Dosage compensation can buffer copy-number variation in wild yeast

James Hose[1], Chris Mun Yong[1], Maria Sardi[1], Zhishi Wang[2], Michael A. Newton[2,3], Audrey P. Gasch[1,3,*]

[1] Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI, 53706
[2] Department of Statistics, University of Wisconsin-Madison, Madison, WI, 53706
[3] Genome Center of Wisconsin, University of Wisconsin-Madison, Madison, WI, 53706

* correspondence should be addressed to agasch@wisc.edu

James Hose, jhose@wisc.edu, (608)265-0863
Chris Mun Yong, myong@wisc.edu, (608)265-0863
Maria Sardi, sardi@wisc.edu (608)265-0863
Zhishi Wang, wangz@stat.wisc.edu, (608) 262-0086
Michael A. Newton, newton@biostat.wisc.edu, (608) 262-0086
Audrey P. Gasch, agasch@wisc.edu, (608)265-0859

IMPACT STATEMENT: Our results suggest that dosage-compensated gene expression facilitates chromosomal aneuploidy, which presents a rapid route to phenotypic evolution in natural yeast isolates.

ABSTRACT

Aneuploidy is linked to myriad diseases but also facilitates organismal evolution. It remains unclear how cells overcome the deleterious effects of aneuploidy until new phenotypes evolve. Although laboratory strains are extremely sensitive to aneuploidy, we show here that aneuploidy is common in wild yeast isolates, which show lower-than-expected expression at many amplified genes. We generated diploid strain panels in which cells carried two, three, or four copies of the affected chromosomes, to show that gene dosage compensation functions at >30% of amplified genes. Genes subject to dosage compensation are under higher expression constraint in wild populations – but they show elevated rates of gene amplification, suggesting that copy number variation is buffered at these genes. We find that aneuploidy provides a clear ecological advantage to oak strain YPS1009, by amplifying a causal gene that escapes dosage compensation. Our work presents a model in which dosage compensation buffers gene amplification through aneuploidy to provide a natural, but likely transient, route to rapid phenotypic evolution.

INTRODUCTION

Susumu Ohno proposed over 40 years ago that gene duplication could provide a major force in the evolution of new gene functions, by relaxing constraint on gene sequences and allowing one or both gene copies to evolve (*1*). The genomic era has largely borne out that hypothesis, and many studies have characterized the outcomes of whole and partial genome amplification (*2*). The immediate consequence of duplication is assumed to be increased expression of the affected genes, and in some cases the increased expression provides a selective advantage (e.g. (*3-5*)). Over longer periods, the relaxed constraint afforded by functional redundancy allows one or both gene copies to evolve (*1*), driving sub- and neo-functionalization (*6, 7*), expression divergence (*8-11*), and network rewiring (*12-14*).

Whole and partial chromosome amplification through aneuploidy is frequently observed in laboratory evolution studies and in drug-resistant fungal pathogens (*15-20*), suggesting that aneuploidy is a rapid route to phenotypic evolution. However, aneuploidy comes with a fitness cost, most famously in cases of human aneuploidies such as Down syndrome (*21*). The reasons for aneuploidy toxicity are not entirely clear but may be due to increased expression from genes that are toxic when overexpressed (*22*). Several studies have used *Saccharomyces cerevisiae* as a model for aneuploidy syndromes, since laboratory strains are extremely sensitive to chromosomal amplification. Laboratory strains with forced aneuploidy are extremely slow growing, regardless of the chromosome amplified (*18, 23*). Transcriptomic and proteomic studies in these strains reported proportionately higher expression from virtually all amplified genes (*18, 23-25*), with a handful of exceptions recently identified at the protein level (*26*). The apparent lack of dosage compensation is consistent with another study by M. Springer *et al.* (*27*), which found that expression at hemizygous genes is not up-regulated to compensate for reduced gene copy. While these studies have generated important results on aneuploidy intolerance in these particular strains, one caveat is that they were all done in laboratory strains, which have lost many features inherent in wild strains (*28-30*). A remaining question is the extent to which aneuploidy occurs in nature and contributes to phenotypic variation in the wild.

Here we report that chromosomal amplification is common in non-laboratory yeast strains, which are inherently tolerant of aneuploidy and display an active mode of gene-dosage compensation at the transcript level, for specific classes of amplified genes. Strikingly, genes subject to dosage compensation are buffered against copy-number variation and thus show elevated rates of gene amplification in natural isolates. Our results raise new implications for the role of aneuploidy in phenotypic evolution and the mechanisms cells use to tolerate it.

RESULTS

We sequenced the genomes of 47 non-laboratory yeast strains, including wild, clinical, and industrial isolates (Supplementary File 1) and assessed the copy number of the 16 yeast chromosomes. Nearly a third of all strains carried whole (12 strains) or partial (2 strains) chromosome amplification (Figure 1A). Three strains harbored multiple aneuploidies, the extreme being insect-associated strain Y2189 that amplified four different chromosomes. Some chromosomes were amplified in multiple unrelated strains: Chromosome III (Chr 3), Chr 9, and Chr 12 were each amplified in different sets of strain pairs, Chr 1 (the smallest yeast chromosome) was amplified in three strains, and Chr 8 was aneuploid in five unrelated isolates. The higher incidence of Chr 8 amplification may reflect a higher frequency of mitotic error, since diploid mutation-accumulation (MA) lines propagated in the near-absence of selection amplify Chr 8 at a higher rate (*31*).

However, several lines of evidence refute the model that these aneuploidies represent deleterious mutations yet to be removed by selection. Unlike diploid MA lines that were at maximum trisomic for specific chromosomes, half of naturally aneuploid strains are tetrasomic (Figure 1A). Furthermore, natural aneuploids showed no significant growth reduction compared to closely related euploid reference strains (p = 0.19, paired T-test), although there was a slight negative correlation between growth rate and extra DNA content (R = -0.3, Figure 1B). This is in stark contrast to aneuploid laboratory strains, which show extreme growth retardation ((*18, 23*) and below). We also found that chromosomal amplification was stable for >200-400 generations in four interrogated wild

strains, whereas the W303 laboratory strain generally loses aneuploidy within 20 generations. To distinguish if the tolerant strains have adapted to aneuploidy or if non-W303 strains can inherently accommodate chromosome amplification, we selected aneuploid derivatives of several naturally euploid parents (see Methods). We found little to no growth defect in derived aneuploid strains (Figure 1C). Thus, *S. cerevisiae* isolates are inherently tolerant of chromosomal amplification, which is common in nature.

A common aneuploidy response is distinct from that in the laboratory strain.
Aneuploid laboratory strains are reported to show proportionately higher expression from virtually all amplified genes, causing proteotoxicity from excess protein production (*18, 23, 25*). We therefore investigated transcriptome profiles through RNA deep sequencing (RNA-seq) in six naturally aneuploid strains normalized to paired euploid reference strains that are closely related (thereby minimizing neutral expression differences unrelated to aneuploidy, see Methods).

Consistent with their near-normal growth rates, naturally aneuploid strains did not activate the environmental stress response as seen in sickly laboratory aneuploids (*23, 24*) (Figure 2A). However, we detected a weak signature common to several aneuploid strains, including the up-regulation of 69 unamplified genes (enriched for oxidoreductases) and reduced expression of 269 unamplified genes (strongly enriched for mitochondrial ribosomal protein genes and genes involved in respiration) in at least three of the six aneuploid strains (Figure 2B). We tested the respiratory capabilities of naturally aneuploid yeast with variable chromosome copy number and found no growth defect on non-fermentable carbon sources (Figure 3). In contrast, a diploid W303 strain trisomic for Chr 12 ('W303_Chr12-3n') displayed an exacerbated growth defect on non-fermentable carbon sources, and we were unable to make the tetrasomic W303_Chr12-4n strain that retained its mitochondrial genome, despite numerous backcrossing attempts (Figure 3). This suggests that differences in mitochondrial function may contribute to differences in aneuploid tolerance across strain backgrounds. Interestingly, up-regulation of oxidoreductases and down-regulation of mitochondrial genes is a hallmark of Down syndrome (*32-35*) (see Discussion).

Many amplified genes display lower-than-expected expression.
Given the significant phenotypic differences in laboratory versus non-laboratory aneuploid strains, we were particularly interested in the expression of amplified genes. We investigated transcript abundance relative to DNA copy for amplified genes, interrogating 2,204 genes spanning eight amplified chromosomes across the six aneuploids. Across all strains, nearly 40% of amplified genes showed lower expression per gene copy compared to the paired euploid (Figure 4A, blue points, see Methods). These were enriched for genes encoding ribosomal proteins (RPs), translation factors, proteins localized to the nucleus or to mitochondria, and other groups (p< 1e-5, Supplementary File 2). The lower-than-expected expression could not be explained by a general repression response to the aneuploidy, since only 39 of the 838 affected genes were part of the common response described above. In contrast, amplified genes with expression proportionate to gene copy (Figure 4A, grey points) showed distinct enrichment for genes encoding proteins localized to the cytoplasm or to membranes, stress defense proteins, and kinases and transferases (p<1e-5, Supplementary File 2). A subset of genes was expressed ≥1.5X higher than expected per gene copy (Figure 4A, magenta points), and these were enriched for genes that influence morphology (p = 1e-5, see more below).

Reduced expression in isogenic strain pairs implicates dosage compensation.
Two models could explain the reduced expression from amplified genes. New mutations, through adaptation or drift, could heritably reduce expression at toxic amplified genes. Alternatively, wild strains may actively down-regulate expression in proportion to gene dose, known as dosage compensation. To distinguish between these possibilities, we generated isogenic aneuploid-euploid strain pairs for three wild strains. We isolated euploid derivatives of strain T73_Chr8-4n (denoting four copies of Chr 8 in the diploid strain) and YJM428_Chr16-4n by serial passaging for many generations. We also used drug-based selection to isolate a mutant of naturally euploid oak-soil strain YPS163 in which Chr 8 was amplified (see Methods). We then conducted duplicate RNA-seq analysis in the isogenic aneuploid-euploid pairs and identified genes on Chr 8 or Chr 16 with lower-than-expected expression, as above. Because the strains are nominally isogenic (see Methods), most expression differences between the strain pairs are an active response to the aneuploidy.

Recapitulating the results shown above, roughly 11 – 36% of amplified genes, depending on the strain, showed lower-than-expected expression (Figure 4B, blue points), while 2 – 4% showed higher-than-expected expression (Figure 4B, magenta points). Few of these genes participate in the common response to chromosome amplification, suggesting a mechanism of dosage compensation. A substantial fraction of the genes scored in the forced YPS163 aneuploid displayed reduced expression, even though this strain had little time to adapt to the aneuploidy. Thus, dosage compensation is likely an inherent trait in *Saccharomyces cerevisiae* that functions at a subset of yeast genes.

Gene classification further refines expression patterns at amplified genes.
To more accurately define genes subject to dosage compensation versus heritable polymorphisms, we next generated isogenic strain panels for oak strain YPS1009 and West African strain NCYC110, in which isogenic diploids carry two, three, or four copies of Chr 12 or Chr 8, respectively (Figure 5A). Expression in each strain within the YPS1009_Chr12 or NCYC110_Chr8 panels was normalized to a closely related euploid reference, YPS163 or NCYC3290, respectively. Comparing expression within the isogenic strain panel allows us to identify genes whose expression does not increase linearly as gene copy increases; comparing expression in the panel strains to the closely related euploid reference identifies heritable expression differences in the two strain backgrounds. A comparable panel was developed for laboratory strain W303 aneuploid for Chr 12. As expected, there was little growth difference across the wild strain panels but a major defect as aneuploidy increased in the W303 laboratory strain (Figure 5B).

We measured mRNA and DNA abundance across each panel relative to the paired euploid references and developed a mixture of linear regression (MLR) model to classify genes based on the slope and intercept of the mRNA-gene copy relationships (Figure 6, see Methods). Genes in Class 1 show proportionate increases in mRNA abundance as gene copy increases across the strain panel, with a slope of 1.0 and a $\log_2$ intercept of 0 that indicates comparable expression in the two euploid strains (Figure 6A). These genes therefore show no evidence of dosage compensation or heritably altered expression. Genes in Class 2 also show a linear relationship between mRNA and DNA copy (slope of 1.0), but have an altered intercept that reflects either constitutively reduced (Class 2a, Figure 6B) or constitutively elevated (Class 2b, Figure 6C) mRNA per gene copy. Thus, Class 2 genes display heritably altered expression but no

evidence of dosage compensation. In contrast, genes in Class 3 show a disproportionate relationship between mRNA abundance and gene copy number. For genes in Class 3a, mRNA does not increase proportionately as gene copy increases across the strain panel, as evidenced by reduced slope of the linear fit (Figure 6D). Analogously, genes in Class 3b show a slope >1, indicating that mRNA abundance is amplified above expectation as gene copy increases. Because strains within each panel are isogenic aside of the aneuploidy, the reduced slope for Class 3a genes is indicative of dosage compensation while the elevated slope for Class 3b genes represents a disproportionate increase in expression.

We applied the MLR model to classify genes based on their maximum posterior probability (see Methods). The results varied to some degree depending on the strain (Table 1). In the YPS1009_Chr 12 strain panel, 7% of genes showed expected mRNA increases across the strain panel and with regard to the reference strain (Class 1). In contrast, 28% of genes showed heritably reduced expression per gene copy (Class 2a) while only 1% showed heritably increased expression (Class 2b). However, nearly a third of Chr 12 genes displayed a phenotype consistent with dosage compensation (Class 3a). In the NCYC110_Chr 8 strain panel, 11% of genes displayed a phenotype consistent with dosage compensation while almost half the Chr 8 genes were classified as having heritably reduced expression.

Genes linked to cell morphology show amplified expression in aneuploid strains

Sixteen to 20% of duplicated genes showed amplified mRNA abundance across the strain panels (Class 3b) or in isogenic aneuploid-euploid pairs (Figure 4B). As a group, these 206 genes were enriched for those important for cell morphology (p = 6e-6). We noted that several of our aneuploid strains displayed differences in flocculence or cell shape compared to euploid controls, as previously reported for aneuploid strains (*36, 37*). We reasoned that this may be an indirect effect that extends to unamplified genes, and thus we identified unamplified genes with linear expression increase proportionate to the copy number of the amplified chromosome. This identified 39 genes in the YPS1009_Chr12 panel and 96 genes in the NCYC110_Chr8 panel (false discovery rate 0.05, see Methods). When pooled together, and especially when combined with the 206 genes with amplified expression described above, there was strong enrichment for proteins localized to the membrane (p = 5e-4) and for those important for proper morphology (1e-8). Genes encoding cell-surface proteins are known to display altered expression due to changes in cell size/shape induced by ploidy effects (*37*), which may explain their altered expression here.

Dosage compensation occurs at specific functional groups with some variation across strains.

To further investigate genes subject to dosage compensation, we combined Class 3a genes identified in the strain panels with genes displaying lower-than-expected expression in the aneuploid-euploid strain pairs (Figure 4A-B, see Methods), for a total of 245 genes. As a group, these genes were strongly enriched for translation factors (p = 5e-7), cytosolic RPs (p = 1e-5), ribosome biogenesis genes (p = 1e-4), iron/copper transporters (p = 4e-5), guanyl-nucleotide exchange factors (p = 7e-4), and genes encoding mitochondrial proteins (p = 3e-5). Once again, few of these genes (<7%) were repressed in other aneuploid strains, demonstrating that this is not an aneuploidy-stress response.

Three of the isogenic aneuploid-euploid strain groups involved amplification of Chr 8, allowing us to assess strain-specific effects on dosage compensation. Of the genes with lower-than-expected expression in either the haploid YPS163_Chr8-disome or diploid T73_Chr8-4n strain, 79% showed lower-than-expected expression in the NCYC110_Chr8 strain panel. While an over-abundance of these genes were scored as dosage compensated in Class 3a (p = 7e-3), a surprising number of genes called compensated in the paired analysis were classified as having heritably reduced expression without dosage compensation (Class 2a) in the NCYC110_Chr8 panel. One possibility is that the genes are mis-classified or are affected by both heritably reduced expression and a mode of dosage compensation – while the latter was true for a handful of the NCYC110 genes, most displayed high membership in Class 2a and had a slope and fit that was inconsistent with dosage compensation. These results raise the possibility that there exist strain-specific differences in the genes that are subject to dosage compensation (see Discussion).

Dosage compensation is mediated by multiple mechanisms.

In the case of several interrogated RP genes, dosage compensation is most likely due to feedback control (Figure 7). When cloned onto a low-copy plasmid with flanking intergenic regions, transcripts *RPL15A* and *RPL22A* did not increase despite increasing DNA copy (Figure 7A and B). This was true regardless of allele, ploidy or strain background, indicating that the dosage response of RP genes functions in laboratory-strain backgrounds. In contrast, the dosage response was not seen when several mitochondrial genes from Chr 12 were duplicated in isolation, since mRNA for the genes increased ~2X when the genes were amplified (Figure 7C and D). To further investigate the mechanism, we deleted the right arm of two of the four copies of acrocentric Chr 12 in the YPS1009_Chr12-4n strain (Figure 8A), thereby relieving aneuploidy at ~889 kb and 80% of the Chr 12 genes. For several mitochondrial genes that remained amplified in this strain, the dosage effect was lost, while for other genes expression remained lower than expected (Figure 8B). These results suggest a more complicated mechanism that will require further experiments to dissect.

We were unable to fit the MLR model to the W303 panel: the W303_Chr12-4n strain had very abnormal expression, as described above. Surprisingly, however, expression in the W303_Chr12-3n strain was highly correlated to the YPS1009_Chr12-3n strain: for two-thirds of the Class 3a genes, expression in W303_Chr12-3n was within 10% of YPS1009_Chr12-3n values. Thus the dosage compensation phenotype appeared in effect at some genes in the W303_Chr12-3n strain but was lost or obscured in the W303_Chr12-4n strain. One clear exception was the group of eleven mitochondrial genes on Chr 12 that showed lower-than-expected expression in the YPS1009_Chr12 panel but proportionately higher (or amplified) expression in the W303_Chr12 panel (Figure 8C). These results once again point to a difference in mitochondrial function in W303 compared to other strains.

Genes subject to dosage compensation are buffered against copy number variation

We sought to identify common features of dosage-compensated genes that may explain their tighter expression control. One prediction is that dosage compensation occurs at genes that are most toxic when over-expressed. Indeed, the combined set of 245 dosage-compensated genes is enriched for genes that are deleterious in very high copy in the laboratory strain (p = 0.009, (*38*)). A second prediction is posited by the Balance Hypothesis (*39*), which asserts that expression of multi-subunit protein complexes may

be more tightly controlled to maintain protein stoichiometry. We found weak enrichment for proteins in multi-subunit complexes (p = 0.03 (*40*)), but the significance eroded if ribosomal proteins were removed from the analysis. The group of dosage-compensated genes displayed slightly higher transcript abundance and higher RNA polymerase occupancy than the average gene (*41*), but the trends did not hold if genes related to translation were removed.

Dosage-compensated genes were enriched for genes that are toxic when highly abundant, raising the possibility that their expression may also be under greater evolutionary constraint. To investigate this, we compared the variance in gene expression seen in natural isolates subject to mutation and selection (the genetic variance, V_g) (*42*) to mutational variance (V_m) from MA lines (*43*). Genes that are under the highest constraint will have negative $\log_2(V_g/V_m)$ ratios reflecting that expression variation is being removed by purifying selection. Genes subject to dosage compensation and genes affected by heritably reduced expression show higher constraint compared to all genes, as expected for genes that are toxic when over-expressed (Figure 9A). The effect was also true for amplified genes that display proportionate expression upon amplification, which is likely a byproduct of our classification (since we effectively eliminated genes with variable expression across strains and replicates). The trends were consistent when genes belonging to all enriched functional groups were removed from the analysis.

An active mode of dosage compensation is predicted to constrain expression in the face of underlying gene amplification, thereby buffering copy-number variation (CNV). To test this hypothesis, we cataloged gene amplifications measured by array-CGH in 103 strains (*44-46*). In fact, dosage-compensated genes show considerably higher CNV than genes that display proportionate mRNA increase upon gene amplification (p =4e-8, Figure 9B). Somewhat surprisingly, genes with heritably reduced expression also showed relaxed constraint in CNV, which may reflect that some of these genes are dosage compensated in other strain backgrounds (see Discussion). Once again, the trends were not driven by enriched functional groups. We devised a CNV-buffering score, B_v, for each gene as the phylogeny-weighted sum of gene copy number across strains divided by V_g/V_m measured for that gene (see Methods) – genes with the strongest buffering capacity will therefore have the largest scores. Genes subject to dosage compensation show significantly higher B_v scores compared to all genes and compared to amplified genes with proportionately higher expression (p = 1.7e-4). Amplified genes with heritably reduced expression also showed higher buffering scores albeit lower than dosage-compensated genes (see Discussion). Together, these results show that genes subject to dosage-compensated expression can be buffered against CNV in *S. cerevisiae* populations.

Chr 12 amplification provides a selective advantage for an ecological trait.

The ability to buffer the toxicity of gene amplification could facilitate rapid evolution through aneuploidy. Aneuploidy in laboratory strains can be advantageous under adverse conditions, when non-dosage compensated defense genes are amplified in expression (*15, 16, 20, 47*); but the extent to which this occurs in nature is not known. We and others have previously shown that strains of *S. cerevisiae* and *S. paradoxus* from wintry climates have undergone selection to maintain freeze-thaw (FT) tolerance (*48, 49*). We noted that the *AQY2* water-transporter gene underlying FT tolerance (*49*) resides on Chr 12, which is amplified in New Jersey-oak strain YPS1009_Chr12-3n. *AQY2* escapes dosage compensation, resulting in higher *AQY2* expression across the

strain panel (Figure 10A). Indeed, we found that FT resistance improves with increasing Chr 12 copy number in YPS1009 (Figure 10B). The strong signatures of selection that we previously observed at the *AQY2* gene (*49*) and the clear relevance of the trait suggests that Chr 12 amplification is advantageous to YPS1009 in nature.

DISCUSSION

Our results provide new insight into a long-standing conundrum: how do cells tolerate duplication of toxic genes long enough for new phenotypes to evolve? Our results show that up to 40% of amplified genes in naturally aneuploid strains show lower-than-expected expression. The reduced expression is in part due to heritable polymorphisms that down-regulate expression from amplified genes, some of which may have been selected for during the adaptation to aneuploidy. But reduced expression of other amplified genes – up to 30-36% in some strains – appears to be actively regulated in proportion to increased gene dosage. The prevalence of the dosage compensation found here is likely true for other species and may explain the lack of expression increase previously noted for amplified autosomal genes in *Drosophila* (*50, 51*) and humans (*21, 52-54*).

The clear enrichment of specific functional classes – including genes encoding RP and mitochondrial proteins – points to specific targeted processes. While we did not find overwhelming support for the Balance Hypothesis, dosage compensated genes include several multi-subunit protein complexes and pathways. In the case of RP genes, dosage compensation most likely occurs via feedback to modulate mRNA abundance (Figure 7). Feedback is known to occur for several RP genes, including excess, unassembled L32 that binds a stem loop structure in its own transcript and inhibits intron splicing, perhaps triggering mRNA degradation (*55, 56*). Our results imply that feedback pertains more broadly to other RP genes, including *RPL15A* (Figure 7A) – notably, this transcript lacks an intron and thus must utilize a mechanism distinct from *RPL32*. The mechanism of dosage compensation at nuclear-encoded mitochondrial genes is less clear; however, several mitochondrial proteins are known to regulate translation or stability of other transcripts in shared pathways. For example, unassembled Cox1 protein can suppress translation of the *COX1* transcript (*57, 58*), while several TCA-cycle enzymes double as RNA-binding proteins that report on TCA activity (*59, 60*). A mode of dosage compensation that functions at the level of transcript abundance may provide an additional level of regulation to control protein stoichiometry.

Natural isolates are generally tolerant to aneuploidy, whether or not they have had time to adapt to the extra chromosome. We did observe a common expression response in different aneuploid strains, including the up-regulation of oxidoreductases and the down-regulation of genes involved in mitochondrial translation and respiration. Interestingly, these trends are also seen in Down syndrome (DS) cells, which are extremely sensitive to reactive-oxygen species (ROS) and, perhaps consequently, down-regulate mitochondrial respiration that in turn limits ROS production (*33*). We found no defect in respiratory capacity in naturally aneuploid yeast, and so the implications of this response in yeast are not clear. Human patients display significant variability in their sensitivity to DS, suggesting genetic effects on DS pathologies (*61, 62*). Our results also implicate strain-specific responses to Chr 8 amplification, in particular in the genes subject to dosage compensation versus heritably reduced expression. While some of the Class 2a genes could be mis-classified, these results raise the possibility of natural genetic variation in dosage compensation.

While the expression of dosage-compensated genes is controlled in proportion to gene copy, many genes escape dosage compensation in response to aneuploidy and can contribute significantly to phenotypic variation. Whole and partial chromosome amplification is common in laboratory selection experiments (*15, 20, 63*) and is particularly prevalent in laboratory mutants that have an extreme fitness defect (*16, 64, 65*). Yet the *Saccharomycotina* karyotype is extremely stable, holding at eight or sixteen chromosomes in most pre- and post-genome duplication species, respectively. This suggests that aneuploidy may serve as a transient intermediate, one that can be readily generated – and readily lost – depending on the selective pressures. A. Yona *et al.* showed that selection for increased stress tolerance in the laboratory produced aneuploid strains that eventually gave way to other solutions even when the selective pressure persisted (*20*). Our results show that aneuploidy is frequent and tolerated in nature and may provide an important route toward natural genetic variation.

MATERIALS AND METHODS

Strains and Growth Conditions

S. cerevisiae strains listed in Supplementary File 1 (diploid unless otherwise noted) were grown at 30°C in batch culture in YPD (1% yeast extract, 2% peptone and 2% dextrose) medium. We chose this growth regime as opposed to chemostat cultures because our strains show little growth differences across the strain panels and also to avoid confounding effects in interpreting gene expression. Given the instability of chromosomal amplification in W303, aneuploid W303 strains were grown under appropriate selection to maintain the aneuploidy and shifted to YPD batch culture growth at 30°C depending on the experiment (outlined below). Growth rates shown in Figure 1B were obtained on a TECAN infinite M200 PRO instrument (Tecan Austria GmbH), growing cells in 96-well plates without shaking and scoring doubling times using the program GrowthRates (*66*). Growth rates shown in Figure 1B for each aneuploid strain were normalized to the average doubling time for euploid strains in the same ecological group (*e.g.* clinical, natural, oak, vineyard, industrial, sake groups). Doubling times shown in Figure 1C, Figure 3, and Figure 5 were determined from batch cultures.

Forcing Aneuploidy in Euploid Parents

We followed the protocol outlined by Chen et al. (*47*) to isolate aneuploid derivatives. Haploid strains (derived from S288c (DBY7286), oak strain YPS163 or vineyard strain KCY40) were grown in YPD + 20ug/mL radicicol (AG Scientific) for 24 hours to promote chromosome instability and induction of aneuploidy. Cells were then washed 3X with YPD, grown in YPD for 24 hours, and plated on YPD + 8-32ug/mL fluconazole to select for Chr 8 aneuploidy, since amplification of *ERG11* on Chr 8 was shown to confer fluconazole resistance (*47*). Minimal inhibitory concentrations on fluconazole at which 90% of cells (MIC90) die were determined for each strain: 32, 8 and 16ug/mL for DBY7286, YPS163 and KCY40, respectively. Cells with Chr 8 amplification were initially screened via qPCR of select Chr 8 genes and confirmed by aCGH.

Generation of Aneuploid Isogenic Strain Panels

We generated isogenic aneuploid-euploid strain pairs from two strains, T73_Chr8-4n and YJM428_Chr16-4n (derived from a spore of the original parent, YJM428_Chr12-3n_Chr16-4n in which Chr 12 amplification was lost). T73_Chr 8-4n was passaged in liquid YPD culture for ~275 generations before a euploid derivative was identified by qPCR screening. YJM428_Chr16-4n was passaged for ~160 generations to produce a

triploid strain that was then dissected and underwent mating-type switching to generate a diploid, euploid YJM428_Chr16-2n derivative.

We also generated two wild strain panels from YPS1009 (also referred to as YPS1009_Chr12-3n for clarity in the text) and NCYC110 (aka NCYC110_Chr8-4n), and one laboratory strain panel for diploid W303_Chr12-3n (generated by mating a haploid disomic strain, generously provided by A. Amon, to the euploid W303). Strain panels for W303_Chr12 and YPS1009_Chr12 were generated by sporulating the trisomic parents and selecting haploid spores with either one or no extra copies of Chr 12. The YPS1009 spores then diploidized after mating-type switching to form isogenic diploid strains having two (2n, i.e. euploid) or four (4n) copies Chr 12. The haploid W303_Chr12 spores were crossed to appropriate W303 strains to generate isogenic diploids with two, three, or four copies of Chr 12. Homozygous strain NCYC110_Chr8-4n was passaged in liquid YPD for ~427 generations before an isolate with only three Chr 8 copies was identified by qPCR screening. (A similar procedure was performed for YPS1009_Chr12-4n and W303_Chr12-4n to determine the stability of the aneuploidies). A completely euploid derivative of NCYC110 was isolated shortly after at 444 generations. Aneuploidy of the respective chromosomes in each of the strain panels was initially verified via array comparative genomic hybridization (aCGH) and subsequently via genomic sequencing, with periodic confirmation by qPCR to ensure aneuploidy maintenance.

Genomic Sequencing

Genomic sequencing was performed on genomic DNA isolated with a Genomic-tip 20/G Kit (Qiagen 10223). 1 ug genomic DNA in a total volume of 50uL in a 0.5mL microcentrifuge tube was fragmented with a Diagenode Bioruptor sonication device (Diagenode) to a peak fragment size of 300-400bp, and 1 ug was used as an input into Illumina's TruSeq DNA Library Prep (Illumina FC-121-2001). Ligation products were purified with the E-Gel SizeSelect System (Life Technologies G6612ST). All cleanup steps in the genomic library prep were performed with Axy Prep MAG PCR Cleanup beads (Corning MAG-PCR-CL-250). Genomic libraries were sequenced on Illumina's HiSeq 2000 or HiSeq 2500 System, generating single end 100bp reads. Sequencing reads were processed with Trimmomatic version 0.30 (*67*) and reads were mapped to a reference genome with strain-specific SNPs (for YPS163, YPS1009, NCYC110, NCYC3290, and W303) or to the S288c reference 58 for all other strains, using BWA version 0.7.3 (*68*). HTseq version 0.5.4 (*69*) summed read counts per gene, which were then normalized for gene size and the number of reads generated per library via RPKM. Genomic DNA was sequenced in duplicates for all strain panels. Aneuploid strains were identified as triploid or tetraploid if the median RPKM signal across the chromosome was 0.4-0.7 (to call triploids) or 0.7->1.0 (to call tetraploids). Genome sequences are available in the NIH Sequence Read Archive (SRA) under accession SRP047341.

To identify potential polymorphisms in nominally isogenic aneuploid-euploid pairs, SNPs in each parental strain were called with GATK (*70*) and substituted into the S288c genome reference, which was then used as the reference for remapping of all related strains. GATK was used to call SNPs in each remapped strain. We found one homozygous SNP in NCYC110_Chr8-3n and a different homozygous SNP in NCYC110_Chr8-2n that were each identified in both replicates of the strain sequences. We performed a similar procedure for singleton genome sequences from YJM428_Chr16-2n and T73_Chr8-2n and identified 13 or 40 homozygous SNPs; these numbers are in the range of false-positive identifications found in a single replicate, as assessed from the NCYC110_Chr8 analysis above. We conclude from this analysis that

there are relatively few legitimate SNPs in passaged strains and that the majority of expression differences are a direct response to the aneuploidy.

RNA Sequencing (RNA-seq)

Cells were harvested for 3 min at 3000 r.p.m., after which time the cells were flash frozen in liquid nitrogen and maintained at -80°C until RNA extraction. Select samples (for T73_Chr8, YPS163_Chr8, and YJM428_Chr16 strain pairs and one replicate of the NCYC110_Chr8 strain panel) were collected and mixed at a 10:1 cellular ratio with *Schizosaccharomyces pombe* PR100, which had been grown in YES medium for >7 generations to OD_{600} ~0.5, killed with 0.125V ice cold quench solution (5% acid phenol in 100% EtOH), and collected. Cells were carefully counted on a hemocytometer to estimate cell counts. Total RNA was extracted with hot phenol as previously described (*71*), DNase-treated at 37°C for 30 min with TURBO DNase (Life Technologies AM2238), and then precipitated at -20°C in 2.5M LiCl for 30 min. rRNA depletion of the DNase-treated total RNA and subsequent cDNA library preparation were performed with ScriptSeq Complete Kit H/M/R (Epicentre BHMR1224), Index PCR Primers (Epicentre SSIP1234) and FailSafe PCR Enzyme Mix (Epicentre FSE51100). rRNA-depleted RNA was purified with a RNeasy MinElute Cleanup Kit (Qiagen 74204), while cDNA was purified with Axy Prep MAG PCR Cleanup beads (Corning MAG-PCR-CL-250). cDNA libraries were sequenced on Illumina's HiSeq 2000 System (UW-Madison DNA Sequencing Facility), generating single-end 100bp reads. Sequencing reads were processed with Trimmomatic version 0.30 (*67*) and reads were mapped to a reference genome with strain-specific SNPs (for YPS163, YPS1009, NCYC110, NCYC3290, and W303), to the S288c reference concatenated with the *Sz. pombe* genome (for doped samples listed above), or to the S288c reference 58 for all other strains, using BWA version 0.7.3(*68*). HTseq version 0.5.4(*69*) was used to obtain read counts per gene. Sequencing was done in biological triplicate for strain panels or biological duplicate for all strains except sake strains, which were done as singlets, with paired growth and strain collection for each replicate. For samples spiked with *Sz. pombe* controls, reads were normalized by a scaling factor such that the slope of the *Sz. pombe* reads across samples was 1.0, or normalized by traditional RPKM. The normalization methods produced data that was virtually indistinguishable; because RPKM-normalized data agreed better across biological replicates, RPKM normalization was used for all analyses described here. All RNA-seq data are available at NIH GEO accession GSE61532.

Transcriptome profiling was done for six aneuploid strains and paired controls (including YJM428 and YJM308, Y2189 and Y2209, K9 or K1 and K10, NCYC110 and NYCY3290, and YPS1009 and YPS163). In the case of K9/K1, NCYC110, and YPS1009 the control was chosen based on the genetically closest known relative at the time of the analysis based on phylogenetic comparisons (*28, 72, 73*). Controls for clinical isolate YJM428 and natural strain Y2189 were based on ecological group, choosing another clinical or natural isolate, respectively. RPKM values from each aneuploid were divided by RPKM measured from the euploid controls and logged, for both mRNA and DNA samples. For the analysis shown in Figure 4, we randomly chose two of the three replicates of YPS1009_Chr12-3n and NCYC110_Chr8-4n to produce statistical power comparable to the other strains analyzed in duplicate. Genes with lower-than-expected expression per gene copy (Figure 4) were identified as follows:

For each aneuploid chromosome in each strain, we first removed sub-telomeric genes with skewed measurements since these genes frequently show copy-number differences

across strains. We then calculated the chromosome-wide mean and standard deviation of the $\log_2$(aneuploid versus euploid DNA reads), across all genes on the affected chromosome. For each amplified gene being considered, we took the mean of the $\log_2$(aneuploid versus euploid DNA reads) measured for that gene specifically; we used this value minus one standard deviation of the chromosome-wide mean (or two standard deviations in the case of the sake strains) – this value served as a gene-specific cutoff for relative mRNA abundance. Genes with lower-than-expected expression per gene copy were identified if the $\log_2$(aneuploid versus euploid mRNA reads) was less than the filtering cutoff in both replicates (or one replicate for sake strains, which used a more stringent cutoff). This process identified genes whose relative mRNA abundance was lower than the relative DNA abundance at a high confidence interval. Genes whose $\log_2$(aneuploid versus euploid mRNA reads) difference was more than 1.5X lower (0.6 in $\log_2$ space) than the euploid control were excluded from the list, since their expression may be influenced by other effects. This identified 838 unique genes from the six aneuploid strains whose expression was lower-than-expected per gene copy, with 111 genes amplified and affected in >1 strain. We also identified amplified genes whose expression was distinctly not affected by dosage compensation as genes whose relative $\log_2$(mRNA abundance) was within the filtering cutoff for both replicates, excluding genes whose $\log_2$ expression difference was more than 2.5X higher than the euploid control, since their expression may be influenced by other effects. This identified 928 genes whose expression was in the expected range if *no* dosage compensation is at work. A subset of these genes showed expression that was at least 1.5X higher than expected, and these are annotated as magenta points in Figure 4.

Genes that were differentially expressed in aneuploid versus euploid strains (regardless of DNA abundance) were identified with the program edgeR (*74*). Because of limited power for the duplicated datasets, we also included genes whose expression was greater than 1.3X ($\log_2$ of 0.4) different from the euploid control in three of the six strains. Clustering of expression was done using the program Cluster 3.0 (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm) and visualized by Java Treeview (http://jtreeview.sourceforge.net). Enrichment of functional groups was done using the program FunSpec (*75*). Enrichment of proteins in complexes was scored by hypergeometric test, comparing to curated complexes from (*40*); differences in degree from protein interaction networks was scored by T-test comparing the degree compiled by (*41*) from genes in Class 3a versus other gene groups.

Mixture of Linear Regressions (MLR) Model

A novel MLR model was developed to classify genes based on their expression across the two wild-strain panels. Preprocessed mRNA and DNA abundance data were analyzed further in the context of a new mixture of linear regressions (MLR) model in order to cluster genes according to their DNA/mRNA relationship. Within a given strain panel we modeled the relationship between DNA abundance at gene g from sample i, say $X_{g,i}$, and mRNA abundance $Y_{g,i}$, where both are normalized to euploid controls and considered on the logarithmic scale. Briefly, we first filtered genes that exhibited a nonlinear relationship using a likelihood ratio test (Supplementary File 3). The interpretation of these profiles is not clear and therefore they were removed from further consideration. Data from the remaining genes, which exhibited a linear relationship $Y_{g,i} = A_g + B_g X_{g,i} + \text{error}_{g,i}$ were fit to a discrete, random-effects mixture model in order to produce a relationship classifier (*76*). In the proposed MLR model, the five discrete classes correspond to constraints on the intercept α and slope β parameters of the linear regression: (Class 1) $\alpha_g = 0$, $\beta_g = 1$, (Class 2a) $\alpha_g < 0$, $\beta_g=1$, (Class 2b) $\alpha_g>0$, $\beta =1$,

(Class 3a) β <1, (Class 3b) β_g > 1. Class 1 represents genes with no dosage compensation and statistically indistinguishable expression between the 2n strain and paired euploid control. Class 2 represents genes whose expression increases linearly across the strain panel, but whose expression in the aneuploid derivatives is lower (Class 2a) or higher (Class 2b) per gene copy than the paired euploid. In contrast, Class 3 genes show expression changes across the strain panel that are disproportionate to the change in DNA copy, either showing lower (Class 3a) or greater (Class 3b) expression per gene copy as the aneuploidy increases. The inference is stabilized by treating the gene-specific slopes and intercepts as random effects that are constrained by the class structure; this effectively reduces the dimension of the parameter space. Fortuitously, explicit formula are available for the probability density of data within each class. The expectation maximization (EM) algorithm was used to estimate the MLR parameters; subsequently, genes were clustered together if they had a high posterior probability of arising from the same discrete class. Genes were classified based on the maximum posterior probability (Supplementary File 3). Further details are provided in Appendix 1. Unamplified genes in the two strain panels that displayed linear increases in expression in proportion to the amplified chromosome copy number were identified based on the probability of linear fit in R, using Benjamini and Hochberg multiple test correction (*77*).

Gene cloning, Chr 12 arm deletion, and qPCR

W303 and/or YPS1009 alleles of each query gene were cloned onto a KANMX-marked CEN plasmid via homologous recombination in strain BY4741, from which plasmids were isolated and used to transform haploid or diploid versions of YPS1009_Chr12-2n, euploid W303, or euploid BY4743, where noted. Strains were grown in YPD + G418 for at least 3 generations to OD_{600} ~0.3. Cells were collected for genomic-DNA and RNA preps to measure copy number and transcript levels, respectively, as described above. RNA was DNase-digested and subsequently used to synthesize cDNA with an Oligo (dT) primer and Superscript III reverse transcriptase (Invitrogen, Carlsebad, CA). Quantitative PCR (qPCR) was performed with SYBR-Green as previously described (*78*). mRNA or DNA abundance measured for each gene was normalized to an internal control (*ERV25*) and compared to a strain carrying an empty vector.

To test the effect of aneuploidy dose on gene expression, a strain carrying an extra copy of the left-arm of Chr 12 only was generated (referred to a the “mini-4n strain”), by deleting the right arm of Chr 12 from the YPS1009_Chr12-4n strain, as follows: The KANMX cassette was PCR amplified with homology to the promoter region of *GAT3*, which flanks the Chr 12 centromere, or the promoter region of subtelomerically encoded gene YLR460C. The diploid, aneuploid YPS1009_Chr12-4n strain was transformed and G418-resistant colonies were selected and sporulated. One spore showed the expected 2/4 or 0/4 spore survival when dissected onto YPD + G418. PCR was used to confirm integration of the KANMX cassette and aCGH was used to confirm that the diploid cells are aneuploid for the left arm of Chr 12, but not the right arm. Expression levels of the amplified Class 3a genes were assessed by qPCR or DNA-microarray analysis as described above.

Expression constraint and CNV in wild strains

V_g/V_m values were calculated for each gene as follows: genetic variance V_g was taken as the variance in expression across 22 wild strains analyzed in (*42*). V_m was taken as the variance in expression across MA lines generated in the absence of selection by (*43*). The ratio of V_g/V_m represents the effects of mutation + selection as measured in

natural strains versus the effects of mutation in the near-absence of selection as measured in MA lines. CNV data were taken from (*44-46*), which measured gene copy number by aCGH. Strains whose aCGH profiles were correlated >0.9 were represented by a single, randomly chosen strain, to avoid over-representing strains that are highly related. This left 103 strains with aCGH values for each gene. Amplified genes were identified if the relative gene abundance was at least 1.6X higher than the S288c reference. Genes that showed CNV in at least one of the 103 strains were compared in Figure 9B. To calculate B_v, a measure of the buffering capacity of each gene, we summed the number of strains in which the gene was amplified, weighted by the aCGH similarity weights (described in http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm) as a proxy for strain relatedness. The weighting was applied to control for genetic similarities in the strains, to therefore highlight independent gene amplification events. For each gene, this value was divided by its calculated V_g/V_m ratio. Larger values represent a higher propensity for CNV coupled with more stringent expression constraint (i.e. small V_g/V_m).

Freeze-Thaw Tolerance

The YPS1009 isogenic strains and YPS163 were grown in YPD to OD_{600} ~0.3, and 100uL of cells was subjected to dry ice/ethanol bath freezing (< -50°C) or ice as previously described (*49*). Viability was determined using Live/Dead straining (Life Technologies L34952) on a Guava EasyCyte Plus flow cytometer (Guava Technologies, Inc., Hayward, CA). Percentages of cells surviving freeze/thaw are reported for 3 biological replicates.

AUTHOR CONTRIBUTIONS. JH & APG conceived of the experiments. JH generated strain panels, conducted RNA and DNA sequencing and initial analysis, and performed all other experimental studies. MS conducted genome sequencing and growth analysis. CMY generated the mini-4n strain, assisted in mini-4n analysis, and assisted in cloning and qPCR quantification. ZW & MAN conceived of and implemented statistical models and analysis. APG performed all downstream analyses and wrote the manuscript, with input from all authors.

ACKNOLWEDGEMENTS: We thank A. Amon for providing the haploid W303 aneuploid strain, M. Place for informatics support, and D. Gottschling and members of the Gasch Lab for useful discussions. This work was supported by NIH grants R01GM083989 (to APG) and R21HG006568 (to MAN), with partial support from the Department of Energy Great Lakes Bioenergy Research Center (DE–FC02–07ER64494) and fellowships from Morgridge Institute of Research for ZW and NSF for M. Sardi.

Table 1: MLR Gene Classifications

Gene Class	NCYC110 (Chr 8)	YPS1009 (Chr 12)
Non-linear	19 (7%)	32 (7%)
Class 1	16 (6%)	81 (16%)
Class 2a	138 (49%)	137 (28%)
Class 2b	32 (11%)	4 (1%)
Class 3a	30 (11%)	142 (29%)
Class 3b	46 (16%)	96 (20%)
TOTAL	281	492

The number and percentage of genes classified in each group. Functional enrichments are listed in Supplementary File 2.

SUPPLEMENTARY FILES

Supplementary File 1: Strains used in this study

Supplementary File 2: Functional enrichments and p-values for gene classes. All enrichments with p< 2e-3 are shown; enrichments that meet the Bonferroni cutoff (2e-5) are highlighted in red. Enrichments are listed for a) amplified genes with expression proportionate to gene copy (from Figure 4A), b) pooled set of amplified genes with expression proportionate to gene copy (from Figure 9), c) dosage compensated genes (from Figure 9). d) genes in Class 2a, e) genes with higher-than-expected expression (from Figure 4A and 4B plus Class 3b genes).

Supplementary File 3: Genes in each class are listed on file tabs. The first two columns list gene name and annotation, while the third column lists the strain and/or analysis that the gene was identified in. a) Amplified genes with expression proportionate to gene copy. b) Amplified genes with lower-than-expected expression. c) Amplified genes with higher-than-expected expression. d. Genes in each MLR classification group – intercept ("alpha-hat"), slope ("beta-hat"), and posterior probabilities for each class are shown. The mean DNA and mRNA sequencing reads for each gene are also shown.. e. Unamplified genes in the strain panels with linear increases in expression across the panel (FDR = 0.05). Pooled set of amplified genes with expression proportionate to gene copy (from Figure 9). c) Dosage compensated genes (from Figure 9).

REFERENCES

1. S. Ohno, *Evolution by Gene Duplication*. (Springer-Verlag, 1970), pp. 160.
2. O. Jaillon, J. M. Aury, P. Wincker, "Changing by doubling", the impact of Whole Genome Duplications in the evolution of eukaryotes. *Comptes rendus biologies* **332**, 241-253 (2009).
3. S. L. Chang, H. Y. Lai, S. Y. Tung, J. Y. Leu, Dynamic large-scale chromosomal rearrangements fuel rapid adaptation in yeast populations. *PLoS genetics* **9**, e1003232 (2013).
4. L. Sandegren, D. I. Andersson, Bacterial gene amplification: implications for the evolution of antibiotic resistance. *Nature reviews. Microbiology* **7**, 578-588 (2009).
5. C. V. Edi, L. Djogbenou, A. M. Jenkins, K. Regna, M. A. Muskavitch, R. Poupardin, C. M. Jones, J. Essandoh, G. K. Ketoh, M. J. Paine, B. G. Koudou, M. J. Donnelly, H. Ranson, D. Weetman, CYP6 P450 enzymes and ACE-1 duplication produce extreme and multiple insecticide resistance in the malaria mosquito Anopheles gambiae. *PLoS genetics* **10**, e1004236 (2014).
6. M. Lynch, A. Force, The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459-473 (2000).
7. M. Lynch, M. O'Hely, B. Walsh, A. Force, The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**, 1789-1804 (2001).

8. X. Gu, Z. Zhang, W. Huang, Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 707-712 (2005).
9. W. H. Li, J. Yang, X. Gu, Expression divergence between duplicate genes. *Trends Genet* **21**, 602-607 (2005).
10. Y. Wang, X. Wang, A. H. Paterson, Genome and gene duplications and gene expression divergence: a view from plants. *Annals of the New York Academy of Sciences* **1256**, 1-14 (2012).
11. Z. Gu, S. A. Rifkin, K. P. White, W. H. Li, Duplicate genes increase gene expression diversity within and between species. *Nature genetics* **36**, 577-579 (2004).
12. A. Presser, M. B. Elowitz, M. Kellis, R. Kishony, The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 950-954 (2008).
13. L. Freschi, M. Courcelles, P. Thibault, S. W. Michnick, C. R. Landry, Phosphorylation network rewiring by gene duplication. *Molecular systems biology* **7**, 504 (2011).
14. R. De Smet, Y. Van de Peer, Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current opinion in plant biology* **15**, 168-176 (2012).
15. M. J. Dunham, H. Badrane, T. Ferea, J. Adams, P. O. Brown, F. Rosenzweig, D. Botstein, Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 16144-16149 (2002).
16. T. R. Hughes, C. J. Roberts, H. Dai, A. R. Jones, M. R. Meyer, D. Slade, J. Burchard, S. Dow, T. R. Ward, M. J. Kidd, S. H. Friend, M. J. Marton, Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature genetics* **25**, 333-337 (2000).
17. M. Ni, M. Feretzaki, W. Li, A. Floyd-Averette, P. Mieczkowski, F. S. Dietrich, J. Heitman, Unisexual and heterosexual meiotic reproduction generate aneuploidy and phenotypic diversity de novo in the yeast Cryptococcus neoformans. *PLoS biology* **11**, e1001653 (2013).
18. N. Pavelka, G. Rancati, J. Zhu, W. D. Bradford, A. Saraf, L. Florens, B. W. Sanderson, G. L. Hattem, R. Li, Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature* **468**, 321-325 (2010).
19. A. Selmecki, A. Forche, J. Berman, Aneuploidy and isochromosome formation in drug-resistant Candida albicans. *Science (New York, N.Y* **313**, 367-370 (2006).
20. A. H. Yona, Y. S. Manor, R. H. Herbst, G. H. Romano, A. Mitchell, M. Kupiec, Y. Pilpel, O. Dahan, Chromosomal duplication is a transient evolutionary solution to stress. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 21010-21015 (2012).
21. S. E. Antonarakis, R. Lyle, E. T. Dermitzakis, A. Reymond, S. Deutsch, Chromosome 21 and down syndrome: from genomics to pathophysiology. *Nature reviews* **5**, 725-738 (2004).
22. J. J. Siegel, A. Amon, New insights into the troubles of aneuploidy. *Annual review of cell and developmental biology* **28**, 189-214 (2012).
23. E. M. Torres, T. Sokolsky, C. M. Tucker, L. Y. Chan, M. Boselli, M. J. Dunham, A. Amon, Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science (New York, N.Y* **317**, 916-924 (2007).

24. J. M. Sheltzer, E. M. Torres, M. J. Dunham, A. Amon, Transcriptional consequences of aneuploidy. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 12644-12649 (2012).
25. E. M. Torres, N. Dephoure, A. Panneerselvam, C. M. Tucker, C. A. Whittaker, S. P. Gygi, M. J. Dunham, A. Amon, Identification of aneuploidy-tolerating mutations. *Cell* **143**, 71-83 (2010).
26. N. Dephoure, S. Hwang, C. O'Sullivan, S. E. Dodgson, S. P. Gygi, A. Amon, E. M. Torres, Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *eLife*, e03023 (2014).
27. M. Springer, J. S. Weissman, M. W. Kirschner, A general lack of compensation for gene dosage in yeast. *Molecular systems biology* **6**, 368 (2010).
28. D. J. Kvitek, J. L. Will, A. P. Gasch, Variations in stress sensitivity and genomic expression in diverse S. cerevisiae isolates. *PLoS genetics* **4**, e1000223 (2008).
29. J. A. Lewis, I. M. Elkon, M. A. McGee, A. J. Higbee, A. P. Gasch, Exploiting natural variation in Saccharomyces cerevisiae to identify genes for increased ethanol resistance. *Genetics* **186**, 1197-1205 (2010).
30. J. A. Lewis, A. P. Gasch, Natural variation in the yeast glucose-signaling network reveals a new role for the Mig3p transcription factor. *G3 (Bethesda)* **2**, 1607-1612 (2012).
31. Y. O. Zhu, M. L. Siegal, D. W. Hall, D. A. Petrov, Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E2310-2318 (2014).
32. A. Conti, F. Fabbrini, P. D'Agostino, R. Negri, D. Greco, R. Genesio, M. D'Armiento, C. Olla, D. Paladini, M. Zannini, L. Nitsch, Altered expression of mitochondrial and extracellular matrix genes in the heart of human fetuses with chromosome 21 trisomy. *BMC genomics* **8**, 268 (2007).
33. P. Helguera, J. Seiglie, J. Rodriguez, M. Hanna, G. Helguera, J. Busciglio, Adaptive downregulation of mitochondrial function in down syndrome. *Cell metabolism* **17**, 132-140 (2013).
34. C. Lintas, R. Sacco, A. M. Persico, Genome-wide expression studies in autism spectrum disorder, Rett syndrome, and Down syndrome. *Neurobiology of disease* **45**, 57-68 (2012).
35. D. Valenti, L. de Bari, B. De Filippis, A. Henrion-Caude, R. A. Vacca, Mitochondrial dysfunction as a central actor in intellectual disability-related diseases: An overview of Down syndrome, autism, Fragile X and Rett syndrome. *Neuroscience and biobehavioral reviews*, (2014).
36. Z. Tan, M. Hays, G. A. Cromie, E. W. Jeffery, A. C. Scott, V. Ahyong, A. Sirr, A. Skupin, A. M. Dudley, Aneuploidy underlies a multicellular phenotypic switch. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 12367-12372 (2013).
37. C. Y. Wu, P. A. Rolfe, D. K. Gifford, G. R. Fink, Control of transcription by cell size. *PLoS biology* **8**, e1000523 (2010).
38. K. Makanae, R. Kintaka, T. Makino, H. Kitano, H. Moriya, Identification of dosage-sensitive genes in Saccharomyces cerevisiae using the genetic tug-of-war method. *Genome Res* **23**, 300-311 (2013).
39. J. A. Birchler, R. A. Veitia, The gene balance hypothesis: from classical genetics to modern genomics. *The Plant cell* **19**, 395-402 (2007).
40. S. Pu, J. Wong, B. Turner, E. Cho, S. J. Wodak, Up-to-date catalogues of yeast protein complexes. *Nucleic acids research* **37**, 825-831 (2009).
41. D. Chasman, Y. H. Ho, D. B. Berry, C. M. Nemec, M. E. MacGilvray, J. Hose, A. E. Merrill, M. V. Lee, J. L. Will, J. J. Coon, A. Z. Ansari, M. Craven, A. P. Gasch,

Pathway connectivity and signaling coordination in the yeast stress-activated signaling network. *Molecular systems biology* **10**, 759 (2014).
42. D. A. Skelly, G. E. Merrihew, M. Riffle, C. F. Connelly, E. O. Kerr, M. Johansson, D. Jaschob, B. Graczyk, N. J. Shulman, J. Wakefield, S. J. Cooper, S. Fields, W. S. Noble, E. G. Muller, T. N. Davis, M. J. Dunham, M. J. Maccoss, J. M. Akey, Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res* **23**, 1496-1504 (2013).
43. C. R. Landry, B. Lemos, S. A. Rifkin, W. J. Dickinson, D. L. Hartl, Genetic properties influencing the evolvability of gene expression. *Science (New York, N.Y* **317**, 118-121 (2007).
44. L. Carreto, M. F. Eiriz, A. C. Gomes, P. M. Pereira, D. Schuller, M. A. Santos, Comparative genomics of wild type yeast strains unveils important genome diversity. *BMC genomics* **9**, 524 (2008).
45. B. Dunn, R. P. Levine, G. Sherlock, Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. *BMC genomics* **6**, 53 (2005).
46. B. Dunn, C. Richter, D. J. Kvitek, T. Pugh, G. Sherlock, 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 (2012).
47. G. Chen, W. D. Bradford, C. W. Seidel, R. Li, Hsp90 stress potentiates rapid cellular adaptation through induction of aneuploidy. *Nature* **482**, 246-250 (2012).
48. J. B. Leducq, G. Charron, P. Samani, A. K. Dube, K. Sylvester, B. James, P. Almeida, J. P. Sampaio, C. T. Hittinger, G. Bell, C. R. Landry, Local climatic adaptation in a widespread microorganism. *Proceedings. Biological sciences / The Royal Society* **281**, 20132472 (2014).
49. J. L. Will, H. S. Kim, J. Clarke, J. C. Painter, J. C. Fay, A. P. Gasch, Incipient balancing selection through adaptive loss of aquaporins in natural Saccharomyces cerevisiae populations. *PLoS genetics* **6**, e1000893 (2010).
50. J. Zhou, B. Lemos, E. B. Dopman, D. L. Hartl, Copy-number variation: the balance between gene dosage and expression in Drosophila melanogaster. *Genome biology and evolution* **3**, 1014-1024 (2011).
51. Y. Zhang, J. H. Malone, S. K. Powell, V. Periwal, E. Spana, D. M. Macalpine, B. Oliver, Expression in aneuploid Drosophila S2 cells. *PLoS biology* **8**, e1000320 (2010).
52. B. Schuster-Bockler, D. Conrad, A. Bateman, Dosage sensitivity shapes the evolution of copy-number varied regions. *PloS one* **5**, e9474 (2010).
53. C. Woodwark, A. Bateman, The characterisation of three types of genes that overlie copy number variable regions. *PloS one* **6**, e14814 (2011).
54. R. T. Wang, S. Ahn, C. C. Park, A. H. Khan, K. Lange, D. J. Smith, Effects of genome-wide copy number variation on expression in mammalian cells. *BMC genomics* **12**, 562 (2011).
55. M. D. Dabeva, J. R. Warner, The yeast ribosomal protein L32 and its gene. *The Journal of biological chemistry* **262**, 16055-16059 (1987).
56. J. Vilardell, J. R. Warner, Ribosomal protein L32 of Saccharomyces cerevisiae influences both the splicing of its own transcript and the processing of rRNA. *Molecular and cellular biology* **17**, 1959-1965 (1997).
57. D. U. Mick, T. D. Fox, P. Rehling, Inventory control: cytochrome c oxidase assembly regulates mitochondrial translation. *Nature reviews. Molecular cell biology* **12**, 14-20 (2011).

58. A. P. Van Loon, E. Van Eijk, L. A. Grivell, Biosynthesis of the ubiquinol-cytochrome c reductase complex in yeast. Discoordinate synthesis of the 11-kd subunit in response to increased gene copy number. *The EMBO journal* **2**, 1765-1770 (1983).
59. L. de Jong, S. D. Elzinga, M. T. McCammon, L. A. Grivell, H. van der Spek, Increased synthesis and decreased stability of mitochondrial translation products in yeast as a result of loss of mitochondrial (NAD(+))-dependent isocitrate dehydrogenase. *FEBS letters* **483**, 62-66 (2000).
60. S. D. Elzinga, A. L. Bednarz, K. van Oosterum, P. J. Dekker, L. A. Grivell, Yeast mitochondrial NAD(+)-dependent isocitrate dehydrogenase is an RNA-binding protein. *Nucleic acids research* **21**, 5328-5331 (1993).
61. P. Prandini, S. Deutsch, R. Lyle, M. Gagnebin, C. Delucinge Vivier, M. Delorenzi, C. Gehrig, P. Descombes, S. Sherman, F. Dagna Bricarelli, C. Baldo, A. Novelli, B. Dallapiccola, S. E. Antonarakis, Natural gene-expression variation in Down syndrome modulates the outcome of gene-dosage imbalance. *American journal of human genetics* **81**, 252-263 (2007).
62. R. J. Roper, R. H. Reeves, Understanding the basis for Down syndrome phenotypes. *PLoS genetics* **2**, e50 (2006).
63. D. Gresham, M. M. Desai, C. M. Tucker, H. T. Jenq, D. A. Pai, A. Ward, C. G. DeSevo, D. Botstein, M. J. Dunham, The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS genetics* **4**, e1000303 (2008).
64. G. Rancati, N. Pavelka, B. Fleharty, A. Noll, R. Trimble, K. Walton, A. Perera, K. Staehling-Hampton, C. W. Seidel, R. Li, Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. *Cell* **135**, 879-893 (2008).
65. R. Koszul, S. Caburet, B. Dujon, G. Fischer, Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *The EMBO journal* **23**, 234-243 (2004).
66. B. G. Hall, H. Acar, A. Nandipati, M. Barlow, Growth rates made easy. *Molecular biology and evolution* **31**, 232-238 (2014).
67. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**, 2114-2120 (2014).
68. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754-1760 (2009).
69. S. Anders, P. T. Pyl, W. Huber, HTSeq — A Python framework to work with high-throughput sequencing data. *bioRxiv preprint*, (2014).
70. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498 (2011).
71. A. P. Gasch, Yeast genomic expression studies using DNA microarrays. *Methods in enzymology* **350**, 393-414 (2002).
72. J. C. Fay, J. A. Benavides, Evidence for domesticated and wild populations of Saccharomyces cerevisiae. *PLoS genetics* **1**, 66-71 (2005).
73. G. Liti, D. M. Carter, A. M. Moses, J. Warringer, L. Parts, S. A. James, R. P. Davey, I. N. Roberts, A. Burt, V. Koufopanou, I. J. Tsai, C. M. Bergman, D. Bensasson, M. J. O'Kelly, A. van Oudenaarden, D. B. Barton, E. Bailes, A. N. Nguyen, M. Jones, M. A. Quail, I. Goodhead, S. Sims, F. Smith, A. Blomberg, R.

Durbin, E. J. Louis, Population genomics of domestic and wild yeasts. *Nature* **458**, 337-341 (2009).
74. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**, 139-140 (2010).
75. M. D. Robinson, J. Grigull, N. Mohammad, T. R. Hughes, FunSpec: a web-based cluster interpreter for yeast. *BMC bioinformatics* **3**, 35 (2002).
76. G. McLachlan, D. Peel, *Finite Mixture Models*. Wiley Series in Probability and Statistics (Wiley, New York, 2000).
77. Y. Benjamini, Y. Hochberg, Controllign the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* **57**, 289-300 (1995).
78. M. V. Lee, S. E. Topper, S. L. Huberl, J. Hose, C. D. Wenger, J. Coon, A. P. Gasch, A Dynamic Model of Proteome Changes Reveals New Roles for Transcript Alteration in Yeast *Molecular systems biology*, (2011).

FIGURE LEGENDS

Figure 1. Aneuploidy is common in non-laboratory strains. A) Relative DNA copy (log_2 (RPKM)) per gene (rows) across each of the 16 chromosomes for 47 sequenced strains (columns). Strains with 1.5X (grey text) or 2X (blue text) chromosomal copy per haploid genome are annotated by name. **B**) Growth rates of aneuploid strains normalized to the niche-specific growth rate (see Methods) plotted against the additional DNA content in each strain. **C**) Haploid strains with a duplication of Chr 8 (labeled as 'disome') were selected from euploid parents (labeled as 'monosome') including the S288c-derived DBY7286, derived vineyard strain KCY40, and a haploid derivative of oak strain YPS163, see Methods. Doubling times in YPD medium represent the average of four biological replicates; the average difference ('dif') in growth rate is indicated. An asterisk represents statistically significant differences in doubling time (p<6e-4, T-test).

Figure 2. Naturally aneuploid strains show a weak common response to aneuploidy but no activation of the ESR. **A**) Expression of genes in the yeast environmental stress response (ESR) in aneuploid versus genetically related euploid strains. Biological replicates are shown for YJM428, Y2189, YPS1009, and NCYC110. The magnitude of the expression difference (log_2 fold change) is as indicated in the key. **B**) Expression of 69 genes with higher expression (top) and 263 genes with lower expression (bottom) in aneuploid strains versus their paired euploid control.

Figure 3. Naturally aneuploid strains have no respiratory defect. The average and standard deviation of doubling times measured for strains growing in yeast extract-peptone medium supplemented with 2% glucose (fermentable), 2% ethanol, 3% glycerol, or 2% acetate (non-fermentable) is shown for indicated strains. Isogenic, diploid panels of YPS100, NCYC110, and W303 carried variable copy of Chr 12 or Chr 8 (see text). The relative growth rate on non-fermentable carbon sources (normalized to each strain's growth rate on glucose) was not significantly different for most aneuploid strains, with the exception of the W303_Chr12-3n strain that showed a greater growth defect when grown on ethanol or acetate compared to glucose (+ = p<0.07, * = p<0.05, paired T-test). The W303_Chr12-4n did not grow on non-fermentable carbon sources.

Figure 4. A large fraction of amplified genes are expressed lower than expected in aneuploid isolates. The average log_2(aneuploid versus euploid RNA-seq reads) per amplified gene was plotted against average log_2(aneuploid versus euploid DNA seq

reads) measured for that gene, in each of the interrogated aneuploid strains. A) shows genes measured in each wild isolate normalized to a genetically related euploid reference strain and B) shows genes measured in two tetrasomic diploid strains and one disomic haploid strain normalized to isogenic euploids. Genes with lower-than-expected expression are plotted in blue and genes with higher-than-expected expression are plotted in magenta (see Methods). The expected relationship representing proportionate increases in expression (slope of 1.0, intercept of 0) is shown in each panel.

Figure 5. Aneuploid strain panels. A) Strain panels were generated by sporulating YPS1009 (naturally trisomic for Chr 12) or serial passaging of NCYC110 (naturally tetrasomic for Chr 8). **B**) Average and standard deviation of doubling times for strains in each panel and euploid reference strains YPS163 or NCYC3290, normalized to the respective 2n strain in each panel. Asterisks represent significant differences (p<0.01).

Figure 6. Classification of dosage-responsive genes. The $\log_2$(aneuploid versus euploid RNA-seq reads) for each gene were plotted against the $\log_2$(aneuploid versus euploid DNA-seq reads) measured for that gene in the -2n, -3n, and -4n strains within the strain panel normalized to the closely related euploid reference. **A-E**) Average and standard deviation of representative genes in the denoted classes, across three biological replicates. **F**) All 173 Class 3a genes, normalized to the euploid strain from within the respective panel for clarity. The expected relationship representing proportionate increases in expression (slope of 1.0, intercept of 0) is shown in each panel.

Figure 7. Expression response when genes are duplicated on a plasmid. To measure the effects of gene duplication in the absence of aneuploidy, representative Class 3a genes were cloned onto a CEN plasmid and introduced into otherwise euploid strains. As a control, strains were also compared to an empty vector for normalization. mRNA and DNA abundance for the gene were quantified by qPCR as described in Methods. Relative DNA (black) and RNA (grey) abundance for A) *RPL15A*, B) *RPL22A*, C) *COX8*, or D) *MEF1* when introduced in extra copy into each strain. Panels A, C, D show abundance in haploid W303 or the haploid, monosomic (euploid) derivative of YPS1009 ('YPS1009_m') and panel B shows abundance in diploid laboratory strain BY4743 and diploid, euploid derivative YPS1009_Chr12-2n. In all cases, the average and standard deviation of biological triplicates is shown.

Figure 8. Relative mRNA abundance of amplified genes in the mini-4n strain and strain panels. A) Cartoon diagram of the four copies of Chr 12 in the tetraploid YPS1009_Chr12-4n strain and in the "mini_4n" strain where two copies of the Chr 12-right arm are deleted (see Methods for details). **B)** Expression was measured on tiled yeast-genome DNA microarrays. Shown are Class 3a genes that remain amplified in the mini-4n strain and could be quantified by arrays. Relative mRNA abundance was measured in biological triplicate in the YPS1009_Chr12-4n or YPS1009_mini-Chr12-4n strain versus the euploid YPS1009_Chr12-2n strain. Genes *COX17*, *SPA2*, and *SDH2* showed an increase in expression in the mini_4n strain, however expression remained significantly lower than the expected two-fold difference proportionate to the gene amplification. *FRE6* showed little dosage compensation when measured by DNA microarray analysis and correspondingly its expression did not change in the mini_4n strain. **C**) Relative mRNA abundance of eleven genes that are dosage compensated across the YPS1009_Chr12 strain panel (left) but not the W303_Chr12 strain panel (right).

Figure 9. Natural variation in expression constraint and gene copy-number variation. **A**) $\log_2(V_g/V_m)$ as described in the text, **B**) fraction of genes with copy-number variation (CNV) in at least one of 103 strains, and C) log2 of the buffering capacity score, *Bv*, are shown for the gene groups in the key (where number of genes is indicated in parentheses). Amplified genes with proportionately higher expression include Class 1 genes and genes with proportionate expression in strain pairs (Figure 4A, grey points). Amplified genes with heritably reduced expression correspond to genes in Class 2a, minus genes identified as dosage compensated in other strains, while amplified genes subject to dosage compensation represent genes in Class 3a plus genes with lower-than-expected expression as identified in both Figure 4A and 4B. Statistical significance was scored by comparing each group to the total set of yeast genes.

Figure 10. Amplified expression of *AQY2* provides a fitness advantage **A**) *AQY2* expression as described in Figure 6. **B**) Average and standard deviation of percent cell viability after freeze-thaw stress.

Appendix
Main paper: Dosage-responsive gene expression buffers copy-number variation in wild yeast
Authors: J Hose, CM Yong, M Sardi, Z Wang, MA Newton, and AP Gasch

2 Statistical analysis of DNA/RNA association

2.1 Initial Data Processing

The data analysis pipeline for each strain panel involved a series of preprocessing steps to prepare data for the mixture-model-based cluster analysis.

1. *Generation:* The sequencing facility at the UW Biotechnology Center aligned raw reads and produced read counts per yeast gene for both DNA and RNA from a number of samples in each strain panel.

2. *Coverage filter*: We removed genes for which mean (over different experimental cells) RNA counts per gene < 5. This cutoff was defined empirically based on reproducibility across the strain panels.

3. *Zeros:* Where RNA or DNA count was 0, we replaced by 1/2 to avoid downstream log errors.

4. *DNA/RNA Alignment:* Genomic DNA counts were measured in biological duplicate. The replicates were averaged for each strain, and then the average normalized DNA counts per gene were used as the reference point for all three biological RNA measurements for that strain.

5. *Library size adjustment:* For each profile, we summed over all chromosome library sizes and corrected for library size. I.e., we divided counts by profile-specific (whole genome) library size.

6. *Reference normalization:* We divided each experimental profile from the aneroid strains by its associated (via replicate) euploid control profile.

7. *Log transform:* We transformed relative counts to natural logarithm scale.

8. *Aneuploid filter:* We restricted to chromosome 8 genes (for West African strain NCYC110) or to chromosome 12 genes (for the YPS panel).

9. *Nonlinear filter:* We filtered genes showing nonlinear DNA/RNA relationship on the transformed scale (see non-linear filter, below)

10. *Output:* We sent the processed data to the mixture of linear regression (MLR) clustering calculation.

2.2 Non-linear filter

We had 9 DNA measurements (3 repeated measures for 2n, 3n and 4n, respectively) and 9 RNA measurements, for each gene under study and for both strain panels. From above, these measurements were on a logarithmic scale, and normalized to a reference strain. Owing to design of the controls, the DNA values were identical across the three replicate samples within each strain, and

recorded changes in the relative DNA copy number. If RNA/DNA followed a linear regression line (as in Appendix Figure 1), then

$$\frac{\text{B-A}}{\text{C-A}} = \frac{\text{3n-2n}}{\text{4n-2n}}. \tag{1}$$

A non-linear relationship would violate this rule. To test rule (1), we fit an ANOVA model

$$\text{RNA} \sim \text{Aneuploid type (a factor with three levels: 2n, 3n and 4n)}, \tag{2}$$

and observed that 1 is equivalent to $\frac{\text{coeff\_3n}}{\text{coeff\_4n}} = \frac{\text{3n-2n}}{\text{4n-2n}}$. That is, the coefficients of model 2 satisfied $(4n-2n)\text{coeff\_3n} - (3n-2n)\text{coeff\_4n} = 0$. We deployed the test with the help of the R package `car` (Fox and Weisberg, 2011), taking care with the order of the DNA values. The normal-theory p-value is:

```
linearHypothesis(lm(RNA ~ Aneuploid type, c(0, 4n-2n, -(3n-2n), 0)\$'Pr($>$F)'[2]))
```

Genes were flagged as possibly having a non-linear relationship if this p-value was less than 0.05.

2.3 Clustering by Mixture of Linear Regressions

2.3.1 Overview

For a number of yeast genes $\{g\}$ we had measurements on the relative abundance of genomic DNA $x_g = (x_{g,i})$ and matched RNA $y_g = (y_{g,i})$ in cultured cells. Here i indexes the specific measurement, running from $i = 1$ to the number of samples n_g in the strain panel. Having removed genes exhibiting a non-linear DNA/RNA relationship, our goal was to cluster genes according to the pattern of their linear relationship. Phenomena related to dosage compensation may be reflected in this relationship, and by clustering genes we hoped to gain further insight into these phenomena.

We sought to cluster according to several underlying patterns: (1) intercept 0, slope 1, (2) intercept negative, slope 1, (3) slope less than one. These patterns, when considered on the relative measurement scale, refer to three basic regulatory regimes of interest. We developed a method based on treating the data as a mixture of normal components, and within a specific probability model we computed the posterior probability that each gene arose from any of these components, following the prescripts of model-based mixture clustering (McLachlan and Peel, 2000). An approximate Expectation-Maximization (EM) algorithm was constructed to compute these probabilities and thus determine the clustering of genes according to their probable regime of linear DNA/RNA relationship.

2.3.2 Normal mixture model

Firstly, we conditioned on DNA measurements $\{x_g\}$, as is common in regression analysis. We treated the RNA vectors $\{y_g\}$ as the realization of Gaussian random vectors $\{Y_g\}$, whose mean and covariance depended on latent classes in operation for the different genes. To develop this further, we supposed

$$Y_{g,i} = \alpha_g + \beta_g x_{g,i} + \epsilon_{g,i} \qquad \text{for } i = 1, 2, \ldots, n_g \tag{3}$$

where α_g and β_g were the gene-specific intercept and slope, respectively, and where $\epsilon_{g,i}$ was a mean zero Gaussian random error that had variance σ_g^2. Five distinct hypotheses (classes, clusters) were considered:

1. $\alpha_g = 0,\ \beta_g = 1$

2a. $\alpha_g < 0,\ \beta_g = 1$

2b. $\alpha_g > 0,\ \beta_g = 1$

3a. $\beta_g < 1$

3b. $\beta_g > 1$

Cases 2b and 3b were entertained for completeness, though they were expected to be less intepretable than the primary classes of interest: 1, 2a, and 3a. In classes 3a and 3b, the intercept was not constrained. The model was strengthened by not requiring that we fix or estimate parameters $\{\alpha_g, \beta_g\}$. Specifically, we treated these as random effects, governed by suitably constrained Gaussian distributions that respect the class structure. The random-effects distribution was as follows:

1. $\alpha_g = 0,\ \beta_g = 1$

2a. $\alpha_g \sim \text{Normal}\big(0, \sigma_\alpha^2\big)$ restricted to $(-\infty, 0)$, $\quad \beta_g = 1$

2b. $\alpha_g \sim \text{Normal}\big(0, \sigma_\alpha^2\big)$ restricted to $(0, \infty)$, $\quad \beta_g = 1$

3a. $\alpha_g \sim \text{Normal}\big(0, \sigma_\alpha^2\big)$, $\quad \beta_g \sim \text{Normal}\Big(1, \sigma_\beta^2\Big)$ restricted to $(-\infty, 1)$

3b. $\alpha_g \sim \text{Normal}\big(0, \sigma_\alpha^2\big)$, $\quad \beta_g \sim \text{Normal}\Big(1, \sigma_\beta^2\Big)$ restricted to $(1, \infty)$

These class-specific distributions were governed by two variance components σ_α^2 and σ_β^2, which we estimated from the data. They express variation within each class and across genes in the specific values of intercepts and slopes. Formally, we assumed mutual independence among all $\{\alpha_g\}$, $\{\beta_g\}$ and errors $\{\epsilon_{g,i}\}$. The mixture model was fully specified by introducing discrete latent class random variables $\{Z_g\}$ and parameters $\lambda = (\lambda_j)$, to be estimated from the data, defined

$$P(Z_g = j) = \lambda_j \qquad j \in \{1, 2a, 2b, 3a, 3b\}. \tag{4}$$

After parameter estimation, model-based clustering was based on the posterior probabilities

$$P(Z_g = j | x_g, y_g) = \text{constant} \times \lambda_j \times p(y_g | Z_g = j, x_g), \tag{5}$$

with the constant computed by normalizing the probabilities to sum to one. The key ingredients in (5) were the class-specific probability densities $p(y_g|Z_g = j, x_g)$, which measured how well the data were explained by each class j. Note that $p(y_g|Z_g = j, x_g)$ is a multivariate joint density, since y_g is a vector of RNA measurements, and it is also a marginal density, because the latent random effects α_g and β_g were integrated away. An interesting element of the proposed approach was that the densities $p(y_g|Z_g = j, x_g)$ were available analytically, in much the same way as marginal compound Gamma distributions were computed in Newton *et al.* (2004).

Class 1: There was no randomness in the intercept slope in this class, so the RNA vector Y_g satisfied:

$$Y_g \sim \text{Normal}_{n_g}\left(x_g, \sigma_g^2 I\right) \tag{6}$$

where I is the $n_g \times n_g$ identity matrix. Then $p(y_g|Z_g = 1, x_g)$ is the ordinate of the multivariate normal density in (6): up to a constant across classes,

$$\log p(y_g|Z_g = 1, x_g) = -\frac{n_g}{2}\log(\sigma_g^2) - \left(\frac{1}{2\sigma_g^2}\right)(y_g - x_g)^t(y_g - x_g). \tag{7}$$

Classes 2a, 2b: In these classes, it was convenient to introduce a related vector $\tilde{Y}_g$ based on having no constraints on α_g; that is taking $\alpha_g \sim \text{Normal}(0, \sigma_\alpha^2)$. Without the ordering constraint, it was immediate that

$$\tilde{Y}_g \sim \text{Normal}_{n_g}\left(x_g, \sigma_g^2 I + \sigma_\alpha^2 ee^t\right) \tag{8}$$

where e is an $n_g \times 1$ vector of $1's$. If $\tilde{p}_2$ denotes the density of the multivariate normal in (8), then we found

$$\begin{aligned} p(y_g|Z_g = 2a, x_g) &= \tilde{p}_2(y_g|x_g)\,\tilde{P}_2(\alpha_g < 0|y_g, x_g) \\ p(y_g|Z_g = 2b, x_g) &= \tilde{p}_2(y_g|x_g)\,\tilde{P}_2(\alpha_g > 0|y_g, x_g). \end{aligned} \tag{9}$$

Here the probabilities involving $\tilde{P}_2$ refer to conditional probabilities about the random intercept in the *unconstrained* model for α_g. From standard conjugate Bayesian analysis, the unconstrained posterior for α_g was

$$\alpha_g|y_g, x_g \sim \text{Normal}\left(\frac{e^t(y_g - x_g)}{\sigma_g^2\left(\frac{1}{\sigma_\alpha^2} + \frac{n_g}{\sigma_g^2}\right)}, \frac{1}{\frac{1}{\sigma_\alpha^2} + \frac{n_g}{\sigma_g^2}}\right), \tag{10}$$

and so the ordering factors on the right side of (9) were readily computed from univariate Gaussian CDF's. As to the unconstrained ordinate $\tilde{p}_2$, following (7),

$$\log \tilde{p}_2(y_g|x_g) = -\frac{1}{2}\log\det\Sigma_2 - \frac{1}{2}(y_g - x_g)^t\Sigma_2^{-1}(y_g - x_g) \tag{11}$$

where $\Sigma_2 = \sigma_g^2 I + \sigma_\alpha^2 ee^t$ as in (8). Using Woodbury's formula, this log density was readily computed in terms of summary statistics from gene g.

Classes 3a, 3b: Here the constraints were on the slopes β_g, but randomness in both slopes and intercepts must be accounted for. Again we computed ordinates by considering an model in which there were no constraints on β_g. (Call it $\tilde{p}_3$.) In comparison with (8),

$$\tilde{Y}_g \sim \text{Normal}_{n_g}\left(x_g, \Sigma_3 = \sigma_g^2 I + \sigma_\alpha^2 ee^t + \sigma_\beta^2 x_g x_g^t\right). \tag{12}$$

To compute $p(y_g|Z_g = 3a, x_g)$ (for example), we required the normal mass $\tilde{P}_3(\beta < 1|x_g, y_g)$ from the unconstrained model as well as the unconstrained ordinate $\tilde{p}_3(y_g|x_g)$, which was

$$\log \tilde{p}_3(y_g|x_g) = -\frac{1}{2}\log\det\Sigma_3 - \frac{1}{2}(y_g - x_g)^t\Sigma_3^{-1}(y_g - x_g). \tag{13}$$

Convenient formulas for these quantities were available in terms of gene-level statistics.

Appendix Figures

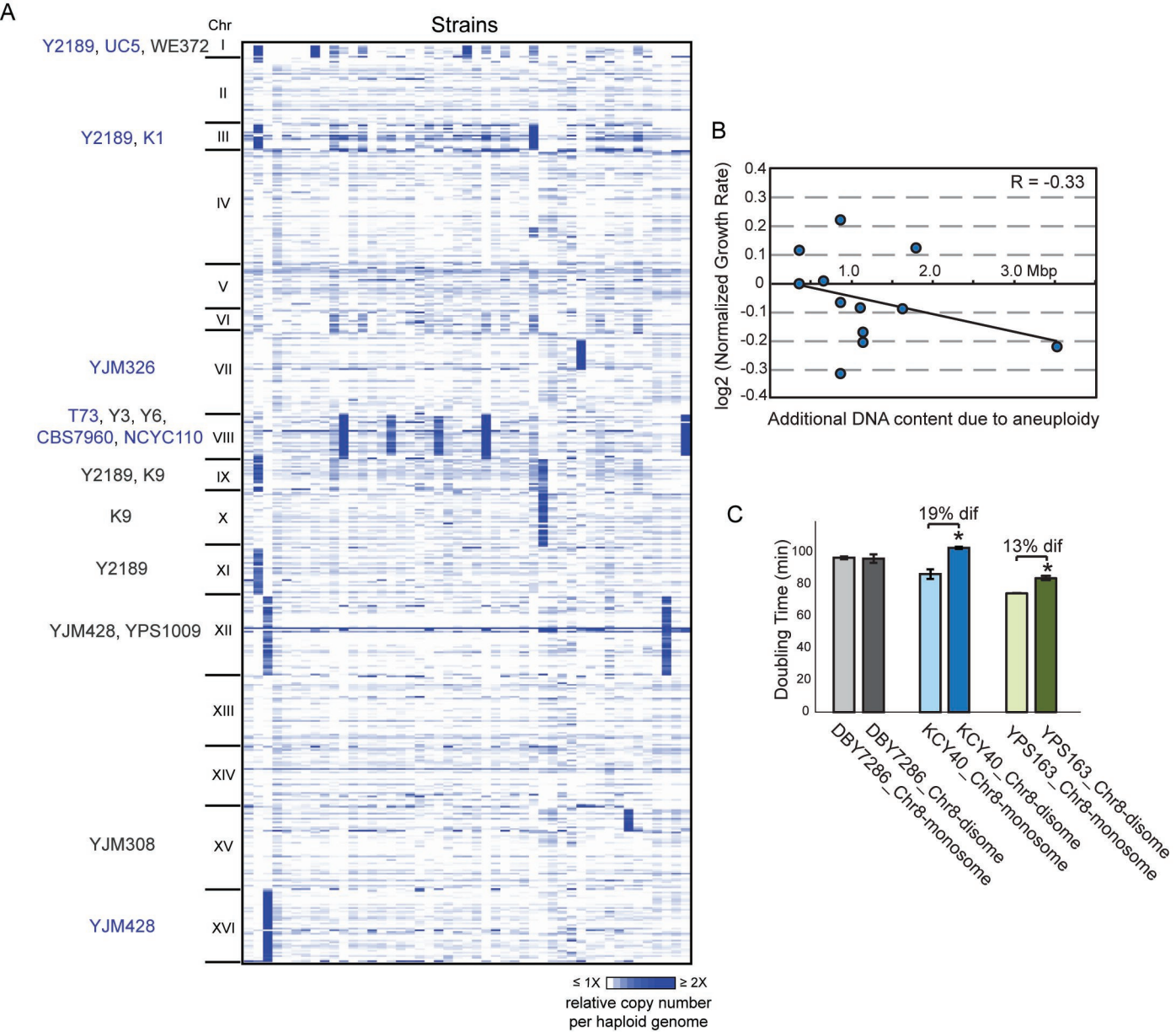
Appendix Figure 1. Relative DNA and RNA values follow a linear regression.

References

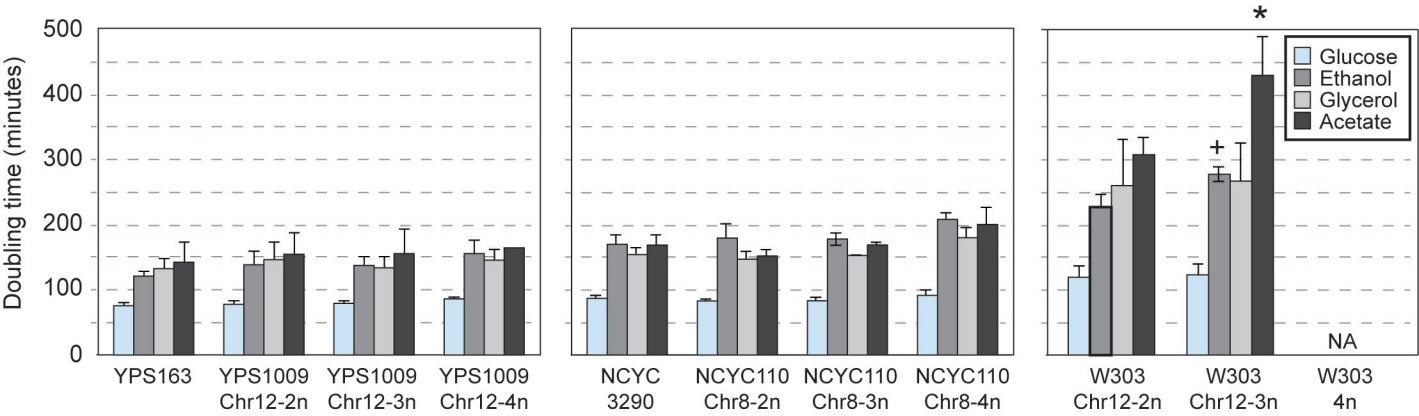
1. Fox, J and Weisberg, S (2011). An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. `URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`

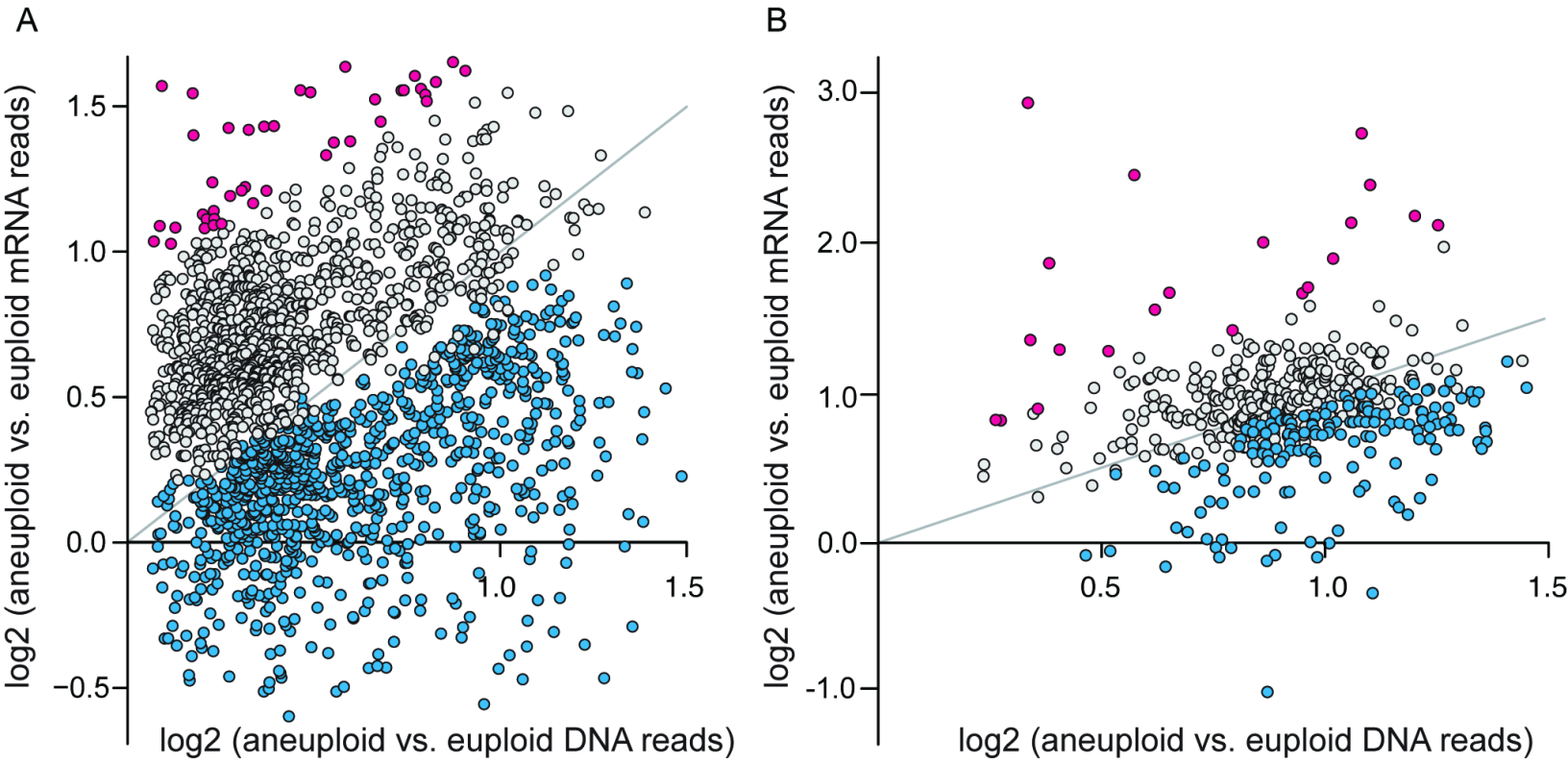
2. McLachlan, GJ and Peel, D (2000). *Finite mixture models*. Wiley, New York.

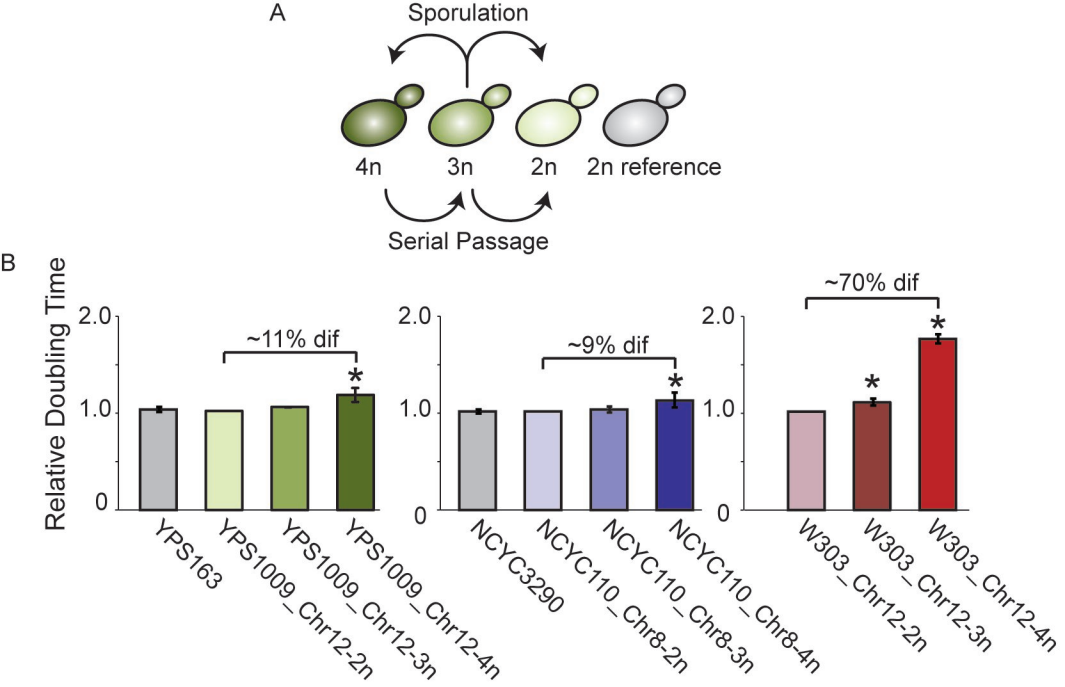
3. Newton, MA, Noueiry, A, Sarkar, D, and Ahlquist P (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, *5*, 155-176.

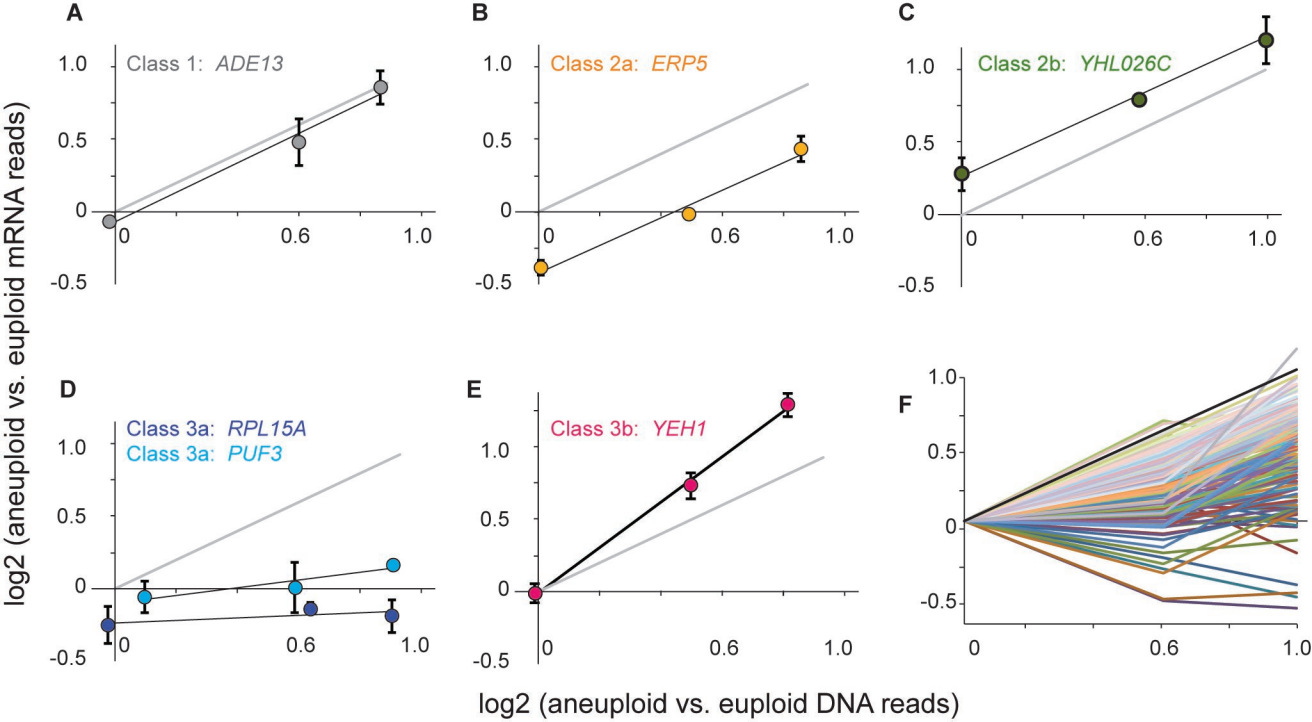

A
Chr
Strains
Y2189, UC5, WE372
I
II
Y2189, K1
III
IV
V
VI
YJM326
VII
T73, Y3, Y6,
CBS7960, NCYC110
VIII
Y2189, K9
IX
K9
X
Y2189
XI
YJM428, YPS1009
XII
XIII
XIV
YJM308
XV
YJM428
XVI
≤ 1X
≥ 2X
relative copy number
per haploid genome
B
R = -0.33
1.0
2.0
3.0 Mbp
log2 (Normalized Growth Rate)
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
Additional DNA content due to aneuploidy
C
19% dif
13% dif
*
*
Doubling Time (min)
100
80
60
40
20
0
DBY7286_Chr8-monosome
DBY7286_Chr8-disome
KCY40_Chr8-monosome
KCY40_Chr8-disome
YPS163_Chr8-monosome
YPS163_Chr8-disome

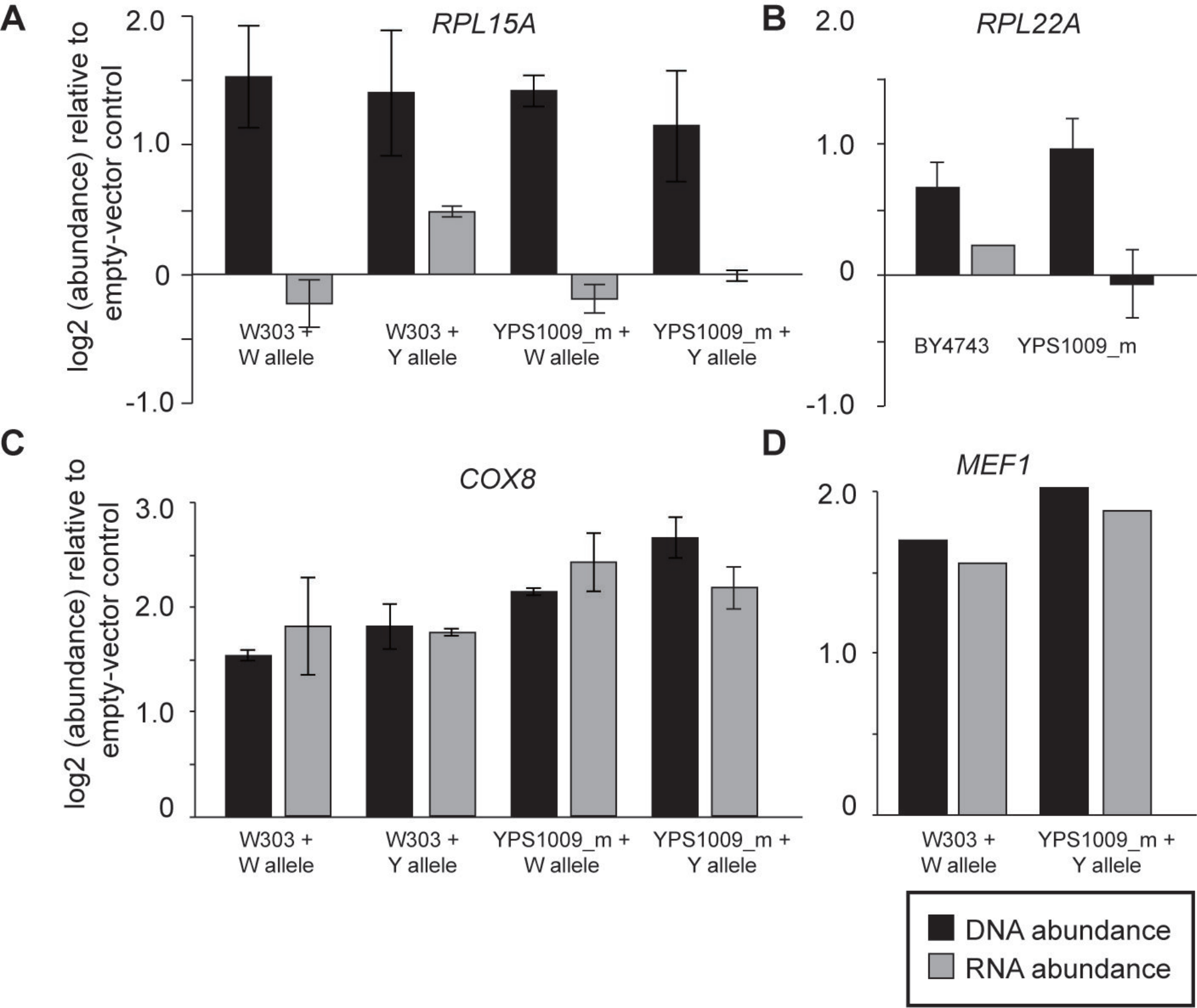
A
B
K1 vs K10
K9 vs K10
YJM428 vs YJM308
Y2189 vs Y2209
YPS1009 vs YPS163
NCYC110 vs NCYC3290
255 ESR genes normally induced by stress
573 ESR genes normally repressed by stress
69 genes induced in ≥ 3 strains
263 genes repressed in ≥ 3 strains
4X lower in aneuploid vs euploid strain
4X higher in aneuploid vs euploid strain

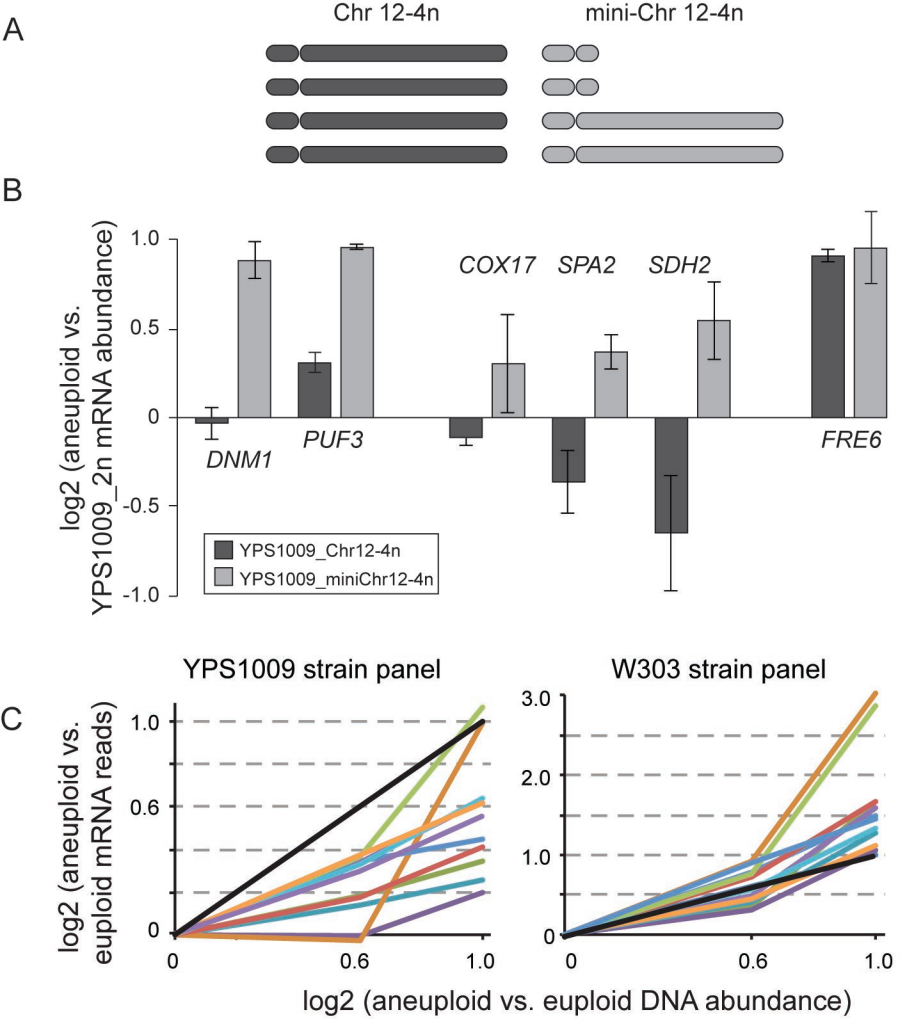

Doubling time (minutes)
500
400
300
200
100
0
Glucose
Ethanol
Glycerol
Acetate
*
+
NA
YPS163
YPS1009 Chr12-2n
YPS1009 Chr12-3n
YPS1009 Chr12-4n
NCYC 3290
NCYC110 Chr8-2n
NCYC110 Chr8-3n
NCYC110 Chr8-4n
W303 Chr12-2n
W303 Chr12-3n
W303 4n

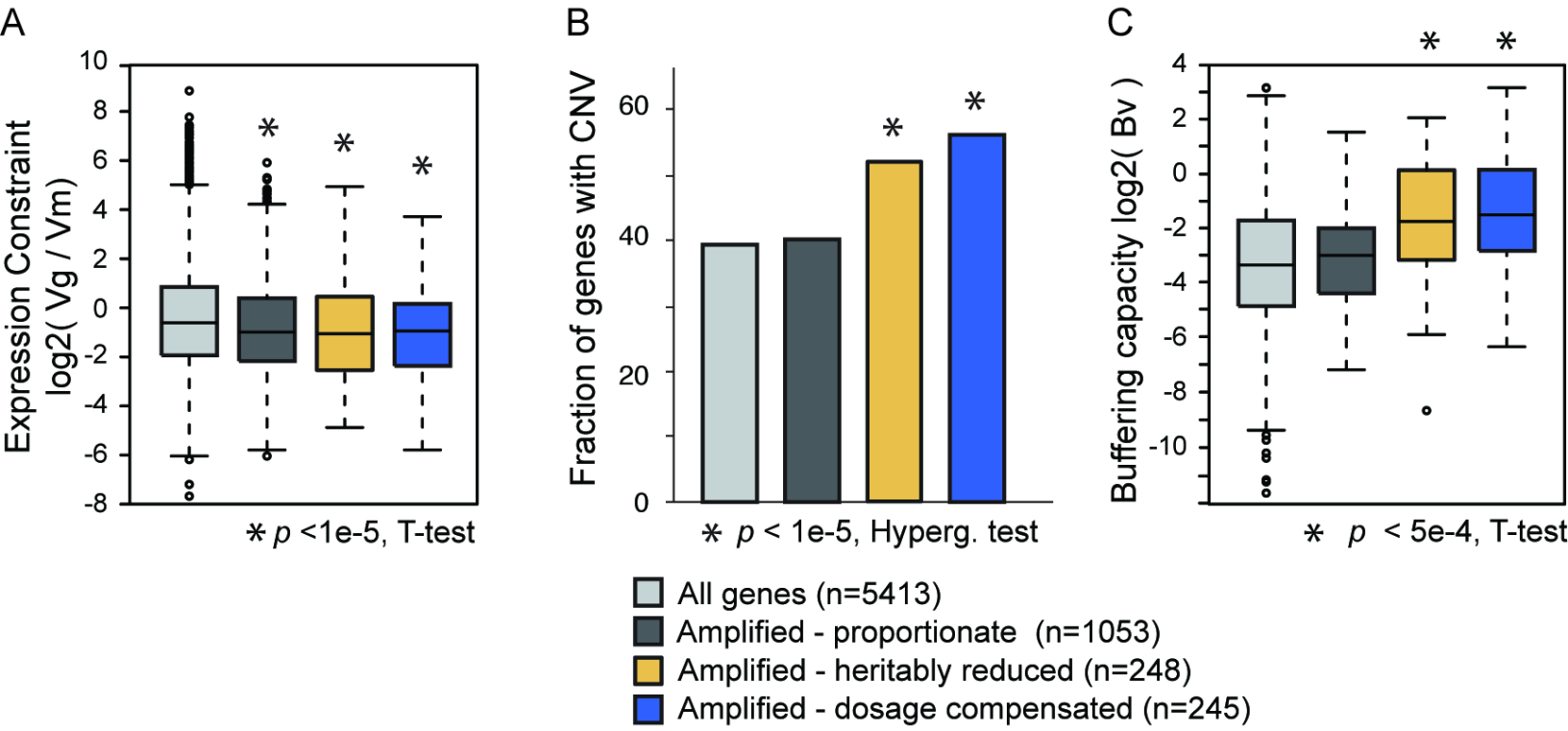

A
log2 (aneuploid vs. euploid mRNA reads)
log2 (aneuploid vs. euploid DNA reads)
1.5
1.0
0.5
0.0
−0.5
1.0
1.5
B
log2 (aneuploid vs. euploid mRNA reads)
log2 (aneuploid vs. euploid DNA reads)
3.0
2.0
1.0
0.0
-1.0
0.5
1.0
1.5

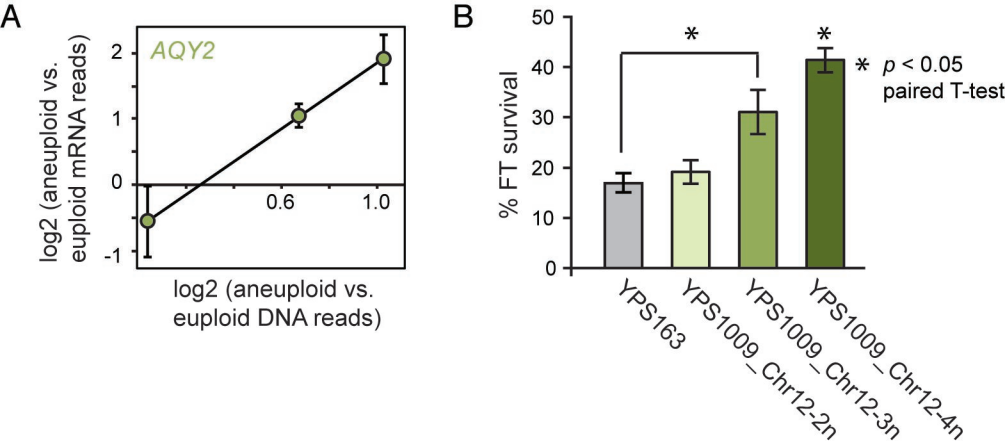

A
Sporulation
4n
3n
2n
2n reference
Serial Passage
B
Relative Doubling Time
2.0
1.0
0
~11% dif
~9% dif
~70% dif
*
YPS163
YPS1009
YPS1009_Chr12-2n
YPS1009_Chr12-3n
YPS1009_Chr12-4n
NCYC3290
NCYC110
NCYC110_Chr8-2n
NCYC110_Chr8-3n
NCYC110_Chr8-4n
W303
W303_Chr12-2n
W303_Chr12-3n
W303_Chr12-4n


A
Class 1: ADE13
B
Class 2a: ERP5
C
Class 2b: YHL026C
D
Class 3a: RPL15A
Class 3a: PUF3
E
Class 3b: YEH1
F
log2 (aneuploid vs. euploid mRNA reads)
log2 (aneuploid vs. euploid DNA reads)


A
RPL15A
log2 (abundance) relative to empty-vector control
2.0
1.0
0
-1.0
W303 + W allele
W303 + Y allele
YPS1009_m + W allele
YPS1009_m + Y allele
B
RPL22A
2.0
1.0
0
-1.0
BY4743
YPS1009_m
C
COX8
log2 (abundance) relative to empty-vector control
3.0
2.0
1.0
0
W303 + W allele
W303 + Y allele
YPS1009_m + W allele
YPS1009_m + Y allele
D
MEF1
2.0
1.0
0
W303 + W allele
YPS1009_m + Y allele
DNA abundance
RNA abundance


A
Chr 12-4n
mini-Chr 12-4n
B
log2 (aneuploid vs. YPS1009_2n mRNA abundance)
1.0
0.5
0
-0.5
-1.0
DNM1
PUF3
COX17
SPA2
SDH2
FRE6
YPS1009_Chr12-4n
YPS1009_miniChr12-4n
C
YPS1009 strain panel
W303 strain panel
log2 (aneuploid vs. euploid mRNA reads)
1.0
0.6
0
3.0
2.0
1.0
0
0.6
1.0
log2 (aneuploid vs. euploid DNA abundance)


A
Expression Constraint
log2(Vg / Vm)
10
8
6
4
2
0
-2
-4
-6
-8
* p <1e-5, T-test
B
Fraction of genes with CNV
60
40
20
0
* p < 1e-5, Hyperg. test
C
Buffering capacity log2(Bv)
4
2
0
-2
-4
-6
-8
-10
* p < 5e-4, T-test
All genes (n=5413)
Amplified - proportionate (n=1053)
Amplified - heritably reduced (n=248)
Amplified - dosage compensated (n=245)


A
AQY2
log2 (aneuploid vs. euploid mRNA reads)
log2 (aneuploid vs. euploid DNA reads)
2
1
0
-1
0.6
1.0
B
% FT survival
50
40
30
20
10
0
*
*
*
p < 0.05
paired T-test
YPS163
YPS1009_Chr12-2n
YPS1009_Chr12-3n
YPS1009_Chr12-4n

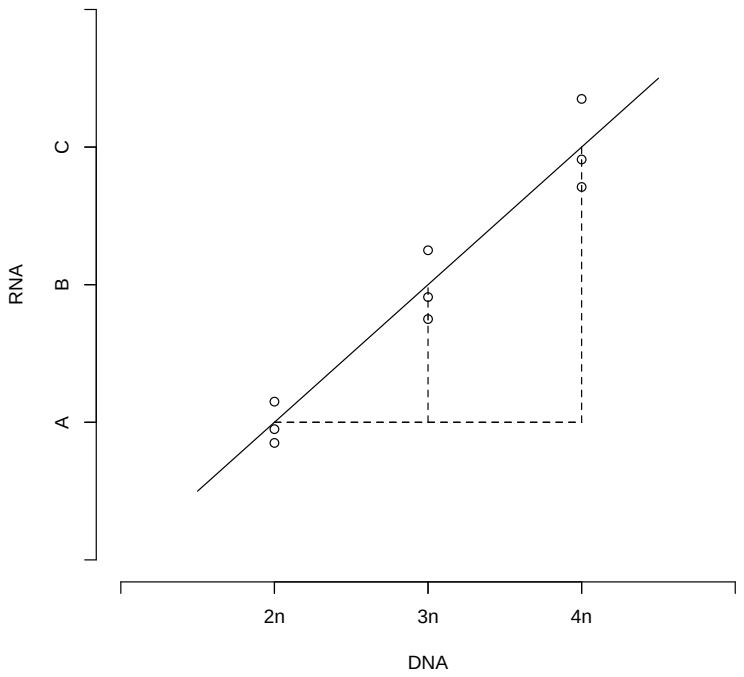


Figure S7: Relative DNA and RNA values follow a linear regression.