Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast

Daniel A. Skelly,[1] Gennifer E. Merrihew,[1] Michael Riffle,[2] Caitlin F. Connelly,[1] Emily O. Kerr,[1] Marnie Johansson,[1] Daniel Jaschob,[2] Beth Graczyk,[2] Nicholas J. Shulman,[2] Jon Wakefield,[3,4] Sara J. Cooper,[1,7] Stanley Fields,[1,5] William S. Noble,[1,6] Eric G.D. Muller,[2] Trisha N. Davis,[2] Maitreya J. Dunham,[1,8] Michael J. MacCoss,[1,8] and Joshua M. Akey[1,8]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; [2]Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA; [3]Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA; [4]Department of Statistics, University of Washington, Seattle, Washington 98195, USA; [5]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA; [6]Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, USA

To better understand the quantitative characteristics and structure of phenotypic diversity, we measured over 14,000 transcript, protein, metabolite, and morphological traits in 22 genetically diverse strains of *Saccharomyces cerevisiae*. More than 50% of all measured traits varied significantly across strains [false discovery rate (FDR) = 5%]. The structure of phenotypic correlations is complex, with 85% of all traits significantly correlated with at least one other phenotype (median = 6, maximum = 328). We show how high-dimensional molecular phenomics data sets can be leveraged to accurately predict phenotypic variation between strains, often with greater precision than afforded by DNA sequence information alone. These results provide new insights into the spectrum and structure of phenotypic diversity and the characteristics influencing the ability to accurately predict phenotypes.

[Supplemental material is available for this article.]

Considerable progress has been made in characterizing genomes, allowing comprehensive insights into patterns of genetic diversity in many organisms (Liti et al. 2009; The 1000 Genomes Project Consortium 2010; Gan et al. 2011; Keane et al. 2011; Tennessen et al. 2012). However, interpreting the functional and phenotypic consequences of genetic variation remains challenging and is exacerbated by the paucity of data on the quantitative characteristics of phenotypes. One approach to bridge the gap between genetic variation and organismal phenotypes is the comprehensive and systematic collection of carefully measured phenotypes, an approach referred to as phenomics (Schork 1997; Freimer and Sabatti 2003; Houle et al. 2010). To date, phenomics studies have often studied a moderate number of phenotypes or have been limited to only a single individual (Warringer et al. 2003, 2011; Kvitek et al. 2008; Ratnakumar et al. 2011; Chen et al. 2012). However, advances in functional genomics technology, instrumentation, and computational biology are providing the necessary tools to extensively phenotype increasingly large collections of individuals.

Here, we describe a comprehensive phenomics data set consisting of over 14,000 molecular and morphological traits collected in 22 genetically diverse yeast strains. More specifically, we measured gene expression, protein and metabolite abundance, and quantitative morphological phenotypes in isolates of *S. cerevisiae* sampled from six continents and a wide variety of microenvironments (Supplemental Table 1). Although previous studies have reported measurements of one or two of these data types in samples derived from a smaller number of strains (e.g., Nogami et al. 2007; Foss et al. 2011; Xu et al. 2011), our study is the most comprehensive data set of multiple molecular and morphological phenotypes measured simultaneously in a large number of strains. These data reveal new insights into the patterns, structure, and determinants of phenotypic variation and provide a powerful resource to enable a deeper understanding of the principles governing the relationship between genotypes and phenotypes.

Results

High-dimensional phenotyping and genome sequencing

To comprehensively measure molecular and morphological phenotypes while mitigating confounding variables, we used a randomized study design and obtained biological replicates for each measured trait (Fig. 1). Previous studies have shown a large, generic gene expression, protein, and metabolite response correlated with even small changes in growth rate (Regenberg et al. 2006; Castrillo et al. 2007; Brauer et al. 2008), which we were concerned might overwhelm other aspects of phenotypic variation. Therefore, to control for growth rate variation among strains and maintain a constant external cellular milieu, we grew strains to steady state in chemostats under phosphate-limited conditions. For each sample, we performed RNA-seq to characterize gene expression and transcript structure (N = 6702 transcripts); chromatography and mass spectrometry to measure protein (N = 6842 peptides) and metabolite (N = 115 metabolites) abundances; and quantitative microscopy to measure morphological phenotypes (N = 398 traits)

[7]Present address: HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, Alabama 35806, USA.
[8]Corresponding authors
E-mail akeyj@uw.edu
E-mail maitreya@uw.edu
E-mail maccoss@uw.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/ 10.1101/gr.155762.113. Freely available online through the *Genome Research* Open Access option.

23:1496–1504 © 2013, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/13; www.genome.org

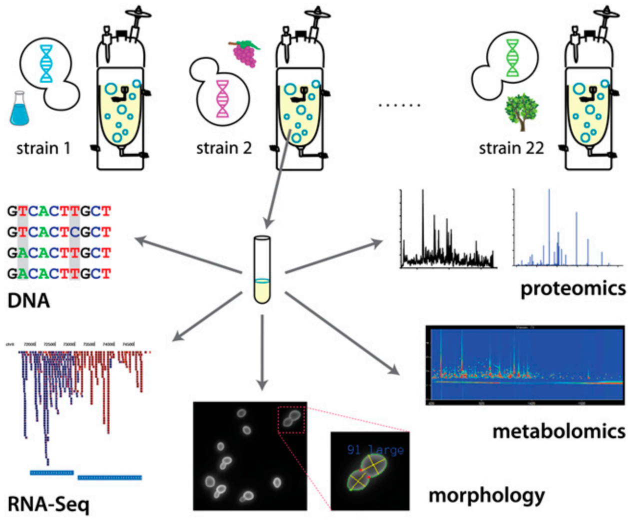

Figure 1. Experimental design facilitates high-dimensional phenotyping. A schematic of the experimental design used to obtain phenotypic data for the 22 strains. Icons next to strains show examples of sources from which strains were isolated (Supplemental Table 1). The schematic outlines the process of obtaining phenotype data for a single strain. In total (aggregated across all strains), we obtained 16 Gb of DNA sequence; 820 million RNA-seq reads; 912,000 mass spectra that we used to infer peptide levels; metabolic measurements of molecules in 40 different biochemical pathways; and 2000 images that together captured 21,000 cells, whose morphological characteristics we tabulated.

(Fig. 1). Across all samples combined, we obtained ~13.2 Gb of uniquely mappable genome sequence (equivalent to more than 1000× coverage of the genome) and 38.6 Gb of uniquely mappable transcript sequence. We conducted extensive quality control and normalized for technical effects (Supplemental Figs. 1, 2; Supplemental Table 2) and found that the data were highly reproducible, with median correlations between biological replicates exceeding 0.97 for RNA-seq and morphological trait measurements, 0.87 for quantitative proteomics, and 0.81 for metabolomics data.

In addition, we resequenced the genome of each strain to high coverage (mean approximately 30×). These data supplemented existing low-coverage Sanger sequence data (Liti et al. 2009) and allowed us to call additional single nucleotide polymorphisms (SNPs), map RNA-seq reads in an unbiased fashion, and identify peptides expected to differ in amino acid sequence between strains. Concordance with previously published Sanger-based sequences (Liti et al. 2009) was >99.5% for all strains. Previous low-coverage sequencing (Liti et al. 2009) reported approximately 230,000 SNPs (20,000–80,000 SNPs per strain) in these strains relative to the *S. cerevisiae* reference sequence (S288c); our high-coverage short-read sequencing yielded an additional 50,000 SNPs (1500–44,000 per strain). An analysis of sequences imputed by Liti et al. (2009) using a phylogenetically motivated approach revealed significant discrepancies with our sequence data (Supplemental Note), highlighting the difficulty of accurately imputing sequence in a model organism with complex and heterogeneous patterns of population structure.

Pervasive phenotypic diversity

We found widespread heritable variation within every class of phenotypic data measured, with >50% of all measured traits varying between strains. Specifically, 74% (4565) of transcript levels, 23% (1553) of peptides, 10% (12) of metabolites, and 64% (255) of morphological traits significantly varied [false discovery rate (FDR) = 5%] across strains. Following Nogami et al. (2007), the morphological traits that we tabulated included both directly measured traits and their coefficient of variation (CV). We found more directly measured traits differed between strains (151/199) than CV traits (104/199).

Among transcript and protein levels, genes that varied most across strains were involved in aerobic respiration and the electron transport chain, with highly significant gene ontology enrichment ($P < 10^{-5}$) for cellular respiration, ATP synthesis coupled proton transport, and mitochondrial respiratory chain complexes (Supplemental Table 4). Indeed, we found consistent differences between strains in the overall activity of central carbon metabolic pathways, reflecting contrasting strategies for energy generation (Fig. 2A). The strong anticorrelation between the activity of genes involved in fermentation versus aerobic respiration was largely, though not entirely, associated with the major phylogenetic division between the strains (Fig. 2B); strains involved in the production of alcoholic beverages as well as their close relatives tended to be more active fermenters.

We examined each differentially abundant metabolite in the context of 162 well-annotated biochemical pathways. Overall, metabolites were significantly correlated (FDR = 5%) with a large number of pathways (mean = 58, standard deviation = 30), consistent with the highly interconnected nature of metabolism. The metabolite ribose (due to the derivatization process, this measurement included both free ribose and ribose-5-phosphate) was significantly correlated (FDR = 5%; $|\rho| > 0.43$) with the largest number of pathways, 96. Ribose-5-phosphate is produced by the pentose phosphate pathway and required for nucleotide biosynthesis, and the activity of these pathways and the corresponding ribose/ribose-5-phosphate levels varied across strains (Fig. 2C). At the morphological level, we observed consistent, heritable differences in traits related to cell size (Fig. 2D; Supplemental Fig. 3).

Identifying large-effect *cis*-regulatory transcript and protein quantitative trait locus (QTL)

To search for genomic variants underlying variation in functional genomics phenotypes, we performed association tests between variants within 500 bp of each gene and its corresponding transcript and peptide levels. Although the number of individuals sampled is small ($N = 22$), simulations indicate that we have moderate to high power to detect large-effect variants (Supplemental Table 5). We focused on common variants near the gene of interest, which presumably act primarily in *cis* to influence transcript and protein levels because complex patterns of population structure in *S. cerevisiae* render genome-wide association studies susceptible to a high type I error rate (Connelly and Akey 2012). Before performing association tests, we controlled for population structure using mixed models and selected tag SNPs with $r^2 > 0.6$ (Supplemental Note). We found 64 significant peptide-SNP associations (from 42 distinct proteins) and 302 significant transcript-SNP associations (FDR = 5%) (Fig. 3A). Genetic variants underlying associations have large effects, explaining on average over half (53%) of the variation in peptide or transcript level. Thus, large-effect transcript and protein QTL are relatively common in natural populations of yeast. Variants associated with transcript or protein levels were found in promoters, 3′ untranslated regions (UTRs), and genes, without a significant enrichment of any location type.

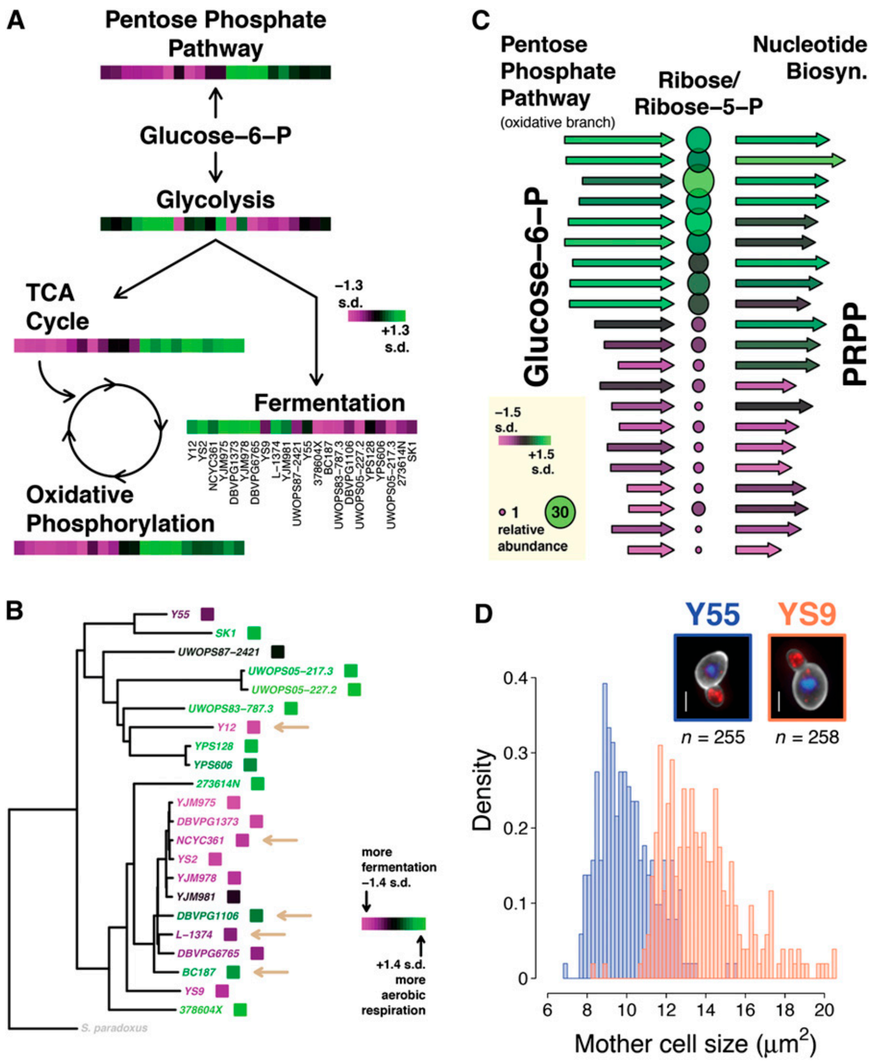

Figure 2. Pervasive heritable phenotypic variation. (*A*) Overview of central carbon metabolism, with heatmaps indicating pathway activity by strain. Transcript and protein data for genes in each pathway was combined, the first principal component extracted, and the numerical sign adjusted to ensure higher numbers corresponded to higher average transcript and protein abundance across the pathway. Order of strains is listed under the fermentation heatmap. (*B*) Phylogeny based on complete genome sequences, with strain names colored according to the key shown. Tan arrows indicate strains used in the fermentation of alcoholic beverages. Pathway activity for strains was calculated as in *A*, using genes in the tricarboxylic acid cycle (TCA cycle) and involved in fermentation. (*C*) Pathways leading to the production of phosphoribosyl pyrophosphate (PRPP) from glucose-6-phosphate, with pathway activity displayed vertically for 21 strains (Supplemental Note). Arrows represent pathway activity calculated as in *A*, with longer arrows/green indicating higher activity and shorter arrows/magenta indicating lower activity. Arrows on *left* indicate pathway activity of the oxidative branch of the pentose phosphate pathway, and on *right* activity of 5-phospho-ribosyl-1(alpha)-pyrophosphate synthetase, the heteromultimeric complex that synthesizes PRPP. Circles represent measurements of ribose/ribose-5-phosphate and are colored and sized accordingly. (*D*) Differences in mother cell size between a small and large strain. Histograms are composed of measurements made on individual cells. *Inset* photos show a typical cell from each strain (with size near the strain mean). White scale bars show ~2 μm. Actin stain is shown in red, DNA stain in blue, and cell wall stain in greyscale in the merged images.

Of the 69 significant transcript associations that also had peptide data, six had at least one significant peptide association. Of these six, five were associated with the same SNP as the transcript association, consistent with variants that affect transcript level and thus indirectly affect protein level. Figure 3B shows one such example in the gene *PPN1*, an endopolyphosphatase involved in phosphate metabolism. The absence of a significant peptide association in 63/69 transcript associations suggests that the heritable basis of regulatory variants influencing transcript and protein levels is largely distinct (Foss et al. 2011; Ghazalpour et al. 2011). Alternatively, the lack of a significant association at both the transcript and protein level may simply reflect a lack of statistical power. To investigate these two hypotheses, we estimated the fraction of truly significant peptide associations among the set of genes with significant transcript associations using a conservative method based on the distribution of *P*-values (Fig. 3C; Storey and Tibshirani 2003). We estimate that 53% of these peptides have a true association; thus, a substantial fraction of large-effect variants that influence transcript levels also affect peptide levels.

Densely connected network structure of phenotypic correlations

To explore the correlation structure among traits, we calculated pairwise correlation coefficients among 8365 phenotypes (collapsing all peptide measurements into a single mean number for each protein) and identified 68,558 correlations, involving a total of 7078 phenotypes, which were significant at an FDR of 5%. Approximately 60% (41,649) of the trait comparisons were positively correlated. The excess of positive correlations is partially attributable to the fact that transcript and protein levels of genes in the same pathway or protein complex tend to be positively correlated (mean ρ = 0.12 across n = 427 pathways and protein complexes), but those from different pathways are equally likely to be negatively as positively correlated (mean ρ = 0.008). Overall, there were strong correlations both within (79%) and between (21%) data types, with a particularly dense set of connections within and between highly correlated transcripts and proteins (Fig. 4A).

Of the 7078 phenotypes correlated to at least one other trait, the mean number of significant correlations to other traits was 19.4 (bootstrap 95% confidence interval 18.6–20.2). On average, transcript levels were correlated with the largest number of other traits (20.4) and metabolite levels the fewest (12.0). The single most highly correlated phenotype was the histidine tRNA synthetase *HTS1* transcript, which was correlated with 328 other phenotypes from all four data types but consisting largely of other transcript levels (N = 293). Among the 50 most highly correlated transcript and 50 most highly correlated protein levels, we observed strong enrichment for genes involved in

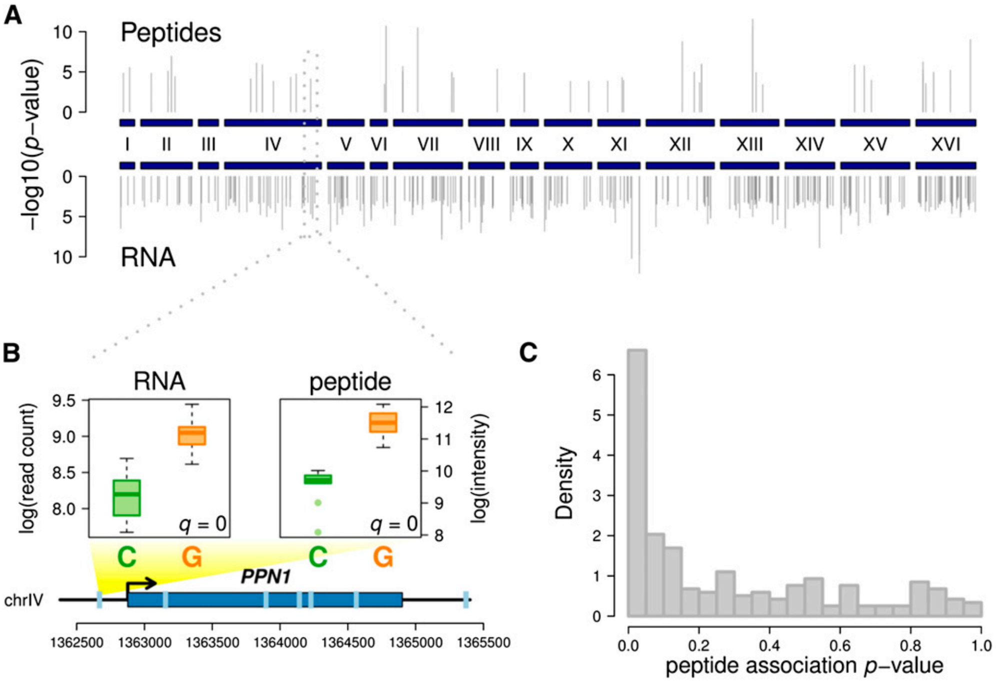

Figure 3. *Cis*-regulatory transcript and peptide QTL. (*A*) Manhattan plot showing results for significant (FDR = 5%) transcript and peptide *cis*-association tests. Gray vertical lines indicate individual tests. Blue boxes and associated roman numerals across the *middle* of the panel indicate the 16 chromosomes of yeast. (*B*) Transcript and peptide levels association with the same polymorphism. Light blue ticks along gene model indicate locations of tag SNPs tested for association. Transcript and peptide levels are significantly associated with the first SNP. Yellow gradient originating at first SNP expands to boxplots of transcript and peptide levels separated by allele; boxes indicate lower quartile, median, and upper quartile, and whiskers extend to half the interquartile range. (*C*) Histogram of *P*-values for 236 peptide associations in 69 genes with significant transcript associations.

energy generation, the mitochondrial respiratory chain, and ATP synthesis (Supplemental Table 6). Transcripts and proteins with *cis*-regulatory QTL were correlated with a significantly higher ($P <$ 0.001 for both tests) number of phenotypes (mean 33.1 and 35.9, respectively) than those without associations (mean 19.7 and 15.3, respectively).

Previous studies in a diverse complement of organisms have reported widely varying levels of RNA-protein correlation (Greenbaum et al. 2003; de Sousa Abreu et al. 2009; Schwanhäusser et al. 2011). These studies have largely measured RNA-protein correlation between different genes within an individual, whereas we sought to measure RNA-protein correlation between individuals (strains) on a gene-by-gene basis (Foss et al. 2011). Our measurements of transcript and protein levels used aliquots of cells taken from the same chemostat sample, minimizing environmental/batch effects that could lower correlations. We found a modest correlation (median 0.33; Spearman's ρ), with 44% (728 of 1636 genes with RNA and protein data) of genes having a significant positive correlation (FDR = 5%). However, restricting the analysis to genes with a significantly differentially expressed transcript and with at least one differentially abundant peptide increased the median Spearman correlation to 0.50 and resulted in ∼85% of genes having a significant RNA-protein correlation (FDR = 5%); (Fig. 4B). Genes with the highest RNA-protein correlations were strongly enriched for TATA box-containing genes (Fig. 4C). TATA-containing genes show greater variability in transcript and protein abundance between strains compared to TATA-less genes (*t*-test, $P < 1 \times 10^{-5}$). Thus, the larger variation among strains in these genes likely dominates measurement variation, resulting in stronger RNA-protein correlations. Alternatively, TATA-containing genes may be subject to less post-transcriptional regulation than TATA-less genes. Nevertheless, the relatively modest overall correlations between transcript and protein levels point to a substantial role for post-translational modifications and protein degradation in the control of steady-state protein abundances (Foss et al. 2011; Vogel and Marcotte 2012).

Integrative phenomics facilitates the prediction of phenotypes

The ability to accurately predict phenotypes would have profound consequences for basic and biomedical science (Zbuk and Eng 2007; Ng et al. 2009; Gonzaga-Jauregui et al. 2012), yet remains a challenging problem. We first predicted all RNA, protein, metabolite, and morphological traits in each single strain using simple models in which predicted phenotypes in the *i*th strain were a linear function of the phenotype in its closest relative (Supplemental Note) and the mean across other strains. These models accurately predict transcript, protein, and metabolite levels (median R^2_{adj} across strains = 0.97, 0.88, and 0.89, respectively) as well as morphological traits (median R^2_{adj} across strains > 0.99). However, although this analysis captures relative differences in abundance *between genes* within individuals, it does not robustly predict variation in abundance *between individuals* for a particular trait (Fig. 5A shows this in the context of gene expression levels).

To address this problem, we leveraged the complex correlation structure of our data set (Fig. 4A) to predict interstrain variation for all 5494 phenotypes that vary significantly between strains (with peptide measurements collapsed into a single mean number for each protein). To perform prediction, we used random forest regression, a statistical technique that allows for complex and

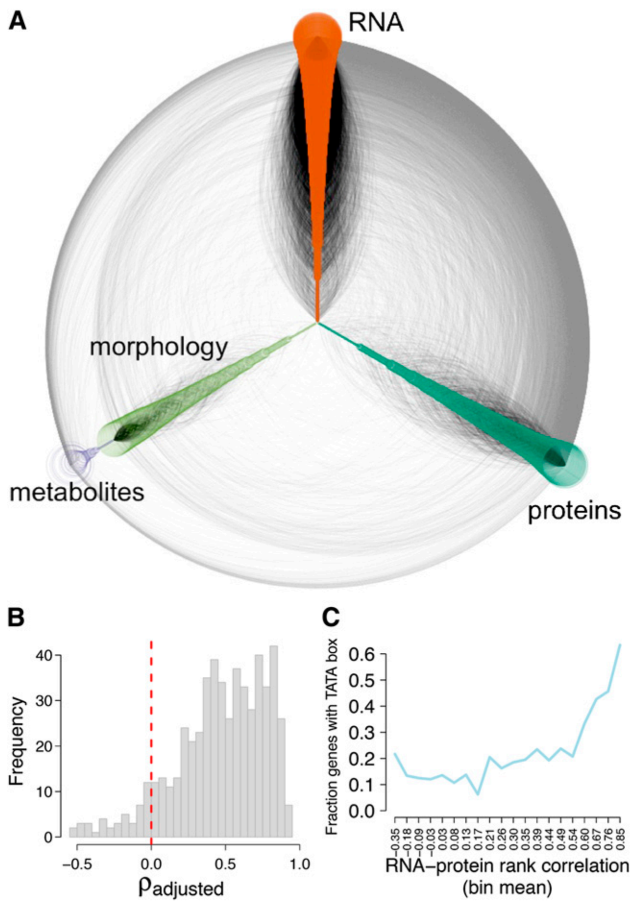

Figure 4. Dense network structure of phenotypic correlations. (*A*) Hive plot showing network composed of highly correlated phenotypic traits. Nodes arrayed along the three axes represent individual phenotypes, colored by data type as indicated. Lines drawn between nodes indicate a significant correlation between the two phenotypes. Black lines indicate connections within the same data type, and gray lines indicate connections between data types. (*B*) RNA-protein correlations for 542 genes with differentially expressed transcripts and at least one differentially abundant peptide. ρ_{adjusted} indicates a correlation calculated by subtracting from the true correlation the mean of correlations calculated from 1000 randomly reshufflings of the data. Vertical red dotted line is drawn at 0. (*C*) Fraction of genes containing a TATA box as a function of RNA-protein correlation. Each point plotted shows the fraction of genes with a TATA box among a bin of approximately 80 genes with similar RNA-protein correlations (bin means are labeled on the *x*-axis).

nonlinear interactions among predictor variables, and measured variation explained for each phenotype using out-of-bag data, which provide an unbiased estimate of error (Breiman 2001). Specifically, we sequentially withheld each strain, recalculated phenotypic correlations, and used only highly correlated (FDR = 5%) phenotypes as predictor variables in separate random forest regression models for each phenotype (for example, orange lines in Fig. 5B show phenotypes highly correlated with the abundance of the Tim11 protein). Across all phenotypes, our predictions can account for a median of 30% of variation (Fig. 5C, black line); and for 28% of phenotypes (1545), our predictions explain at least 50% of variation. For example, we can account for 86% of variation in abundance of the Tim11 protein, which is a subunit of the mitochondrial F1F0 ATPase required for ATP synthesis (Fig. 5C inset, black points). Other well-predicted phenotypes were highly enriched for mitochondrial and ribosomal functions (Supplemental Table 7). However, for 1984 (36%) of phenotypes, our predictions failed to explain >10% of variation, indicating that information beyond values of correlated traits is necessary for robust predictions.

To explore how informative additional predictors could be, we incorporated functional annotation data available for a subset of the phenotypes we measured into our model. Specifically, we considered variation in 1303 transcript and 660 protein levels that differed significantly between strains, using an approach similar to the above with the addition of approximately 1000 heterogeneous predictor variables. Additional predictors included transcript and protein levels of other genes with similar functions, genic characteristics, sequence features, and pathway annotations (Supplemental Fig. 5). We ran the model on data from all transcripts or all proteins simultaneously and found that our predictions explained ~45% of the variation in both transcript (median 46.8%) and protein (median 44.8%) levels, significantly better than the ~36% (median 37.0% and 35.6%, respectively) of variation explained using correlated traits alone (Fig. 5D). In some cases, the difference in prediction accuracy was dramatic: For 98 (6%) phenotypes, predictions using correlated traits alone explained <10% of variation, but the inclusion of additional covariates increased accuracy to >40% of variation explained. Nevertheless, there were some phenotypes we remained unable to predict accurately: For 20% of transcripts and 17% of proteins, our predictions explained less than one-fifth of the variation between strains.

For both transcript and protein predictions in our expanded model, the most informative predictors were the correlation-based predictions (above) and abundance of the opposite data type (RNA or protein) for the gene in question, followed by strain and RNA/protein levels in other closely related strains (Supplemental Table 8). The presence of a TATA box was associated with better-predicted genes; we could explain more than half of the variation on average in transcript (median 54.6%; *N* = 343) and protein levels (median 55.1%; *N* = 211) for TATA box-containing genes (Fig. 5E). Predictions for the same TATA box-containing genes using the method above (without additional predictor variables) explained a median 41.7% of the variation, reinforcing the notion that these predictors can substantially improve prediction accuracy, at least for some subsets of phenotypes. To explore the predictive power of DNA sequence alone, we predicted variation in transcript and protein levels using only sequence and annotation information. Specifically, we used genic characteristics, sequence features, and pathway annotations as predictor variables, and found that we were able to explain far less variation: a median 24.6% across all transcripts and 21.7% across all proteins.

Moreover, we also implemented targeted models for specific pathways and protein complexes whose steps and constituents are well understood. We were able to make highly accurate predictions in some cases, explaining at least 75% of the variation for half or more of measured transcript and protein levels in pathways including ATP synthesis and the electron transport chain, trehalose biosynthesis, glycogen catabolism, and protein levels of the RNA polymerase I complex. We also constructed models to predict metabolites that differed in abundance between strains using genes in biochemical pathways known to involve the metabolite. For some metabolites (e.g., ribose/ribose-5-phosphate, trehalose), we achieved high predictive accuracy, explaining >50% of the variation in metabolite levels (Fig. 5F), but other metabolite levels were poorly predicted, probably due to the influence of numerous pathways on metabolic flux.

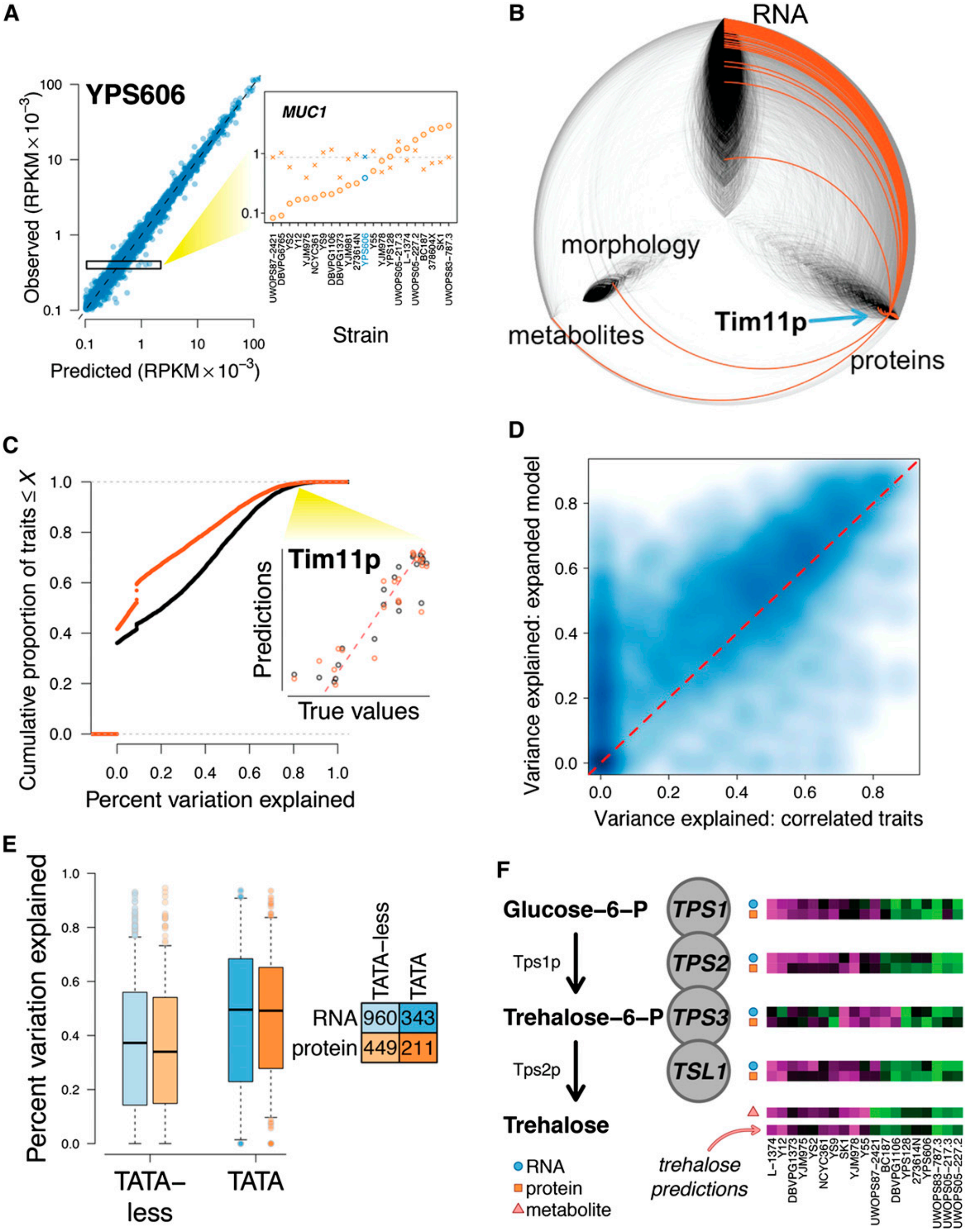

Figure 5. Integrating data to predict phenotypes. (*A*) Simple models can accurately predict gene expression levels when compared between genes (*left*, predictions for $n = 5385$ genes in strain YPS606 are shown), but do not fully capture variation between strains at a specific gene (*right inset*, gene expression levels for *MUC1* are shown for all 22 strains, with YPS606 highlighted in blue; units on *y*-axis are same as at *left*). At the *MUC1* locus, predicted values (X's) are clustered around the mean expression across strains (gray dotted line), but observed values (circles) diverge substantially. Observed RPKM values have been normalized (Supplemental Note). (*B*) Hive plot arranged identically to Figure 4A, with orange edges indicating connections to the node representing abundance of the Tim11 protein (blue arrow). (*C*) Empirical cumulative distribution function (CDF) displaying predictive accuracy for correlation-based phenotype predictions. Black line indicates CDF for predictions made using all phenotypes, and orange line for predictions made using only 1000 tag traits. *Inset* shows predictions for abundance of the Tim11 protein made using all traits (black dots) and tag traits only (orange dots). (*D*) Smooth scatter plot comparing performance of prediction models discussed in the text. Darker blue indicates higher density of points, and lighter blue indicates lower density. Dotted red line is drawn at $y = x$. (*E*) Boxplot indicating percent variation explained for models of transcript and protein levels. Boxes indicate lower quartile, median, and upper quartile, and whiskers extend to half the interquartile range. (*F*) Model for predicting levels of the metabolite trehalose, which is synthesized by the trehalose-phosphate synthase complex. Heatmaps show relative levels of transcript, protein, or metabolite, with blue circles, orange squares, and pink triangles distinguishing between data types. Each heatmap is arranged with the strains ordered *left* to *right* in the order shown at *bottom*.

Just as tag SNPs can be used to capture a large fraction of genetic variation with a small number of SNPs, a relatively small number of phenotypes (which we term tag traits) can capture a significant fraction of phenotypic variability. We implemented a simple greedy algorithm to identify tag traits highly correlated to many other phenotypes. Using only the top 1000 tag traits (12% of the data), we were able to explain at least 50% of variation in 975 (18%) phenotypes that differed significantly between strains (Fig. 5C, orange line). Abundance of the Tim11 protein was well-predicted using tag traits (Fig. 5C, inset, orange points), and well-predicted phenotypes were enriched for largely identical functions as those predicted without the aid of tag traits (Supplemental Table 7). Among the 2569 phenotypes for which the full models explained at least one-third of the variation among strains, tag trait models explained a median 79.1% of variation explained using all traits, indicating that tag traits can make use of a relatively small portion of the data to capture a significant fraction of variability between strains.

Discussion

The extensive phenomics data set described here provides new insights into the structure and characteristics of phenotypic diversity and highlights the information available in deep phenotyping across many strains. We observed pervasive phenotypic diversity, with a substantial fraction of phenotypes varying between strains for each data type. Many of the differences between strains occurred in pathways and networks related to cellular respiration, fermentation, and mitochondrial function, likely driven by both adaptations to ecological niche and interactions with humans (e.g., strains involved in the production of alcoholic beverages) (Fig 2B). In addition, we found several hundred *cis*-regulatory QTL that influence levels of transcripts or peptides, demonstrating that many large-effect regulatory polymorphisms segregate in natural populations of yeast. A substantial proportion of regulatory QTL act at both the transcript and peptide level, suggesting that the genetic basis of transcript and protein levels may overlap to a greater extent than previously thought (Foss et al. 2011; Ghazalpour et al. 2011).

Furthermore, we show that the structure of phenotypic correlations can be exploited to predict variation in phenotypes between strains. Tag traits, which we define as phenotypes belonging to a relatively small collection of traits that can capture a significant fraction of phenotypic variability, may be a useful way to conceptualize the state of a cell within a relatively low-dimensional space. Our limited ability to predict phenotypic variation using DNA sequence alone suggests that simple DNA sequence-based models of variation might benefit from the inclusion of additional strategically chosen functional genomics phenotypes, which has implications for the successful implementation of personal genomics (Chen et al. 2012). Moreover, our data will serve as a useful starting point for transitioning from quantitative, systems-level models of bacterial cells (Karr et al. 2012) to similar models of eukaryotes. As technology development, advances in instrumentation, and algorithmic improvements allow for increasingly comprehensive phenomics studies, a promising future direction will be to extend this approach to multiple environments, where organisms are naturally found. Finally, our data will be a useful community resource, and is available in multiple forms at http://www.yeastrc.org/g2p/ (Supplemental Fig. 6).

Methods

Chemostat growth

We grew strains in chemostats in phosphate-limited media until cultures were deemed to have reached steady state (defined as stabilizing to within 10% of the previous day's density measurements). We grew all strains in at least two chemostats to produce biological replicates. In order to quantify steady state, we used measurements by Klett colorimeter and by spectrophotometer, using the same instruments each time for consistency. To avoid any perturbation, we took sample culture passively at the effluent port. When the culture density stabilized, we harvested the chemostat and used the samples for RNA, protein, metabolite, and microscopy studies. Details of strain preparation, chemostat media, and harvesting procedures are provided in the Supplemental Note.

Phenotyping

For genome sequencing, we grew strains to mid-log phase (OD_{660} 0.8–1.0) in yeast extract peptone dextrose and extracted DNA by the phenol:chloroform:IAA method (Rose et al. 1990). We constructed sequencing libraries as previously described (Tennessen et al. 2012) and performed whole-genome sequencing using the Illumina HiSeq 2000 (50-bp paired-end reads), barcoding individual samples with Illumina's multiplex sample preparation oligonucleotide kit. Details on genotyping, validation of SNP calls using Sanger sequencing, phylogeny construction, and preparing strain-specific reference genomes and peptide databases are provided in the Supplemental Note.

For RNA-seq, we began with frozen cells from samples taken from the chemostats and extracted RNA by the acid phenol method (Chomczynski and Sacchi 1987). We performed poly(A) enrichment [MicroPoly(A) Purist Kit, Ambion] followed by ribosomal depletion (RiboMinus Kit, Invitrogen). We prepared RNA-seq libraries, barcoded samples, and performed sequencing (50-bp single-end reads) according to the manufacturer's recommendations using the ABI SOLiD v4 (SOLiD Whole Transcriptome Analysis Kit, ABI). We randomly allocated all samples across three flowcells and obtained 5–50 million reads per sample.

For quantitative proteomics, we lysed and digested samples (Supplemental Note) and ran on an LTQ-FT mass spectrometer (ThermoFisher), using an equimolar mix of a six protein bovine digest (Michrom Bioresources, Inc.) for quality control (Supplemental Fig. 1). We randomized samples and ran them in replicate. We processed high-resolution mass spectrometry data using Bullseye (Hsieh et al. 2010) to optimize precursor mass information. To identify peptides, we searched the MS/MS data using SEQUEST (Eng et al. 1994) against a FASTA database containing all protein sequences from all the strains. We determined peptide spectrum match false discovery rates using Percolator (Käll et al. 2007) at a *Q*-value threshold of 0.01 and a posterior error probability threshold of 1. We assembled peptides into protein identifications using an in-house implementation of IDPicker (Zhang et al. 2007). To obtain a quantitative measure of peptide abundance, we used the program Topograph (Hsieh et al. 2012). Detailed experimental and algorithmic protocols are provided in the Supplemental Note.

We extracted metabolites (Supplemental Note) and performed derivatizations as previously described (Fowler et al. 2011). All metabolite analysis was done on a Leco Pegasus 4D system (GC×GC-TOFMS). We acquired and processed data using the ChromaTOF software. We eliminated from analysis all samples with an insufficient average signal-to-noise ratio, which resulted in a total of 91 samples representing 23 different strains. We used the software ChromaTOF (Leco) for peak calling and deconvolution and the software package Guineu version 1.0 (Castillo et al. 2011) to align the common metabolites among 91 individual sample files. Detailed experimental protocols and a list of metabolites measured are provided in the Supplemental Note.

For cellular morphology, we fixed and stained cells (Supplemental Note) and mounted on an agarose pad as described (http://www.youtube.com/watch?v=ZrZVbFg9NE8) except that we did not dry the pad before adding cells. We acquired images on a DeltaVision Core using a 100× UPlanApo NA 1.35 objective and the Photometrics CoolSnapHQ camera. We processed images using CalMorph software (Ohya et al. 2005; Supplemental Note). We collected data for approximately 800 cells per strain in two replicates. CalMorph outputs measurements that can be used to construct a total of 501 traits (Nogami et al. 2007), but we discarded any measurements related to image brightness as we considered this unreliable to measure. We did not use any "total stage" traits calculated at the population level since we selected some cells from each cell cycle stage rather than completely at random, leaving a total of 398 traits.

Normalization and data analysis

Overall, we implemented a two-stage model to test for differential abundance of transcripts, proteins, and metabolites. In the first stage, we used a linear model to remove effects due to batch and other factors not of primary interest. We obtained normalized data values by extracting residuals from this model. In the second stage, we considered each gene separately and tested for a strain influence on phenotype using a random effects model with normalized data values from stage one. We used a similar approach to test for differences in morphological traits, modified slightly to appropriately handle measurements from many cells for each trait. We used R (R Development Core Team 2012) for all statistical analysis throughout the paper. Details of this statistical approach are provided in the Supplemental Note.

Identifying large-effect *cis*-regulatory transcript and protein QTL

We conducted simulations to determine the power (Supplemental Table 5) and false positive rate (Supplemental Fig. 4) for our tests of association (Supplemental Note). To map *cis*-regulatory gene and protein expression QTL, for each gene we focused on SNPs located within 500 bp of annotated gene boundaries. We employed the program EMMA (Kang et al. 2008), a mixed model approach that performs well for controlling population structure in this scenario (Connelly and Akey 2012), and performed permutations to calculate *P*-values.

Examining the structure of phenotypic correlations

Before calculating RNA-protein correlations, we averaged the abundance of all peptides mapping to the protein in question and used this as a surrogate for protein level. We averaged transcript and protein levels between biological replicates. We searched for a difference in variability between TATA-containing genes and TATA-less genes because it has been shown that the former category of genes shows greater variability in gene expression than the latter among yeast species (Tirosh et al. 2006).

To produce hive plots, we obtained values for RNA, protein, metabolite, and morphological phenotypes averaged between biological replicates and used nearest neighbor averaging to impute missing values (in the metabolite and morphological data) (Troyanskaya et al. 2001). We combined the resulting 1645 protein, 6207 gene expression, 115 metabolite, and 398 morphology trait values into a single matrix and calculated Spearman's rank correlation coefficient between each pair of phenotypes. We permuted the data set and recalculated correlations, finding that a cutoff of $\rho = 0.7625$ corresponded to a FDR of ~5%. We modified code from the HiveR package, available at http://academic.depauw.edu/~hanson/HiveR/HiveR.html, to create the plots in Figures 4A and 5B.

Phenotypic prediction

We exploited the phenotypic correlation structure to make predictions of interstrain phenotypic variation. Beginning with 8365 phenotypes (above), we sequentially withheld each strain, recalculated pairwise correlations between all phenotypes, and recorded the phenotypes that were highly correlated (FDR = 5%) with each other phenotype. Unless otherwise noted, we constructed models using random forest regression. Algorithmic details for predictive modeling are provided in the Supplemental Note.

We identified tag traits using a greedy algorithm where we first selected the phenotype correlated with the largest number of other phenotypes, then removed the selected phenotype and all phenotypes correlated with it (as they are "tagged" by the selected phenotype). We repeated this process until we acquired the desired number of tag traits.

Data access

All data from this study are available at http://www.yeastrc.org/g2p/. In addition, whole-genome sequence data and gene expression data have been deposited in the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession number SRP018005. Fluorescence microscopy data have been placed in the YRC Image Repository (Riffle and Davis 2010) and may be accessed at http://images.yeastrc.org/g2p. Tracks for visualization of the DNA sequencing and RNA-seq data have been made available as a public track hub at the UCSC Genome Browser (Kent et al. 2002); see http://www.yeastrc.org/g2p/.

Acknowledgments

We thank Deborah Nickerson, Jay Shendure, and Emily Turner for genome sequencing advice; Rachel Youngblood for assistance verifying the phosphate limitation of strains; and John Stamatoyannopoulos and Daniel Bates for assistance with RNA-seq. This project was supported by NIGMS grant P41GM103533 to T.N.D., NIDDK grant 1F32DK080608 to S.J.C., and NIGMS grants R01GM098360 and R01GM094810 to J.M.A. M.J.D. is a Rita Allen Foundation Scholar. S.F. is an investigator of the Howard Hughes Medical Institute.

References

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R, Matese JC, Gresham D, Boer VM, Troyanskaya OG, Botstein D. 2008. Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Mol Biol Cell* **19:** 352–367.

Breiman L. 2001. Random forests. *Mach Learn* **45:** 5–32.

Castillo S, Mattila I, Miettinen J, Orešič M, Hyötyläinen T. 2011. Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry. *Anal Chem* **83:** 3058–3067.

Castrillo JI, Zeef LA, Hoyle DC, Zhang N, Hayes A, Gardner DCJ, Cornell MJ, Petty J, Hakes L, Wardleworth L, et al. 2007. Growth control of the eukaryote cell: A systems biology study in yeast. *J Biol* **6:** 4.

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, et al. 2012. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148:** 1293–1307.

Chomczynski P, Sacchi N. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* **162:** 156–159.

Connelly CF, Akey JM. 2012. On the prospects of whole-genome association mapping in *Saccharomyces cerevisiae*. *Genetics* **191:** 1345–1353.

de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. 2009. Global signatures of protein and mRNA expression levels. *Mol Biosyst* **5:** 1512–1526.

Eng JK, McCormack AL, Yates JR III. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5:** 976–989.

Foss EJ, Radulovic D, Shaffer SA, Goodlett DR, Kruglyak L, Bedalov A. 2011. Genetic variation shapes protein networks mainly through non-transcriptional mechanisms. *PLoS Biol* **9:** e1001144.

Fowler DM, Cooper SJ, Stephany JJ, Hendon N, Nelson S, Fields S. 2011. Suppression of statin effectiveness by copper and zinc in yeast and human cells. *Mol Biosyst* **7:** 533–544.

Freimer N, Sabatti C. 2003. The human phenome project. *Nat Genet* **34:** 15–21.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477:** 419–423.

Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, Mungrue IN, Farber CR, Sinsheimer J, Kang HM, Furlotte N, et al. 2011. Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet* **7:** e1001393.

Gonzaga-Jauregui C, Lupski JR, Gibbs RA. 2012. Human genome sequencing in health and disease. *Annu Rev Med* **63:** 35–61.

Greenbaum D, Colangelo C, Williams K, Gerstein M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* **4:** 117.

Houle D, Govindaraju DR, Omholt S. 2010. Phenomics: The next challenge. *Nat Rev Genet* **11:** 855–866.

Hsieh EJ, Hoopmann MR, MacLean B, MacCoss MJ. 2010. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J Proteome Res* **9:** 1138–1143.

Hsieh EJ, Shulman NJ, Dai D-F, Vincow ES, Karunadharma PP, Pallanck L, Rabinovitch PS, MacCoss MJ. 2012. Topograph, a software platform for precursor enrichment corrected global protein turnover measurements. *Mol Cell Proteomics* **11:** 1468–1474.

Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. 2007. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4:** 923–925.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* **178:** 1709–1723.

Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW. 2012. A whole-cell computational model predicts phenotype from genotype. *Cell* **150:** 389–401.

Keane TM, Goodstadt L, Danacek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477:** 289–294.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996–1006.

Kvitek DJ, Will JL, Gasch AP. 2008. Variations in stress sensitivity and genomic expression in diverse *S. cerevisiae* isolates. *PLoS Genet* **4:** e10000223.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458:** 337–341.

Ng PC, Murray SS, Levy S, Venter JC. 2009. An agenda for personalized medicine. *Nature* **461:** 724–726.

Nogami S, Ohya Y, Yvert G. 2007. Genetic complexity and quantitative trait loci mapping of yeast morphological traits. *PLoS Genet* **3:** e31.

Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, Saito TL, Saka A, Fukuda T, Ishihara S, Oka S, et al. 2005. High-dimensional and large-scale phenotyping of yeast mutants. *Proc Natl Acad Sci* **102:** 19015–19020.

R Development Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Ratnakumar S, Hesketh A, Gkargkas K, Wilson M, Rash BM, Hayes A, Tunnacliffe A, Oliver SG. 2011. Phenomic and transcriptomic analyses reveal that autophagy plays a major role in desiccation tolerance in *Saccharomyces cerevisiae*. *Mol Biosyst* **7:** 139–149.

Regenberg B, Grotkjær T, Winther O, Fausbøll A, Åkesson M, Bro C, Hansen LK, Brunak S, Nielsen J. 2006. Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in *Saccharomyces cerevisiae*. *Genome Biol* **7:** R107.

Riffle M, Davis TN. 2010. The Yeast Resource Center Public Image Repository: A large database of fluorescence microscopy images. *BMC Bioinformatics* **11:** 263.

Rose MD, Winston F, Hieter P. 1990. *Methods in yeast genetics: A laboratory course manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schork NJ. 1997. Genetics of complex disease: Approaches, problems, and solutions. *Am J Respir Crit Care Med* **156:** S103–S109.

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473:** 337–342.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci* **100:** 9440–9445.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337:** 64–69.

Tirosh I, Weinberger A, Carmi M, Barkai N. 2006. A genetic signature of interspecies variations in gene expression. *Nat Genet* **38:** 830–834.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17:** 520–525.

Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13:** 227–232.

Warringer J, Ericson E, Fernendez L, Nerman O, Blomberg A. 2003. High-resolution yeast phenomics resolves different physiological features in the saline response. *Proc Natl Acad Sci* **100:** 15724–15729.

Warringer J, Zörgö E, Cubillos FA, Zia A, Gjuvsland A, Simpson JT, Forsmark A, Durbin R, Omholt SW, Louis EJ, et al. 2011. Trait variation in yeast is defined by population history. *PLoS Genet* **7:** e1002111.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457:** 1033–1037.

Xu Z, Wei W, Gagneur J, Clauder-Münster S, Smolik M, Huber W, Steinmetz LM. 2011. Antisense expression increases gene expression variability and locus interdependency. *Mol Syst Biol* **7:** 468.

Zbuk KM, Eng C. 2007. Cancer phenomics: *RET* and *PTEN* as illustrative models. *Nat Rev Cancer* **7:** 35–44.

Zhang B, Chambers MC, Tabb DL. 2007. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res* **6:** 3549–3557.

Received February 1, 2013; accepted in revised form May 20, 2013.