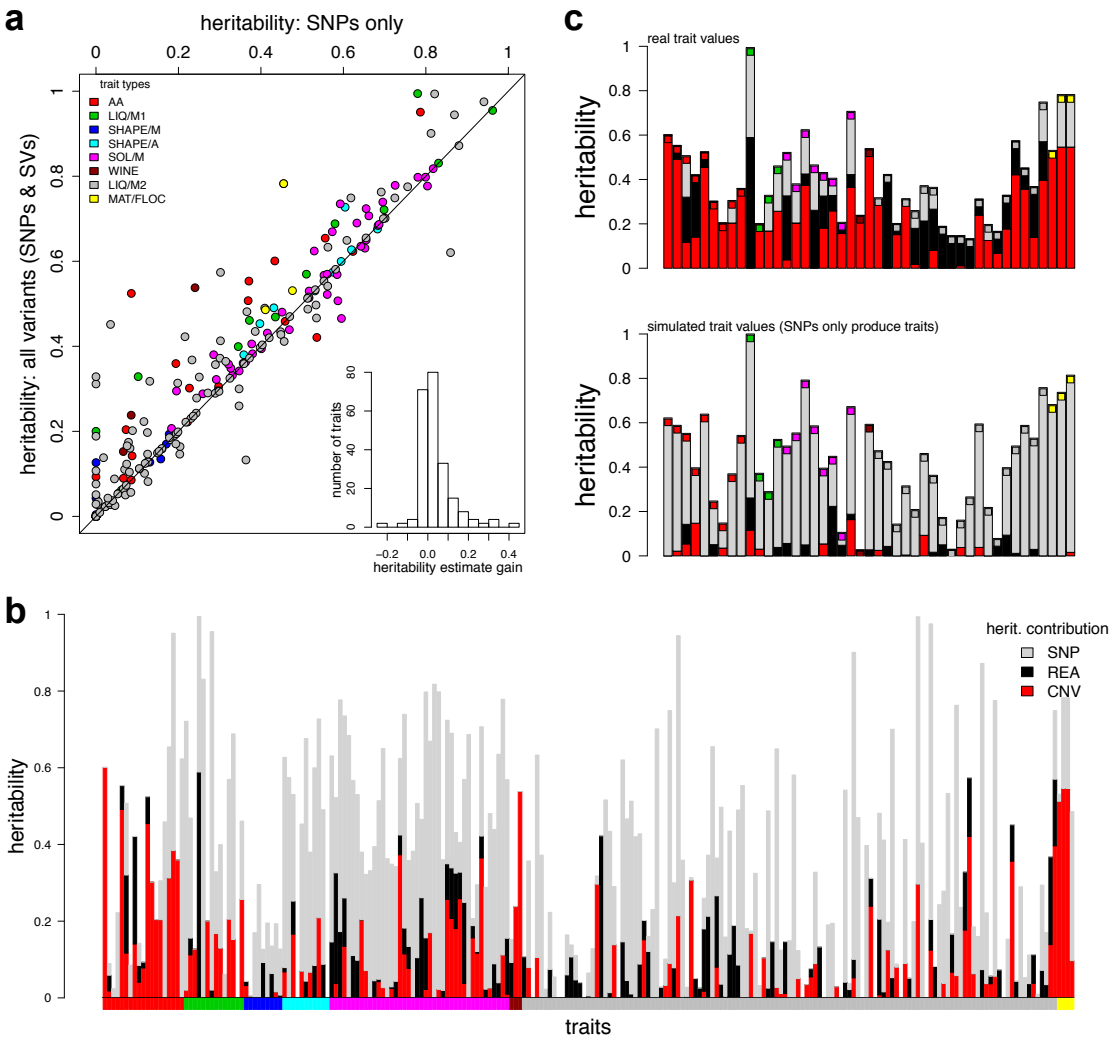


Figure 4. SVs contribute to quantitative traits. (**a**) Heritability estimates are improved by the addition of SVs. Heritability estimates for 227 traits (**Supplementary Table 5**), using only SNP data (x axis) range from 0 to 96% (median 29%). Adding SV calls (y axis) increases the estimates (median 34%), with estimates for some traits being improved up to a gain of 43% (histogram inset). The diagonal line shows where estimates after adding SVs are the same as those without (x=y). Inset: the distribution of the 'gain' in heritability after adding SV calls (median 0.4%, maximum 43%). Points are colored by trait types, according to legend top left. (**b**) The contributions of SNPs (grey), CNVs (red) and rearrangements (black) to heritability varied considerably between traits. Coloured

bars along the *x* axis indicate the trait types. heritability estimates are in **Supplementary Table 5**. The panel below bars indicates trait types as in the legend for part (a). (**c**) Top panel, for some traits, SVs explained more of the trait variation than SNPs. Boxes are colored as legend in part (a). Lower panel, analysis of simulated data generated with assumption that only SNPs cause traits indicates that the contribution of SVs to trait variance is unlikely to be due to linkage.

Structural variations contribute to intrinsic reproductive isolation

Crosses between *S. pombe* strains produce between <1% and 90% viable offspring[6,18]. We have previously shown that spore viability correlates inversely with the number of SNPs between the parental strains[17]. This intrinsic reproductive isolation may be due to the accumulation of Dobzhansky-Muller incompatibilities (variants that are neutral in one population, but incompatible when combined)[33,34]. However, genetically distant strains also accumulate SVs, which are known to lower hybrid viability and drive reproductive isolation[9]. In *S. pombe,* engineered inversions and translocations reduce spore viability by ~40% [6]. At present the impact of naturally occurring rearrangements, sequence divergence, and incompatible alleles in speciation within budding yeast is unclear[12-14,35,36].

To analyse intrinsic reproductive isolation in our population based on naturally occurring SVs, we examined the relationship between viability, SNPs and SVs. Both SV-distance (number of unshared SVs between parents) and SNP-distance inversely correlated with hybrid viability (Kendall correlation coefficients, SVs: $\tau = -0.26$, $P = 5.6 \times 10^{-3}$, SNPs: $\tau = -0.35$, $P = 1.6 \times 10^{-4}$) **(Supplementary Figure 7)**. While inversions and

translocations are known to lower hybrid viability as they affect chromosome pairing and segregation during meiosis[6,18,37], CNVs are not expected to influence spore viability. Consistent with this view, there was no significant correlation between CNVs and viability (rearrangements, $\tau = -0.36$, $P = 2.0 \times 10^{-4}$; CNVs, $\tau = -0.10$, $P = 0.28$).

As the numbers of SNP and rearrangement differences between mating parents are themselves correlated ($\tau = 0.53$, $P = 1.3 \times 10^{-8}$), we also estimated the influence of each factor alone using partial correlations. When either SNPs or rearrangements were controlled for, both remained significantly correlated with offspring viability ($P = 0.04$, $P = 0.02$, respectively) (**Figure 5)**. Taken together, these analyses indicate that both rearrangements and SNPs contribute to reproductive isolation, but CNVs do not.

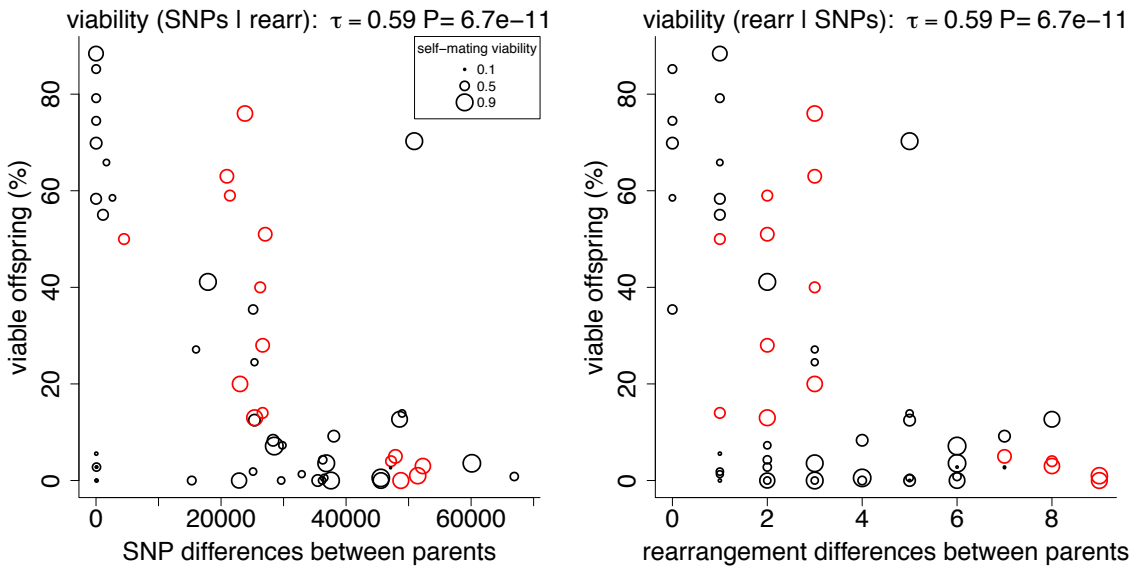


Figure 5. Both SNPs and rearrangements contribute to intrinsic reproductive isolation. Spore viability was measured from 58 different crosses from Jeffares, et al. [17] (black) or Avelar, et al. [6] (red), with each circle in the plots representing one cross. An

361 additive linear model incorporating both SNP and rearrangement differences showed
362 highly significant correlations with viability ($P = 1.2x10^{-6}$, $r^2 = 0.39$). Both genetic
363 distances measured using SNPs and rearrangements (inversions and translocations)
364 significantly correlated with viability when controlling for the other factor (Kendall
365 partial rank order correlations with viability SNPs|rearrangements $\tau = -0.19$, $P = 0.038$;
366 rearrangements|SNPs $\tau = -0.22$, $P = 0.016$). Some strains produce low viability spores
367 even when self-mated with their own genotype. The lowest self-mating viability of each
368 strain pair is indicated by circle size (see legend, smaller circles indicate lower self-
369 mating viability) to illustrate that low-viability outliers tend to include such cases (see
370 **Supplementary Table 8** for details).
371
372
373 **DISCUSSION**
374 Here we present the first genome- and population-wide catalog of SVs among *S. pombe*
375 strains. To account for the high discrepancy of available methods[25], we applied a
376 consensus approach to identify SVs (SURVIVOR), followed by rigorous filtering and
377 manual inspection of all calls. We focused on high specificity (the correctness of the
378 inferred SV) rather than high sensitivity (attempting to detect all SVs).
379 Our previous analyses of these strains, conducted without SV data[17], attributed
380 both trait variations and reproductive isolation to SNPs and/or small indels. Here we
381 show that the small number of SVs we describe make substantial contributions to both of
382 these factors. We demonstrate that CNVs (duplications and deletions) contribute
383 significantly to our ability to describe quantitative traits, whereas variants that rearrange

384 the order of the genome (inversions and translocations) produce much weaker effects on
385 traits. In contrast, CNVs have no detectable influence on reproductive isolation, while
386 rearrangements contribute substantially to reproductive isolation, similar to other
387 species[10,38].

388 We show that CNVs and, to a lesser extent, rearrangements can produce
389 substantial contributions to trait variation. These CNVs subtly alter the expression of
390 genes within and beyond the duplications, and contribute considerably to quantitative
391 traits. Within small populations, CNVs may produce larger effects on traits in the short
392 term than SNPs, since their effect sizes can be substantial (SVs significant in GWAS
393 have a mean effect size of 16% in this study). Within budding yeast, clearly measured
394 effects of alterations to gene order in the DAL metabolic cluster[39] and the lethality of
395 some engineered rearrangements[40] indicates that rearrangements can also effect
396 phenotypic changes. Given the evidence for extensive ploidy and aneuploidy variation
397 with budding yeasts, including clinical and industrial budding yeasts[29,41,42], SVs can be
398 expected to have considerable impacts on phenotypic variation these fungi.

399 In this context, it is striking that CNVs appear to be transient within the clonal
400 populations that we studied. Our analysis is consistent with experimental studies with
401 budding yeast, indicating that both rearrangements and CNVs may be gained or lost at
402 rates in excess of point mutations. For example, frequent gain of duplications has been
403 observed in laboratory cultures of *S. pombe*, where spontaneous duplications suppress
404 *cdc2* mutants at least 100 times more frequently than point mutations. These suppressor
405 strains lose their duplications with equal frequency[43], indicating reversion of alleles.
406 Similarly, duplications frequently occur during experimental evolution with budding

yeast[44]. This instability is likely facilitated by repeated elements, which are unstable within both budding and fission yeast genomes[45-48], which is also supported by the enrichment of SVs in our population near retrotransposon LTRs (**Supplementary Figure 8**). Though we do not examine the stability of rearrangements, there is also evidence for their instability. Transposon-mediated rearrangements are highly dynamic in laboratory cultures during selection[49,50], and show elevated mutation rates at subtelomeric regions[51].

This analysis also has relevance for human diseases, since *de novo* CNV formation in the human genome occurs at a rate of approximately one CNV/10 generations[52], and CNVs are known to contribute to a wide variety of diseases[4]. Indeed, both the population genetics and the effects of SVs within *S. pombe* seem similar to human, in that CNVs are associated with stoichiometric changes on gene expression, and SVs are in weak linkage with SNPs[53,54], and therefore may be badly tagged by SNPs in GWAS studies. We show that CNVs and rearrangements in fission yeast not only rapidly emerge, but substantially contribute to quantitative traits independent of weakly linked SNPs. These findings highlight the need to identify SVs when describing traits using GWAS, and indicate that a failure to call SVs can lead to an overestimation of the impact of SNPs to traits or contribute to the problem that large proportions of the heritable component of trait variation are not discovered in GWAS (the 'missing heritability'). We observed a clear example of this effect in two winemaking traits, where heritability was entirely due to SVs.

In summary, we show that different types of SVs are transient within populations of fission yeast, where they alter gene expression, impact phenotypes and can lead to

reproductive isolation.

METHODS

Performance assessment of SV callers using simulated data

To identify filtering parameters for DELLY, LUMPY and Pindel for the *S. pombe* genome, we simulated seven datasets (s1-s7) of 40x coverage with a range of different SV types and sizes (**Supplemental Table 7**). The simulated read sets contained sequencing errors (0.4%), SNPs and indels (0.1%) within the range of actual data from *S. pombe* strains and between 30 and 170 SVs. These data sets were produced by modifying the reference genome using our in-house software (SURVIVOR, described below), and simulating reads from this genome with Mason software [55].

After mapping the reads and calling SVs, we evaluated the calls. We defined a SV correctly predicted if: i) the simulated and reported SV were of the same type (e.g. duplication), ii) were predicted to be on same chromosome, and iii) their start and stop locations were with 1 kb. We then defined caller-specific thresholds to optimize the sensitivity and false discovery rate (FDR) for each caller. FDRs on the simulated data were low: DELLY (average 0.13), LUMPY (average 0.06) and Pindel (average 0.04).

Selecting calls that were present in at least two callers further reduced the FDR (average of 0.01). DELLY had the highest sensitivity (average 0.75), followed by SURVIVOR (average 0.70), LUMPY (average 0.62) and Pindel (0.55). We further used simulated data to assess the sensitivity and FDR of our predictions. cn.mops was evaluated with a 2 kb distance for start and stop coordinates. Our cn.mops parameters were designed to identify large (above 12 kb) events and thus did not identify any SVs

453 simulated for s1-s6. Details of simulations and caller efficacy are provided in
454 **Supplementary Table 9**.
455
456 **SURVIVOR (StructURal Variant majorIty VOte) Software Tool**
457 We developed the SURVIVOR tool kit for assessing SVs for short read data that contains
458 several modules. The first module simulates SVs given a reference genome file (fasta)
459 and the number and size ranges for each SV (insertions, deletions, duplications,
460 inversions and translocations). After reading in the reference genome, SURVIVOR
461 randomly selects the locations and size of SV following the provided parameters.
462 Subsequently, SURVIVOR alters the reference genome accordingly and prints the so
463 altered genome. In addition, SURVIVOR provides an extended bed file to report the
464 locations of the simulated SVs.
465 The second module evaluates SV calls based on a variant call format (VCF) file [56]
466 and any known list of SVs. A SV was identified as correct if i) they were of same type
467 (e.g. deletion); ii) they were reported on same chromosome, and iii) the start and stop
468 coordinates of the simulated and identified SV were within 1 kb (user definable).
469 The third module of SURVIVOR was used to filter and combine the calls from
470 three VCF files. In our case, these files were the results of DELLY, LUMPY and Pindel.
471 This module includes methods to convert the method-specific output formats to a VCF
472 format. SVs were filtered out if they were unique to one of the three VCF files. Two SVs
473 were defined as overlapping if they occur on the same chromosome, their start and stop
474 coordinates were within 1 kb, and they were of the same type. In the end, SURVIVOR

475 produced one VCF file containing the so filtered calls. SURVIVOR is available at
476 github.com/fritzsedlazeck/SURVIVOR.
477
478 **Read mapping and detection of structural variants**
479 Illumina paired-end sequencing data for 161 *S. pombe* strains were collected as described
480 in Jeffares, et al. [17], with the addition of Leupold's reference 975 h^+ (JB32) and excluding
481 JB374 (known to be a gene-knockout version of the reference strain, see below).
482 Leupold's 968 h^{90} and Leupold's 972 h^- were included as JB50 and JB22, respectively
483 (**Supplementary Table 6**). For all strains, reads were mapped using NextGenMap
484 (version 0.4.12)[57] with the following parameter (-X 1000000) to the *S. pombe* reference
485 genome (version ASM294v2.22). Reads with 20 base pairs or more clipped were
486 extracted using the script *split_unmapped_to_fasta.pl* included in the LUMPY package
487 (version 0.2.9)[25] and were then mapped using YAHA (version 0.1.83)[58] to generate split-
488 read alignments. The two mapped files were merged using Picard-tools (version 1.105)
489 (http://broadinstitute.github.io/picard), and all strains were then down-sampled to 40x
490 coverage using Samtools (version 0.1.18)[59].
491 Subsequently, DELLY (version 0.5.9, parameters: " –q 20 -r")[26], LUMPY
492 (version 0.2.9, recommended parameter settings)[25] and Pindel (version 0.2.5a8, default
493 parameter)[27] were used to independently identify SVs in the 161 strains using our
494 SURVIVOR software. This included merging any variants of the same type (duplication,
495 deletion *etc*) whose start and end coordinates where within 1 kb. Merging was justified by
496 the finding that most allele calls were close to the defined call (only 5% of start or end
497 positions were >300nt from the defined consensus boundary). We then retained all

498 variants predicted by at least two methods. These SVs calls were genotyped using
499 DELLY.

500 To identify further CNVs, we ran cn.MOPS [24] with parameters tuned to collect
501 large duplications/deletions as follows: read counts were collected from bam alignment
502 files (as above) with *getReadCountsFromBAM* and WL=2000, and CNVs predicted using
503 *haplocn.mops* with minWidth= 6, all other parameters as default. Hence, the minimum
504 variant size detected was 12 kb. CNV were predicted for each strain independently by
505 comparing the alternative strain to the two reference strains (JB22, JB32) and four
506 reference-like strains that differed from the reference by less than 200 SNPs (JB1179,
507 JB1168, JB937, JB936).

508 After CNV calling, allele calling was achieved by comparing counts of coverage
509 in 100bp windows for the two reference strains (JB22, JB32) to each alternate strain
510 using custom R scripts. Alleles were called as non-reference duplications if the one-sided
511 Wilcoxon rank sum test p-values for both JB22 and JB32 vs alternate strain were less
512 than $1x10^{-10}$ (showing a difference in coverage) and the ratio of alternate/reference
513 coverage (for both JB22 and JB32) was >1.8 (duplications), or <0.2 (deletions). Manual
514 inspection of coverage plots showed that the vast majority of the allele calls were in
515 accordance with what we discerned by eye. These R scripts were also used to examine
516 CNVs predicted to be segregating within clusters (clonal populations). All such CNVs
517 were examined in all clusters that contained at least one non-reference allele call
518 (**Supplementary Table 10**).

519 Finally, we manually mapped two large duplications that did not satisfy these
520 criteria (DUP.I:2950001..3190000, 240kb and DUP.I:5050001..5560000, 510kb – both

singletons in JB1207), but were clearly visible in chromosome-scale read coverage plots **(Supplementary Figure 9).**

Reduction of false discovery rate

This filtering produced 315 variant calls. However, because 31 of these 315 (~10%) were called within the two reference strains (JB22, JB32), we expected that this set still contained false positives. To further reduce the false positive rate, we looked for parameters that would reduce calls made in reference strains (JB22 and JB32) but not reduce calls in strains more distantly related to the reference (JB1177, JB916 and JB894 that have 68223, 60087 and 67860 SNP differences to reference[17]). The reasoning was that we expected to locate few variants in the reference, and more variants in the more distantly related strains. This analysis showed that paired end support, repeats and mapping quality were of primary value.

We therefore discarded all SVs that had a paired end support of 10 or less. In addition, we ignored SVs that appeared in low mapping quality regions (i.e. regions where reads with MQ=0 map) or those where both start and end coordinates overlapped with previously identified retrotransposon LTRs[17].

Finally, to ensure a high specificity call set, these filtered SVs were manually curated using IGV [60] (**Supplementary Tables 11,12**). We assigned each SVs a score (0: not reliable, 1: unclear, 2: reliable based on inspection of alignments through IGV). We utilized different visualizations from IGV to identify regions were pairs of the reads mapped to different loci, for example, which we interpreted as possible artefacts. Overall, we investigated whether the alignments of the breakpoints and reads in close proximity

had a reliable mapping in terms of mapping quality and clearness of the distortions of the pairs. Only calls passing this manual curation as reliable (score 2) were included in the final data set of 113 variants utilized for all further analyses. These filtering and manual curation steps reduced our variant calls substantially, from 315 to 113. At this stage only 1/113 (~1%) of these variants was called within the two standard reference strains (Leupolds's *h+* and *h-,* JB22 and JB32 in our collection).

PCR validation

PCR analysis was performed to confirm 10 of the 11 inversions and all 15 translocations from the curated data set. One inversion was too small to examine by PCR (INV.AB325691:6644..6784, 140 nt). Primers were designed using Primer3[61] to amplify both the reference and alternate alleles. PCR was carried out with each primer set using a selection of strains that our genotype calls predict to include at least one alternate allele and at least one reference allele (usually 6 strains). Products were scored according to product size and presence/absence (**Supplementary Tables 13,14**).

Inversions: 9/10 variants were at least partially verified by either reference or alternate allele PCR (3 variants were verified by both reference and alternate PCRs), and 7/10 inversions also received support from BLAST (see below). Translocations: 10/15 were at least partially verified by either reference or alternate allele PCR (5/15 variants were verified by both reference and alternate PCRs). One additional translocation received support from BLAST (see below), meaning that 11/15 translocations were supported by PCR and/or BLAST. Three of the four translocations that could not be verified were probably nuclear copies of mitochondrial genes (NUMTs) [62], because one

breakpoint was mapped to the mitochondrial genome. Details of the 113 curated variants are presented in Supplementary Table 15.

Validation by BLAST of *de novo* assemblies

We further assessed the quality of the predicted breakpoints for the inversions and translocations by comparing them to the previously created *de novo* assemblies for each of the 161 strains [17]. To this end, we created blast databases for the scaffolds of each strain that were >1kb. We then created the predicted sequence for 1 kb around each junction of the validated 10 inversions and 15 translocations. These sequences were used to search the blast databases using BLAST+ with --gapopen 1 --gapextend 1 parameters. We accepted any blast hsp with a length >800 bp as supporting the junction (because these must contain at least 300 bp at each side of the break point). Four inversions and three translocations gained support from these searches (Supplementary File Tables2-PCR.xlsx).

Knockout strain control

Our sample of sequenced strains included one strain (JB374) that is known to contain deletions of the *his3* and *ura4* genes. Our variant calling and validation methods identified only two variants in this strain, both deletions that corresponded to the positions of these genes, as below:

his3 gene location is chromosome II, 1489773-1488036, deletion detected at II:1488228-1489646.

589 *ura4* gene location is chromosome III, 115589-116726, deletion detected at III:115342-
590 117145.
591 This strain was not included in the further analyses of the SVs.
592
593 **Microarray expression analysis**
594 Cells were grown in YES (Formedium, UK) and harvested at OD_{600} =0,5. RNA was
595 isolated followed by cDNA labeling [63]. Agilent 8 x 15K custom-made *S. pombe*
596 expression microarrays were used. Hybridization, normalization and subsequent washes
597 were performed according to the manufacturer's protocols. The obtained data were
598 scanned and extracted using GenePix and processed for quality control and normalization
599 using in-house developed R scripts. Subsequent analysis of normalized data was
600 performed using R. Microarray data have been submitted to ArrayExpress (accession
601 number E-MTAB-4019). Genes were considered as induced if their expression signal
602 after normalization was >1.9, and repressed if <0.51.
603
604 **Time to most recent common ancestor (TMRCA) estimates**
605 Previously, based on the genetic distances between these strains and the 'dated tip' dating
606 method implemented in BEAST [64], we have estimated the divergence times between all
607 161 *S. pombe* strains sequenced [17]. To determine the TMRCA for pairs of strains, we re-
608 examined the BEAST outputs using FigTree to obtain the medium and 95% confidence
609 intervals.
610
611 **SNP and indel calling**

612 SNPs were called as described [17]. Insertions and deletions (indels) were called in 160
613 strains using stampy-mapped, indel-realigned bams as described previously [17]. We
614 accepted indels that were called by both the Genome Analysis Toolkit HaplotypeCaller [65]
615 and Freebayes[66], and then genotyped all these calls with Freebayes.
616 Briefly, indels were called on each strains bam with HaplotypeCaller, and filtered
617 for call quality >30 and mapping quality >30 (bcftools filter --include 'QUAL>30 &&
618 MQ>30'). Separately, indels were called on each strains bam with Freebayes, and filtered
619 for call quality >30. All Freebayes vcf files were merged, accepting only positions called
620 by both Freebayes and HaplotypeCaller. These indels were then genotyped with
621 Freebayes using a merged bam (containing reads from all strains), using the --variant-
622 input flag for Freebayes to genotyped only the union calls. Finally indels were filtered for
623 by score, mean reference mapping quality and mean alternate mapping quality >30
624 (bcftools filter --include 'QUAL>30 && MQM>30 & MQMR>30'). These methods
625 identified 32,268 indels. Only 50 of these segregated between Leupold's h^- reference
626 (JB22) and Leupold's h^{90} reference (JB50), whereas 12109 indels segregated between the
627 JB22 reference and the divergent strain JB916.
628
629 **Heredity and GWAS**
630 We selected 53 traits that contained at least values from 100 strains[17], and so included
631 multiple individuals from within clonal populations (growth rates on 42 different solid
632 media and 11 cell shape characters measured with automated image analysis). Trait
633 values were normalized using a rank-based transformation in R, for each trait vector y,
634 normal.y =qnorm(rank(y)/(1+length(y))). Total heritability, and the contribution of SNPs,

635 CNVs and rearrangements were estimated using LDAK (version 5.94) [67], with kinship
636 matrices derived from all SNPs, 146 CNVs, and 15 rearrangements. All genotypes,
637 including CNVs were encoded as binary values (1 or 0) for heritability and GWAS. To
638 assess whether the contribution of CNVs could be primarily due to linkage with causal
639 SNPs, we simulated trait data using the --make-phenos function of LDAK with the
640 relatedness matrix from all SNPs, assuming that all variants contributed to the trait (--
641 num-causals -1). We made one simulated trait data set per trait, for each of the 53 traits,
642 with total heritability defined as predicted from the real data. We then estimated the
643 heritability using LDAK, including the joint matrix of SNPs, CNVs and rearrangements.
644 To assess the extent to which the contribution of SNPs to heritability was overestimated,
645 we performed another simulation using the relatedness matrix from the 87 segregating
646 CNVs alone, and then estimated the contribution of SNPs, CNVs and rearrangements in
647 this simulated data as above.
648 Genome-wide associations were performed with LDAK (version 5) using default
649 parameters. To account for the unequal relatedness of strains, we used a kinship matrix
650 derived from all 172,368 SNPs called previouslyJeffares, et al. [17]. Association analysis
651 was used to find associations between traits, testing SVs, SNPs and indels with a minor
652 allele count ≥5. Analysis was run separately for 68 SVs, 139,396 SNPs and 22,058 indels
653 (each used the kinship derived from all SNPs). We examined the same 53 traits as for the
654 heritability analysis (above). For each trait, we carried out 1000 permutations of trait
655 data, and define the 5th percentile of these permutations as the trait-specific P-value
656 threshold.
657

Model details for Heritability and GWAS Analysis

To estimate the heritability contribution of SNPs, we computed a kinship matrix (K_{SNP}) using all 172,368 SNPs that we had discovered in our previous published analysis[17] (elements of this matrix represent pairwise allelic correlations across all SNPs)[67] , onto which we regressed the phenotypic values assuming the following model:

$$Y \sim N\left(0, K_{SNP}\ \sigma^{2}_{SNP} + \ \sigma^{2}_{e} I\right)$$

We estimated the two variance components, σ^{2}_{SNP} and σ^{2}_{e}, using REML (restricted maximum likelihood), based on which our estimates of the heritability of SNPs is

$$\frac{\sigma^{2}_{SNP}}{\sigma^{2}_{SNP} + \sigma^{2}_{e}}$$

To estimate the heritability of CNVs and rearrangements, we repeated this analysis using instead K_{CNV} then K_{REA}, computed using only 146 segregating CNVs and 15 segregating rearrangements, respectively.

We additionally considered the model

$$Y \sim N(0, K_{SNP}\ \sigma^{2}_{SNP} + K_{CNV}\ \sigma^{2}_{CNV} + K_{REA}\ \sigma^{2}_{\Re A} + \ \sigma^{2}_{e} I),$$

Having estimated the four variance components, again using REML, the relative contributions of SNPs, CNVs and rearrangements are, respectively,

$\frac{\sigma^{2}_{SNP}}{S}$, $\frac{\sigma^{2}_{CNV}}{S}$ and $\frac{\sigma^{2}_{REA}}{S}$

where $S = \ \sigma^{2}_{SNP} + \ \sigma^{2}_{CNV} + \ \sigma^{2}_{REA}$

678

679 To test the specificity of this analysis, we generated phenotypes for which only one
680 predictor type contributed (e.g., only SNPs), then analyzed using the individual and joint
681 models above, which allowed us to assess how accurately we can distinguish between
682 contributions of different predictor types.

683 For the mixed model association analysis, we used the same the SNP kinship
684 matrix. As the predictors (variants that we examined for effects on a trait), we chose to
685 analyse SNPs, indels and SVs with a minor allele count ≥5 (68 SVs, 139396 SNPs and
686 22,058 indels).

687 Then for each predictor X_j we considered the model

688 $Y \sim N(\beta_j X_j K_{SNP} \sigma^2_{SNP} + \sigma^2_e I)$,

689 where β_j is the effect size of predictor X_j

690

691 Having solved using REML, we used a likelihood ratio test (comparing to the null model
692 ($\beta_j = 0$) to assess whether β_j is significantly non-zero. Each of these analyses used the
693 kinship derived from all SNPs.

694

695

696 **Offspring viability and genetic distance**

697 Cross spore viability data and self-mating viability were collected from previous analyses
698 [6,17]. The number of differences between each pair was calculated using vcftools vcf-
699 subset [56], and correlations were estimated using R, with the ppcor package. When
700 calculating the number of CNVs differences between strains, we altered our criteria for

'different' variants (to merge variants whose starts and ends where within 1 kb), and merged CNVs if their overlap was >50% and their allele calls were the same.

Transience analysis

For each CNV, we extracted all SNPs from 20 kb upstream and 20 kb downstream. 86/87 CNVs showed variation in these regions (DUP.MT:1..19382 was the only CNV with no corresponding SNPs). We then used these concatenated SNPs to build a local SNP-based tree with FastTree (version 2.1.9)[68]. To build a CNV-based tree from the copy number variation in each CNV region, we used a neighbour-joining tree estimation based on the Euclidean distances between strains.

The total branch length of the CNV-based tree was strongly correlated (Spearman rank correlation ρ=0.90, P <0.001) with the standard deviation of copy number variation (Supplementary Figure 4). We therefore used this standard deviation to define a relative rate of transience for each cluster, $\sigma_{rc} = \sigma_{ic}/\sigma_{oc}$ where σ_{ic} and σ_{oc} are the within cluster and without cluster standard deviations respectively, meaning that CNVs which were highly relatively transient within a given cluster would have high values of σ_{rc}. This was used to select the three CNVs visualised in Figure 2c. See Supplementary Table 2 for all values of σ_{rc}, Supplementary Figure 4 for visualization as heatmap). Visualisations of all 86/87 CNVs with their SNP-based phylogenies are available at: https://figshare.com/projects/fission_yeast_structural_variation/15798.

Circle plots were used to visualize the variation in copy number over the SNP-based phylogeny for each CNV using Anvi'o (version 2.0.3)[69].

724

References

1 Chen, C. *et al.* A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. *BMC Genomics* **13**, 733, doi:10.1186/1471-2164-13-733 (2012).

2 Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361-366, doi:10.1038/nature12818 (2014).

3 Wang, Y. *et al.* Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nat Genet* **47**, 944-948, doi:10.1038/ng.3346 (2015).

4 Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**, 451-481, doi:10.1146/annurev.genom.9.081307.164217 (2009).

5 Zhang, H. *et al.* Gene copy-number variation in haploid and diploid strains of the yeast Saccharomyces cerevisiae. *Genetics* **193**, 785-801, doi:10.1534/genetics.112.146522 (2013).

6 Avelar, A. T., Perfeito, L., Gordo, I. & Ferreira, M. G. Genome architecture is a selectable trait that can be maintained by antagonistic pleiotropy. *Nat Commun* **4**, 2235, doi:10.1038/ncomms3235 (2013).

7 Brown, W. R. *et al.* A Geographically Diverse Collection of Schizosaccharomyces pombe Isolates Shows Limited Phenotypic Variation but Extensive Karyotypic Diversity. *G3 (Bethesda)* **1**, 615-626, doi:10.1534/g3.111.001123 (2011).

8 McGaugh, S. E. & Noor, M. A. Genomic impacts of chromosomal inversions in parapatric Drosophila species. *Philos Trans R Soc Lond B Biol Sci* **367**, 422-429, doi:10.1098/rstb.2011.0250 (2012).

9 Rieseberg, L. H. Chromosomal rearrangements and speciation. *Trends Ecol Evol* **16**, 351-358 (2001).

10 Ayala, D., Guerrero, R. F. & Kirkpatrick, M. Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution* **67**, 946-958, doi:10.1111/j.1558-5646.2012.01836.x (2013).

11 Stevison, L. S., Hoehn, K. B. & Noor, M. A. Effects of inversions on within- and between-species recombination and divergence. *Genome Biol Evol* **3**, 830-841, doi:10.1093/gbe/evr081 (2011).

12 Liti, G., Barton, D. B. & Louis, E. J. Sequence diversity, reproductive isolation and species concepts in Saccharomyces. *Genetics* **174**, 839-850, doi:10.1534/genetics.106.062166 (2006).

13 Hou, J., Friedrich, A., de Montigny, J. & Schacherer, J. Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in Saccharomyces cerevisiae. *Curr Biol* **24**, 1153-1159, doi:10.1016/j.cub.2014.03.063 (2014).

14 Leducq, J. B. *et al.* Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat Microbiol* **1**, 15003, doi:10.1038/nmicrobiol.2015.3 (2016).

15 Ortiz-Barrientos, D., Engelstadter, J. & Rieseberg, L. H. Recombination Rate Evolution and the Origin of Species. *Trends Ecol Evol* **31**, 226-236, doi:10.1016/j.tree.2015.12.016 (2016).

16 Fawcett, J. A. *et al.* Population genomics of the fission yeast Schizosaccharomyces pombe. *PLoS One* **9**, e104241, doi:10.1371/journal.pone.0104241 (2014).

17 Jeffares, D. C. *et al.* The genomic and phenotypic diversity of Schizosaccharomyces pombe. *Nat Genet* **47**, 235-241, doi:10.1038/ng.3215 (2015).

18 Zanders, S. E. *et al.* Genome rearrangements and pervasive meiotic drive cause hybrid infertility in fission yeast. *Elife* **3**, e02630, doi:10.7554/eLife.02630 (2014).

19 Wood, V. *et al.* The genome sequence of Schizosaccharomyces pombe. *Nature* **415**, 871-880, doi:10.1038/nature724 (2002).

20 Sabatinos, S. A. & Forsburg, S. L. Molecular genetics of Schizosaccharomyces pombe. *Methods Enzymol* **470**, 759-795, doi:10.1016/S0076-6879(10)70032-X (2010).

21 Kim, D. U. *et al.* Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol* **28**, 617-623, doi:10.1038/nbt.1628 (2010).

22 Marguerat, S. *et al.* Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* **151**, 671-683, doi:10.1016/j.cell.2012.09.019 (2012).

23 Ryan, C. J. *et al.* Hierarchical modularity and the evolution of genetic interactomes across species. *Mol Cell* **46**, 691-704, doi:10.1016/j.molcel.2012.05.028 (2012).

24 Klambauer, G. *et al.* cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res* **40**, e69, doi:10.1093/nar/gks003 (2012).

25 Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84, doi:10.1186/gb-2014-15-6-r84 (2014).

26 Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339, doi:10.1093/bioinformatics/bts378 (2012).

27 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).

28 Bergstrom, A. *et al.* A high-definition view of functional genetic variation from natural yeast genomes. *Mol Biol Evol* **31**, 872-888, doi:10.1093/molbev/msu037 (2014).

29 Dunn, B., Richter, C., Kvitek, D. J., Pugh, T. & Sherlock, G. Analysis of the Saccharomyces cerevisiae pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Res* **22**, 908-924, doi:10.1101/gr.130310.111 (2012).

30 Bitton, D. A. *et al.* AnGeLi: A Tool for the Analysis of Gene Lists from Fission Yeast. *Front Genet* **6**, 330, doi:10.3389/fgene.2015.00330 (2015).

31 Chikashige, Y. *et al.* Gene expression and distribution of Swi6 in partial aneuploids of the fission yeast Schizosaccharomyces pombe. *Cell Struct Funct* **32**, 149-161 (2007).

32 Benito, A. *et al.* Selected Schizosaccharomyces pombe Strains Have Characteristics That Are Beneficial for Winemaking. *PLoS One* **11**, e0151102, doi:10.1371/journal.pone.0151102 (2016).

33 Dobzhansky, T. On the Sterility of the Interracial Hybrids in Drosophila Pseudoobscura. *Proc Natl Acad Sci U S A* **19**, 397-403 (1933).

34 Muller, H. J. Reversibility in Evolution Considered from the Standpoint of Genetics. *Biological Reviews* **14**, 261–280, doi:10.1111/j.1469-185X.1939.tb00934.x (1939).

35 Fischer, G., Rocha, E. P., Brunet, F., Vergassola, M. & Dujon, B. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet* **2**, e32, doi:10.1371/journal.pgen.0020032 (2006).

36 Gordon, J. L., Byrne, K. P. & Wolfe, K. H. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. *PLoS Genet* **5**, e1000485, doi:10.1371/journal.pgen.1000485 (2009).

37 Delneri, D. *et al.* Engineering evolution to study speciation in yeasts. *Nature* **422**, 68-72, doi:10.1038/nature01418 (2003).

38 Noor, M. A., Grams, K. L., Bertucci, L. A. & Reiland, J. Chromosomal inversions and the reproductive isolation of species. *Proc Natl Acad Sci U S A* **98**, 12084-12088, doi:10.1073/pnas.221274498 (2001).

39 Naseeb, S. & Delneri, D. Impact of chromosomal inversions on the yeast DAL cluster. *PLoS One* **7**, e42022, doi:10.1371/journal.pone.0042022 (2012).

40 Naseeb, S. *et al.* Widespread Impact of Chromosomal Inversions on Gene Expression Uncovers Robustness via Phenotypic Buffering. *Mol Biol Evol* **33**, 1679-1696, doi:10.1093/molbev/msw045 (2016).

41 Zhu, Y. O., Sherlock, G. & Petrov, D. A. Whole Genome Analysis of 132 Clinical Saccharomyces cerevisiae Strains Reveals Extensive Ploidy Variation. *G3 (Bethesda)* **6**, 2421-2434, doi:10.1534/g3.116.029397 (2016).

42 Gallone, B. *et al.* Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell* **166**, 1397-1410 e1316, doi:10.1016/j.cell.2016.08.020 (2016).

43 Carr, A. M., MacNeill, S. A., Hayles, J. & Nurse, P. Molecular cloning and sequence analysis of mutant alleles of the fission yeast cdc2 protein kinase gene: implications for cdc2+ protein structure and function. *Mol Gen Genet* **218**, 41-49 (1989).

44 Dunham, M. J. *et al.* Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **99**, 16144-16149, doi:10.1073/pnas.242624799 (2002).

45 Chan, J. E. & Kolodner, R. D. A genetic and structural study of genome rearrangements mediated by high copy repeat Ty1 elements. *PLoS Genet* **7**, e1002089, doi:10.1371/journal.pgen.1002089 (2011).

46 Coulon, S. *et al.* Slx1-Slx4 are subunits of a structure-specific endonuclease that maintains ribosomal DNA in fission yeast. *Mol Biol Cell* **15**, 71-80, doi:10.1091/mbc.E03-08-0586 (2004).

47 Gadaleta, M. C. *et al.* Swi1Timeless Prevents Repeat Instability at Fission Yeast Telomeres. *PLoS Genet* **12**, e1005943, doi:10.1371/journal.pgen.1005943 (2016).

48 Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K. J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213-1216, doi:10.1126/science.1170097 (2009).

49 Chang, S. L., Lai, H. Y., Tung, S. Y. & Leu, J. Y. Dynamic large-scale chromosomal rearrangements fuel rapid adaptation in yeast populations. *PLoS Genet* **9**, e1003232, doi:10.1371/journal.pgen.1003232 (2013).

50 Gresham, D. *et al.* The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet* **4**, e1000303, doi:10.1371/journal.pgen.1000303 (2008).

51 Nishant, K. T. *et al.* The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. *PLoS Genet* **6**, e1001109, doi:10.1371/journal.pgen.1001109 (2010).

52 Itsara, A. *et al.* De novo rates and selection of large copy number variation. *Genome Res* **20**, 1469-1481, doi:10.1101/gr.107680.110 (2010).

53 Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-853, doi:10.1126/science.1136678 (2007).

54 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).

55 Holtgrewe, M. Mason-A Read Simulator for Second Generation Sequencing Data. (Institut für Mathematik und Informatik, Freie Universität Berlin, 2010).

56 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

57 Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790-2791, doi:10.1093/bioinformatics/btt468 (2013).

58 Faust, G. G. & Hall, I. M. YAHA: fast and flexible long-read alignment with optimal breakpoint detection. *Bioinformatics* **28**, 2417-2424, doi:10.1093/bioinformatics/bts456 (2012).

59 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

60 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

61 Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**, e115, doi:10.1093/nar/gks596 (2012).

62 Lenglez, S., Hermand, D. & Decottignies, A. Genome-wide mapping of nuclear mitochondrial DNA sequences links DNA replication origins to chromosomal double-strand break formation in Schizosaccharomyces pombe. *Genome Res* **20**, 1250-1261, doi:10.1101/gr.104513.109 (2010).

63 Lyne, R. *et al.* Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics* **4**, 27, doi:10.1186/1471-2164-4-27 (2003).

64 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, doi:10.1093/molbev/mss075 (2012).

65 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

66 Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv* **1207.3907** (2012).

67 Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* **91**, 1011-1021, doi:10.1016/j.ajhg.2012.10.010 (2012).

68 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).

69 Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319, doi:10.7717/peerj.1319 (2015).

Supplementary data

SNP, indel and SVs calls, genotypes and copy numbers are available on figshare at:

https://figshare.com/projects/fission_yeast_structural_variation/15798

Array data is available at ArrayExpress, accession number: E-MTAB-4019.

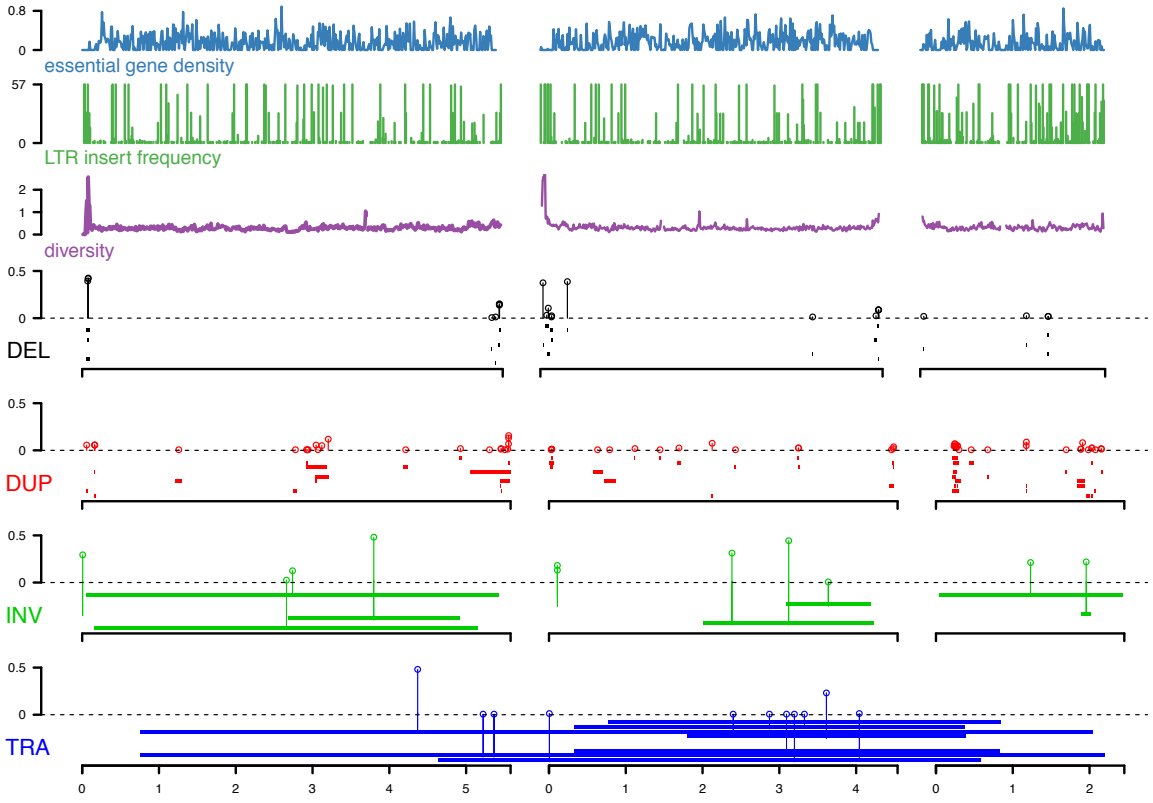
Acknowledgments

We thank Günter Klambauer for advice on cn.MOPS and Michael C. Schatz for helpful discussions and comments on the manuscript. F.S. was supported through National Science Foundation awards (DBI-1350041) and National Institutes of Health award (R01-HG006677). D.J., M.H., C.R. were supported by a Wellcome Trust Senior Investigator Award to J.B. (grant 095598/Z/11/Z). J.B. was supported by a Royal Society Wolfson Research Merit Award.

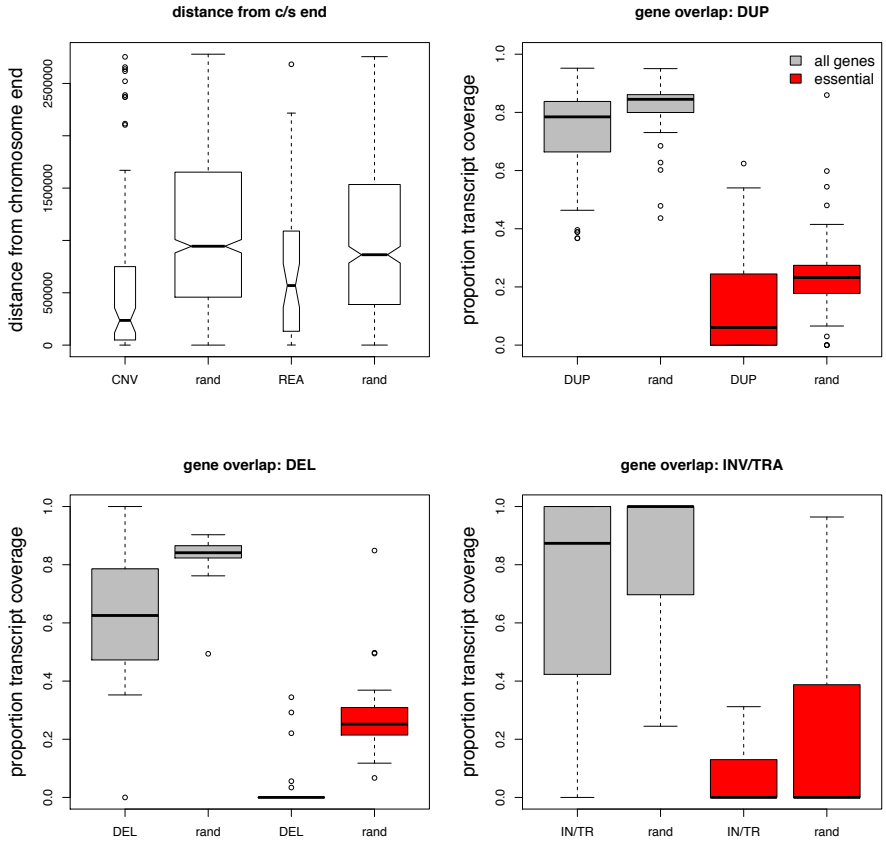
Author contributions

DJ, FS, CD and JB conceived and developed the study. DJ, LS, CJ and FS conducted the bioinformatics analysis. DJ designed the laboratory work. FS designed and implemented SURVIVOR. DS contributed to analysis of heritability and GWAS. CR and MH produced the expression array analysis. MH conducted PCR validation of variants. JB provided the bulk of funding for personnel and research costs. DJ, FS, CJ, LS, FB, CD and JB wrote the manuscript.

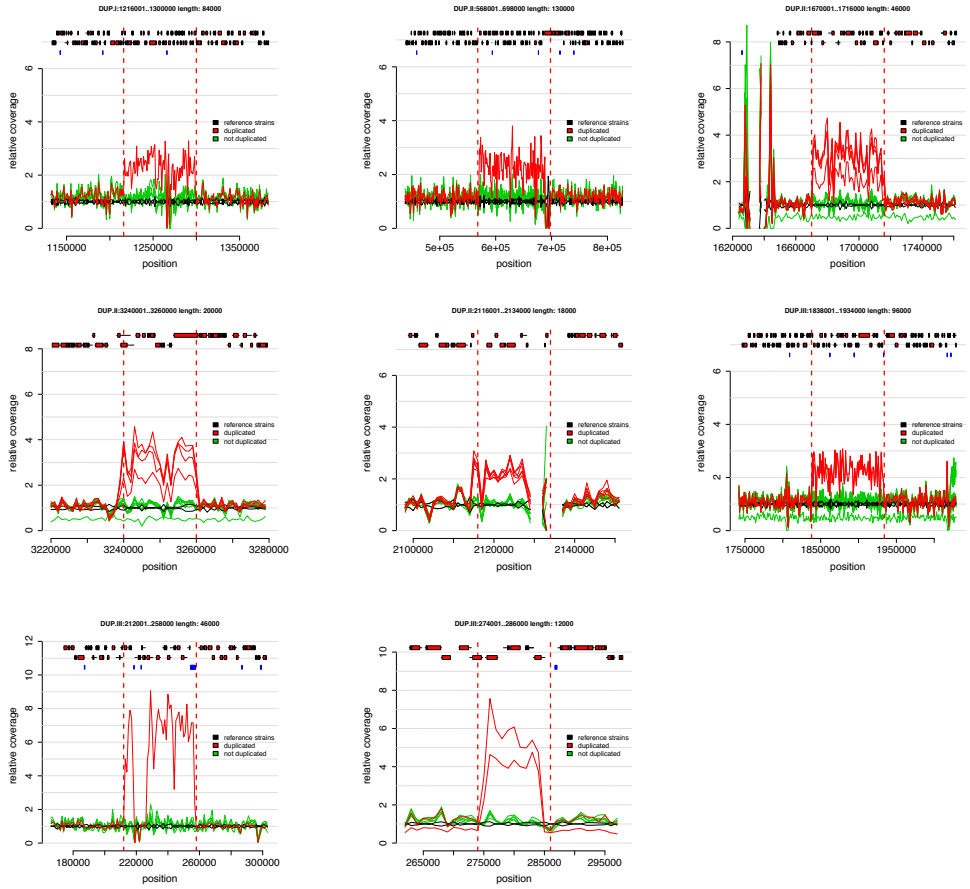
Supplementary Figures



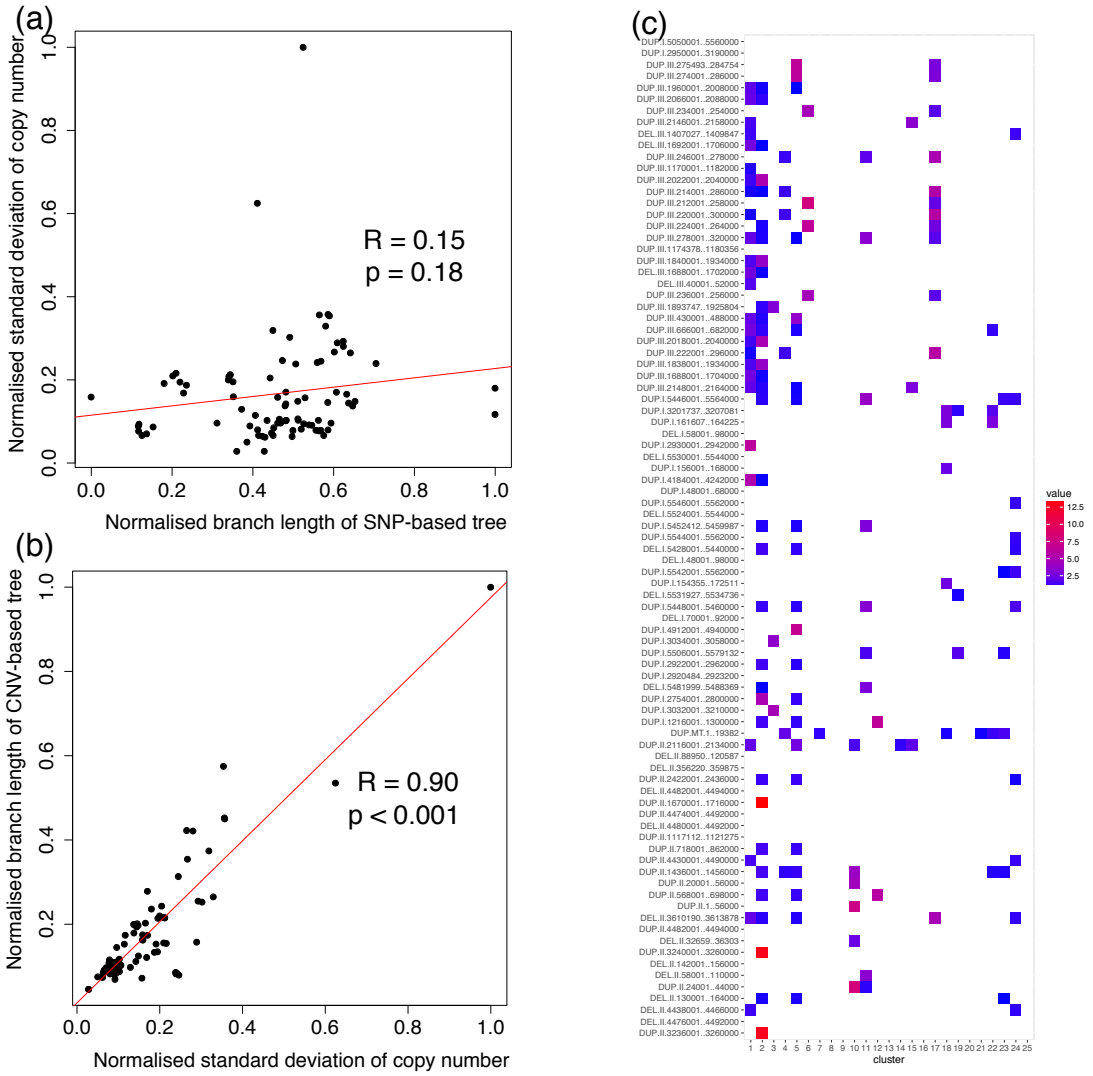
Supplementary Figure 1. Locations and minor allele frequencies of all structural variants in curated data set. Each of the three chromosomes is indicated by black bar, with scale (in megabases) at bottom. From top (same data as Fig 1): density of essential genes (blue), locations of *Tf*-type retrotransposons (green), and diversity (π, average pairwise diversity from SNPs, purple). Bar heights for deletions and duplications are proportional to minor allele frequency, the scale for retrotransposons is the frequency of the insertion in the 57 non-clonal strains. Diversity and retrotransposon were calculated from 57 non-clonal strains as described in Jeffares, et al. [17]. Below, we show different types of SVs: deletions (black), duplications (red), inversions (green) and translocations (blue). The vertical lines terminating with open circles above dotted lines emit from the mid-point of each SV and indicate the minor allele frequencies in the population of 161 strains.



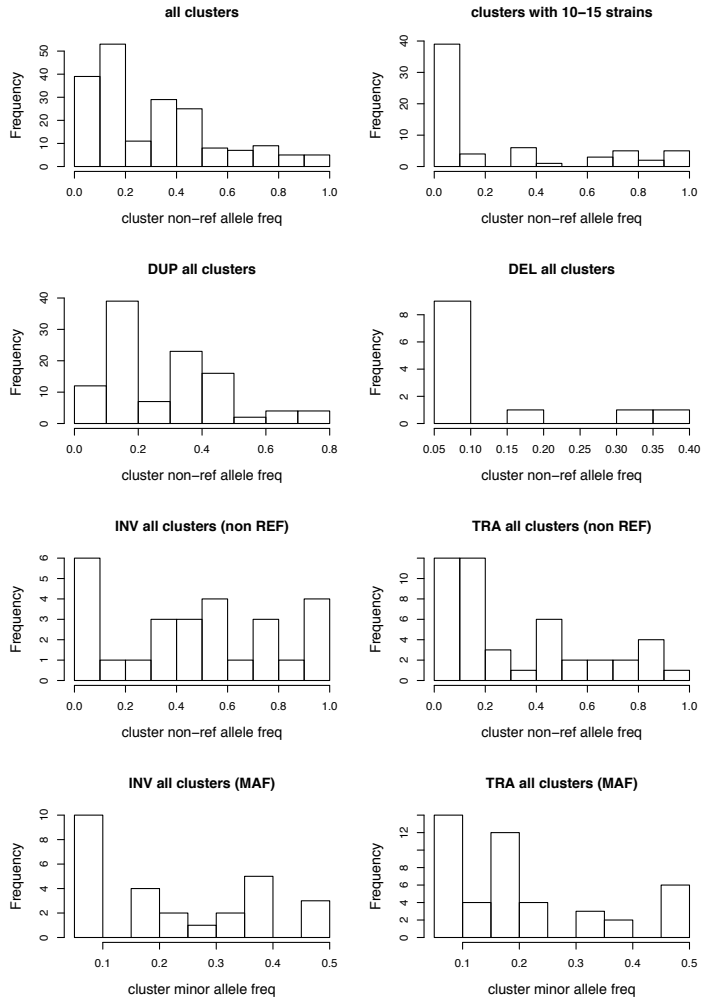
Supplementary Figure 2. Structural variations are biased towards chromosome ends and to low gene density regions. Top left panel, both CNVs and rearrangements are biased towards the ends of chromosomes. CNVs; median distance to chromosome ends 236 kb compared to chromosome- and size-matched random sites 944 kb, Wilcoxon rank sum test P = 1.3 x 10^{-11}, rearrangements median distance 569 kb *vs,* matched random 863 kb, Wilcoxon test P = 0.03). All other panels, calculated proportion of each duplication and deletion that contained all protein-coding or essential genes. Box plots show the distributions of these proportions for all genes (grey), and proportion of coverage by essential genes (red), compared to the null distribution (rand). All comparisons were significantly less than the null distributions (Wilcoxon rank sum test, P-values <1.6 x 10^{-4}). The same analysis was performed with the junctions of inversions and translocations, by calculating the transcript coverage in the region 500 bp up- and down-stream of the predicted start and end junctions. These rearrangements are slightly biased away from genes (P = 1.9x10^{-3}), but not significantly biased away from essential genes (P >0.05). The null distributions were determined by selecting 10 regions for each actual variant/junction that were the same size, and were placed in random positions on the same chromosome and calculating the gene coverage of these regions. Essential genes were those with the Fission Yeast Phenotype Ontology term defined as FYPO:0002061 ("inviable") in PomBase.



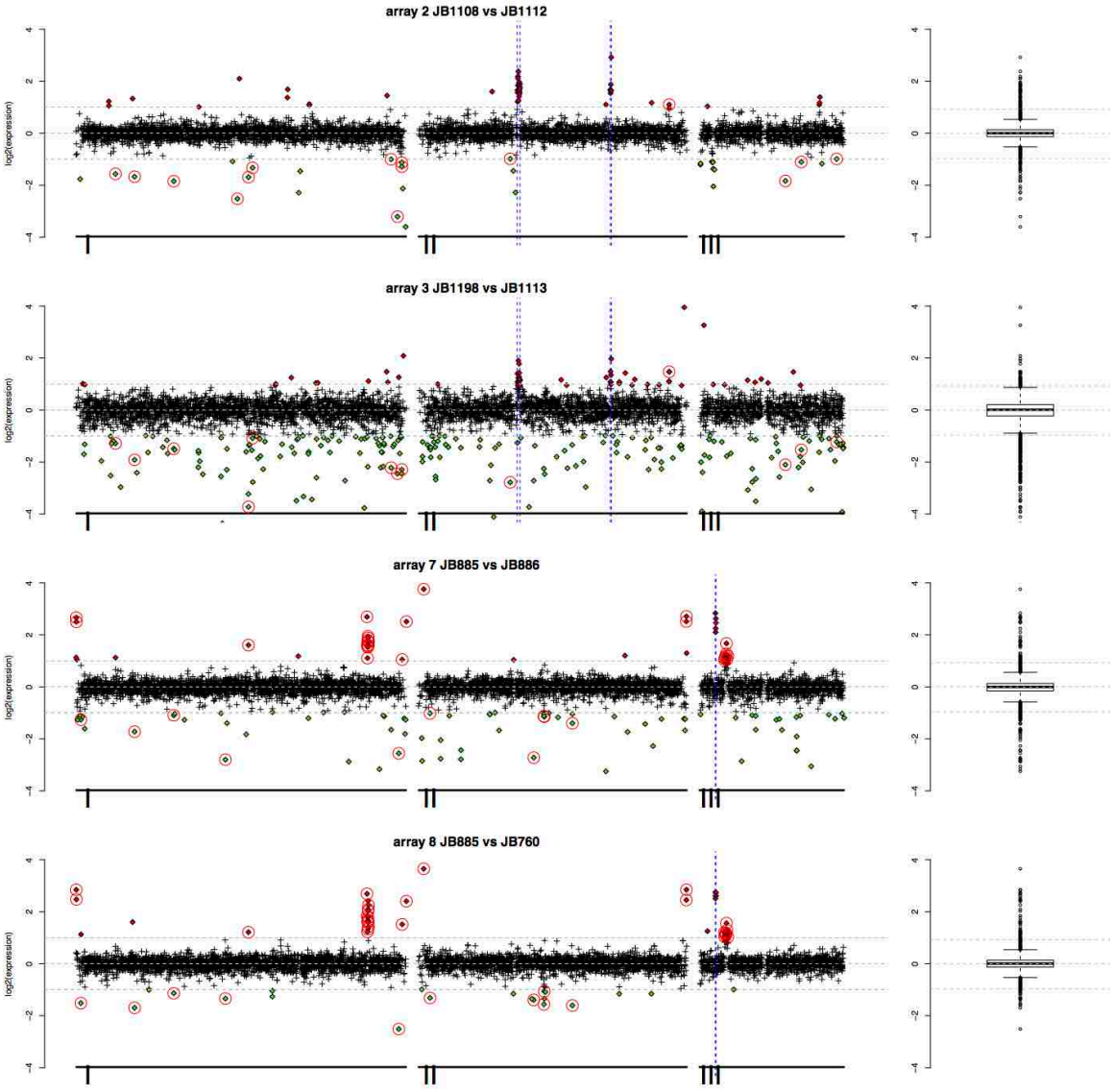
Supplementary Figure 3. Duplications that segregate within closely related strains. Plots show the average coverage in 1 kb non-overlapping windows for strains with a duplication (red) and all closely related strains without duplication (green); all these strains differ by <150 SNPs. The coverage of the two standard reference strains (h^+ and h^-) is shown in black. Top row, from left: variant DUP.I:1216001..1300000 (cluster 12, from Japan in 57), DUP.II:568001..698000 (cluster 12), DUP.II:1670001..1716000 (cluster 2, unknown origin), second row DUP.II:3240001..3260000 (cluster 2), DUP.II:2116001..2134000 (cluster 1, includes reference strain from French grapes in 1947), DUP.III:1838001..1934000 (cluster 2, various locations 1921-22). Bottom row: DUP.III:212001..258000 (cluster 6, Jamaica/USA), and DUP.III:274001..286000 (cluster 5, Sicily 1966). Genes are shown on top of plots with exons as red rectangles and retrotransposon LTRs as blue rectangles.



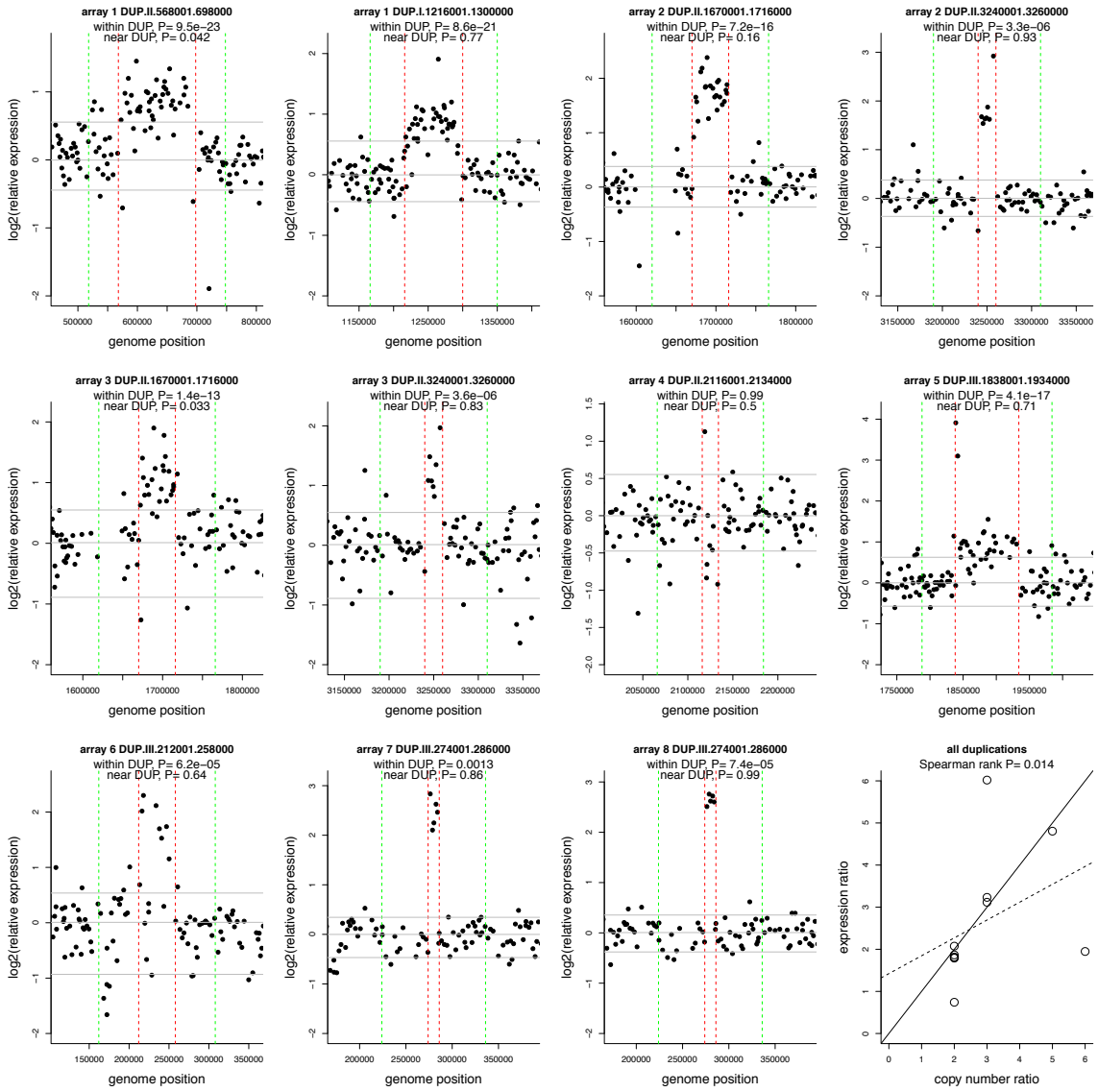
Supplementary Figure 4. Relative standard deviation of copy number variation within clusters. (a) Standard deviation of copy number for a CNV across the dataset is only weakly correlated with the total branch length of SNP-based phylogeny from the 20kb up- and down-stream phylogeny. **(b)** Standard deviation is highly correlated with the branch length of a CNV-based neighbour-joining tree. **(c)** The relative standard deviation of each CNV within each identified cluster of strains (<150 SNPs apart) relative to its change in rest of the dataset. For clarity, all relative standard deviations <1 are shown as white.



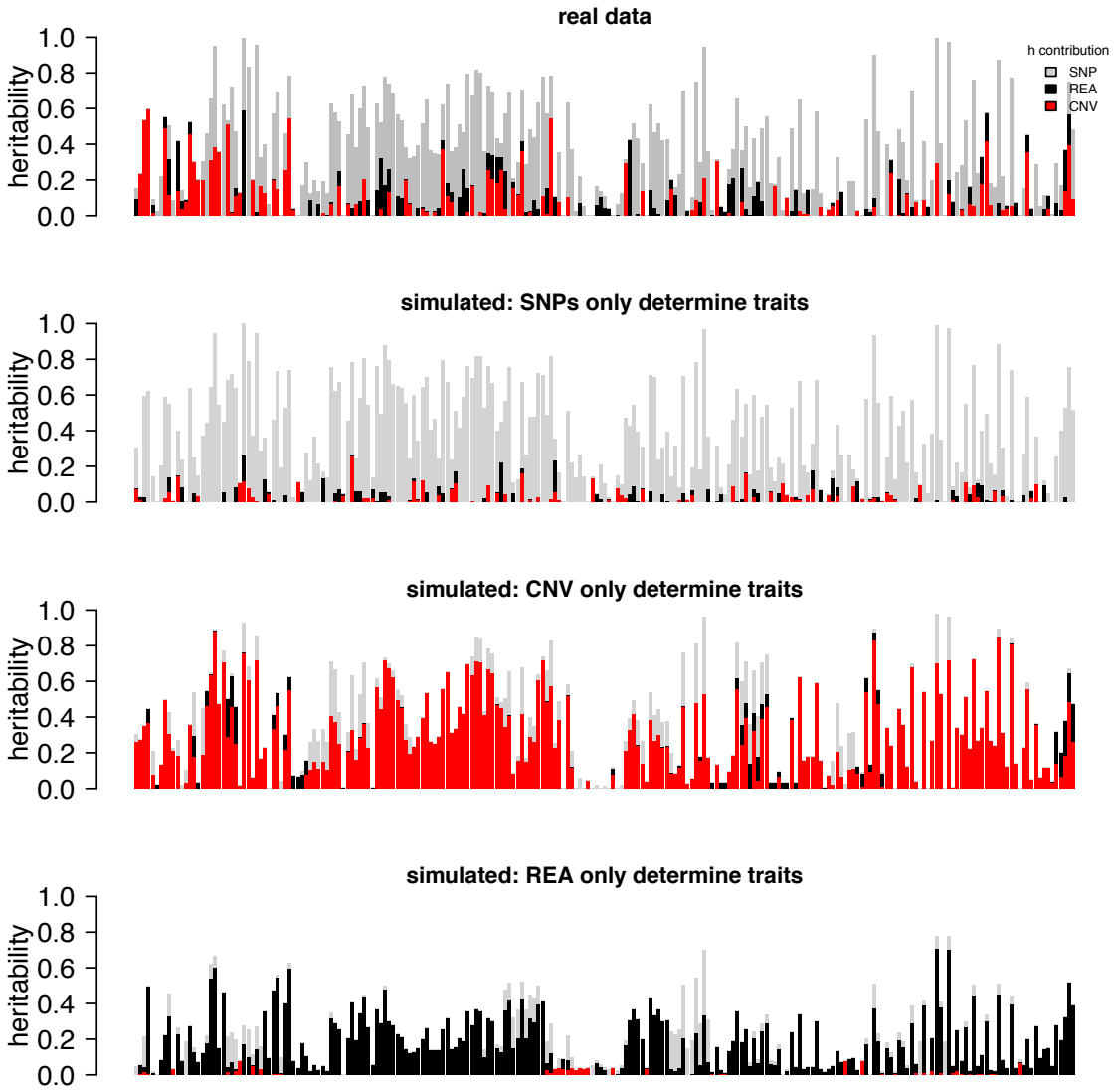
Supplementary Figure 5. Copy number variants are usually rare alleles within clonal populations. Clonal clusters, or clonal populations all differ by < 150 SNPs. In rows, from top left; we show the within-cluster frequency of the non-reference allele for all SVs, which is skewed to rare alleles. Limiting this analysis to cluster with 10 to 15 strains highlights the low frequency of non-reference alleles. Second row; CNVs (duplications and deletions) are skewed to rare alleles, because the non-reference allele is usually the derived allele. Third row; inversions and translocations are not skewed to the non-reference allele, but here non-reference alleles are not necessarily the derived allele. Bottom row; the *minor allele* of inversions and translocations, however, is skewed to rare alleles.



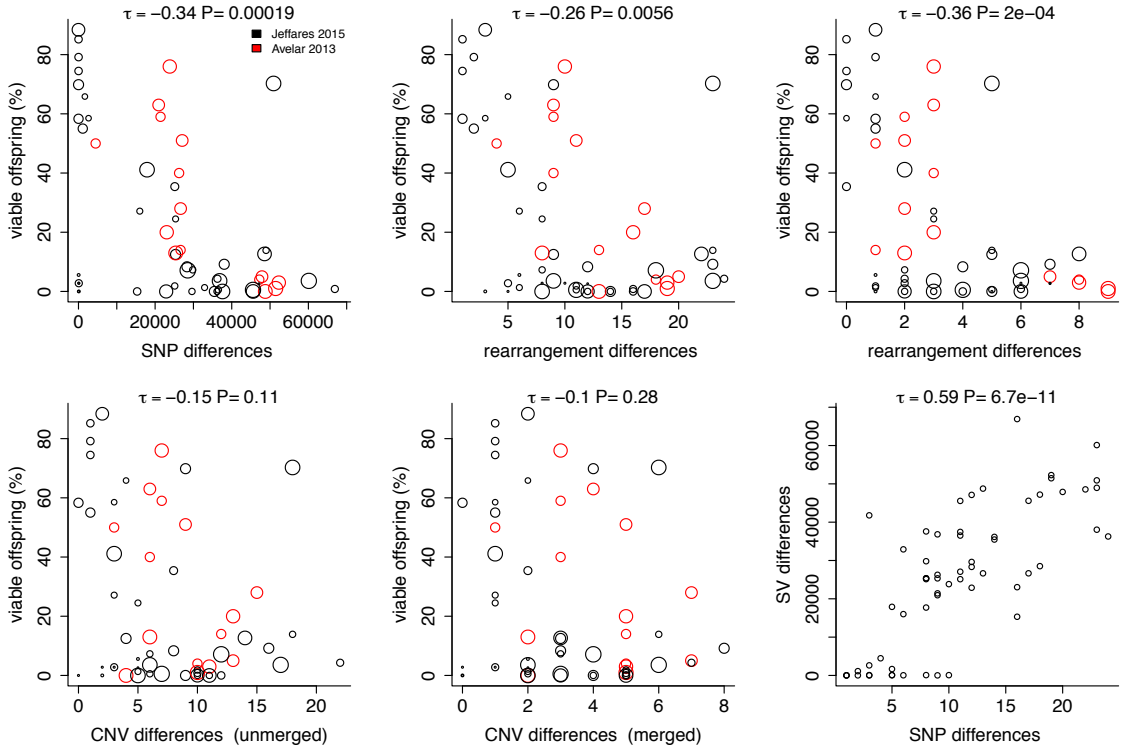
Supplementary Figure 6. Chromosome-scale view of gene expression changes. The relative gene expression levels (strain1/strain2) for arrays 2 and 3, and arrays 7 and 8 are shown with their positions on the three chromosomes. Filled circles indicates genes that we consider to be up-regulated (red) or repressed (green). Those highlighted with open red circles are consistently altered in both arrays (either 2+3, or 7+8). The blue lines show where the segregating duplications are. Box plots at right show the spread of data.



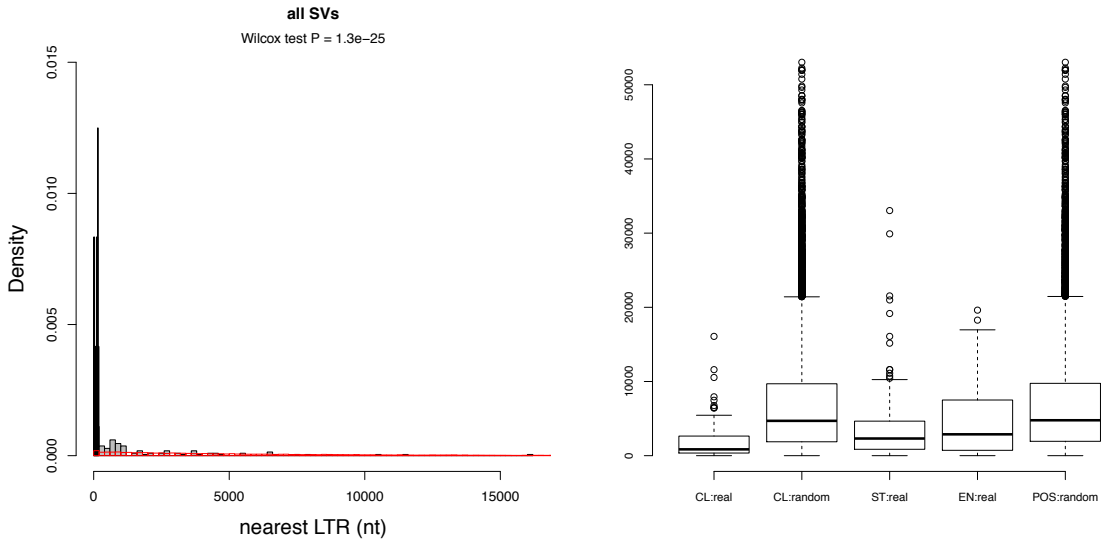
Supplementary Figure 7. No significant increase in gene expression immediately adjacent to duplications. For each duplication examined with DNA arrays, we show the relative expression (strain 1 *vs* strain 2) near the duplication. P-values show the support for the genes within the duplication (red vertical lines), or the 50 kb adjacent to the duplication (green vertical lines) being more highly expressed than all other genes in the chromosome (one-sided Wilcoxon rank sum tests). The grey horizontal lines show the 5th, 50th and 95th percentiles for gene expression data on the chromosome. The bottom right panel shows that the median increase in expression level within a duplication correlates with the increase in genomic copy number. The solid back line shows the expected increase for the 1:1 correspondence between genomic copy number and relative expression (the line $y=x$), and the dashed line shows the linear model for the data. Copy number and relative expression change are correlated (Spearman rank correlation ρ = 0.71 and P = 0.014).



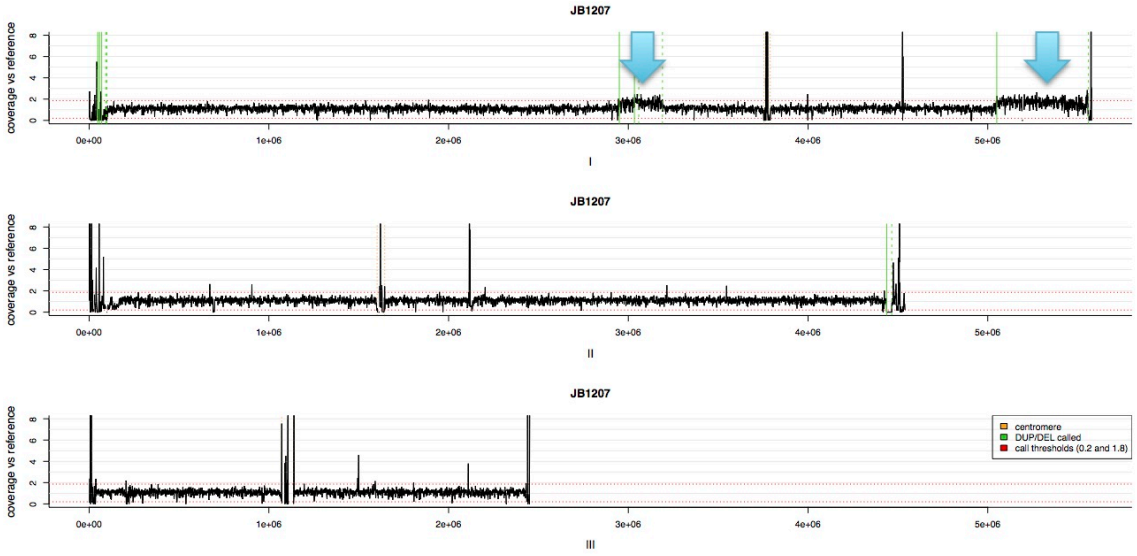
Supplementary Figure 6. Contributions of SNPs, CNVs and rearrangements to traits. Top panel: for 227 traits, we show the total heritability estimated by the combination of 243,289 SNPs (green), 87 CNVs (red), and 26 rearrangements (grey). We then simulated data that was entirely due to the effects of SNPs (second panel), entirely due to the effects of CNVs (next panel) or entirely due to the effects of rearrangements (lower). In the second panel (entirely due to the effects of SNPs), the contribution of CNVs and rearrangements are artefacts, but these are relatively minor. This analysis indicates that the estimates are not strongly affected by linkage.



Supplementary Figure 7. Correlations between spore viability, parental SNP-genetic distance and parental SV-genetic distance. Spore viability was measured for 58 crosses in total, including data from both Jeffares, et al. [17] (black) and Avelar, et al. [6] (red), with each circle representing one cross. Unmerged CNV differences count any CNV as being different between parents when either start or end coordinates are more than 1 kb apart. Because this definition can cause us to count largely overlapping events as 'different', we also counted 'merged' differences where two CNVs were considered different only if their overlap was >50% of the total of both variants. This approach will exclude nested CNVs. CNV-genetic distance is not significantly correlated with viability in either case.



Supplementary Figure 8. SVs are enriched close to retrotransposon LTRs. For all SVs, we computed the closest distance of start or end coordinates to any LTR discovered previously [17]. As a control, we compute the closest distance of 10 random coordinates on the same chromosome. Left: the distributions of distances for real SVs (grey), those that are within 200nt (black) or random coordinates (red). Right: using the same analysis, we show the closest distance of real SVs (CL: real), and random coordinates (C: random). We also show that both start and end coordinates of SVs (ST:real, EN:real) are closer than random positions (POS:random).



Supplementary Figure 9. Chromosome-scale read coverage plots for three chromosomes of strain JB1207. Coverage is calculated relative to the reference strain (JB22 in our collection). Two large duplications that did not satisfy the criteria used to detect CNVs with cn.MOPs are indicated with blue arrows (DUP.I:2950001..3190000, 240kb and DUP.I:5050001..5560000, 510kb).