Transient structural variations alter gene expression and quantitative traits in *Schizosaccharomyces pombe*.

Authors

Daniel C. Jeffares[1,2], Clemency Jolly[1], Mimoza Hoti[1,2], Doug Speed[2], Charalampos Rallis[1,2,3], 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

The effects of structural variants on phenotypic diversity and evolution are poorly understood. We recently described the genetic and phenotypic variation among fission yeast strains and showed that genome-wide association studies are informative in this model. Here we extend this work by systematically identifying structural variations and investigating their consequences. We establish a curated catalog of copy number variants (CNVs) and rearrangements, including inversions and translocations. We find that SVs frequently vary within clonal populations and are weakly tagged by SNPs, consistent with rapid turnover. We show that CNVs produce measurable effects on gene expression both within and outside the duplicated regions. CNVs contribute to quantitative traits such as cell shape, cell growth under diverse conditions, sugar utilization in winemaking and antibiotic resistance, whereas rearrangements are strongly associated with reproductive isolation. Collectively, these findings have broad implications for evolution and for our understanding of quantitative traits and complex human diseases.

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

INTRODUCTION

Structural variations (SVs), such as deletions, duplications, insertions, inversions or translocations, affect organismal phenotypes and complex diseases (Klopocki and Mundlos, 2011, Stefansson et al., 2014, Walters et al., 2013). SVs also influence reproductive isolation within populations of yeast species (Avelar et al., 2013, Brown et al., 2011, Rieseberg, 2001) as well as flies and mosquitoes (Ayala et al., 2013, McGaugh and Noor, 2012). We, and others have recently begun to develop the fission yeast *Schizosaccharomyces pombe* as a model for population genomics and quantitative trait analysis (Avelar et al., 2013, Brown et al., 2011, Fawcett et al., 2014, Jeffares et al., 2015, Zanders et al., 2014). This model organism combines the advantages of a small, well-annotated haploid genome (Wood et al., 2002), abundant tools for genetic manipulation and high-throughput phenotyping (Sabatinos and Forsburg, 2010), and considerable resources of genome-scale and gene-centric data (Kim et al., 2010, Marguerat et al., 2012, Ryan et al., 2012).

One facet of diversity studies in this species has been the discovery and analysis of a small number of inversions and reciprocal translocations (Avelar et al., 2013, Zanders et al., 2014, Brown et al., 2011). Given this evidence for SVs in this species, we recognized that a systematic survey of SVs would progress our understanding of their biological influence. Here, we build upon the recent availability of 161 fission yeast genomes (Jeffares et al., 2015), representing the known population of this species worldwide, to undertake such a survey. We show that rearrangements contribute towards reproductive isolation, while copy number variants (CNVs) alter gene expression levels, contribute significantly to quantitative traits, and are transient within clonal populations.

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) (Klambauer et al., 2012, Layer et al., 2014, Rausch et al., 2012, Ye et al., 2009) to existing short-read data (Jeffares et al., 2015), 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 previously annotated long terminal repeats (LTRs) (Jeffares et al., 2015). Finally, we manually vetted all the remaining SVs by visual inspection of read

alignments in multiple strains for all 315 candidates. This meticulous approach aimed to ensure a high quality call set, to mitigate against the high uncertainty associated with SV calling (Layer et al., 2014).

This curation produced a set of 113 SVs, comprising 23 deletions, 64 duplications, 11 inversions and 15 translocations **(Figure 1a)**. Reassuringly, when our variant calling methods were applied 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 (Methods).

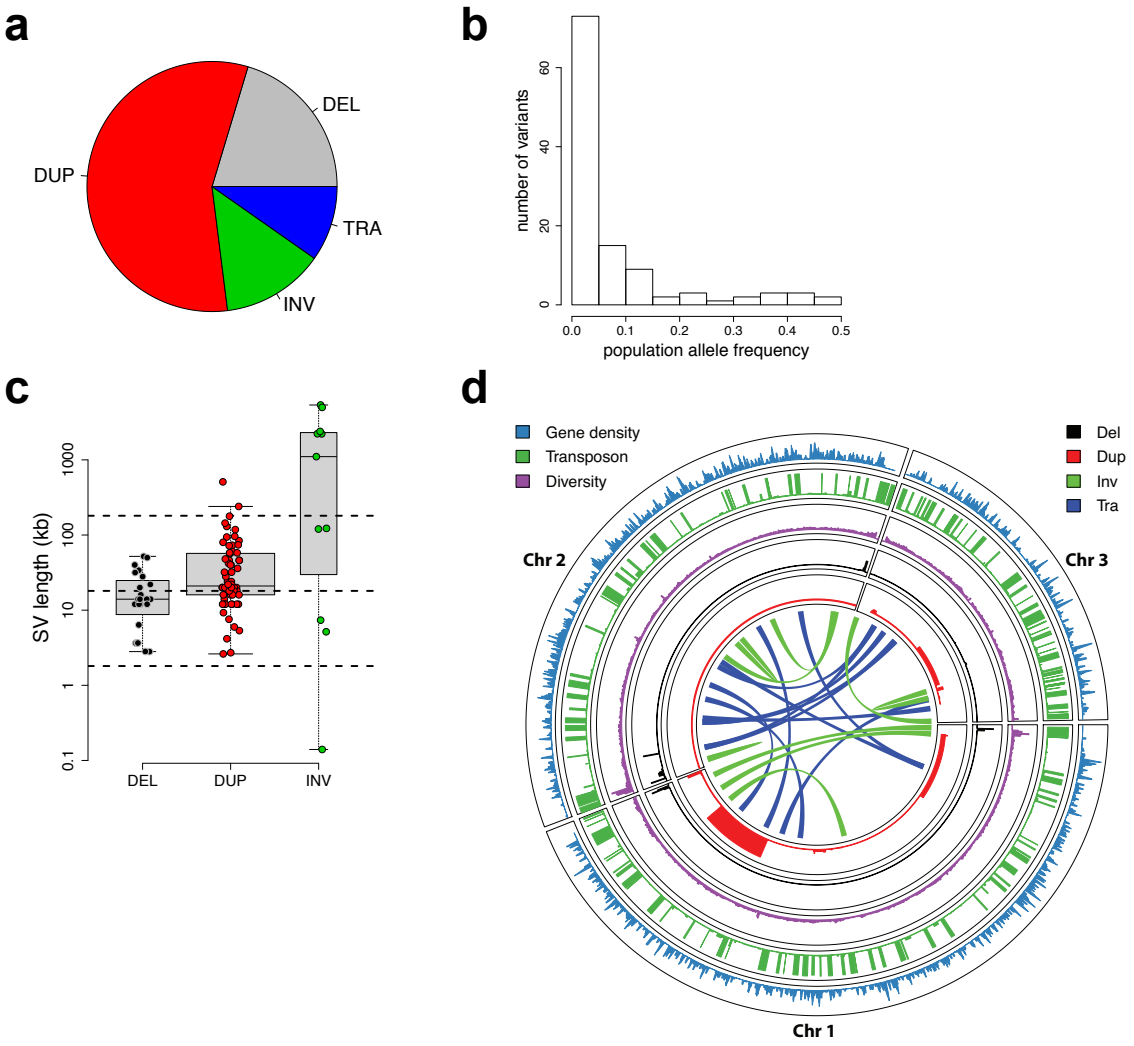


Figure 1. Characteristics of SVs in *S. pombe*. (**a**) Relative proportions of SVs identified. Duplications (DUP) were the most abundant SVs, followed by deletions (DEL), inversions (INV), and translocations (TRA). (**b**) Population allele frequency distribution of SVs, showing the frequencies of less 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 very large, spanning large portions of chromosomes (1.04 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. (2015).

Most SVs were present at low frequencies, with 28% discovered in only one of the strains analyzed **(Figure 1b)**. The deletions were generally small (median length 595bp, **Figure 1c)**, duplications showed a median length of 20 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').

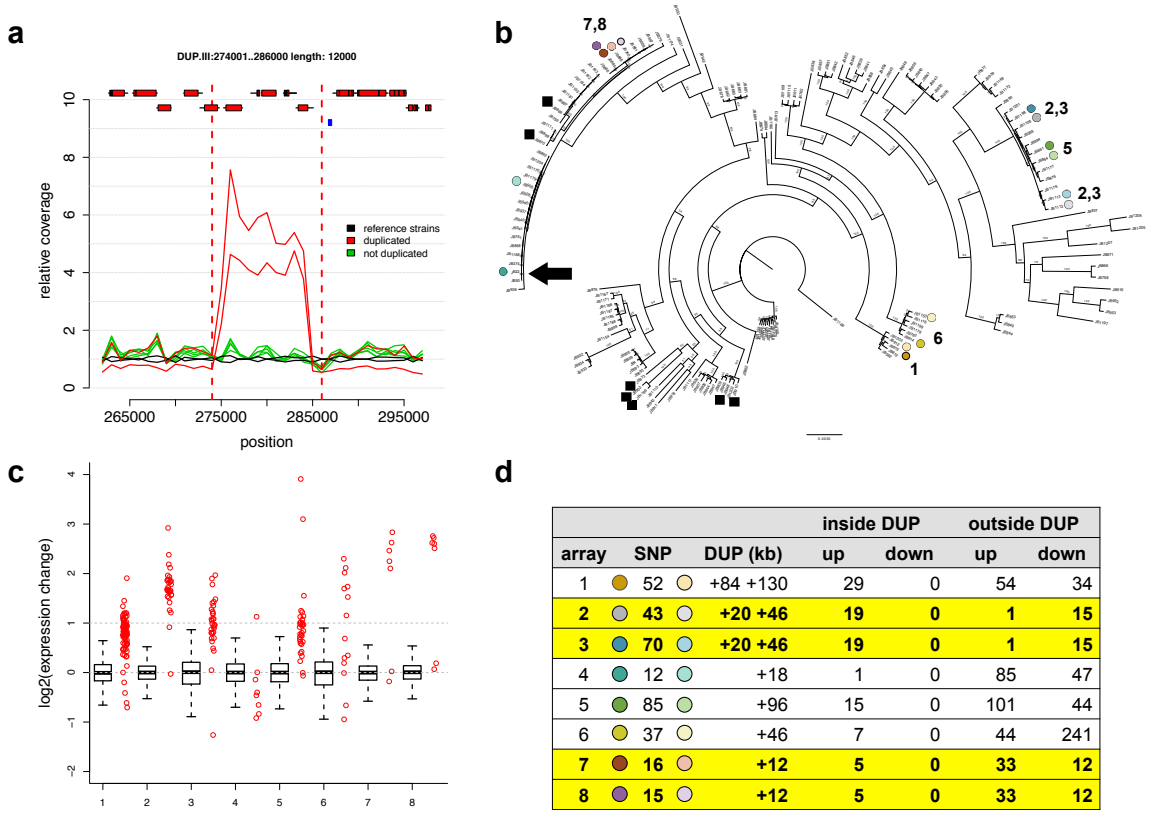
Deletions and duplications and strongly biased towards the ends of chromosomes (**Figure 1d, Supplementary Figure 1**), which are characterized by high genetic diversity, frequent transposon insertions, and a paucity of essential genes (Jeffares et al., 2015). All SVs preferentially occurred in positions of low gene density and showed a strong tendency to not overlap with 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 (Bitton et al., 2015). 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 DNA polymerase II occupancy and were less likely to contain genes that are known to produce abnormal cell phenotypes ($P= 1.8x10^{-5}$ and $3x10^{-5}$, respectively). These analyses are all broadly consistent with a paucity of CNVs in genes that encode essential mitotic functions. Rearrangements disrupted only a few genes and showed no significant enrichments.

Duplications are transient within clonal populations

Our previous work identified 25 clusters of near-clonal strains, which differed by <150 SNPs within each cluster (Jeffares et al., 2015). We expect that these clusters reflect either repeat depositions of strains differing only at few sites (e.g. mating-type variants of reference strains h^{90} and h^{-} differ by 14 SNPs) or natural populations of strains collected from the same location. Such 'clonal populations' reflect products of mitotic propagation from a very recent common ancestor, without any outbreeding. We therefore expected that SVs should be largely shared within these clonal populations.

Surprisingly, our genotype predictions indicated that most SVs present in clonal populations were segregating, i.e. were not fixed within the clonal population (68/95 SVs, 72%). Furthermore, we observed instances of the same SVs that were present in two or more different clonal populations that were not fixed within any clonal population. These SVs could be either incorrect allele calls in some strains, or alternatively, recent events that have emerged during mitotic propagation. To distinguish between these two

147 scenarios, we re-examined the read coverage of all 49 CNVs present within at least one
148 clonal population. Since translocations and inversions were more challenging to
149 accurately genotype, we did not re-examine these variants. This analysis verified that 40
150 of these 49 CNVs (37 duplications, 3 deletions) were clearly segregating within at least
151 one clonal cluster **(Figure 2a, Supplementary Figure 3)**. For example, one clonal
152 population of seven closely related strains, collected together in 1966 from grape must in
153 Sicily, have an average pairwise difference of only 19 SNPs (diversity $\pi = 1.5x10^{-6}$).
154 Notably, this collection showed four non-overlapping segregating duplications. This
155 striking finding indicates that CNVs can arise or disappear frequently during evolution.
156
157 **Transient duplications affect gene expression**
158 Partial aneuploidies of 500-700 kb in the *S. pombe* reference strain are known to alter
159 gene expression levels within and, to some extent, outside of the duplicated region
160 (Chikashige et al., 2007). The naturally occurring duplications we described are typically
161 smaller (median length: 46 kb), including an average of 6.5 genes. To examine whether
162 naturally occurring CNVs have similar effects on gene expression, we examined eight
163 pairs of closely related strains (<150 SNPs among each pair) that contained at least one
164 unshared duplication **(Figure 2b, Supplementary Figure 3, Supplementary Table 2)**.
165 Several of these strain pairs have been isolated from the same substrate at the same time,
166 and all pairs are estimated to have diverged approximately 50 to 65 years ago
167 (**Supplementary Table 2**). We assayed transcript expression from log phase cultures
168 using DNA microarrays, each time comparing a duplicated to a non-duplicated strain
169 from within the same clonal population. In seven of the eight strain pairs, the expression
170 levels of genes within duplications were significantly induced, although the degree of
171 expression changes between genes was variable **(Figure 2c)**. The increased transcript
172 levels correlated with the increased genomic copy numbers, so that higher copy numbers
173 produced correspondingly more transcripts (**Supplementary Figure 4**). No changes in
174 gene expression were evident immediately adjacent to the duplications (**Supplementary**
175 **Figure 4)**, suggesting that the local chromatin state was not strongly altered by the
176 CNVs.
177 Some genes outside the duplicated regions also showed altered expression levels
178 **(Figure 2d, Supplementary Table 3)**. For example, two strain pairs differ by a single 12
179 kb duplication. Here, five of seven genes within the duplication showed induced
180 expression, while 45 genes outside the duplicated region also showed consistently altered
181 expression levels (38 protein-coding genes, 7 non-coding RNAs) (**Figure 2d**, arrays 7
182 and 8). As environmental growth conditions were tightly controlled, these changes in
183 gene expression probably reflect indirect and compensatory effects of the initial
184 perturbation caused by the duplication (**Supplementary Figure 5**). We conclude that
185 these evolutionary unstable duplications reproducibly affect the expression of distinct sets
186 of genes and thus have the potential to influence cellular function and phenotypes.



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

Figure 2. 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 2** 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. Black squares indicate the presence of a 32 kb duplication that is associated with Brefeldin resistance (see below). (**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 $<1.5\times10^{-3}$). (**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 length X and length 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 influence quantitative traits

To test whether SVs affect phenotypes, we examined the contributions of SNPs, CNVs and rearrangements to 53 quantitative traits, including 11 cell shape parameters and colony size on solid media assaying 42 stress and nutrient conditions (Jeffares et al., 2015). 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).

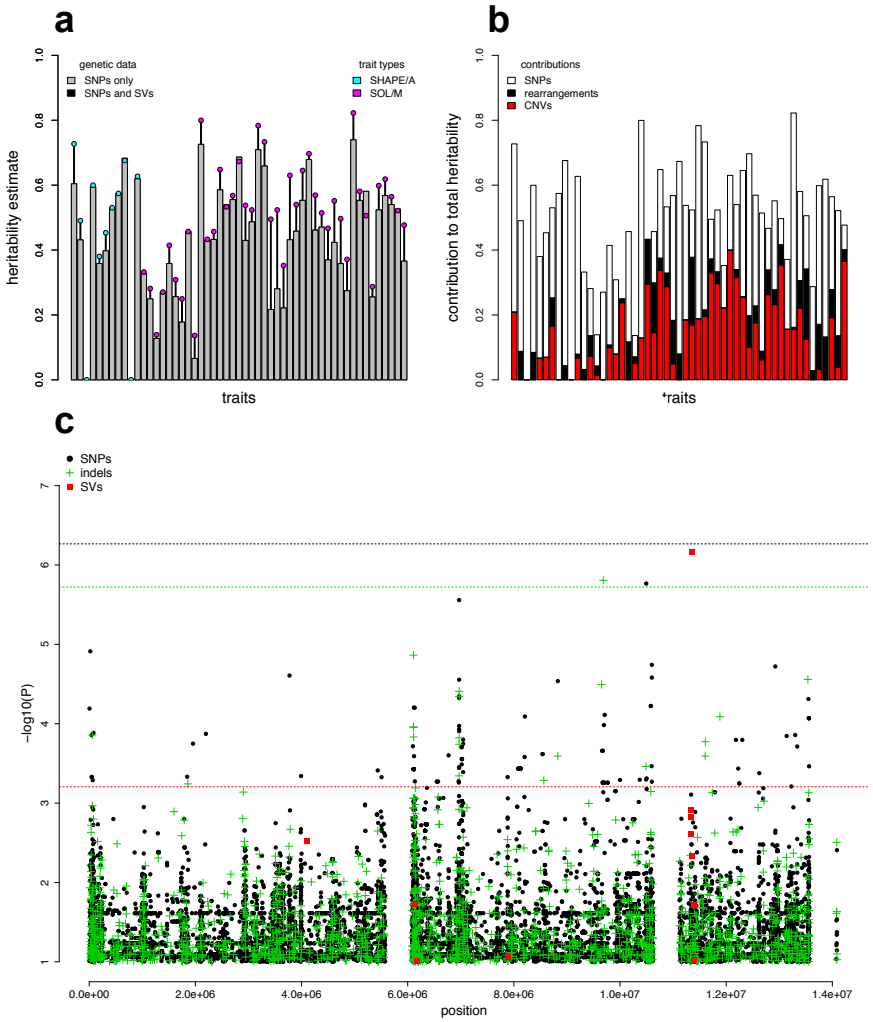
When we determined heritability using only SNP data, estimates varied between 0% and 74%. After adding CNVs and rearrangements to SNPs in a composite model, the estimated overall heritability increased for nearly all traits, in some cases by more than 2-fold (e.g., resistance to Rapamycin) (**Fig 3a**). 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 (**Fig 3b**). On average, SNPs explained 30% of trait variance, CNVs 14%, and rearrangements 5% (**Supplementary Figure 6, 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**).

As some of the strains have recently been shown to have fermentation properties that may be beneficial for winemaking (Benito et al., 2016), we examined three traits related to wine fermentations (glucose/fructose utilization, malic acid degradation, acetic acid content, all measured in grape must). Remarkably, the heritability of these wine-fermentation traits was almost entirely due to SVs, with negligible contributions from SNPs (**Supplementary Figure 7, Supplementary Table 6**). For glucose/fructose utilization, the CNVs accounted for the entire heritability of 0.53 (**Supplementary Figure 8**). Since many of these strains have been collected from fermentations (**Supplementary Table 7**), the strong influence of CNVs may represent recent strong selection and adaptation to fermentation conditions that has occurred via recent CNV acquisition.

To locate specific SVs that affected these traits, we performed mixed model genome-wide association studies, using all 68 SVs with minor allele counts >5 as well as 139,396 SNPs and 22,058 indels. Trait-specific significance thresholds for 5% familywise error rates were computed via permutation analysis, and were approximately

248 10^{-4} (SVs), 10^{-6} (SNPs and indels). Five SVs were significantly associated with traits (3
249 duplications, 1 deletion, 1 translocation) (**Supplementary Table 8**). The median effect
250 size of these SVs was 9% (range 6-13%). The strongest signal was from a 32kb
251 duplication that covered 11 protein-coding genes and was significantly associated with 15
252 growth traits. This duplication is segregating (not fixed) within three clonal populations
253 (**Fig. 2b, Supplementary Figure 9**), often as part of a larger duplication. It was the most
254 significantly associated variant (from SNPs, indels and SVs) for growth in the antibiotic
255 Brefeldin, where it contributes 15% of the trait variance (**Fig 3c**). Three of the 11
256 duplicated genes encode transmembrane proteins, any of which could contribute to the
257 trait. Since some fungi produce Brefeldin to inhibit competitive growth, this duplication
258 is a striking example of a transient CNV that could provide a strong selective advantage
259 in a natural environment.
260 Our analysis of heritability showed that SNPs are generally able to capture most
261 of the genetic contribution of SVs (**Fig. 3**). To examine whether trait-influencing SVs
262 will be effectively detected by tagging SNPs in in this population, we examined the
263 linkage of all 113 SVs with SNPs. We found that only 63 of these SVs (55%) are in
264 strong linkage (r^2 >0.6), leaving 45% of the SVs weakly linked (**Supplementary Table**
265 **16**). This lack of linkage is consistent with SVs being transient, rather than persisting
266 within haplotypes. Such weakly-linked SVs will not be tagged well by SNPs (as shown
267 in **Fig. 3c**). Collectively, these analyses indicate that SVs, most notably CNVs, contribute
268 substantially to quantitative traits and suggests that GWAS analyses conducted without
269 genotyping SVs could fail to capture these genetic factors.
270

271



272
273

274 **Figure 3. CNVs contribute to quantitative traits.** (**a**) Heritability estimates obtained
275 using SNPs alone (grey bars) are generally lower than estimates obtained using SNPs,
276 CNVs and rearrangements (black lines extending above bars), consistent with CNVs and
277 rearrangements contributing to traits. The types of traits described are indicated by the
278 filled circles, including automated shape parameters (SHAPE/A), and growth rates on
279 solid media (SOL/M). (**b**) Contributions of CNVs (red), rearrangements (black) and
280 SNPs (open bars) to total heritability are shown with a stacked bar plot. CNVs in
281 particular make substantial contributions that will appear be associated with non-causal
282 SNPs if the causal CNVs have not been mapped in the genetic data. (**c**) Manhattan plot
283 for all variants vs. resistance to Brefeldin. Only variants with association P-values <0.1
284 are shown, including SNPs (black filled circles), indels (green crosses), and SVs (red
285 squares). Variants from the three chromosomes are shown, from left to right, followed by
286 variants from the mitochondrial genome. The 32 kb duplication on chromosome III is the
287 most significant variant. The horizontal dashed lines show the significance threshold

288 derived from 1000 permutations for SNPs (black), indels (green) and SVs (red). Each is
289 set to ensure a 5% family-wise error rate. The trait measured in this case was colony size
290 in the presence of 80 μM Brefeldin (as a ratio of colony size without Brefeldin). Similar
291 results, including a significant P-value for the 32kb duplication were found with 40 μM
292 and 120 μM Brefeldin (**Supplementary Table 8**).
293
294
295 **Structural variations contribute to intrinsic reproductive isolation**
296 Crosses between *S. pombe* strains produce between 90% and < 1% viable offspring
297 (Avelar et al., 2013, Zanders et al., 2014). We have previously shown that spore viability
298 correlates inversely with the number of SNPs between the parental strains (Jeffares et al.,
299 2015). This intrinsic reproductive isolation may be due to the accumulation of
300 Dobzhansky-Muller incompatibilities (variants that are neutral in one population, but
301 incompatible when combined) (Dobzhansky, 1933, Muller, 1939). However, genetically
302 distant strains also accumulate SVs, which are known to lower hybrid viability and drive
303 reproductive isolation (Rieseberg, 2001). In *S. pombe,* engineered inversions and
304 translocations reduce spore viability by ~40% (Avelar et al., 2013).
305 To analyse intrinsic reproductive isolation in more detail, we examined the
306 relationship between viability, SNPs and SVs. Both SV-distance (number of unshared
307 SVs between parents) and SNP-distance inversely correlated with hybrid viability
308 (Kendall correlation coefficients, SVs: $\tau = -0.26$, $P = 5.6 \times 10^{-3}$, SNPs: $\tau = -0.35$, $P =$
309 1.6×10^{-4}) (**Supplementary Figure 10**). While inversions and translocations are known to
310 lower hybrid viability as they affect chromosome pairing and segregation during meiosis
311 (Avelar et al., 2013, Delneri et al., 2003, Zanders et al., 2014), CNVs are not expected to
312 influence spore viability. Consistent with this view, we found that rearrangements
313 explained spore viability better than CNVs (rearrangements, $\tau = -0.36$, $P = 2.0\times10^{-4}$;
314 CNVs, $\tau = -0.10$, $P = 0.28$).
315 As the numbers of SNP and rearrangement differences between mating parents
316 are themselves correlated ($\tau = 0.53$, $P = 1.3\times10^{-8}$), we also estimated the influence of each
317 factor alone using partial correlations. When either SNPs or rearrangements were
318 controlled for, both remained significantly correlated with offspring viability ($P = 0.04$, P
319 $= 0.02$, respectively) (**Figure 4**). Taken together, these analyses indicate that both
320 rearrangements and SNPs contribute to reproductive isolation, but CNVs do not.
321

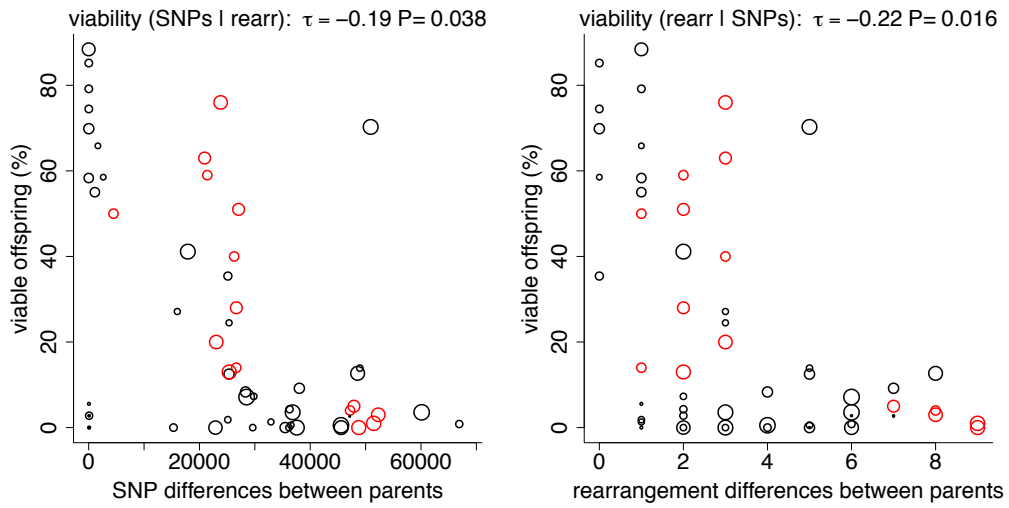


Figure 4. Both SNPs and rearrangements contribute to intrinsic reproductive isolation. Spore viability was measured from 58 different crosses from Jeffares et al. (2015) (black) or Avelar et al. (2013) (red), with each circle in the plots representing one cross. An additive linear model incorporating both SNP and rearrangement differences showed highly significant correlations with viability ($P = 1.2x10^{-6}$, $r^2 = 0.39$). Both genetic distances measured using SNPs and rearrangements (inversions and translocations) significantly correlated with viability when controlling for the other factor (Kendall partial rank order correlations with viability SNPs|rearrangements $\tau = -0.20$, $P = 0.03$; rearrangements|SNPs $\tau = -0.22$, $P = 0.02$). Some strains produce low viability spores even when self-mated with their own genotype. The lowest self-mating viability of each strain pair is indicated by circle size (smaller circles indicate lower self-mating viability) to illustrate that low-viability outliers tend to include such cases (see **Supplementary Table 9** for details).

DISCUSSION

Here we present the first genome- and population-wide catalog of SVs among *S. pombe* strains. To account for the high discrepancy of available methods (Layer et al., 2014), we applied a consensus approach to identify SVs (SURVIVOR), followed by rigorous filtering and manual inspection of all calls. We focused on high specificity (the correctness of the inferred SV) rather than high sensitivity (attempting to detect all SVs).

Our previous analyses of these strains, conducted without SV data (Jeffares et al., 2015), attributed both trait variations and reproductive isolation to SNPs and/or small indels. Here we show that the small number of SVs we describe make substantial contributions to both of these factors. We demonstrate that CNVs (duplications and deletions) contribute significantly to our ability to describe quantitative traits, whereas variants that rearrange the order of the genome (inversions and translocations) produce much weaker effects on traits. In contrast, CNVs have no detectable influence on

351 reproductive isolation, while rearrangements contribute substantially to reproductive
352 isolation, similar to other species (Ayala et al., 2013, Noor et al., 2001).
353 A surprising aspect of our analysis is that duplications are generated and/or lost
354 frequently within clonal populations. For example, within seven strains that differ by as
355 few as 19 SNPs, we discovered four segregating duplications. A similar rapid occurrence
356 of duplications in *S. pombe* has been observed in laboratory conditions, where
357 spontaneous duplications suppress *cdc2* mutants at least 100 times more frequently than
358 SNPs, and these suppressor strains lose their duplications with equal frequency (Carr et
359 al., 1989). Similarly, duplications frequently occur during experimental evolution with
360 budding yeast (Dunham et al., 2002). Consistent with the transience of these variants,
361 they are frequently not well tagged by SNPs. These CNVs subtly alter the expression of
362 genes within and beyond the duplications, and contribute considerably to quantitative
363 traits. Within small populations, CNVs may produce larger effects on traits in the short
364 term than SNPs, as demonstrated by the 32 kb duplication that is associated with
365 resistance to the Brefeldin A (an antibiotic produced by entophytic fungi).
366 This analysis has relevance for human diseases, since *de novo* CNV formation in
367 the human genome occurs at a measurable rate (approximately one CNV/10 generations
368 (Itsara et al., 2010)), and CNVs are known to contribute to a wide variety of diseases
369 (Zhang et al., 2009). Indeed, both the population genetics and the effects of SVs within *S.*
370 *pombe* seem similar to human, in that CNVs are associated with stoichiometric changes
371 on gene expression, and SVs are in weak linkage with SNPs (Stranger et al., 2007,
372 Sudmant et al., 2015), and therefore may be badly tagged by SNPs in GWAS studies. We
373 show that CNVs in fission yeast contribute to quantitative traits and are also weakly
374 linked to SNPs. These findings highlight the need to identify SVs when describing traits
375 using GWAS, and indicate that a failure to call SVs can lead to an overestimation of the
376 impact of SNPs to traits or contribute to the problem that large proportions of the
377 heritable component of trait variation are not discovered in GWAS (the 'missing
378 heritability').
379 In summary, we show that different types of SVs have distinct influences at the
380 phenotype level and that a small number of SVs produce profound effects on the biology
381 of this species.

382 **MATERIAL AND METHODS**

383 **Performance assessment of SV callers using simulated data**

384 To identify filtering parameters for DELLY, LUMPY and Pindel for the *S. pombe*
385 genome, we simulated seven datasets (s1-s7) of 40x coverage with a range of different
386 SV types and sizes (**Supplemental Table 7**). The simulated read sets contained
387 sequencing errors (0.4%), SNPs and indels (0.1%) within the range of actual data from *S.*
388 *pombe* strains and between 30 and 170 SVs. These data sets were produced by modifying
389 the reference genome using our in-house software (SURVIVOR, described below), and
390 simulating reads from this genome with Mason software (Holtgrewe, 2010).
391 After mapping the reads and calling SVs, we evaluated the calls. We defined a SV
392 correctly predicted if: i) the simulated and reported SV were of the same type (e.g.
393 duplication), ii) were predicted to be on same chromosome, and iii) their start and stop
394 locations were with 1 kb. We then defined caller-specific thresholds to optimize the
395 sensitivity and false discovery rate (FDR) for each caller. FDRs on the simulated data
396 were low: DELLY (average 0.13), LUMPY (average 0.06) and Pindel (average 0.04).
397 Selecting calls that were present in at least two callers further reduced the FDR
398 (average of 0.01). DELLY had the highest sensitivity (average 0.75), followed by
399 SURVIVOR (average 0.70), LUMPY (average 0.62) and Pindel (0.55). We further used
400 simulated data to assess the sensitivity and FDR of our predictions. cn.mops was
401 evaluated with a 2 kb distance for start and stop coordinates. Our cn.mops parameters
402 were designed to identify large (above 12 kb) events and thus did not identify any SVs
403 simulated for s1-s6. Details of simulations and caller efficacy are provided in
404 **Supplementary Table 10**.
405

406 **SURVIVOR (StructURal Variant majorIty VOte) Software Tool**

407 We developed the SURVIVOR tool kit for assessing SVs for short read data that contains
408 several modules. The first module simulates SVs given a reference genome file (fasta)
409 and the number and size ranges for each SV (insertions, deletions, duplications,
410 inversions and translocations). After reading in the reference genome, SURVIVOR
411 randomly selects the locations and size of SV following the provided parameters.
412 Subsequently, SURVIVOR alters the reference genome accordingly and prints the so
413 altered genome. In addition, SURVIVOR provides an extended bed file to report the
414 locations of the simulated SVs.
415 The second module evaluates SV calls based on a variant call format (VCF) file
416 (Danecek et al., 2011) and any known list of SVs. A SV was identified as correct if i)
417 they were of same type (e.g. deletion); ii) they were reported on same chromosome, and
418 iii) the start and stop coordinates of the simulated and identified SV were within 1 kb
419 (user definable).
420 The third module of SURVIVOR was used to filter and combine the calls from
421 three VCF files. In our case, these files were the results of DELLY, LUMPY and Pindel.

422 This module includes methods to convert the method-specific output formats to a VCF
423 format. SVs were filtered out if they were unique to one of the three VCF files. Two SVs
424 were defined as overlapping if they occur on the same chromosome, their start and stop
425 coordinates were within 1 kb, and they were of the same type. In the end, SURVIVOR
426 produced one VCF file containing the so filtered calls. SURVIVOR is available at
427 github.com/fritzsedlazeck/SURVIVOR.
428
429 **Read mapping and detection of structural variants**
430 Illumina paired-end sequencing data for 161 *S. pombe* strains were collected as described
431 in Jeffares et al. (2015), with the addition of Leupold's reference 975 h^+ (JB32) and
432 excluding JB374 (known to be a gene-knockout version of the reference strain, see
433 below). Leupold's 968 h^{90} and Leupold's 972 h^- were included as JB50 and JB22,
434 respectively (**Supplementary Table 7**). For all strains, reads were mapped using
435 NextGenMap (version 0.4.12)(Sedlazeck et al., 2013) with the following parameter (-X
436 1000000) to the *S. pombe* reference genome (version ASM294v2.22). Reads with 20 base
437 pairs or more clipped were extracted using the script *split_unmapped_to_fasta.pl*
438 included in the LUMPY package (version 0.2.9)(Layer et al., 2014) and were then
439 mapped using YAHA (version 0.1.83)(Faust and Hall, 2012) to generate split-read
440 alignments. The two mapped files were merged using Picard-tools (version 1.105)
441 (http://broadinstitute.github.io/picard), and all strains were then down-sampled to 40x
442 coverage using Samtools (version 0.1.18)(Li et al., 2009).
443 Subsequently, DELLY (version 0.5.9, parameters: " –q 20 -r")(Rausch et al.,
444 2012), LUMPY (version 0.2.9, recommended parameter settings)(Layer et al., 2014) and
445 Pindel (version 0.2.5a8, default parameter)(Ye et al., 2009) were used to independently
446 identify SVs in the 161 strains using our SURVIVOR software. We then retained all
447 variants predicted by at least two methods. These SVs calls were genotyped using
448 DELLY.
449 To identify further CNVs, we ran cn.MOPS (Klambauer et al., 2012) with
450 parameters tuned to collect large duplications/deletions as follows: read counts were
451 collected from bam alignment files (as above) with *getReadCountsFromBAM* and
452 WL=2000, and CNVs predicted using *haplocn.mops* with minWidth= 6, all other
453 parameters as default. Hence, the minimum variant size detected was 12 kb. CNV were
454 predicted for each strain independently by comparing the alternative strain to the two
455 reference strains (JB22, JB32) and four reference-like strains that differed from the
456 reference by less than 200 SNPs (JB1179, JB1168, JB937, JB936).
457 After CNV calling, allele calling was achieved by comparing counts of coverage
458 in 100bp windows for the two reference strains (JB22, JB32) to each alternate strain
459 using custom R scripts. Alleles were called as non-reference duplications if the one-sided
460 Wilcoxon rank sum test p-values for both JB22 and JB32 vs alternate strain were less
461 than $1x10^{-10}$ (showing a difference in coverage) and the ratio of alternate/reference

462 coverage (for both JB22 and JB32) was >1.8 (duplications), or <0.2 (deletions). Manual
463 inspection of coverage plots showed that the vast majority of the allele calls were in
464 accordance with what we discerned by eye. These R scripts were also used to examine
465 CNVs predicted to be segregating within clusters (clonal populations). All such CNVs
466 were examined in all clusters that contained at least one non-reference allele call
467 (**Supplementary Table 11**).
468 Finally, we manually mapped two large duplications that did not satisfy these
469 criteria (DUP.I:2950001..3190000, 240kb and DUP.I:5050001..5560000, 510kb – both
470 singletons in JB1207), but were clearly visible in chromosome-scale read coverage plots
471 **(Supplementary Figure 11).**
472
473 **Reduction of false discovery rate**
474 This filtering produced 315 variant calls. However, because 31 of these 315 (~10%) were
475 called within the two reference strains (JB22, JB32), we expected that this set still
476 contained false positives. To further reduce the false positive rate, we looked for
477 parameters that would reduce calls made in reference strains (JB22 and JB32) but not
478 reduce calls in strains more distantly related to the reference (JB1177, JB916 and JB894
479 that have 68223, 60087 and 67860 SNP differences to reference(Jeffares et al., 2015)).
480 The reasoning was that we expected to locate few variants in the reference, and more
481 variants in the more distantly related strains. This analysis showed that paired end
482 support, repeats and mapping quality were of primary value.
483 We therefore discarded all SVs that had a paired end support of 10 or less. In
484 addition, we ignored SVs that appeared in low mapping quality regions (i.e. regions
485 where reads with MQ=0 map) or overlapped with previously identified retrotransposon
486 LTRs(Jeffares et al., 2015).
487 Finally, to ensure a high specificity call set, these filtered SVs were manually
488 curated using IGV (Thorvaldsdottir et al., 2013) (**Supplementary Tables 12,13**). We
489 assigned each SVs a score (0: not reliable, 1: unclear, 2: reliable based on inspection of
490 alignments through IGV). Only calls passing this manual curation as reliable (score 2)
491 were included in the final data set of 113 variants utilized for all further analyses.
492 These filtering and manual curation steps reduced our variant calls substantially, from
493 315 to 113. At this stage only 1/113 (~1%) of these variants was called within the two
494 reference strains (JB22, JB32).
495
496 **PCR validation**
497 PCR analysis was performed to confirm 10 of the 11 inversions and all 15 translocations
498 from the curated data set. One inversion was too small to examine by PCR
499 (INV.AB325691:6644..6784, 140 nt). Primers were designed using Primer3(Untergasser
500 et al., 2012) to amplify both the reference and alternate alleles. PCR was carried out with
501 each primer set using a selection of strains that our genotype calls predict to include at

502 least one alternate allele and at least one reference allele (usually 6 strains). Products
503 were scored according to product size and presence/absence (**Supplementary Tables 14,**
504 **15**).
505 Inversions: 9/10 variants were at least partially verified by either reference or
506 alternate allele PCR (3 variants were verified by both reference and alternate PCRs), and
507 7/10 inversions also received support from BLAST (see below). Translocations: 10/15
508 were at least partially verified by either reference or alternate allele PCR (5/15 variants
509 were verified by both reference and alternate PCRs). One additional translocation
510 received support from BLAST (see below), meaning that 11/15 translocations were
511 supported by PCR and/or BLAST. Three of the four translocations that could not be
512 verified were probably nuclear copies of mitochondrial genes (NUMTs) (Lenglez et al.,
513 2010), because one breakpoint was mapped to the mitochondrial genome.
514
515 **Validation by BLAST of *de novo* assemblies**
516 We further assessed the quality of the predicted breakpoints for the inversions and
517 translocations by comparing them to the previously created *de novo* assemblies for each
518 of the 161 strains (Jeffares et al., 2015). To this end, we created blast databases for the
519 scaffolds of each strain that were >1kb. We then created the predicted sequence for 1 kb
520 around each junction of the validated 10 inversions and 15 translocations. These
521 sequences were used to search the blast databases using BLAST+ with --gapopen 1 --
522 gapextend 1 parameters. We accepted any blast hsp with a length >800 bp as supporting
523 the junction (because these must contain at least 300 bp at each side of the break point).
524 Four inversions and three translocations gained support from these searches
525 (Supplementary File Tables2-PCR.xlsx).
526
527 **Knockout strain control**
528 Our sample of sequenced strains included one strain (JB374) that is known to contain
529 deletions of the *his3* and *ura4* genes. Our variant calling and validation methods
530 identified only two variants in this strain, both deletions that corresponded to the
531 positions of these genes, as below:
532 *his3* gene location is chromosome II, 1489773-1488036, deletion detected at II:1488228-
533 1489646.
534 *ura4* gene location is chromosome III, 115589-116726, deletion detected at III:115342-
535 117145.
536 This strain was not included in the further analyses of the SVs.
537
538 **Microarray expression analysis**
539 Cells were grown in YES (Formedium, UK) and harvested at OD_{600} =0,5. RNA was
540 isolated followed by cDNA labeling (Lyne et al., 2003). Agilent 8 x 15K custom-made *S.*
541 *pombe* expression microarrays were used. Hybridization, normalization and subsequent

washes were performed according to the manufacturer's protocols. The obtained data were scanned and extracted using GenePix and processed for quality control and normalization using in-house developed R scripts. Subsequent analysis of normalized data was performed using R. Microarray data have been submitted to ArrayExpress (accession number E-MTAB-4019). Genes were considered as induced if their expression signal after normalization was >1.9, and repressed if <0.51.

Time to most recent common ancestor (TMRCA) estimates

Previously, based on the genetic distances between these strains and the 'dated tip' dating method implemented in BEAST (Drummond et al., 2012), we have estimated the divergence times between all 161 *S. pombe* strains sequenced (Jeffares et al., 2015). To determine the TMRCA for pairs of strains, we re-examined the BEAST outputs using FigTree to obtain the medium and 95% confidence intervals.

SNP and indel calling

SNPs were called as described (Jeffares et al., 2015). Insertions and deletions (indels) were called in 160 strains using stampy-mapped, indel-realigned bams as described previously (Jeffares et al., 2015). We accepted indels that were called by both the Genome Analysis Toolkit HaplotypeCaller (DePristo et al., 2011) and Freebayes(Garrison and Marth, 2012), and then genotyped all these calls with Freebayes.

Briefly, indels were called on each strains bam with HaplotypeCaller, and filtered for call quality >30 and mapping quality >30 (bcftools filter --include 'QUAL>30 && MQ>30'). Separately, indels were called on each strains bam with Freebayes, and filtered for call quality >30. All Freebayes vcf files were merged, accepting only positions called by both Freebayes and HaplotypeCaller. These indels were then genotyped with Freebayes using a merged bam (containing reads from all strains), using the --variant-input flag for Freebayes to genotyped only the union calls. Finally indels were filtered for by score, mean reference mapping quality and mean alternate mapping quality >30 (bcftools filter --include 'QUAL>30 && MQM>30 & MQMR>30'). These methods identified 32,268 indels. Only 50 of these segregated between Leupold's h^- reference (JB22) and Leupold's h^{90} reference (JB50), whereas 12109 indels segregated between the JB22 reference and the divergent strain JB916.

Heredity and GWAS

We selected 53 traits that contained at least values from 100 strains(Jeffares et al., 2015), and so included multiple individuals from within clonal populations (growth rates on 42 different solid media and 11 cell shape characters measured with automated image analysis). Trait values were normalized using a rank-based transformation in R, for each trait vector y, normal.y =qnorm(rank(y)/(1+length(y))). Total heritability, and the contribution of SNPs, CNVs and rearrangements were estimated using LDAK (version

582 5.94) (Speed et al., 2012), with kinship matrices derived from all SNPs, 87 CNVs, and 26
583 rearrangements. To assess whether the contribution of CNVs could be primarily due to
584 linkage with causal SNPs, we simulated trait data using the --make-phenos function of
585 LDAK with the relatedness matrix from all SNPs, assuming that all variants contributed
586 to the trait (--num-causals -1). We made one simulated trait data set per trait, for each of
587 the 53 traits, with total heritability defined as predicted from the real data. We then
588 estimated the heritability using LDAK, including the joint matrix of SNPs, CNVs and
589 rearrangements. To assess the extent to which the contribution of SNPs to heritability was
590 overestimated, we performed another simulation using the relatedness matrix from the 87
591 segregating CNVs alone, and then estimated the contribution of SNPs, CNVs and
592 rearrangements in this simulated data as above.
593 Genome-wide associations were performed with LDAK (version 5) using default
594 parameters, using a mixed model derived from kinship of all SNPs called previously
595 Jeffares et al. (2015). Association analysis was run separately for 68 SVs with a minor
596 allele count >5, for 139,396 SNPs and for 22,058 indels, both minor allele counts >5. We
597 examined the same 53 traits as for the heritability analysis (above). For each trait, we
598 carried out 1000 permutations of trait data, and define the 5th percentile of these
599 permutations as the trait-specific P-value threshold.
600
601 **Offspring viability and genetic distance**
602 Cross spore viability data and self-mating viability were collected from previous analyses
603 (Avelar et al., 2013, Jeffares et al., 2015). The number of differences between each pair
604 was calculated using vcftools vcf-subset (Danecek et al., 2011), and correlations were
605 estimated using R, with the ppcor package. When calculating the number of CNVs
606 differences between strains, we altered our criteria for 'different' variants (to merge
607 variants whose starts and ends where within 1 kb), and merged CNVs if their overlap was
608 >50% and their allele calls were the same.
609

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.

References

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

Ayala, D., Guerrero, R. F. & Kirkpatrick, M. 2013. Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution,* 67**,** 946-58.

Benito, A., Jeffares, D., Palomero, F., Calderon, F., Bai, F. Y., Bahler, J. & Benito, S. 2016. Selected Schizosaccharomyces pombe Strains Have Characteristics That Are Beneficial for Winemaking. *PLoS One,* 11**,** e0151102.

Bitton, D. A., Schubert, F., Dey, S., Okoniewski, M., Smith, G. C., Khadayate, S., Pancaldi, V., Wood, V. & Bahler, J. 2015. AnGeLi: A Tool for the Analysis of Gene Lists from Fission Yeast. *Front Genet,* 6**,** 330.

Brown, W. R., Liti, G., Rosa, C., James, S., Roberts, I., Robert, V., Jolly, N., Tang, W., Baumann, P., Green, C., Schlegel, K., Young, J., Hirchaud, F., Leek, S., Thomas, G., Blomberg, A. & Warringer, J. 2011. A Geographically Diverse Collection of Schizosaccharomyces pombe Isolates Shows Limited Phenotypic Variation but Extensive Karyotypic Diversity. *G3 (Bethesda),* 1**,** 615-26.

Carr, A. M., Macneill, S. A., Hayles, J. & Nurse, P. 1989. 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**,** 41-9.

Chikashige, Y., Tsutsumi, C., Okamasa, K., Yamane, M., Nakayama, J., Niwa, O., Haraguchi, T. & Hiraoka, Y. 2007. Gene expression and distribution of Swi6 in partial aneuploids of the fission yeast Schizosaccharomyces pombe. *Cell Struct Funct,* 32**,** 149-61.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., Mcvean, G., Durbin, R. & Genomes Project Analysis, G. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-8.

Delneri, D., Colson, I., Grammenoudi, S., Roberts, I. N., Louis, E. J. & Oliver, S. G. 2003. Engineering evolution to study speciation in yeasts. *Nature,* 422**,** 68-72.

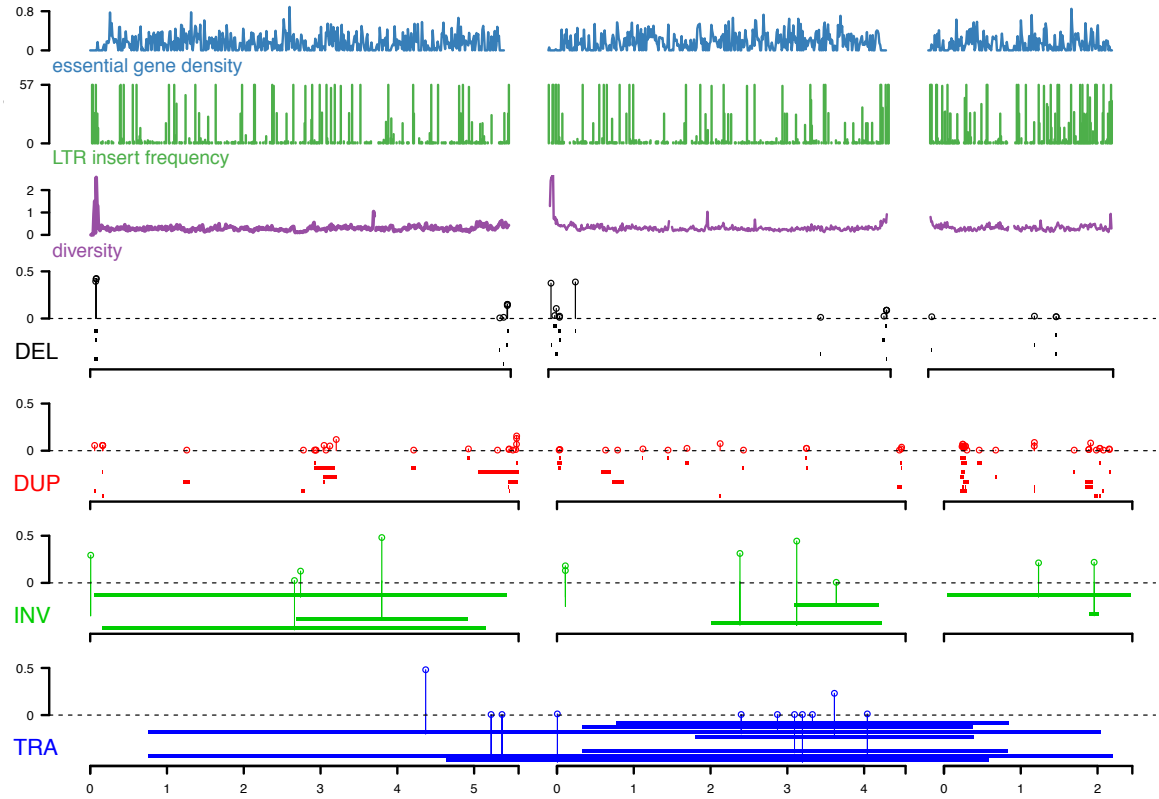
649 Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C.,
650 Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., Mckenna, A.,
651 Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B.,
652 Altshuler, D. & Daly, M. J. 2011. A framework for variation discovery and
653 genotyping using next-generation DNA sequencing data. *Nat Genet,* 43**,** 491-8.
654 Dobzhansky, T. 1933. On the Sterility of the Interracial Hybrids in Drosophila
655 Pseudoobscura. *Proc Natl Acad Sci U S A,* 19**,** 397-403.
656 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. 2012. Bayesian phylogenetics
657 with BEAUti and the BEAST 1.7. *Mol Biol Evol,* 29**,** 1969-73.
658 Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F. &
659 Botstein, D. 2002. Characteristic genome rearrangements in experimental
660 evolution of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A,* 99**,** 16144-9.
661 Faust, G. G. & Hall, I. M. 2012. YAHA: fast and flexible long-read alignment with
662 optimal breakpoint detection. *Bioinformatics,* 28**,** 2417-24.
663 Fawcett, J. A., Iida, T., Takuno, S., Sugino, R. P., Kado, T., Kugou, K., Mura, S.,
664 Kobayashi, T., Ohta, K., Nakayama, J. & Innan, H. 2014. Population genomics of
665 the fission yeast Schizosaccharomyces pombe. *PLoS One,* 9**,** e104241.
666 Garrison, E. & Marth, G. 2012. Haplotype-based variant detection from short-read
667 sequencing. *arXiv preprint arXiv,* 1207.3907
668 Holtgrewe, M. 2010. Mason-A Read Simulator for Second Generation Sequencing Data.
669 Institut für Mathematik und Informatik, Freie Universität Berlin.
670 Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J. & Eichler, E.
671 E. 2010. De novo rates and selection of large copy number variation. *Genome*
672 *Res,* 20**,** 1469-81.
673 Jeffares, D. C., Rallis, C., Rieux, A., Speed, D., Prevorovsky, M., Mourier, T.,
674 Marsellach, F. X., Iqbal, Z., Lau, W., Cheng, T. M., Pracana, R., Mulleder, M.,
675 Lawson, J. L., Chessel, A., Bala, S., Hellenthal, G., O'fallon, B., Keane, T.,
676 Simpson, J. T., Bischof, L., Tomiczek, B., Bitton, D. A., Sideri, T., Codlin, S.,
677 Hellberg, J. E., Van Trigt, L., Jeffery, L., Li, J. J., Atkinson, S., Thodberg, M.,
678 Febrer, M., Mclay, K., Drou, N., Brown, W., Hayles, J., Carazo Salas, R. E.,
679 Ralser, M., Maniatis, N., Balding, D. J., Balloux, F., Durbin, R. & Bahler, J. 2015.
680 The genomic and phenotypic diversity of Schizosaccharomyces pombe. *Nat*
681 *Genet,* 47**,** 235-41.
682 Kim, D. U., Hayles, J., Kim, D., Wood, V., Park, H. O., Won, M., Yoo, H. S., Duhig, T.,
683 Nam, M., Palmer, G., Han, S., Jeffery, L., Baek, S. T., Lee, H., Shim, Y. S., Lee,
684 M., Kim, L., Heo, K. S., Noh, E. J., Lee, A. R., Jang, Y. J., Chung, K. S., Choi, S.
685 J., Park, J. Y., Park, Y., Kim, H. M., Park, S. K., Park, H. J., Kang, E. J., Kim, H.
686 B., Kang, H. S., Park, H. M., Kim, K., Song, K., Song, K. B., Nurse, P. & Hoe, K.
687 L. 2010. Analysis of a genome-wide set of gene deletions in the fission yeast
688 Schizosaccharomyces pombe. *Nat Biotechnol,* 28**,** 617-23.

689 Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterecker, A., Bodenhofer,
690 U. & Hochreiter, S. 2012. cn.MOPS: mixture of Poissons for discovering copy
691 number variations in next-generation sequencing data with a low false discovery
692 rate. *Nucleic Acids Res,* 40**,** e69.
693 Klopocki, E. & Mundlos, S. 2011. Copy-number variations, noncoding sequences, and
694 human phenotypes. *Annu Rev Genomics Hum Genet,* 12**,** 53-72.
695 Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. 2014. LUMPY: a probabilistic
696 framework for structural variant discovery. *Genome Biol,* 15**,** R84.
697 Lenglez, S., Hermand, D. & Decottignies, A. 2010. Genome-wide mapping of nuclear
698 mitochondrial DNA sequences links DNA replication origins to chromosomal
699 double-strand break formation in Schizosaccharomyces pombe. *Genome Res,* 20**,**
700 1250-61.
701 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
702 Abecasis, G., Durbin, R. & Genome Project Data Processing, S. 2009. The
703 Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.
704 Lyne, R., Burns, G., Mata, J., Penkett, C. J., Rustici, G., Chen, D., Langford, C., Vetrie,
705 D. & Bahler, J. 2003. Whole-genome microarrays of fission yeast: characteristics,
706 accuracy, reproducibility, and processing of array data. *BMC Genomics,* 4**,** 27.
707 Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R. & Bahler, J. 2012.
708 Quantitative analysis of fission yeast transcriptomes and proteomes in
709 proliferating and quiescent cells. *Cell,* 151**,** 671-83.
710 Mcgaugh, S. E. & Noor, M. A. 2012. Genomic impacts of chromosomal inversions in
711 parapatric Drosophila species. *Philos Trans R Soc Lond B Biol Sci,* 367**,** 422-9.
712 Muller, H. J. 1939. Reversibility in Evolution Considered from the Standpoint of
713 Genetics. *Biological Reviews,* 14**,** 261–280.
714 Noor, M. A., Grams, K. L., Bertucci, L. A. & Reiland, J. 2001. Chromosomal inversions
715 and the reproductive isolation of species. *Proc Natl Acad Sci U S A,* 98**,** 12084-8.
716 Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V. & Korbel, J. O. 2012.
717 DELLY: structural variant discovery by integrated paired-end and split-read
718 analysis. *Bioinformatics,* 28**,** i333-i339.
719 Rieseberg, L. H. 2001. Chromosomal rearrangements and speciation. *Trends Ecol Evol,*
720 16**,** 351-358.
721 Ryan, C. J., Roguev, A., Patrick, K., Xu, J., Jahari, H., Tong, Z., Beltrao, P., Shales, M.,
722 Qu, H., Collins, S. R., Kliegman, J. I., Jiang, L., Kuo, D., Tosti, E., Kim, H. S.,
723 Edelmann, W., Keogh, M. C., Greene, D., Tang, C., Cunningham, P., Shokat, K.
724 M., Cagney, G., Svensson, J. P., Guthrie, C., Espenshade, P. J., Ideker, T. &
725 Krogan, N. J. 2012. Hierarchical modularity and the evolution of genetic
726 interactomes across species. *Mol Cell,* 46**,** 691-704.
727 Sabatinos, S. A. & Forsburg, S. L. 2010. Molecular genetics of Schizosaccharomyces
728 pombe. *Methods Enzymol,* 470**,** 759-95.

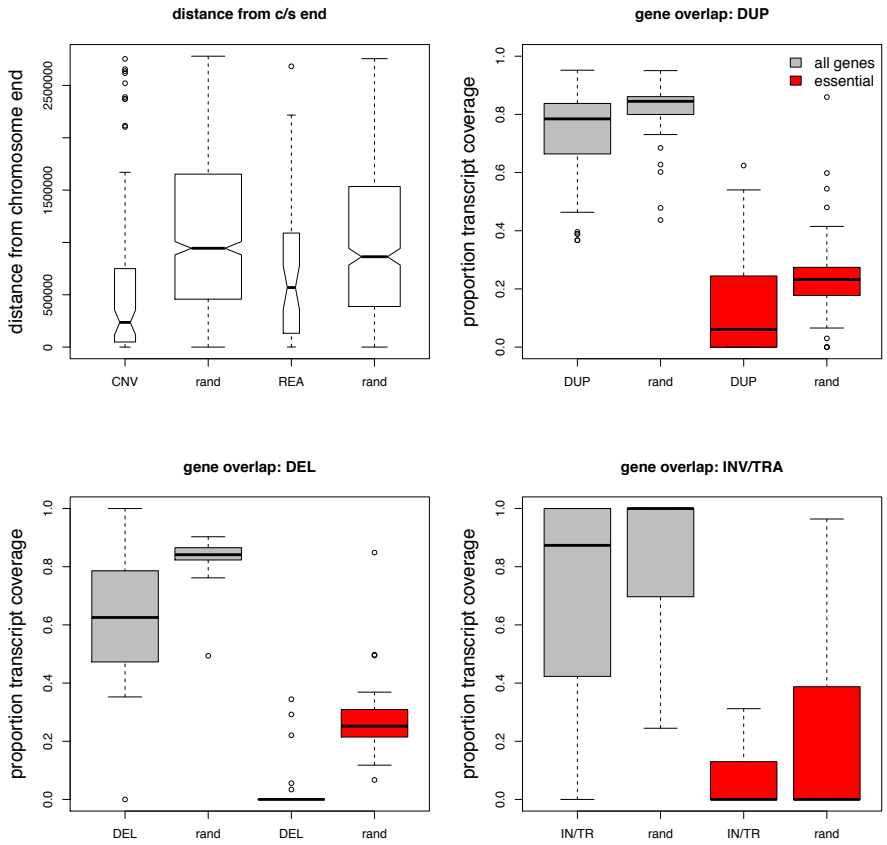
729 Sedlazeck, F. J., Rescheneder, P. & Von Haeseler, A. 2013. NextGenMap: fast and
730 accurate read mapping in highly polymorphic genomes. *Bioinformatics,* 29**,** 2790-
731 1.
732 Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. 2012. Improved heritability
733 estimation from genome-wide SNPs. *Am J Hum Genet,* 91**,** 1011-21.
734 Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K.,
735 Arnarsdottir, S., Bjornsdottir, G., Walters, G. B., Jonsdottir, G. A., Doyle, O. M.,
736 Tost, H., Grimm, O., Kristjansdottir, S., Snorrason, H., Davidsdottir, S. R.,
737 Gudmundsson, L. J., Jonsson, G. F., Stefansdottir, B., Helgadottir, I., Haraldsson,
738 M., Jonsdottir, B., Thygesen, J. H., Schwarz, A. J., Didriksen, M., Stensbol, T. B.,
739 Brammer, M., Kapur, S., Halldorsson, J. G., Hreidarsson, S., Saemundsen, E.,
740 Sigurdsson, E. & Stefansson, K. 2014. CNVs conferring risk of autism or
741 schizophrenia affect cognition in controls. *Nature,* 505**,** 361-6.
742 Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon,
743 R., Bird, C. P., De Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W.,
744 Tavare, S., Deloukas, P., Hurles, M. E. & Dermitzakis, E. T. 2007. Relative
745 impact of nucleotide and copy number variation on gene expression phenotypes.
746 *Science,* 315**,** 848-53.
747 Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J.,
748 Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A.,
749 Stutz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G.,
750 Chen, K., Malig, M., Chaisson, M. J., Walter, K., Meiers, S., Kashin, S., Garrison,
751 E., Auton, A., Lam, H. Y., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T.,
752 Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan,
753 X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E. W., Mccarthy, S.,
754 Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M.,
755 Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B.,
756 Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A.,
757 Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-
758 Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., Mccarroll, S. A.,
759 Genomes Project, C., Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine,
760 S. E., Lee, C., Eichler, E. E. & Korbel, J. O. 2015. An integrated map of structural
761 variation in 2,504 human genomes. *Nature,* 526**,** 75-81.
762 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. 2013. Integrative Genomics Viewer
763 (IGV): high-performance genomics data visualization and exploration. *Brief*
764 *Bioinform,* 14**,** 178-92.
765 Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. &
766 Rozen, S. G. 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Res,*
767 40**,** e115.

768 Walters, R. G., Coin, L. J., Ruokonen, A., De Smith, A. J., El-Sayed Moustafa, J. S.,
769 Jacquemont, S., Elliott, P., Esko, T., Hartikainen, A. L., Laitinen, J., Mannik, K.,
770 Martinet, D., Meyre, D., Nauck, M., Schurmann, C., Sladek, R., Thorleifsson, G.,
771 Thorsteinsdottir, U., Valsesia, A., Waeber, G., Zufferey, F., Balkau, B., Pattou, F.,
772 Metspalu, A., Volzke, H., Vollenweider, P., Stefansson, K., Jarvelin, M. R.,
773 Beckmann, J. S., Froguel, P. & Blakemore, A. I. 2013. Rare genomic structural
774 variants in complex disease: lessons from the replication of associations with
775 obesity. *PLoS One,* 8**,** e58048.
776 Wood, V., Gwilliam, R., Rajandream, M. A., Lyne, M., Lyne, R., Stewart, A., Sgouros,
777 J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown,
778 D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin,
779 A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris,
780 D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E.
781 J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., Mcdonald, S.,
782 Mclean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell,
783 C., Oliver, K., O'neil, S., Pearson, D., Quail, M. A., Rabbinowitsch, E.,
784 Rutherford, K., Rutter, S., Saunders, D., Seeger, K., Sharp, S., Skelton, J.,
785 Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R. G.,
786 Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G.,
787 Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M.,
788 Schafer, M., Muller-Auer, S., Gabel, C., Fuchs, M., Dusterhoft, A., Fritzc, C.,
789 Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H.,
790 Reinhardt, R., Pohl, T. M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R.,
791 Purnelle, B., Goffeau, A., Cadieu, E., Dreano, S., Gloux, S., et al. 2002. The
792 genome sequence of Schizosaccharomyces pombe. *Nature,* 415**,** 871-80.
793 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. 2009. Pindel: a pattern growth
794 approach to detect break points of large deletions and medium sized insertions
795 from paired-end short reads. *Bioinformatics,* 25**,** 2865-71.
796 Zanders, S. E., Eickbush, M. T., Yu, J. S., Kang, J. W., Fowler, K. R., Smith, G. R. &
797 Malik, H. S. 2014. Genome rearrangements and pervasive meiotic drive cause
798 hybrid infertility in fission yeast. *Elife,* 3**,** e02630.
799 Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. 2009. Copy number variation in human
800 health, disease, and evolution. *Annu Rev Genomics Hum Genet,* 10**,** 451-81.
801
802
803

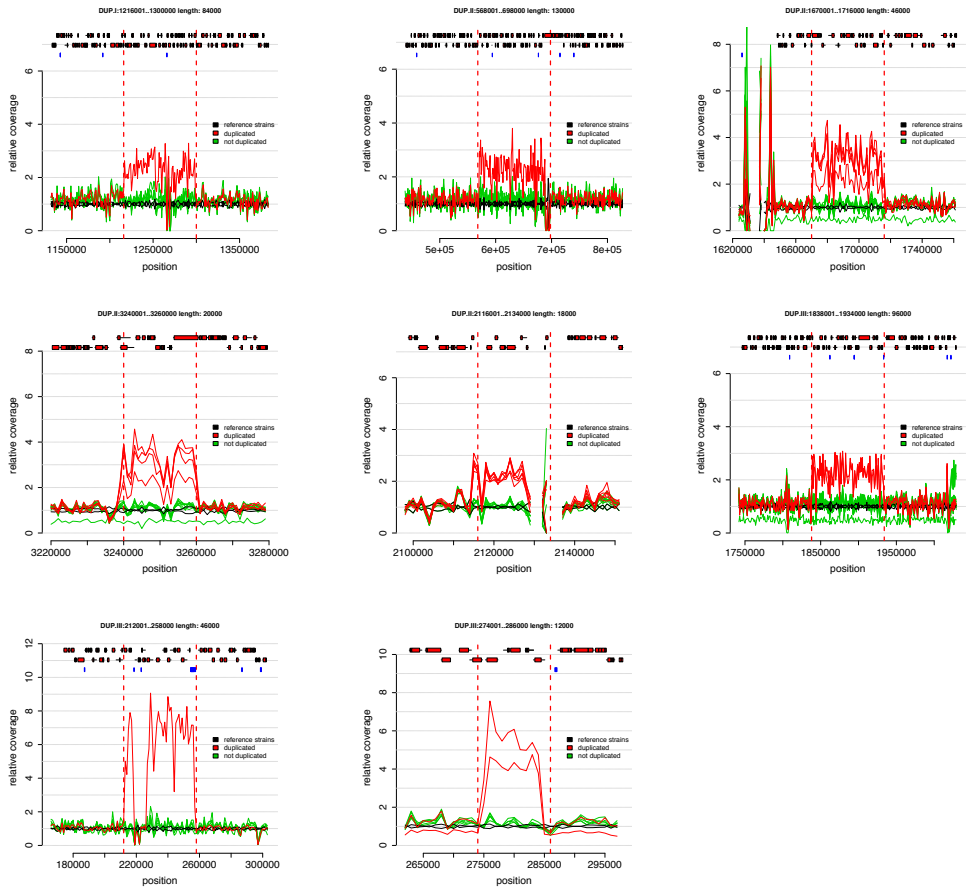
Supplementary Figures



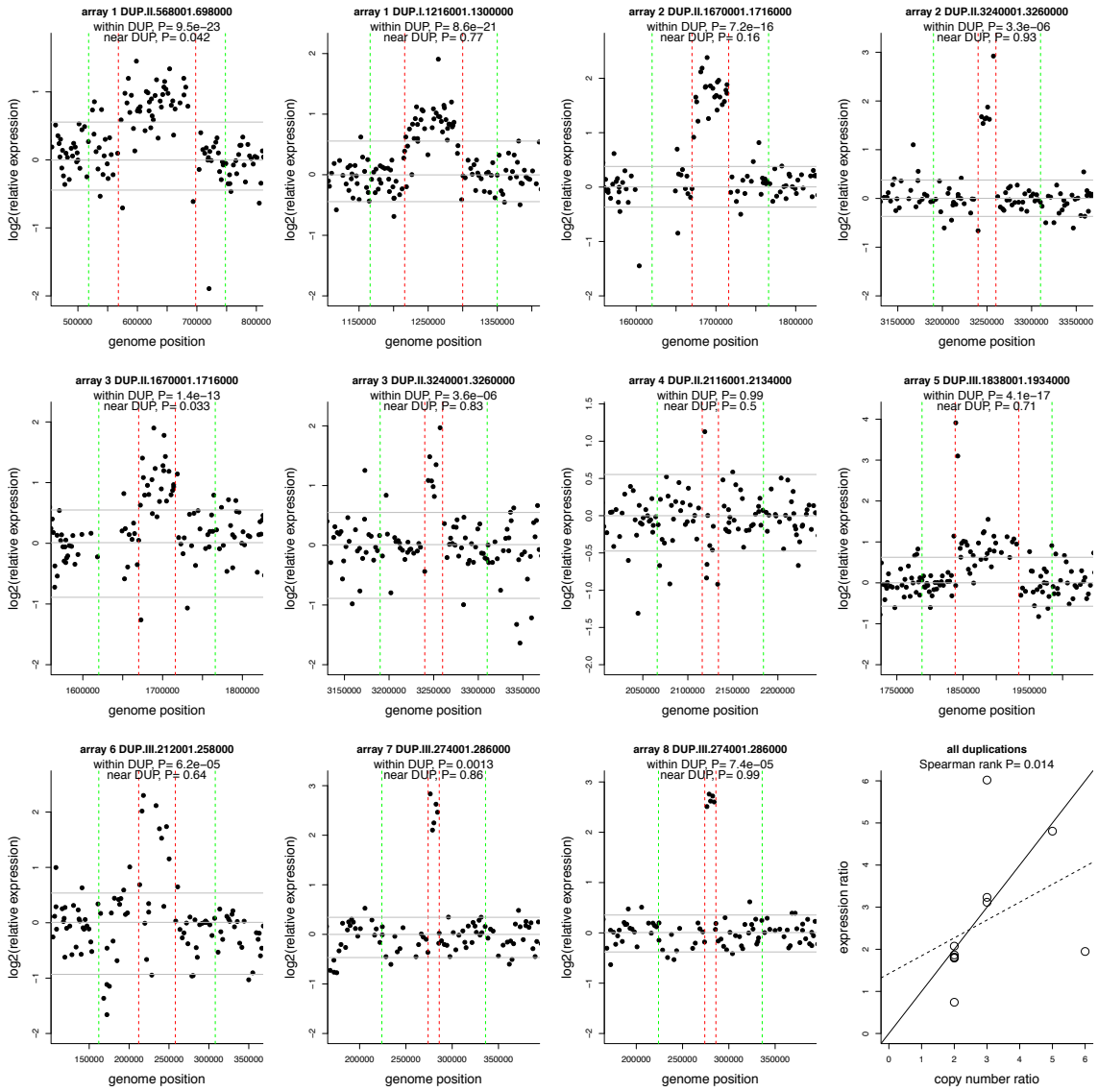
Supplementary Figure 1. Locations and minor allele frequencies of all structural variants in curated data set. Each of the three chromosomes is indicated by black bar, with scale (in megabases) at bottom. From top (same data as Fig 1): density of essential genes (blue), locations of *Tf*-type retrotransposons (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. (2015). 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.



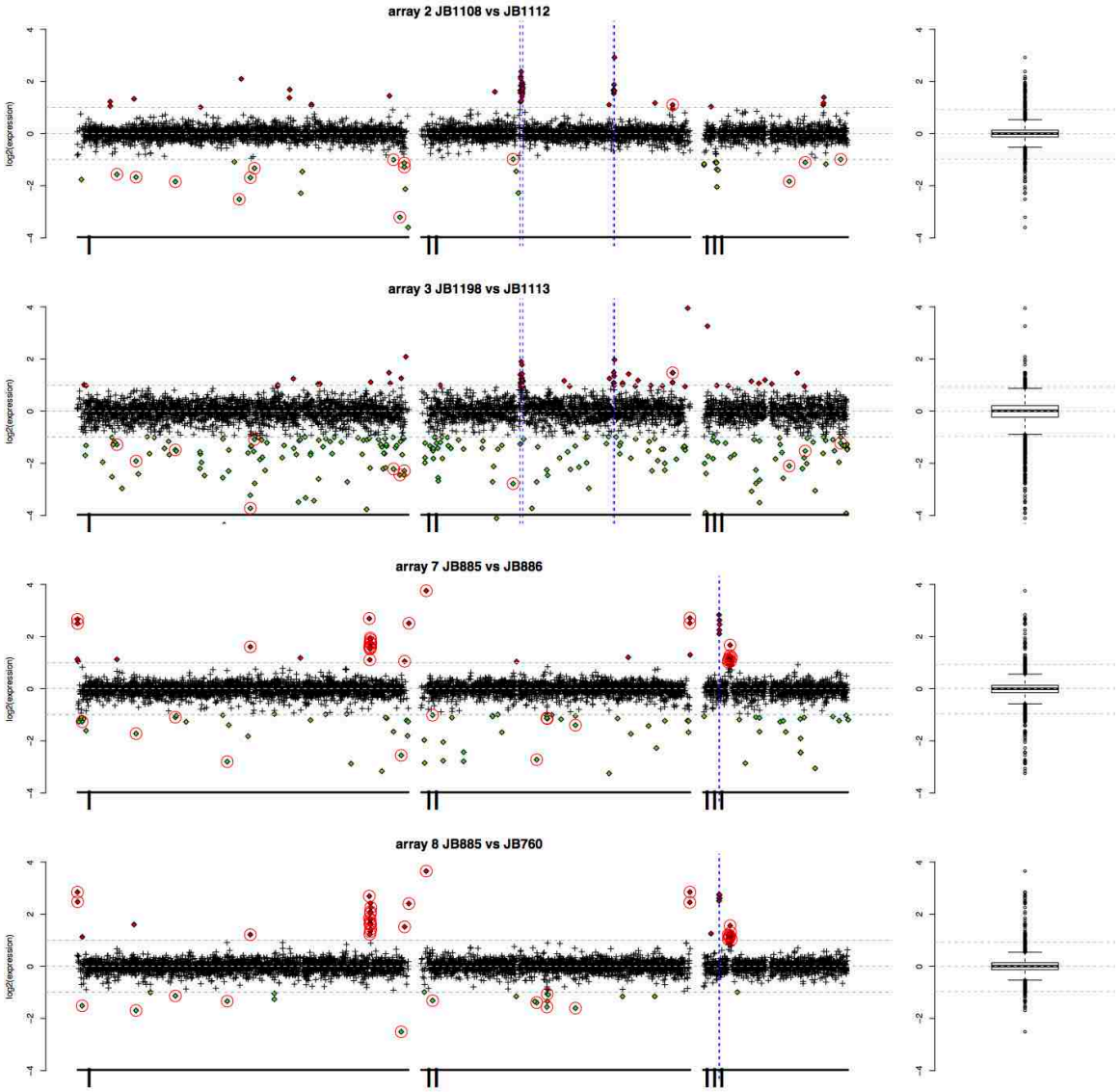
Supplementary Figure 2. Structural variations are biased towards chromosome ends and to low gene density regions. Top left panel, both CNVs and rearrangements (REA) are closer to the nearest chromosome end than the null distribution (rand), (Wilcoxon rank sum test, P-values 1.3×10^{-11} and 0.03, respectively). 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 $<1.6 \times 10^{-4}$). 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 ($P = 1.9 \times 10^{-3}$), 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.

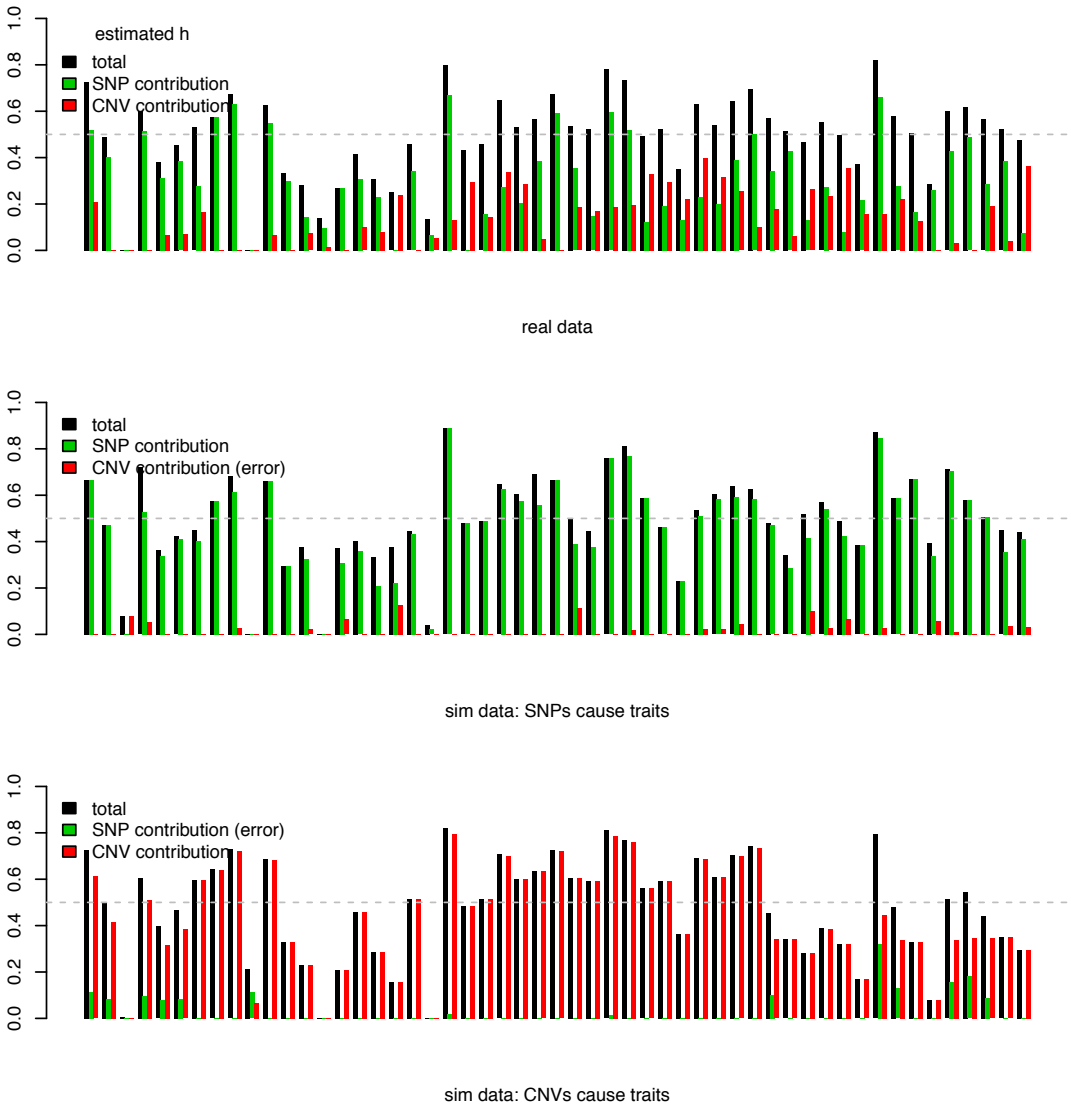


Supplementary Figure 4. No significant increase in gene expression immediately adjacent to duplications. For each duplication examined with DNA arrays, we show the relative expression (strain 1 *vs* strain 2) near the duplication. P-values show the support for the genes within the duplication (red vertical 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.

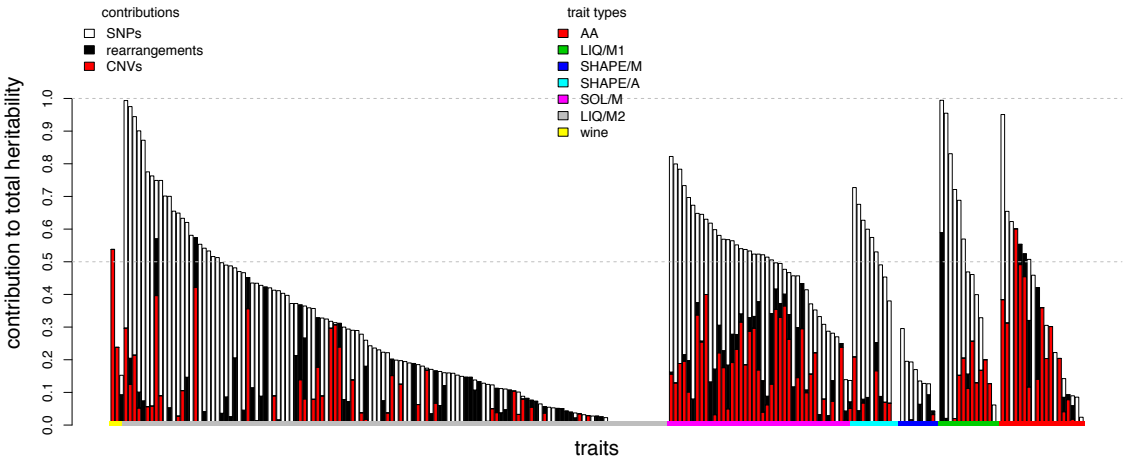


868

869 **Supplementary Figure 5. Chromosome-scale view of gene expression changes.** The
870 relative gene expression levels (strain1/strain2) for arrays 2 and 3, and arrays 7 and 8 are
871 shown with their positions on the three chromosomes. Filled circles indicates genes that
872 we consider to be upregulated (red) or repressed (green). Those highlighted with open red
873 circles are consistently altered in both arrays (either 2+3, or 7+8). The blue lines show
874 where the segregating duplications are. Box plots at right show the spread of data.

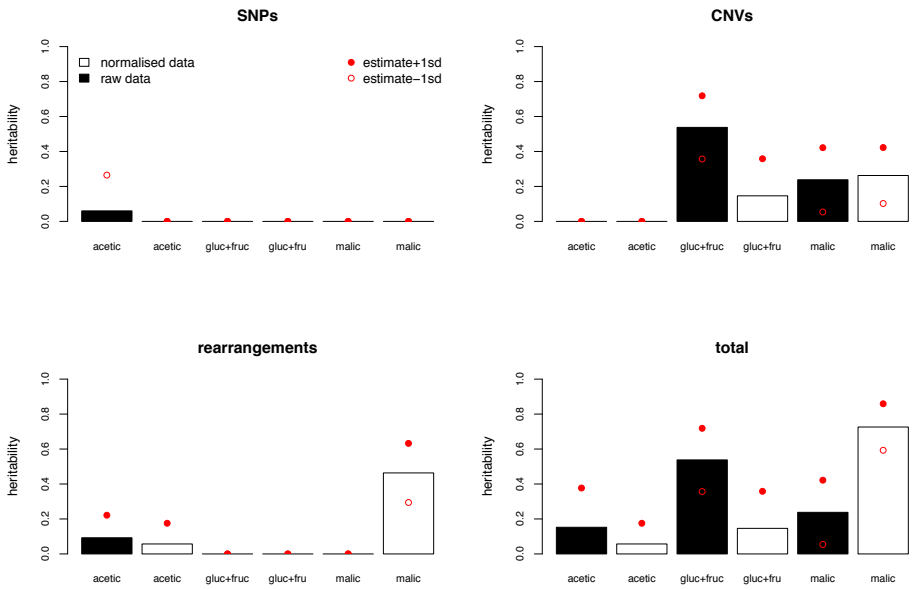


Supplementary Figure 6. Contributions of SNPs and CNVs to traits. Top panel: for 53 traits, we show the total heritability estimated by the combination of 243,289 SNPs (green), 87 CNVs (red), and 26 rearrangements (grey). We then simulated data that was entirely due to the effects of SNPs (middle panel), or entirely due to the effects of CNVs (lower panel). In the middle panel, all estimates of the, usually minor, contribution of CNVs is artefactual. In the lower panel all estimates of the contribution of SNPs is artefactual. Again, these estimated are usually minor.



885
886
887 **Supplementary Figure 7. Contributions to of CNVs, rearrangements and SNPs to**
888 **quantitative traits.** For 223 traits we show the contribution of CNVs (red),
889 rearrangements (black) and SNPs (open bars) to the heritability. Traits are sorted by trait
890 type (indicated by coloured bars on *x* axis), and then by total heritability estimate.
891 Categories are wine traits (yellow), growth on liquid media (LIQ/M2; grey), growth on
892 solid media (SOL/M; magenta), automated shape parameters (SHAPE/A; cyan), manual
893 parameters (SHAPE/M; blue), growth on liquid media from this study (LIQ/M1; green)
894 and amino acid concentrations (AA; red). All traits are from(Jeffares et al., 2015), except
895 LIQ/M2(Brown et al., 2011) and wine(Benito et al., 2016).
896

897



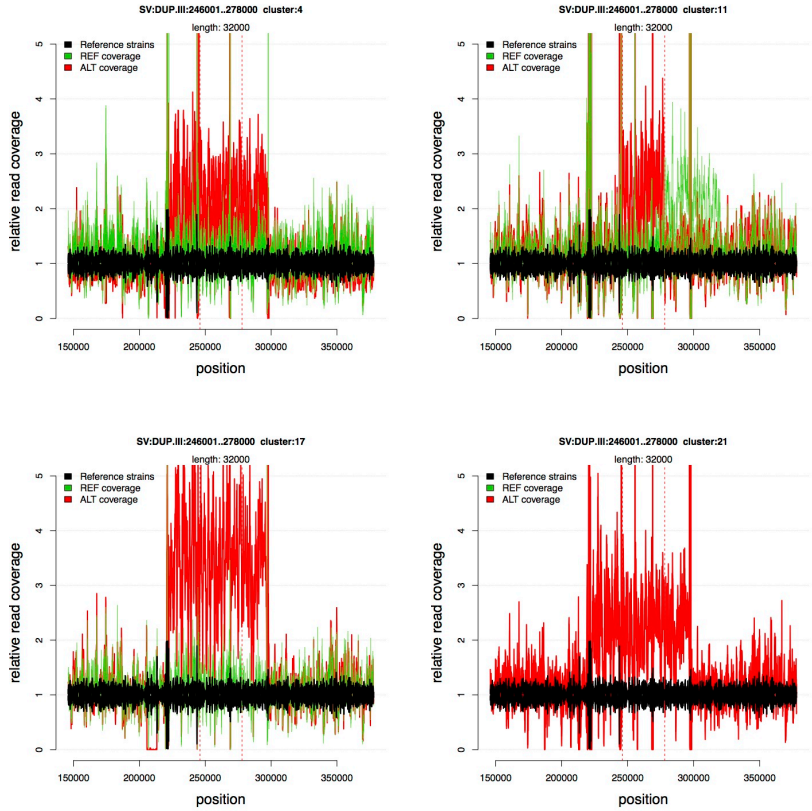
898

899 **Supplementary Figure 8. Details of the heritability of winemaking traits.** Previously,
900 we examined the potential of the 57 non-clonal strains for winemaking(Benito et al.,
901 2016). Three parameters of fermentations in grape must, that are known to influence wine
902 quality, were measured for all strains; the acetic acid content (acetic), the residual glucose
903 + fructose concentration (gluc+fruc), and the percentage of malic acid degradation
904 (malic). Bar plots show the estimated contribution of SNPs, CNVs and rearrangements to
905 these traits, using both normalized and raw (non-normalized) data. The final plot shows
906 the total heritability. Filled red circles show one standard deviation above the estimate,
907 open red circles show one standard deviation below. While estimates vary with
908 normalized/raw data, the SNP contributions are always low compared to SVs, and are not
909 significantly greater than zero for residual glucose + fructose concentration and malic
910 acid degradation.

911

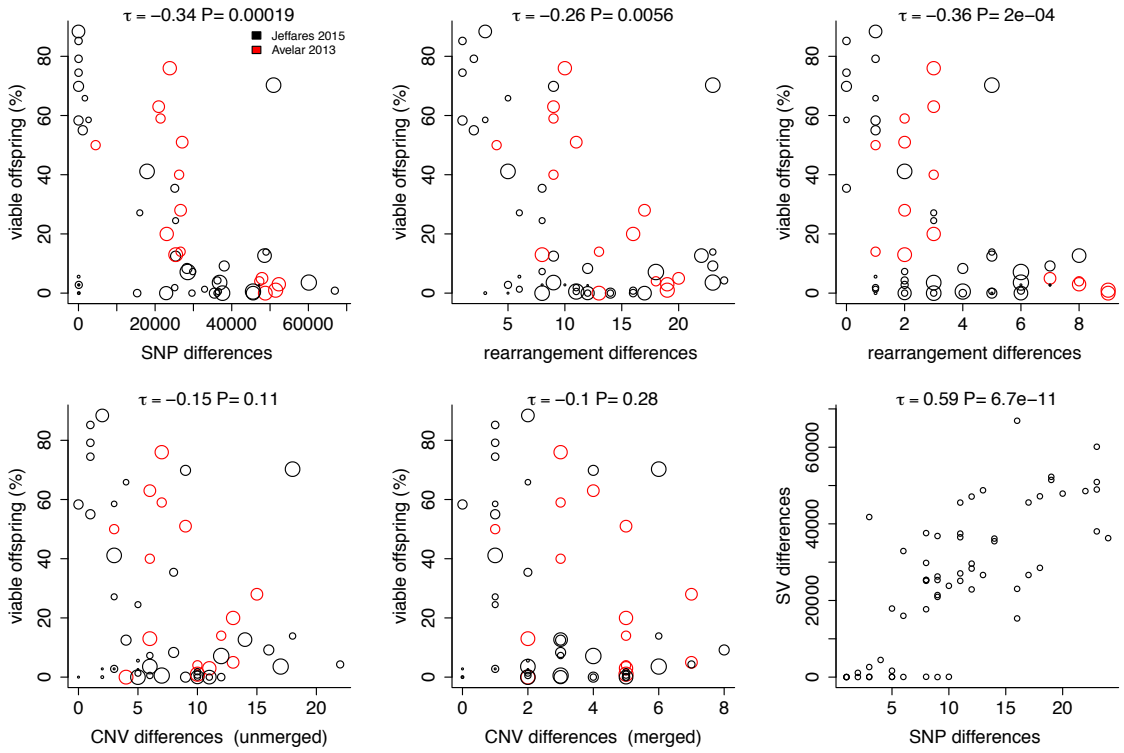
912

913

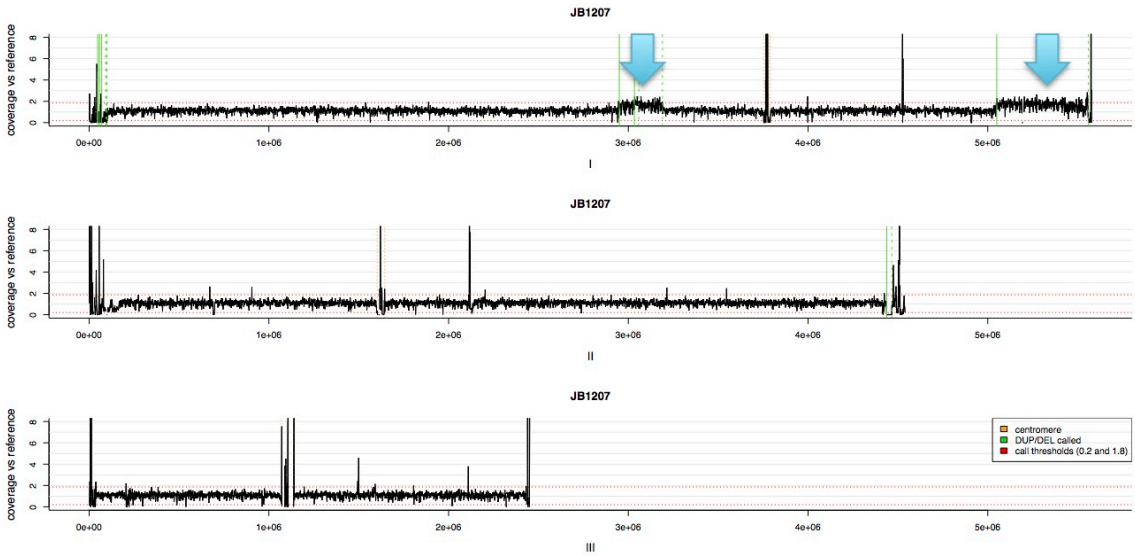


914

915 **Supplementary Figure 9. The 32kb duplication associated with resistance to**
916 **Brefeldin.** Panels show relative coverage in 100bp windows (vs. reference genome) for
917 strains within clonal populations ('clonal clusters'). The coverage for strains that contain
918 the duplication are plotted in red, and those that do not are plotted in green. Reference
919 strain coverage (h^- and h^+) is shown in black. The red vertical dashed line indicates the
920 boundaries of the 32kb duplication – note that some clusters contain a larger duplicated
921 region. Clockwise from top left we show clusters 4, 11, 17 and 21 (see **Supplementary**
922 **Table 7** for cluster members).
923
924
925



Supplementary Figure 10. 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. (2015) (black) and Avelar et al. (2013) (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 11. 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).