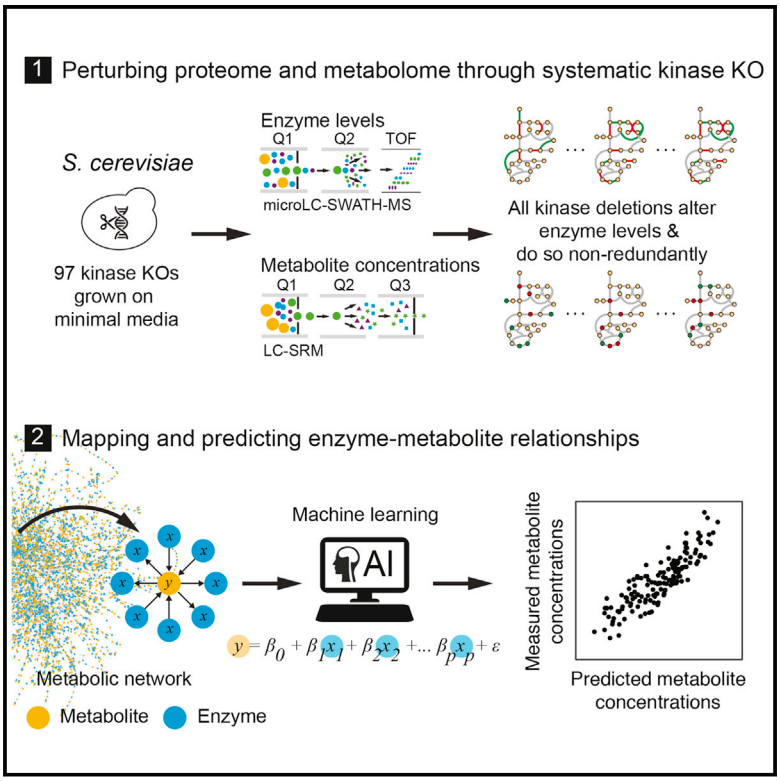
Article

Cell Systems

Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts

Graphical Abstract



Highlights

- The proteome of kinase knockouts is dominated by enzyme abundance changes
- The enzyme expression profiles of kinase knockouts are non-redundant
- Metabolism is regulated by many expression changes acting in concert
- Machine learning accurately predicts the metabolome from enzyme abundance

Authors

Aleksej Zelezniak, Jakob Vowinckel, Floriana Capuano, ..., Bernd Klaus, Markus A. Keller, Markus Ralser

Correspondence

markus.ralser@crick.ac.uk

In Brief

Predicting metabolomes from gene expression data is a key challenge in understanding the genotype-phenotype relationship. Studying the enzyme expression proteome in kinase knockouts, we reveal the importance of a so far overlooked metabolism-regulatory mechanism. Enzyme expression changes are impacting on metabolite levels through many changes acting in concert. We show that one can map regulatory enzyme expression patterns using machine learning and use them to predict the metabolome of kinase-deficient cells on the basis of their enzyme expression proteome. Our study quantifies the role of enzyme abundance in the regulation of metabolism and by doing so reveals the potential of machine learning in gaining understanding about complex metabolism regulation.

Zelezniak et al., 2018, Cell Systems 7, 269–283
September 26, 2018
https://doi.org/10.1016/j.cels.2018.08.001

CellPress

Cell Systems
Article

Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts

Aleksej Zelezniak,[1,2,3,4] Jakob Vowinckel,[2,5] Floriana Capuano,[2] Christoph B. Messner,[1] Vadim Demichev,[1,2] Nicole Polowsky,[2] Michael Mülleder,[1,2] Stephan Kamrad,[1,7] Bernd Klaus,[6] Markus A. Keller,[2,8] and Markus Ralser[1,2,9,10,*]
[1]The Francis Crick Institute, Molecular Biology of Metabolism laboratory, London, UK
[2]Department of Biochemistry and Cambridge Systems Biology Centre, University of Cambridge, Cambridge, UK
[3]Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden
[4]Science for Life Laboratory, KTH – Royal Institute of Technology, Stockholm, Sweden
[5]Biognosys AG, Schlieren, Switzerland
[6]Centre for Statistical Data Analysis, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
[7]Department of Genetics, Evolution and Environment, University College London, London, UK
[8]Medical University of Innsbruck, Innsbruck, Austria
[9]Department of Biochemistry, Charité Universitaetsmedizin Berlin, Berlin, Germany
[10]Lead Contact
*Correspondence: markus.ralser@crick.ac.uk
https://doi.org/10.1016/j.cels.2018.08.001

SUMMARY

A challenge in solving the genotype-to-phenotype relationship is to predict a cell's metabolome, believed to correlate poorly with gene expression. Using comparative quantitative proteomics, we found that differential protein expression in 97 *Saccharomyces cerevisiae* kinase deletion strains is non-redundant and dominated by abundance changes in metabolic enzymes. Associating differential enzyme expression landscapes to corresponding metabolomes using network models provided reasoning for poor proteome-metabolome correlations; differential protein expression redistributes flux control between many enzymes acting in concert, a mechanism not captured by one-to-one correlation statistics. Mapping these regulatory patterns using machine learning enabled the prediction of metabolite concentrations, as well as identification of candidate genes important for the regulation of metabolism. Overall, our study reveals that a large part of metabolism regulation is explained through coordinated enzyme expression changes. Our quantitative data indicate that this mechanism explains more than half of metabolism regulation and underlies the interdependency between enzyme levels and metabolism, which renders the metabolome a predictable phenotype.

INTRODUCTION

Despite the fact that metabolism is intensively studied, one still debates about how much of metabolic regulation is explained by metabolic self-regulation and by regulation of enzyme activity and how much is dependent on enzyme abundance changes. The current literature is split, in essence, between two seemingly contrasting observations. On the one hand, available quantitative models can explain only a minor fraction of metabolite concentrations on the basis of gene expression data (Fendt et al., 2010; Kresnowati et al., 2006; Zelezniak et al., 2014). Moreover, metabolite concentrations seem to correlate much better with metabolic fluxes than with enzyme expression levels (Chubukov et al., 2013; Hackett et al., 2016; Millard et al., 2017). These results seem to suggest that the post-translational regulation, metabolic self-regulation, and allostery are dominant in metabolism regulation.

On the other hand, however, large fractions of the transcriptome respond to changes in metabolism (Alam et al., 2016; Bradley et al., 2009; Chechik et al., 2008; Kresnowati et al., 2006; Murray et al., 2007; Tu et al., 2005; Urbanczyk-Wochniak et al., 2003). The expression changes are centered on metabolites that change in concentration (Patil and Nielsen, 2005; Zelezniak et al., 2010), while systematically recorded transcriptomes and proteomes of metabolically perturbed yeast correlate with metabolic flux distributions (Alam et al., 2016). Hence, despite poor correlation values between individual enzyme levels and metabolism, changes in metabolism seem tightly intertwined with gene expression changes. Indeed, all metabolism-regulating transcriptional and signaling networks identified to date, such as AMP-activated protein kinase (AMPK) (Mihaylova and Shaw, 2011), mechamTOR (González and Hall, 2017), or GCN2/4 (Zaborske et al., 2010), trigger metabolic gene expression changes.

A potential explanation for this apparent paradox could be provided by the nature of enzyme-metabolite relationships. Reaction mechanisms (Braakman and Smith, 2013), the self-regulatory nature of metabolic networks (Alam et al., 2017; Chubukov et al., 2013; Hackett et al., 2016; Millard et al., 2017), post-translational regulation (Daran-Lapujade et al., 2007; Gonçalves et al., 2017; Nilsson et al., 2017; Oliveira et al., 2012), and the topological organization of metabolism that routes evolutionarily in the underlying chemistry (Burgard et al., 2004; Keller et al., 2015; Zelezniak et al., 2014) all dictate that the relationship between enzyme function and metabolites is both multifactorial and dynamic.

We selected a genome-spanning collection of 97 kinase gene deletion ("knockout") *Saccharomyces cerevisiae* strains, known to exhibit differences in metabolism (Bodenmiller et al., 2010; Schulz et al., 2014; van Wageningen et al., 2010; Winzeler et al., 1999), noting that the gene expression changes in these strains remained uninvestigated in the context of their metabolome. A recently developed high-throughput proteomic platform (Vowinckel et al., 2018) was used to quantify enzyme expression and link enzyme expression changes to metabolite concentrations measured.

All kinase deletions triggered enzyme expression changes. Moreover, enzyme abundance changes dominated quantitatively over other differentially expressed functional protein categories in the kinase knockout proteomes. Using metabolic control analysis (MCA), we then revealed the importance of largely overlooked mechanisms in metabolic regulation. The proteomic changes detected were so broad that metabolic control shifts between different sets of enzymes. As a consequence, metabolic regulation becomes sensitive to global changes in gene expression, rather than being correlated to individual enzymes. To capture the multifactorial relationships, we developed a data-driven framework based on machine learning (ML). Training the algorithms on the basis of the metabolic network topology, we achieved the quantitative prediction of entire cellular metabolomes, thereby quantifying the role of enzyme abundance changes in metabolism regulation.

RESULTS

Kinase Knockout Proteomes Are Dominated by Differential Enzyme Expression

S. cerevisiae kinase gene knockout strains (Winzeler et al., 1999) were rendered prototrophic by introducing the pHLUM minichromosome (Mülleder et al., 2012) and cultivated in the absence of amino acid supplementation (STAR Methods; Figure S1 for growth rates). The measurement of the 97 proteomes mounted to 397 whole-proteome samples (triplicates plus controls) processed using the data-independent acquisition method SWATH-MS (Gillet et al., 2012) and the workflow optimized for achieving high quantification precision at large sample numbers (Vowinckel et al., 2018). The median coefficient of variation of protein abundance obtained was 19% (Figures 1A and S2). Cut-off values for differential protein expression were determined experimentally and defined as a 40% change, and we used a Benjamini-Hochberg (BH) adjusted p value cutoff of 0.01 (STAR Methods). To confirm that differentially expressed genes were specific to kinase deletions, a subset of 10 strains was mated to a wild-type (WT) strain or to a complementary kinase-knockout; in all cases, the proteomes in which the kinases were reintroduced centered closer to the WT proteomes and were different from homozygous mutants (Figure S3).

To capture enzyme expression values, we processed all SWATH proteomes using a spectral library generated from a soluble yeast protein extract and obtained a matrix that connects the 97 kinase deletions to the abundance of 286 metabolic enzymes (median q value < 0.01, hereafter called the "kinase-metabolic enzyme matrix"). These represent over 75% of the metabolic reactions that are coupled to biomass growth (STAR Methods; Figure 1B) and capture cytoplasmic metabolism close to completion (Figure 1C).

Each proteome was characterized by strong differential enzyme expression (Figure 1D). By comparing the kinase-metabolic enzyme matrix of each knockout strain to the full SWATH proteomes, we observed that 39% of all detected protein expression changes were attributable to metabolic enzymes. On average, a kinase deletion affected the abundance of 56 metabolic enzymes; the minimum was 7 enzymes differentially expressed upon deletion of *DBF2*, and the maximum was 140 enzymes upon deletion of *MEK1* (Figure 1E). Expressed in absolute protein copy numbers, up to 25% of the total cell protein abundance is affected by kinase deletions acting on enzyme abundance (Figure 1E). It is unlikely that these changes reflect a common pleiotropic mechanism. For example, although yeast growth rate itself is understood to control gene expression (Gasch et al., 2000), less than 10% of the total proteome changes in our kinase knockout strains could be explained by changes in growth rate (Figure S4). Moreover, there was no strong correlation ($r = 0.22$, $p = 0.04$) between the total number of differentially expressed proteins and the fraction of differentially expressed metabolic enzyme genes (Figure 1F).

We then compared protein expression levels to microarray-based transcriptional profiles (van Wageningen et al., 2010). The transcriptional profiles correlated significantly with the enzyme expression proteomes (Figure 1G; STAR Methods; Figure S5). Most likely, as the strains in the microarray-study were cultivated in amino-acid-supplemented media, the absolute correlation values were lower than previous studies in which yeast cells are grown under the exact same condition (Alam et al., 2016; Lahtvee et al., 2017; Marguerat et al., 2012). The significant correlation nonetheless indicates that transcriptional regulation is implicated in the protein abundance changes as detected. This analysis further revealed that enzymes differentially expressed are enriched among the highly expressed genes, while in the low abundant fraction of the transcriptome, differential enzyme expression is also significant but less prevalent (Figure 1F; STAR Methods). In parallel, a weak but significant correlation was obtained between protein degradation rates (Christiano et al., 2014) and the likelihood of an enzyme to be differentially expressed (Figure S6). Kinase deletions hence affect enzyme abundance both via hierarchical regulation, as well as via mechanisms that affect protein turnover.

Enzyme Expression Signatures Reveal a High Degree of Specificity in Kinase Function

In a few cases, we observed a significant overlap between the enzyme proteomes, which seems to suggest common biological function. For example, deletion of MAPK kinases *HOG1* and *KSS1*, which share upstream signaling components (Saito and Tatebayashi, 2004), caused enzyme proteomes that did overlap in 25% and 33% of up- and down-regulated enzymes, respectively. Moreover, kinases of the same protein family were significantly more likely to also affect similar enzyme targets (one-way ANOVA $p = 0.0092$). For instance, the Ca2+/calmodulin-dependent protein kinase (CamK) and Casein kinase I (CKI) families revealed significant co-regulation in enzyme abundances ($p = 0.0091$) (Figure S10).

For most kinase deletions, however, the precise proteome data revealed a high degree of specificity. Moreover, the proteomes suggest that enzyme expression regulation is too

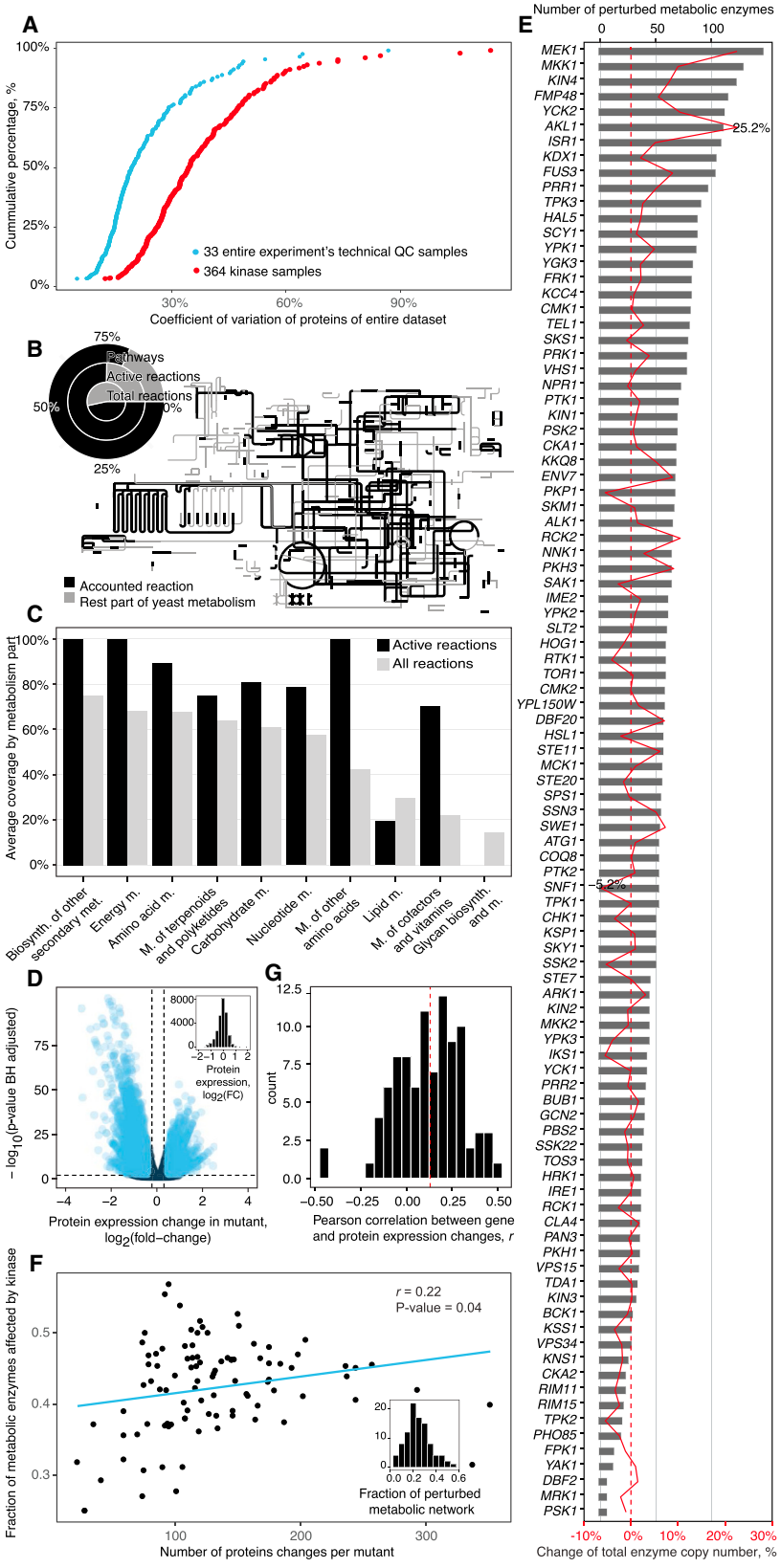


Figure 1. A Deletion of Each of the 97 Non-essential Yeast Protein Kinases Triggers Broad and Quantitatively Strong Changes in Metabolic Enzyme Expression

(A) Biological versus technical variability in a large-scale proteomic experiment. The coefficient of variation (CV) of enzymes at whole-process technical and biological levels. Cyan dots indicate CVs of a standardized proteome digest (quality control [QC] sample) that was used to monitor instrument performance over a 4-month acquisition period. QCs were used to normalize for batch effects, as well as to determine adequate cutoff values for determining differential protein expression. See also Figure S2 and STAR Methods.

(B) Projection of quantified enzymes on the KEGG metabolic pathway map using iPath (Yamada et al., 2011) illustrates a connected network coverage, indicating comprehensive coverage of the active metabolic reactions by the proteome data. The black lines represent reactions catalyzed by at least one quantified enzyme; gray lines represent enzymatic reactions for which no enzyme was quantified. Circle plot: obtained coverage in comparison to all metabolic pathways' theoretically active reactions (reactions that couple to biomass growth) in yeast as determined by flux-coupling analysis (Burgard et al., 2004) and compared to all KEGG-annotated reactions of the yeast metabolic network.

(C) MicroLC-SWATH-MS proteomes capture large parts of the active enzymome. The representation of KEGG metabolic pathways by enzymes quantified in each proteome, shown as average coverage of metabolic pathways per KEGG metabolism category (KEGG BRITE hierarchy level B). A reaction was considered covered if >1 enzyme with the corresponding EC number was quantified.

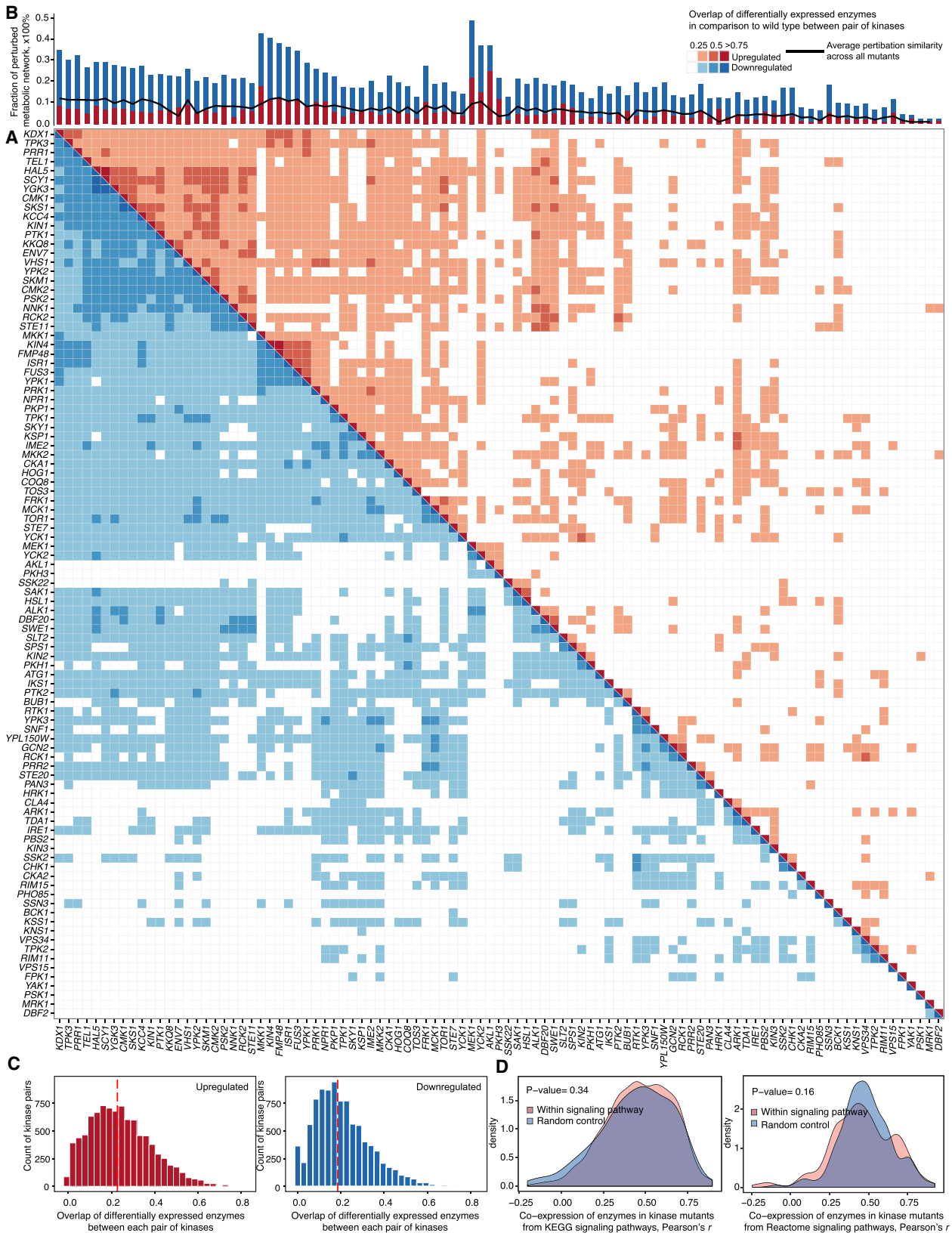
(D) Each of the 97 kinase deletions affects enzyme expression levels (volcano plot). Differential enzyme expression in all mutants is compared to the parental strain. Cutoffs were determined using repeated measurements on the control sample (STAR Methods) and determined as a fold change cutoff > $|\log_2(1.4/0.714)|$, Benjamini-Hochberg (Benjamini and Hochberg, 1995) adjusted $p < 0.01$, cyan colors indicate differentially expressed enzymes. Inset: the distribution of fold change values between mutants and parental strains.

(E) The total number of metabolic enzymes affected by kinase deletions illustrated for each kinase. Red line: influence of the individual kinase deletion in relation to the total enzyme copy number in percent. Copy number changes were obtained by calibrating the proteome data according to the absolute values of protein expression (Kulak et al., 2014) (Details are given in the STAR Methods section).

(F) Enzyme abundance changes account for a major fraction of all differentially expressed proteins as quantified in the kinase knockouts, and the relative contribution of enzymes has a low correlation with the total size of the proteomic perturbation. The y axis represents the fraction of the differentially expressed metabolic enzymes out of all quantified proteins. Inset: kinase deletions affect up to 49% of all quantified enzymes as denoted by the total of the metabolic network, summing up in all strains to 39% of the measured impact of the total kinome on protein expression.

(G) Correlation of metabolic enzymes between proteome and transcriptomes (van Wageningen et al., 2010) expressed as fold changes.

See also Figure S5.



(legend on next page)

complex to be explained by linear signaling pathways. A Jaccard distance calculated between each kinase pair's enzyme expression signature, as well as hierarchical clustering using complete linkage agglomeration (Figure 2A), revealed that 98% of kinase pairs have less than 50% overlap in differential enzyme expression. On average, two kinase enzyme proteomes overlap by less than 12% (Figure 2B). If expressed as a Pearson's correlation, three-quarters of the proteome changes in the typical kinase deletion were specific (Figure S7). A sensitivity analysis ruled out a thresholding artifact; indeed, with more conservative thresholds, the specificity of kinase proteomes is revealed more robustly (Figure S7).

Consistently, the signaling pathway annotations as assembled in both KEGG and Reactome databases (Fabregat et al., 2016; Kanehisa et al., 2016) could not explain enzyme co-expression and indeed were not more predictive about enzyme co-expression as random networks (identical Pearson's correlation coefficient, Wilcoxon rank-sum test, $p > 0.05$ [Figures 2D and S8]). This result was corroborated by comparing the overlaps of differentially expressed enzymes. A borderline significant association with the Reactome database pathways was explained because in the database, one pair of paralogous serine/threonine kinases (*YPK1* and *YPK2*, overlapping in 31% of their enzyme expression changes) is associated with 42% of signaling pathways (Wilcoxon rank-sum test, $p = 0.03$; Figure S9). When this pair is removed, no significant correlation of signaling pathway associations and enzyme co-expression was observed.

Enzyme Expression Affects Steady-State Metabolite Pools

MCA was then used to assess how the observed rearrangements in enzyme levels interact with central metabolism. We generated a specific glycolytic model for each kinase knockout by adjusting enzyme concentrations in a highly curated glycolytic model (Smallbone et al., 2013) according to our measurements. Flux (C^JE) and concentration control (C^SE) coefficients were determined to measure the relative steady-state change in the global system variables, i.e., flux (J) or metabolite concentration (S), in response to differential enzyme expression (E) (Kacser and Burns, 1973). Differential enzyme expression altered the overall flux control coefficients (FCCs) by more than 50%, for two-thirds of glycolytic enzymes in 78% of the kinase knockout strains. Similarly, the enzyme abundance changes as measured, altered the overall concentration control coefficients by more than 50% (Figure 3A). Differential enzyme expression does hence redistribute the control over glycolytic flux between different metabolic enzymes. We illustrate this situation for the metabolic flux going through alcohol dehydrogenase (ADH_ADH1, reaction abbreviation were kept as in Smallbone et al. [2013]). In the WT situation, the highest control over ethanol production is attributable to glucose phosphorylation by *hexokinase 2* (*HXK2*) (Figure 3B). Due to differential enzyme levels, the flux control shifts to other enzymes in the mutants (Figures 3B and S11), altering steady state by more than 2-fold in 48% of the kinase knockouts. The model predicts that in 55% of the kinase mutants, this re-shuffling affects metabolite concentrations (Figure 3A insets).

A principal-component analysis (PCA) of the FCCs yielded four distinct clusters. The cluster division was mainly attributable to the control of *HXK2*, phosphofructokinase 2 *(PFK2)*, and *ADH1* on glycolysis and energy metabolism (Figure S12), of which glucose phosphorylation by *HXK2* was the most dominating (*HXK_GLK*1 flux) (Figures 3C, inset, and S12). This result is consistent with experimental observations. *HXK2* is a known regulator of *GLK1*, alternatively expressed under different carbon sources (Rodríguez et al., 2001). Indeed, the ratio of *HXK2*/*GLK1* expression differs between the clusters (Figure 3D). In cluster 2, *HXK* was more than two times lower expressed as *GLK1*. Despite being the strongest contributor, however, the ratio of *HXK2*/*GLK1* expression alone is not sufficient to explain the differences, underlining that even in central metabolism, differential enzyme expression acting in concert is required to explain metabolic regulation (Figure 3E). Therefore, we simulated the impact of multifactorial enzyme expression changes on glycolytic flux. Altering the expression level of as little as 7 enzymes, as detected in the kinase knockouts, can change the median of all control coefficients by up to 100% (Figure 3H). Taken together, these analyses predict that differential enzyme expression affects central metabolism significantly and mainly by redistributing flux control between different sets of enzymes.

To test the predictions, we used liquid chromatography-selective reaction monitoring (LC-SRM) to quantify ATP, ADP, and AMP; glycolytic and pentose phosphate pathway (PPP) intermediates; as well as amino acids and Krebs cycle metabolites (Figure S13; STAR Methods). On this set of central metabolites, 34 of the 97 kinase knockouts exhibited one or more strong concentration changes (± 2 SDs from mean concentration levels; Figure S14). These measured concentrations correlated significantly with the predictions (Figures 3F and 3G).

Predicting the Metabolome from the Enzyme Expression Data

Next, we asked whether proteomic data could be used to explain the variation metabolite concentrations also at the scale of the

Figure 2. The Deletion of Each Yeast Kinase Triggers a Unique Reconfiguration of Enzyme Expression in the Cell

(A) Similarity and overlap between enzyme expression proteomes obtained upon kinase deletion in *S. cerevisiae*. Each cell represents the overlap in the compendium of differentially expressed enzymes (relative to the parental strain BY4741-pHLUM) between any pair of kinase knockouts. An enzyme is considered differentially expressed if the fold change $> |\log_2(1.4/0.714)|$, BH adj. $p < 0.01$. The matrix distinguishes between upregulated (red, upper right part of the matrix) and downregulated (blue, lower left part) enzymes. For illustration purposes, rows and columns are clustered according to the Jaccard distance between the proteomes, disregarding the directionality of the expression changes. The overlap between each pair of proteomes is shown as Jaccard similarity.

(B) The fraction of differentially expressed metabolic enzymes in comparison to total differential protein expression in all kinase mutants (bar chart). The absolute average similarity of kinase deletion enzyme proteomes, across all kinase mutants, is depicted as a black line. The typical kinase deletion causes a unique enzyme expression signature, with a median dissimilarity between kinase proteome pairs of 88% (average overlap between enzymes differentially expressed = 12%).

(C) The typical overlap of perturbed enzyme proteomes in kinases mutants is not more than ~25% (dotted median line).

(D) Enzyme expression changes ($\log_2$-fold change) are not better explained by the signaling pathway annotations as obtained from KEGG or Reactome databases compared to randomly assembled pathways. More comparisons are provided in Figures S9 and S10.

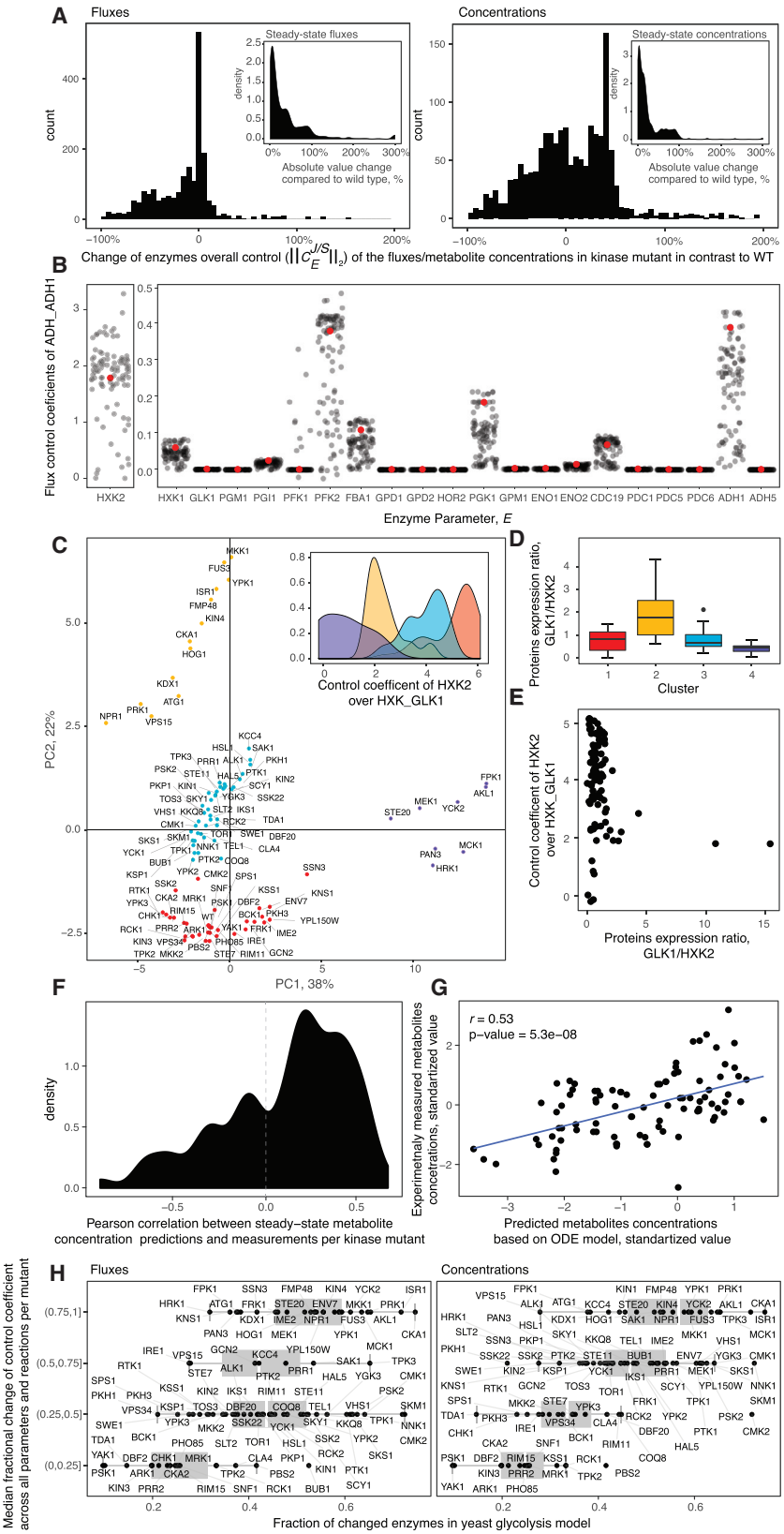


Figure 3. Enzyme Expression Affects Steady-State Metabolism through Redistributing Flux Control

(A) Overall control coefficients of concentrations (CCC) and fluxes (FCC) are changed in kinase deletion strains comparing to WT due to the differential expression of multiple enzymes. The overall FCCs were calculated as described in Millard et al. (2017), i.e., taking for every enzyme the second norm over all its concentrations and FCCs that were parameterized on it (STAR Methods). Insets: simulated steady-state changes of fluxes and metabolite levels in kinase mutants in comparison to WT.

(B) FCCs (C^J_E) over alcohol dehydrogenase (EC 1.1.1.1) reaction (y axis) by corresponding glycolytic enzymes (x axis) upon adjusting protein expression levels in a yeast glycolysis model as measured in each kinase knockout. Red dots indicate the WT strain values. To preserve the original scales, the control coefficients for *HXK2* are plotted on a separate y axis. Differential enzyme expression substantially redistributes control coefficients in multiple kinases to different enzymes.

(C) Principal-component analysis (PCA) of FCCs for every kinase gene deletion mutant reveals a distinct set of expression patterns that influences control over glycolysis. FCCs are not scaled (See also Figure S12). Axes labels represent the percentage of total variance explained by each of the PCs. Colors represent established flux regulatory clusters (STAR Methods). Cluster separation is mainly driven (inset) by control of *HXK2* on *GLK1* reaction.

(D) Within each flux regulatory cluster, large differences between the *GLK1/HXK2* expression ratio are observed. Corresponding p values for each pair of clusters using Wilcoxon rank-sum test (1 versus 2 p = 5.4e−05; 1 versus 3 p = 1.5e−02; 1 versus 4 p = 6.01e−01; 2 versus 3 p = 6.35e−05; 2 versus 4 p = 2.01e−03; 3 versus 4 p = 3.39e−01).

(E) Flux control is a systemic property that depends on the coordinated expression of multiple enzymes. Even the most dominant single contributor (*GLK1/HXK2* ratio, [x axis]) alone cannot explain the variation of flux control coefficients (y axis) as a result of differential enzyme expression.

(F) Measured metabolite concentrations correlate with steady-state predictions by the enzyme-level adjusted kinetic models.

(G) Correlation of model predictions and experimentally measured metabolite concentrations in the top 10 kinase mutants from (F).

(H) The systems-nature of metabolism control: differential expression of a few individual pathway enzymes is sufficient to induce a redistribution of flux control among a broad set of enzymes. Fractions of differentially changed enzymes from the model are plotted on the x axis. The y axis shows the median change of control coefficient for each parameter comparing to the parental strain divided into 4 groups. Group (0.75, 1) has coefficients with the median change up to >100% in comparison to WT.

metabolic network. First, we generated a network graph connecting the metabolites to enzymes according to a genome-scale reconstruction of the yeast metabolic network (Herrgård et al., 2008). The 46 metabolites quantified connect as a substrate, product, or cofactor to 192 enzymes (= 1st order neighbors). Each of these metabolite-enzyme relationships was expressed as a multiple linear regression (MLR) problem (Figure 4A). Then, we used exhaustive feature selection and ranked all possible models ($>10^{12}$) according to minimal Akaike information criterion (AIC) (Akaike, 1974). To minimize the risk of overfitting, we repeated the procedure 1,000 times for each metabolite using random subsets of the data and retained the top 5 most frequently identified metabolite-enzymes relationships from all random subsets (Figure S15), ranked them according to the best agreement between predicted and measured metabolite concentration, and calculated their explanatory power (as adjusted R^2). The models with the best fit were diagnosed for outliers, influential observations, residual structure, and the presence of autocorrelation. All models that violated this set of criteria were discarded. The statistical models obtained in this way show that changes in enzyme abundance do explain metabolite concentration (Figure 4B). Fructose-6-phosphate, glutamate, glutamine, ATP, ADP, and AMP levels as estimated from enzyme abundance correlate with their experimentally measured concentrations (adj. $R^2 > 0.4$; Figure 4B). We illustrate the model performance for ATP, ADP, AMP (Figure 4C), and glutamine, the metabolites for which the simple linear regression model provided the best results (Figures 4D and 4E). Thus, when accounting for metabolic network topology and multifactorial relationships, enzyme expression is informative of metabolite concentrations.

By being applied over the actual metabolic network topology, the feature selection approach identifies the predictive enzymes. For example, glutamine is connected to 28 of the quantified enzymes. Multiple feature selection identified 9 of them (*GLN4, CPA2, URA2, HIS7, ADE4, ADE6, ASN2, URA7*, and *GLT1*) to be significant contributors to its concentration (Figure 4E, in order of weight in the model), and together, their differential expression explains 68% (adj. R^2) of the experimentally detected glutamine concentration changes (Figure 4D). The strongest predictor of glutamine levels was *GLN4*, the glutamine aminoacyl-tRNA synthetase, indeed already known to be important for glutamine regulation (Murray et al., 1998). Of note, MLR identified aminoacyl-tRNA synthetases as the strongest predictors also for aspartate, glycine, proline, and tyrosine (Figure S15). MLR hence confirmed that tRNA loading is a major factor in amino acid concentration regulation (Wegrzyn and Wegrzyn, 2008; Whitney et al., 2007) and revealed that it is quantitatively one of the strongest single contributors to amino acid concentrations in general.

Other illustrative examples are ATP, ADP, and AMP (Figure 4C) that are among the most connected metabolites (Zomorrodi and Maranas, 2010). From the 88 ATP, ADP, or AMP metabolizing enzymes quantified, 33 were found predictive about their levels (Figure S15). This list contains many of the high-flux enzymes associated with the cellular energy charge, including *HXK2*, subunits of the electron transport chain ATPase (Complex V, *ATP2*) or the vacuolar ATPase (*VMA1*), a major consumer of cellular ATP (Beyenbach and Wieczorek, 2006).

Predictive and non-predictive enzymes were not different in their average abundance (Figure S16) but were so in saturation (substrate concentration > K_M) according to *in vitro* determined K_M values obtained from the Braunschweig Enzyme Database (BRENDA) (Chang et al., 2015). Considering the cell-average metabolite concentrations as measured in our study, enzymes identified by MLR were more than three times closer to the saturation than all other enzymes connected to the same substrates (Figure 3F; Wilcoxon rank-sum test, $p < 10^{-16}$). 45% of these enzymes appear saturated. Moreover, 40% of metabolites were associated with at least one enzyme with a K_M value at least ten times below the metabolite concentration. Amino acid metabolizing enzymes were >8-fold (comparing medians) (Wilcoxon rank-sum test, $p < 10^{-16}$) more saturated than predictors for other metabolites (Figure 4G). In accordance with the MLR analysis (Figure S15), we find that aminoacyl-tRNA synthetases were among the most saturated enzymes (Figure 4H).

In order to make use of the full metabolic enzyme expression matrix to predict metabolite concentrations, we implemented a pipeline that makes use of 12 ML algorithms. This analysis is summarized graphically in Figure 5A and detailed in the STAR Methods. In brief, we reduce the dimensionality of the proteome dataset, divide the data into training and testing sets using cross-validation to obtain best predictive regression model for each metabolite. Our ML approach predicted metabolite concentrations that on average correlated with the measured concentration values with a cross-validated R^2 value of 0.55. The highest predictability from enzyme abundance was revealed for tryptophan, ornithine, and citrulline, for which the predictions correlated with cross-validated R^2 of 0.75, 0.75, and 0.73, respectively, with their experimentally measured concentrations (Figure 5B).

The size of the metabolite's network neighborhood had only minimal influence on predictability. Several of the metabolites (ATP, 2-oxoglutarate, tryptophan, glutamine, and methionine) remained exclusively predicted by their directly metabolizing enzymes, (Figure 5C). For all other metabolites, predictability increased upon incorporation of the second-order neighbors but not anymore upon further network expansion (Figure 5C). Metabolite concentrations are therefore most sensitive to enzyme abundance changes occurring in their immediate neighborhood (Figures 5D and 5E). Of note, the best overall performing ML algorithm on our dataset was ridge regression with greedy variable selection (Zhang, 2011) (Figures 5B and S17). Finally, we tested the power of the ML model to predict not only the individual metabolites but also entire metabolomes. For this, we repeated the whole procedure ninety-seven times, using leave-one-out cross-validation for the entire dataset. The metabolomes predicted agreed with the metabolomes as measured experimentally (Figure 5F): 70% of absolute metabolite concentrations were predicted with less than 25% relative error (Figure 5G).

To test the validity of the predictions also on an independently generated dataset, we made use of amino acid concentrations that have been determined upon the systematic deletion of all non-essential yeast genes (Mülleder et al., 2016a). We compared the range of amino acid concentration changes, measured upon the deletion of (1) the subset of non-essential enzymes for which ML had attributed an important regulatory role (defined as >50% maximum weight of predictor variable; see STAR Methods) and

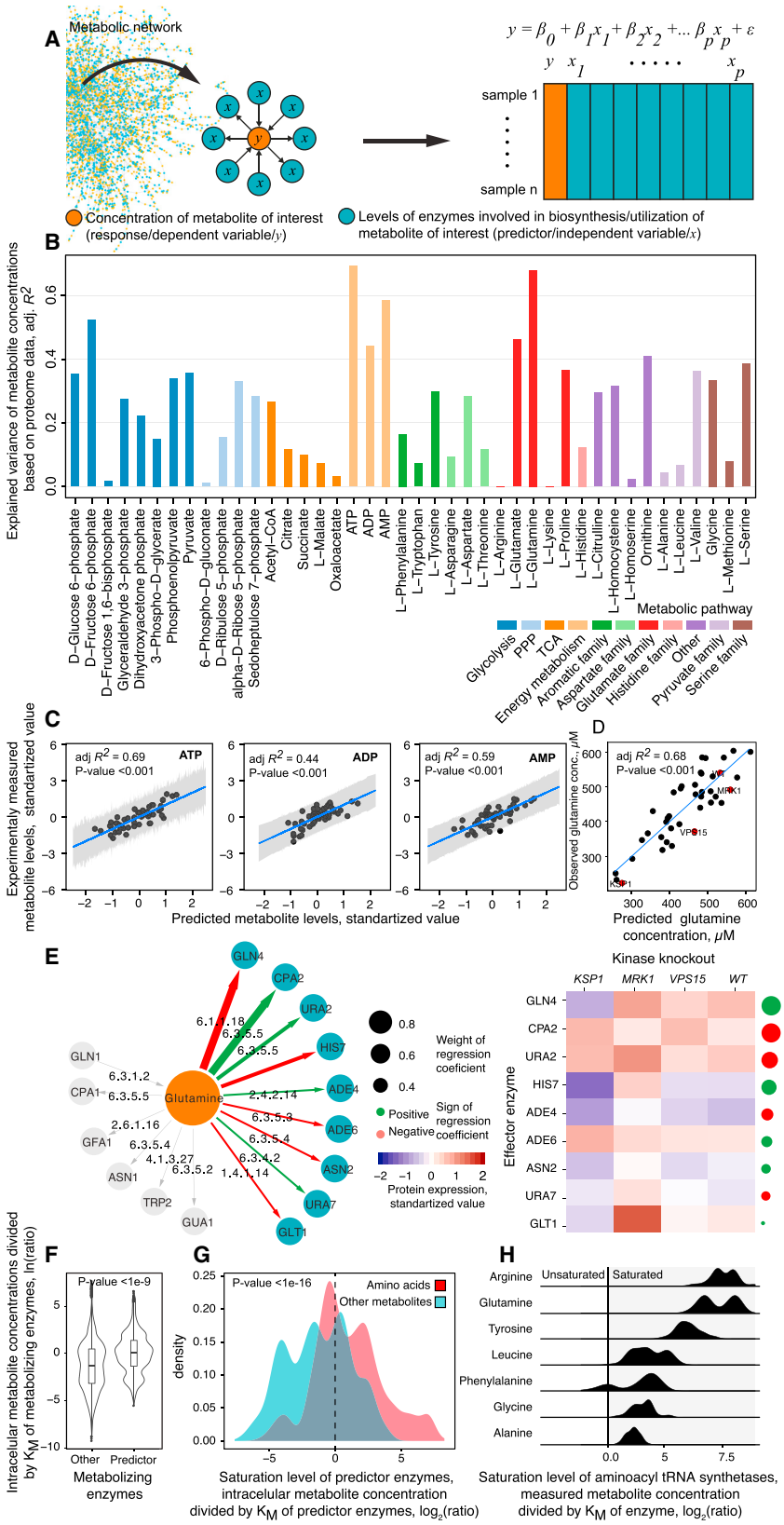


Figure 4. Multiple Linear Regression Identifies Multivariate Metabolite-Enzyme Relationships That Are Informative about Metabolite Concentration

(A) Scheme: multiple linear regression (MLR) applied over the metabolic network topology to connect enzyme levels with metabolite concentrations. Metabolite concentrations (y) are expressed as a function of expression levels (x) of the closest enzyme neighbors in the metabolic network. Informative multivariate relationships between enzyme and metabolite concentrations are identified by exhaustive feature selection by computing all possible linear models and ranking them according to minimal Akaike information criterion (STAR Methods).

(B) MLR reveals multivariate enzyme-metabolite relationships that explain metabolite concentrations in kinase knockouts. The bar plots indicate the coefficient of determination (adjusted R^2) between predicted and experimentally determined metabolite concentrations across the kinase deletion strains. See also Figure S15.

(C) The correlation of predicted and measured ATP, ADP, and AMP levels across kinase knockouts. x axis: predicted concentration from enzyme expression profiles, y axis: concentration as measured by liquid chromatography-selective reaction monitoring (LC-SRM).

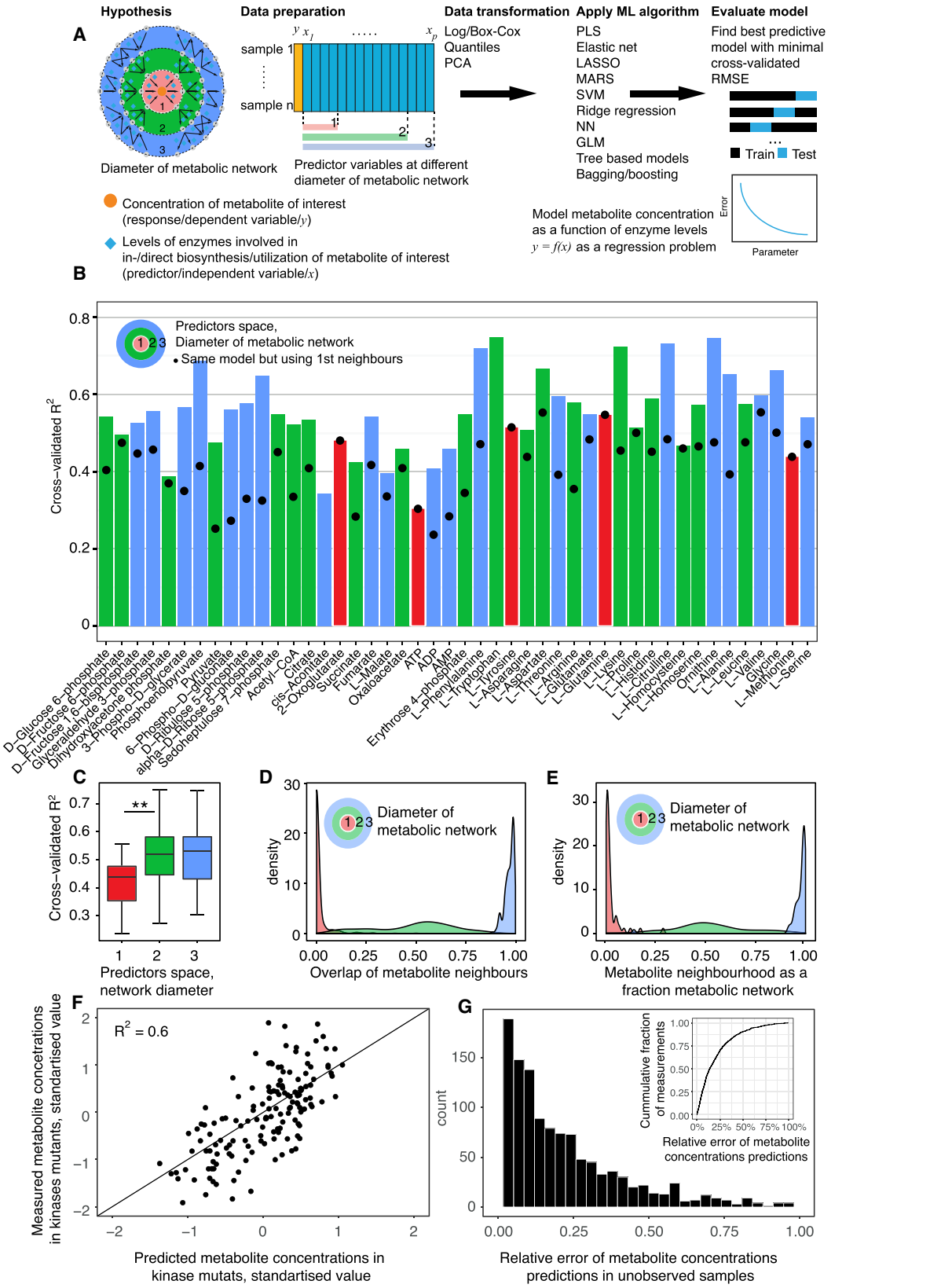
(D) The predicted and experimentally measured glutamine concentrations in kinase deletions correlate with an adjusted R^2 = 0.68. Red dots highlight examples of enzyme expression patterns from (E) for representative in quartile of glutamine concentrations.

(E) Left: graphical illustration of the 9 (out of 15) glutamine-metabolizing enzymes that are associated by the MLR approach to glutamine concentration. Right: as glutamine participates in multiple metabolic reactions, a correlation of the expression level of one glutamine-metabolizing enzyme at a time, as applied in many multi-omic studies, would fail to detect any correlation between enzyme expression and metabolism.

(F) Enzymes that influence metabolite concentrations across kinase knockouts are more likely saturated compared to other enzymes connected to the same metabolites; K_M values, as obtained from BRENDA (Chang et al., 2015), are compared to the concentration of the metabolites as measured in our study by LC-SRM. The level of saturation is expressed as a ratio between metabolite concentration and the enzyme's K_M value.

(G) Enzymes that affect amino acid concentrations are more saturated compared to other enzymes associated with the rest of the metabolites.

(H) Aminoacyl-tRNA synthetases, which are predictive of multiple amino acid concentrations, are typically saturated based on their *in vitro* kinetics.



(legend on next page)

(2) the rest of the enzymes that metabolize the same amino acids. ML had correctly identified genes whose deletion affected affect amino acid concentrations (Figure 6A, Bartlett's test, symbols * and ** correspondingly denote $p < 0.05$ and $p < 0.01$, respectively; Figure 6B, Wilcoxon rank-sum test, $p = 1.6e{-}06$). Hence, on the basis of enzyme abundance, ML is able to estimate entire metabolomes as well as identify genes important for the cell's metabolic phenotype.

Interpreting the Machine Learning Models to Draw Genotype-Phenotype Maps

We made use of the metabolic network topology to interpret the ML models and to gain insights into the biological mechanisms. Each metabolite was connected to the top 5 loading enzymes of highly predictive principal component features (90% of highest predictors weight) to reveal enzymes with the most active role (Figure 6C). The obtained graph reveals that some of the most active regulators, including *glucokinase (GLK1), phosphoglycerate mutase (PGM2), glyceraldehyde 3-phosphate dehydrogenase* (*GAPDH* gene *TDH3*), and *argininosuccinate synthetase (ARG1)*, exert distal regulation over multiple metabolite concentrations. In particular, glucokinase *GLK1* levels are associated with concentration changes in many metabolites, including aspartate, leucine, and glycine. *GLK1* was positively associated ($r = 0.38$, $p = 0.00511$) with the tricarboxylic acid (TCA) cycle metabolite oxaloacetate, indicating the coordinated regulation of glycolysis and the TCA cycle. In contrast, *GLK1* expression was negatively associated with aspartate, glycine, and threonine concentrations ($r = -0.49$; -0.46; -0.38, $p = 0.000762$; 0.00158; 0.0115, respectively). Hierarchical metabolite regulation by *GLK1* and *PGM2* expression has also been independently identified using Bayesian analysis (Bradley et al., 2009). Another example is the identification of *GAPDH*, in which abundance or activity changes have been shown to regulate the PPP to achieve yeast redox balance (Grüning et al., 2011). Indeed, a PCA of all metabolic changes detected reveals PPP metabolites to be the strongest separator; changes in the PPP are the most frequent metabolic response in kinase knockouts (Figure S14).

Finally, we assessed to what extent enzyme expression patterns are recurrent to explain changes in the metabolome. Comparing the profiles of predictive enzyme expression changes, we obtain a wide range of regulation patterns, ranging from specific to general patterns. For example, the enzyme expression landscape leading to a concentration change in dihydroxyacetone phosphate, phosphoenolpyruvate, leucine, and acetyl-CoA were substantially specific to each of the kinase knockouts (Figures 6D and S18). On the other hand, the enzyme expression landscape associated with concentration changes in amino acids such as tyrosine, methionine, and ornithine was observed in multiple knockouts (Figures 6D and S18)

DISCUSSION

Here, we address an apparent contradiction in the current literature about the regulation of metabolism. On the one hand, many investigations attribute an important role to gene expression in the regulation of metabolism. On the other hand, the prediction of metabolomes from gene expression data has so far proven challenging, and several multi-omic studies reported low correlation values of enzyme expression and metabolite levels as well as fluxes (Chubukov et al., 2013; Daran-Lapujade et al., 2007; Fendt et al., 2010; Millard et al., 2017). So how can gene expression regulation be of utmost importance for metabolism and at the same not be correlated with metabolism and not explain metabolite levels?

The reason for this discrepancy is implied by a Gedankenexperiment, in which metabolite levels are calculated upon changing enzyme abundance values in a kinetic model of glycolysis (Smallbone et al., 2013). In this hypothetical simulation, all metabolite concentration changes are caused by enzyme abundance changes. Yet, a typical correlation analysis would have yielded low scores (cophenetic correlation coefficient = 0.35) between enzyme abundance and metabolite concentration (Figure S19). In contrast, the metabolite concentrations were highly correlated with the calculated fluxes (cophenetic correlation coefficient > 0.8, Figure S20). In other words, even in this theoretical simulation in which metabolite concentration changes are fully caused by

Figure 5. Machine Learning Regression Predicts the Concentration of Metabolite Pools from Enzyme Abundance

(A) Scheme: mapping the dependency of metabolite concentrations on enzyme expression levels by incorporating the structure of the metabolic network in a genome-scale application of machine learning (ML). Different data transformation techniques and twelve ML algorithms were applied over the metabolic network topology, and the obtained models were ranked according to their ability to predict metabolite concentrations from the enzym abundance (expressed as minimal cross-validated root-mean-square error [RMSE]). In comparison to MLR (Figure 4), the inclusion of ML enabled network expansion to the 2nd and 3rd order neighbors, upon which enzyme expression changes across the full metabolic network are incorporated (E).

(B) ML enables the predictions of metabolite concentrations in the kinase knockouts on the basis of the enzyme abundances measured. Shown is the correlation of measured metabolite concentrations in relation to the predicted metabolite concentrations, expressed as 10-fold cross-validated R^2. The median cross-validated R^2 is 0.549, implying that at least half of metabolite concentration changes are explained by changes in enzyme abundance. The dots indicate the predictive power achieved with the directly metabolizing enzymes; the color indicates whether maximal predictability was reached upon including 1st, 2nd or 3rd order enzyme neighbors.

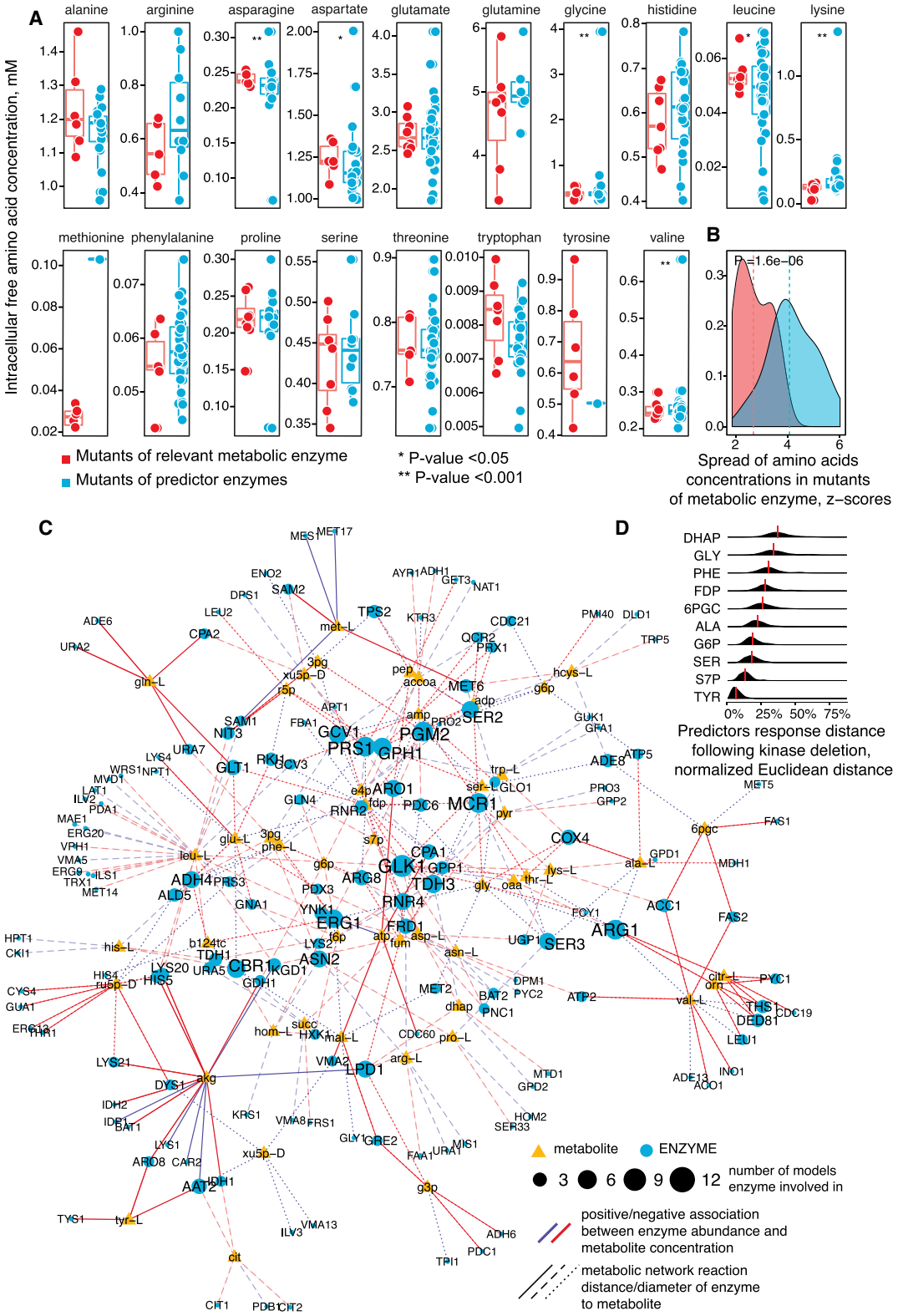
(C) For most metabolites, the predictive power is concentrated within the directly metabolizing enzymes (1st order neighbors) or is partially improved upon incorporating also the 2nd order neighbors. Ruling out overfitting, the predictions did not improve upon further expansion of the predictor variable space to the full metabolic network. ** = Wilcoxon rank sum test p value < 0.01.

(D) The commonality of enzyme predictors for the different metabolites, accounting for network diameter, reveals a spectrum of enzyme expression signatures that can regulate metabolite abundance.

(E) The total fraction of enzymes associated with metabolite concentrations accounting for network distance.

(F) Metabolic phenotype (all metabolites per mutant) predictions by ML in unobserved kinase knockout strains on the basis of their quantitative proteome. The phenotype prediction is based on individual metabolite models; the top 30 predicted kinase metabolomes are shown.

(G) Distribution of relative errors (in %) in the prediction compared to experimental measurements of metabolite concentrations in all kinases knockout strains; ML predicts metabolite concentrations accurately.



(legend on next page)

enzyme abundance changes, consistent with previous reports (Hackett et al., 2016; Millard et al., 2017), the one-to-one mapping would report low correlation values between enzyme abundances and metabolite concentrations (Figures S21 and S22). Taken the other way, a low one-to-one correlation between enzyme and metabolite abundance is not due to metabolite concentrations that would behave independently to changes in enzyme expression; the low values are the consequence of the more complex, multifactorial relationships that describe the interdependence of enzyme abundance and metabolite levels.

Our results show that metabolic gene expression regulation is achieved through many enzyme expression changes acting in concert. Once these multifactorial relationships are identified—in our study through the use of multivariate statistical learning—enzyme expression landscapes become predictive about the cellular metabolome, even at the network scale, that is currently not to be covered by mechanistic models as they are available, i.e., for individual metabolic pathways, such as glycolysis. To our knowledge, this is the first successful attempt to predict a complex, quantitative metabolic phenotype from enzyme expression without taking into account predetermined enzyme reaction mechanisms, phosphorylation states, or kinetics. Applied over the topological organization of the metabolic network, the predictive models are further rendered interpretable, which, as we have shown, enables to draw genotype-phenotype maps. Taken together, these results demonstrate that enzyme expression landscapes are regulated to control metabolite concentrations and as a consequence, fluxes. On average, the metabolome predictions achieved on the basis of enzyme levels correlated with experimental values with a cross-validated R^2 of 0.55. This suggests that more than half of metabolite concentration regulation, at least as observed in kinase knockouts, is attributable to changes in enzyme abundance.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Strains and Culture
- METHOD DETAILS
  - Metabolomics
  - Proteomics
  - Enzyme Expression Analysis
  - Flux Coupling Analysis
  - Metabolic Control Analysis
  - Statistical Modelling
  - Metabolite Concentration Regression Modelling
  - Enzyme Saturation
  - Note on the Relationship between Enzyme Expression Changes in the High and Low Abundant Fraction of the Proteome
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes twenty-two figures and can be found with this article online at https://doi.org/10.1016/j.cels.2018.08.001.

ACKNOWLEDGMENTS

We thank Sergej Andrejev (European Molecular Biology Laboratory, Heidelberg) for providing binaries of FCA implementation. We thank Peter Thorpe for providing yeast kinase mutants *MATα* strains. This work was supported by the Francis Crick Institute, which receives its core funding from the Cancer Research UK (FC001134), the UK Medical Research Council (FC001134), and the Wellcome Trust (FC001134) and received specific funding from the Wellcome Trust (RG 200829/Z/16/Z) and the ERC (starting grant 260809). A.Z. was funded by the EMBO long-term fellowship (ALTF-969 2014), which was co-funded by the European Commission (LTFCOFUND2013, GA-2013-609409) support from Marie Curie Actions.

AUTHOR CONTRIBUTIONS

Conceptualization, A.Z. and M.R.; Methodology, A.Z.; Investigation, A.Z., J.V., F.C., M.M., C.B.M., M.K., N.P., and S.K.; Formal Analysis, A.Z., J.V., S.K., C.B.M., V.D., M.M., and M.K.; Writing—Original Draft, A.Z. and M.R.; Writing—Review & Editing, A.Z. and M.R.; Project Administration, A.Z. and M.R.; Visualization, A.Z.; Data Curation, A.Z.; Funding Acquisition, M.R.; Resources, M.R.; Supervision, A.Z., M.R., and B.K.

Figure 6. Machine Learning Trained over the Metabolic Network Topology Reveals Genes and Metabolites Important for Metabolite Concentration Regulation

(A) Enzymes whose abundance predicts metabolite concentrations in kinase knockouts cause metabolite concentration changes when deleted in a completely independent dataset (Mülleder et al., 2016a).

(B) Summary of (A): the overall range of metabolite concentration changes is broader upon the deletion of enzymes associated with concentration changes, as it is upon the deletion of all other enzymes that convert the same metabolites.

(C) Enzyme metabolite graph depicting hub proteins in the prediction of the yeast cell metabolome. Nodes represent metabolites (triangles) that are predictable using relevant enzyme abundances (circles). Edges represent positive and negative association represented by Pearson's correlation between metabolite and enzymes levels. For visualization purposes, we retained only the most important enzymes (normalized weight of variable >90%, with up to 5 enzymes with highest absolute loading per component).

(D) The concentration of several hub metabolites is affected by a spectrum of enzyme expression signatures, while for some metabolites only specific expression signatures were observed. More distant values (upper density plots) illustrate situations where a (kinase-deletion) unique combination of enzyme expression changes affects a particular metabolite. Contrarily, lower distances illustrate cases where multiple kinase deletions affect a metabolite via the same set of enzyme expression changes. The GAPDH substrate DHAP was the metabolite controlled by the highest number of divergent mechanisms, while tyrosine was the most uniformly regulated metabolite (for illustration purposes, only every 5th metabolite is depicted; the full figure is provided in Figure S18). To compare predictor responses between metabolites, the levels of associated enzymes were standardized (to zero mean and unit variance). The Euclidean distance of standardized enzyme expression was computed pairwise between each kinase mutant and normalized to 100% by the most distant kinase pair. Red vertical lines denote the median value for each enzyme. Abbreviations: amino acids are given in three letter IUPAC code; DHAP, dihydroxyacetone phosphate; FDP, Fructose 1,6 bisphosphate; 6PGC, 6-phosphogluconate; G6P, glucose 6-phosphate; S7P, sedoheptulose 7-P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 20, 2017
Revised: May 29, 2018
Accepted: July 31, 2018
Published: September 5, 2018

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Automat. Contr. *19*, 716–723.

Alam, M.T., Zelezniak, A., Mülleder, M., Shliaha, P., Schwarz, R., Capuano, F., Vowinckel, J., Radmanesfahar, E., Krüger, A., Calvani, E., et al. (2016). The metabolic background is a global player in Saccharomyces gene expression epistasis. Nat. Microbiol. *1*, 15030.

Alam, M.T., Olin-Sandoval, V., Stincone, A., Keller, M.A., Zelezniak, A., Luisi, B.F., and Ralser, M. (2017). The self-inhibitory nature of metabolic networks and its alleviation through compartmentalization. Nat. Commun. *8*, 16018.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B. Stat. Methodol. *57*, 289–300.

Beyenbach, K.W., and Wieczorek, H. (2006). The V-type H+ ATPase: molecular structure and function, physiological roles and regulation. J. Exp. Biol. *209*, 577–589.

Bodenmiller, B., Wanka, S., Kraft, C., Urban, J., Campbell, D., Pedrioli, P.G., Gerrits, B., Picotti, P., Lam, H., Vitek, O., et al. (2010). Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. Sci. Signal. *3*, rs4.

Braakman, R., and Smith, E. (2013). The compositional and evolutionary logic of metabolism. Phys. Biol. *10*, 011001.

Bradley, P.H., Brauer, M.J., Rabinowitz, J.D., and Troyanskaya, O.G. (2009). Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae. PLoS Comput. Biol. *5*, e1000270.

Breitkreutz, A., Choi, H., Sharom, J.R., Boucher, L., Neduva, V., Larsen, B., Lin, Z.Y., Breitkreutz, B.J., Stark, C., Liu, G., et al. (2010). A global protein kinase and phosphatase interaction network in yeast. Science *328*, 1043–1046.

Buescher, J.M., Moco, S., Sauer, U., and Zamboni, N. (2010). Ultrahigh performance liquid chromatography-tandem mass spectrometry method for fast and robust quantification of anionic and aromatic metabolites. Anal. Chem. *82*, 4403–4412.

Burgard, A.P., Nikolaev, E.V., Schilling, C.H., and Maranas, C.D. (2004). Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res. *14*, 301–312.

Chang, A., Schomburg, I., Placzek, S., Jeske, L., Ulbrich, M., Xiao, M., Sensen, C.W., and Schomburg, D. (2015). BRENDA in 2015: exciting developments in its 25th year of existence. Nucleic Acids Res. *43*, D439–D446.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set. J. Stat. Softw. *61*, 1–36.

Chechik, G., Oh, E., Rando, O., Weissman, J., Regev, A., and Koller, D. (2008). Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. Nat. Biotechnol. *26*, 1251–1259.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., et al. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. *40*, D700–D705.

Christiano, R., Nagaraj, N., Fröhlich, F., and Walther, T.C. (2014). Global proteome turnover analyses of the Yeasts S. cerevisiae and S. pombe. Cell Rep. *9*, 1959–1965.

Chubukov, V., Uhr, M., Le Chat, L., Kleijn, R.J., Jules, M., Link, H., Aymerich, S., Stelling, J., and Sauer, U. (2013). Transcriptional regulation is insufficient to explain substrate-induced flux changes in Bacillus subtilis. Mol. Syst. Biol. *9*, 709.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. Inter. J. Complex. Syst. 1695. https://pdfs.semanticscholar.org/1d27/44b83519657f5f2610698a8ddd177ced4f5c.pdf.

Daran-Lapujade, P., Rossell, S., van Gulik, W.M., Luttik, M.A.H., de Groot, M.J.L., Slijper, M., Heck, A.J.R., Daran, J.M., de Winde, J.H., Westerhoff, H.V., et al. (2007). The fluxes through glycolytic enzymes in Saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels. Proc. Natl. Acad. Sci. USA *104*, 15753–15758.

Demichev, V., Messner, C.B., Lilley, K.S., and Ralser, M. (2018). DIA-NN: deep neural networks substantially improve the identification performance of data-independent acquisition (DIA) in proteomics. bioRXiv. https://doi.org/10.1101/282699.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2016). The Reactome pathway KnowledgeBase. Nucleic Acids Res. *44*, D481–D487.

Faraway, J.J. (2016). Linear Models with R, Second Edition (CRC Press).

Fendt, S.M., Buescher, J.M., Rudroff, F., Picotti, P., Zamboni, N., and Sauer, U. (2010). Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. Mol. Syst. Biol. *6*, 356.

Fox, J., and Weisberg, S. (2011). An R Companion to Applied Regression (SAGE).

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell. *11*, 4241–4257.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell. Proteomics *11*, O111.016717.

Gonçalves, E., Raguz Nakic, Z., Zampieri, M., Wagih, O., Ochoa, D., Sauer, U., Beltrao, P., and Saez-Rodriguez, J. (2017). Systematic analysis of transcriptional and post-transcriptional regulation of metabolism in yeast. PLoS Comput. Biol. *13*, e1005297.

González, A., and Hall, M.N. (2017). Nutrient sensing and TOR signaling in yeast and mammals. EMBO J. *36*, 397–408.

Grüning, N.M., Rinnerthaler, M., Bluemlein, K., Mülleder, M., Wamelink, M.M.C., Lehrach, H., Jakobs, C., Breitenbach, M., and Ralser, M. (2011). Pyruvate kinase triggers a metabolic feedback loop that controls redox metabolism in respiring cells. Cell Metab. *14*, 415–427.

Hackett, S.R., Zanotelli, V.R.T., Xu, W., Goya, J., Park, J.O., Perlman, D.H., Gibney, P.A., Botstein, D., Storey, J.D., and Rabinowitz, J.D. (2016). Systems-level analysis of mechanisms regulating yeast metabolic flux. Science *354*, aaf2786.

Herrgård, M.J., Swainston, N., Dobson, P., Dunn, W.B., Arga, K.Y., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., et al. (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat. Biotechnol. *26*, 1155–1160.

Honaker, J., King, G., and Blackwell, M. (2011). Amelia II A Program for Missing Data. J. Stat. Softw. *45*, 1–47.

Kacser, H., and Burns, J.A. (1973). The control of flux. Symp. Soc. Exp. Biol. *27*, 65–104.

Kahm, M., Hasenbrink, G., Lichtenberg-Fraté, H., Ludwig, J., and Kschischo, M. (2010). Grofit: fitting biological growth curves with R. J. Stat. Softw. *33*, 1–21.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44*, D457–D462.

Keller, M.A., Turchyn, A.V., and Ralser, M. (2014). Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. Mol. Syst. Biol. *10*, 725.

Keller, M.A., Piedrafita, G., and Ralser, M. (2015). The widespread role of non-enzymatic reactions in cellular metabolism. Curr. Opin. Biotechnol. *34*, 153–161.

Kresnowati, M.T., van Winden, W.A., Almering, M.J.H., ten Pierick, A., Ras, C., Knijnenburg, T.A., Daran-Lapujade, P., Pronk, J.T., Heijnen, J.J., and Daran, J.M. (2006). When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. Mol. Syst. Biol. *2*, 49.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 1–26.

Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat. Methods *11*, 319–324.

Lahtvee, P.J., Sánchez, B.J., Smialowska, A., Kasvandik, S., Elsemman, I.E., Gatto, F., and Nielsen, J. (2017). Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. Cell Syst. *4*, 495–504.e5.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

Lumley, T., and Miller, A. (2004). Leaps: Regression Subset Selection, R Package Version 2 (R Foundation for Statistical Computing).

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and Bähler, J. (2012). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell *151*, 671–683.

Mihaylova, M.M., and Shaw, R.J. (2011). The AMPK signalling pathway coordinates cell growth, autophagy and metabolism. Nat. Cell Biol. *13*, 1016–1023.

Millard, P., Smallbone, K., and Mendes, P. (2017). Metabolic regulation is sufficient for global and robust coordination of glucose uptake, catabolism, energy production and growth in Escherichia coli. PLoS Comput. Biol. *13*, e1005396.

Mülleder, M., Capuano, F., Pir, P., Christen, S., Sauer, U., Oliver, S.G., and Ralser, M. (2012). A prototrophic deletion mutant collection for yeast metabolomics and systems biology. Nat. Biotechnol. *30*, 1176–1178.

Mülleder, M., Calvani, E., Alam, M.T., Wang, R.K., Eckerstorfer, F., Zelezniak, A., and Ralser, M. (2016a). Functional metabolomics describes the yeast biosynthetic Regulome. Cell *167*, 553–565.e12.

Mülleder, M., Campbell, K., Matsarskaia, O., Eckerstorfer, F., and Ralser, M. (2016b). Saccharomyces cerevisiae single-copy plasmids for auxotrophy compensation, multiple marker selection, and for designing metabolically cooperating communities. F1000Res *5*, 2351.

Murray, D.B., Beckmann, M., and Kitano, H. (2007). Regulation of yeast oscillatory dynamics. Proc. Natl. Acad. Sci. USA *104*, 2241–2246.

Murray, L.E., Rowley, N., Dawes, I.W., Johnston, G.C., and Singer, R.A. (1998). A yeast glutamine tRNA signals nitrogen status for regulation of dimorphic growth and sporulation. Proc. Natl. Acad. Sci. USA *95*, 8619–8624.

Nilsson, A., Nielsen, J., and Palsson, B.O. (2017). Metabolic models of protein allocation call for the Kinetome. Cell Syst. *5*, 538–541.

Nygaard, V., Rødland, E.A., and Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics *17*, 29–39.

Oliveira, A.P., Ludwig, C., Picotti, P., Kogadeeva, M., Aebersold, R., and Sauer, U. (2012). Regulation of yeast central metabolism by enzyme phosphorylation. Mol. Syst. Biol. *8*, 623.

Patil, K.R., and Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc. Natl. Acad. Sci. USA *102*, 2685–2689.

Petrezselyova, S., Zahradka, J., and Sychrova, H. (2010). Saccharomyces cerevisiae BY4741 and W303-1A laboratory strains differ in salt tolerance. Fungal. Biol. *114*, 144–150.

R Core Team. (2015). R: A language and environment for statistical computing (R Foundation for Statistical Computing).

Rodríguez, A., De La Cera, T., Herrero, P., and Moreno, F. (2001). The hexokinase 2 protein regulates the expression of the GLK1, HXK1 and HXK2 genes of Saccharomyces cerevisiae. Biochem. J. *355*, 625–631.

Saito, H., and Tatebayashi, K. (2004). Regulation of the osmoregulatory HOG MAPK cascade in yeast. J. Biochem. *136*, 267–272.

Sakia, R.M. (1992). The Box-Cox transformation technique: a review. Statistician *41*, 169–178.

Schomburg, I., Hofmann, O., Baensch, C., Chang, A., and Schomburg, D. (2000). Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. Gene Funct. Dis. *1*, 109–118.

Schubert, O.T., Gillet, L.C., Collins, B.C., Navarro, P., Rosenberger, G., Wolski, W.E., Lam, H., Amodei, D., Mallick, P., MacLean, B., et al. (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. Nat. Protoc. *10*, 426–441.

Schulz, J.C., Zampieri, M., Wanka, S., von Mering, C., and Sauer, U. (2014). Large-scale functional analysis of the roles of phosphorylation in yeast metabolic pathways. Sci. Signal. *7*, rs6.

Sharifpoor, S., van Dyk, D., Costanzo, M., Baryshnikova, A., Friesen, H., Douglas, A.C., Youn, J.Y., VanderSluis, B., Myers, C.L., Papp, B., et al. (2012). Functional wiring of the yeast kinome revealed by global analysis of genetic network motifs. Genome Res. *22*, 791–801.

Smallbone, K., Messiha, H.L., Carroll, K.M., Winder, C.L., Malys, N., Dunn, W.B., Murabito, E., Swainston, N., Dada, J.O., Khan, F., et al. (2013). A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes. FEBS Lett. *587*, 2832–2841.

Smyth, G.K. (2005). Limma: linear Models for microarray Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V.J. Carey, W. Huber, R.A. Irizarry, and S. Dudoit, eds. (Springer), pp. 397–420.

Somogyi, E.T., Bouteiller, J.M., Glazier, J.A., König, M., Medley, J.K., Swat, M.H., and Sauro, H.M. (2015). libRoadRunner: a high performance SBML simulation and analysis library. Bioinformatics *31*, 3315–3321.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452.

Tu, B.P., Kudlicki, A., Rowicka, M., and McKnight, S.L. (2005). Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. Science *310*, 1152–1158.

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L., and Fernie, A.R. (2003). Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. EMBO Rep. *4*, 989–993.

Van Hoek, P., Van Dijken, J.P., and Pronk, J.T. (1998). Effect of specific growth rate on fermentative capacity of Baker's yeast. Appl. Environ. Microbiol. *64*, 4226–4233.

Vizcaíno, J.A., Csordas, A., Del-Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., et al. (2016). 2016 Update of the PRIDE database and its related tools. Nucleic Acids Res. *44*, 11033.

Vowinckel, J., Capuano, F., Campbell, K., Deery, M.J., Lilley, K.S., and Ralser, M. (2013). The beauty of being (label)-free: sample preparation methods for SWATH-MS and next-generation targeted proteomics. F1000Res *2*, 272.

Vowinckel, J., Zelezniak, A., Bruderer, R., Mülleder, M., Reiter, L., and Ralser, M. (2018). Cost-effective generation of precise label-free quantitative proteomes in high-throughput by microLC and data-independent acquisition. Sci. Rep *8*, 4346.

van Wageningen, S., Kemmeren, P., Lijnzaad, P., Margaritis, T., Benschop, J.J., de Castro, I.J., van Leenen, D., Groot Koerkamp, M.J., Ko, C.W., Miles, A.J., et al. (2010). Functional overlap and regulatory links shape genetic interactions between signaling pathways. Cell *143*, 991–1004.

Wegrzyn, G., and Wegrzyn, A. (2008). Is tRNA only a translation factor or also a regulator of other processes? J. Appl. Genet. *49*, 115–122.

Whitney, M.L., Hurto, R.L., Shaheen, H.H., and Hopper, A.K. (2007). Rapid and reversible nuclear accumulation of cytoplasmic tRNA in response to nutrient availability. Mol. Biol. Cell. *18*, 2678–2686.

Wickham, H., and Grolemund, G. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data ("O'Reilly Media, Inc.").

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science *285*, 901–906.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., and Bork, P. (2011). iPath2.0: Interactive pathway explorer. Nucleic Acids Res. *39*, W412–W415.

Zaborske, J.M., Wu, X., Wek, R.C., and Pan, T. (2010). Selective control of amino acid metabolism by the GCN2 eIF2 kinase pathway in Saccharomyces cerevisiae. BMC Biochem. *11*, 29.

Zeileis, A., and Hothorn, T. (2002) Diagnostic checking in regression relationships. ftp://ftp.auckland.ac.nz/pub/software/CRAN/doc/vignettes/lmtest/lmtest-intro.pdf.

Zelezniak, A., Pers, T.H., Soares, S., Patti, M.E., and Patil, K.R. (2010). Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes. PLoS Comput. Biol. *6*, e1000729.

Zelezniak, A., Sheridan, S., and Patil, K.R. (2014). Contribution of network connectivity in determining the relationship between gene expression and metabolite concentration changes. PLoS Comput. Biol. *10*, e1003572.

Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse representations IEEE Trans. Inform. Theory *57*, 4689–4708.

Zomorrodi, A.R., and Maranas, C.D. (2010). Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data. BMC Syst. Biol. *4*, 178.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Acetonitrile UPLC grade	Greyhound BIOSOLVE	Cat# Bio-012041
Ultra-pure water, ULC-MS grade	Greyhound BIOSOLVE	Cat# 23214125
Methanol Absolute ULC-MS grade	Greyhound BIOSOLVE	Cat# BIO-13684102
Ammonium formate	Fluka	Cat# 14266
Formic acid	Fluka	Cat# O6454
L-Amino acids analytical standard	Sigma-Aldrich	Cat# LAA21
Standards for glycolysis, TCA and PPP intermediates	Sigma-Aldrich	Cat# G8270, G7879, F3627, F6803, G5251, P8877, 79470, 79470, P7127, P2256, P7877, 83899, R7750, 15732, 78832, W302600, A3412, 75892, S3674, 47910, 27606, O4126
Standards for cofactors	Sigma-Aldrich	Cat# N8229, N7004, N5130, 93220, A2754, A2383, 01930, A2056
Critical Commercial Assays		
Retention time peptides Biognosys iRT kit	https://biognosys.com/	Ki-3002-1
Deposited Data		
Raw proteome data	This study	PRIDE: PXD010529
Processed proteome and metabolome data	This study	[https://doi.org/10.5281/zenodo.1320288]
S. cerevisiae genome-scale metabolic reconstruction	Herrgård et al. (2008)	N/A
S. cerevisiae kinetic glycolysis model	Smallbone et al. (2013)	BioModels: MODEL1303260018
Km values from BRENDA database	Chang et al. (2015); Schomburg et al. (2000)	https://www.brenda-enzymes.org/
Curated genetic information, gene ontology, literature and phenotype annotation from the Saccharomyces Genome Database (SGD)	Cherry et al. (2012)	http://www.yeastgenome.org/
Yeast kinase signaling pathway annotations from Reactome	Fabregat et al. (2016)	https://reactome.org
Yeast kinase signaling pathway annotations from KEGG	Kanehisa et al. (2016)	https://www.kegg.jp/kegg/rest/keggapi.html
Amino acid concentrations in metabolic enzyme knockouts	Müelleder et al. (2016a)	https://doi.org/10.17632/bnzdhd6ck8.1
Yeast protein-protein interaction network	Szklarczyk et al. (2015)	https://string-db.org/
Yeast protein degradation rates	Christiano et al. (2014)	https://ars.els-cdn.com/content/image/1-s2.0-S2211124714009346-mmc3.xlsx
Experimental Models: Organisms/Strains		
Prototrophic *Saccharomyces cerevisiae* kinase deletion collection (*MAT*a, prototrophy restored episomally)	Winzeler et al. (1999); Müelleder et al. (2012)	http://www.euroscarf.de/
Selected *Saccharomyces cerevisiae* kinase deletion strains (*MAT*α, prototrophy restored episomally)	Winzeler et al. (1999)	http://www.euroscarf.de/
Recombinant DNA		
Plasmid: pHLUM	Müelleder et al. (2016a)	In addgene.org: #40276
Plasmid: pHLU	Müelleder et al. (2016a)	In addgene.org: #64181
Plasmid: pHLUK	Müelleder et al. (2016a)	In addgene.org: #64167

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Scripts to reproduce main figures	This study	https://github.com/zelezniak-lab/kinase_metabolism
Proteomics data analysis Spectronaut (8.0.9600)	Biognosys	Sw-3001
Proteomics data analysis via Deep Neural Networks, DIA-NN	Demichev et al. (2018)	https://github.com/vdemichev/DiaNN
caret R package (6.0-78) for regression modeling	Kuhn (2008)	http://topepo.github.io/caret/index.html
LP solver IBM ILOG CPLEX Optimization Studio 12.7.1 for flux coupling analysis	IBM	CJ1HQML
libRoadRunner (1.4.8) for metabolic control analysis	Somogyi et al. (2015)	https://github.com/sys-bio/roadrunner
MassHunter software suite for metabolite analysis	Agilent Technologies	N/A
grofit R package (1.0) for growth curves analysis	Kahm et al. (2010)	http://CRAN.R-project.org/package=grofit
sva R package (3.26.0) for batch correction of data	Leek et al. (2012)	https://doi.org/10.18129/B9.bioc.sva

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Markus Ralser (Markus.Ralser@crick.ac.uk)

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Strains and Culture

Yeast strains used in this study were obtained from our published prototrophic gene deletion collection (Mülleder et al., 2012). Kinases were identified following the strategy of Sharifpoor et al. (2012), expanded by the annotation in the yeast kinome and yeast genome database (Breitkreutz et al., 2010; Cherry et al., 2012) and including genes associated to Gene Ontology term 0004672 (protein kinase activity). 97 of the strains grew in triplicates (n=3) in minimal medium without a substantial growth defect (Figure S1), were pre-cultured overnight in 10 ml minimal medium, at 30°C, and diluted them to an OD_{600} of 0.2 in 30 ml main culture.

Growth was monitored, and the strains were sampled by cold methanol quenching at an OD_{600} 1.5 +/- 0.1, before the cultures enter the diauxic shift, for metabolic and proteomic analysis. The growth curves were fitted using non-parametric (without growth law assumption) spline model as implemented in R growFit package (Kahm et al., 2010). Exponential growth rate was estimated as maximal slope of the growth curve (Figure S1).

To generate heterozygous and homozygous diploid strains of the kinase mutants we inoculated the respective *MATa* and *MATα* strains in 150 μl YPGlucose (2%) medium and incubated them overnight before 2 consecutive selection steps on synthetic complete medium (SC) lacking lysine and methionine. For heterozygous strains a *HIS3* deletion strain of the opposite mating type was selected. Prototrophy was restored in the haploid parents and the diploid progeny by transformation with a single copy plasmids containing the required genes (pHLUK, pHLUMv2 and pHLU (Mülleder et al., 2016b). All 5 versions of 10 randomly chosen kinase mutants (the two parental haploid MatA and Matα, 2 heterozygous and 1 homozygous diploid) strains were grown on synthetic minimal (SM) and collected in exponential phase.

METHOD DETAILS

Metabolomics

Free intracellular metabolite pools were quantified by liquid chromatography - selective reaction monitoring (LC-SRM) by protocols described previously. The method used to obtain Dataset 1 (Figure S13) is described in Keller et al. (2014) for the quantification of glycolytic and pentose phosphate pathway metabolites and was expanded with additional transitions for ATP, ADP and AMP. Analytes were separated by gradient elution using 10% and 50% acetonitrile, containing 750 mg l^{-1} octylammonium acetate as solvents A and B at a flow rate of 0.6ml/min and column temperature of 20°C. The gradient program was as follows: 5% B for 3.5 min, then ramping to 70% B within 2.5 min, followed by washing with 80% B for 0.5 min and re-equilibration at 5% B for 0.5 min, resulting in a total cycle time of 7.5 min on a Zorbax SB-C8 Rapid Resolution HD, 2.1x100mm, 1.8 Micron (Agilent) column.

For Dataset 2 and 3 we adapted chromatographic parameters from Buescher et al. (2010) and added a SRM set that has previously been established by individually optimising ion optics and fragmentation settings using commercially available standards on an 6460 or 6470 Triple Quadrupole Mass Spectrometer (Agilent) coupled to UPLC (1290 Infinity, Agilent).

In Dataset 2 analytes were separated by gradient elution using 10 mM TBA 15 mM acetic acid in water and 5% methanol as solvents A and B at a flow rate of 0.5 ml/min and column temperature of 30°C. The gradient program was as follows: 0% B for 4.5 min, increase to 20% B (5 min), 70% B (9.5 min), 90% B (10 min), kept constant until 12 min and returned to initial conditions at 12.5 min followed by 1.5 min equilibration, resulting in a total cycle time of 14 min on a Zorbax Eclipse Plus C18 2.1x50 mm, 1.8 μm column (Agilent).

In Dataset 3 (Figure S13), free amino acids were separated by hydrophilic interaction liquid chromatography (HILIC) using an ACQUITY UPLC BEH amide column (130Å, 1.7 μm, 2.1 mm X 100 mm) by gradient elution at a constant flow rate of 0.9 ml/min and a column temperature of 25°C. Eluents A and B were prepared at 10 mM ammonium formate, 0.176% formic acid and in 95/5/5 acetonitrile/MeOH/water and in 50/50 acetonitrile/water respectively, all of UPLC grade. Chromatographic conditions for the gradient elution were following: solvent A was kept for 0.7 min at 85% before a steady decrease to 5% A until 2.55 min. A was kept at 5% for 0.05 min before returning to the initial conditions of 85% A within 0.05 min. This was followed by and an equilibration step until 3.25 min before injection of the next sample. All metabolites were identified by matching retention time and fragmentation pattern with the commercially obtained standards and were quantified by external calibration (except Dataset 1) with standards prepared at serial dilution from 500 μM to 100 nM.

Dataset 1 was created from the same cells as grown for the proteomic experiments. Metabolomics datasets 2 and 3 were obtained by re-growing 3 independent cultures from strains with highly variable metabolite concentrations based on dataset 1 and a previous genome-scale metabolism study (Mülleder et al., 2016a). Mass spectrometry signals for all metabolites were acquired in dynamic SRM mode in Masshunter software. All preprocessed metabolomics data (integrated SRM transition peaks after external calibration (where applicable)) were corrected for batch effects using *ComBat* approach as implemented in *sva* (Leek et al., 2012) R package. For visualization purposes (Figure S14) missing metabolite concentrations were imputed using *amelia* approach (Honaker et al., 2011).

Proteomics

The proteomic method has been published in parallel (Vowinckel et al., 2018). In brief, tryptic digests for the analysis by SWATH-MS were prepared by the RapiGest method as described previously (Vowinckel et al., 2013), and analysed on on a Tandem Quadrupole Time-of-Flight mass spectrometer (SCIEX TripleTOF5600) coupled to DuoSpray Ion Source (SCIEX) and Eksigent 425 HPLC system running in microflow mode. Before the injection into the mass spectrometer, total protein concentrations were adjusted by dilution. SWATH assay libraries were built following Schubert et al. (2015) by pre-fractionation of the tryptic digest. Unless otherwise indicated, SWATH data quantification was performed in Spectronaut (Biognosys, v. 8.0.9600). Post-processing was conducted in R (R Core Team, 2015) by first removing precursors from all samples where the median peak group Q-value was > 0.01 obtained from mProphet algorithm as implemented in Spectronaut. For label free quantification, we considered only the precursors originating from uniquely mapping peptides. Next, we chose peptides by correlation quantity following our approach as developed previously (Alam et al., 2016; Vowinckel et al., 2018). This strategy assumes that the best quantitation-informative peptides, as they are derived from the same protein, correlate in their abundance. Pearson's correlation coefficients were calculated between each pair of peptides (summed precursor's MS2 peak areas) belonging to the same protein across all samples. Peptides displaying overall low correlation (<0.3) were removed from subsequent analysis. This selection therefore excludes non-specific peptides or precursors which are not linearly responsive for other reasons, e.g. due to post-translational modifications. Furthermore, to account for confounding effects related to acquisition dates, we performed batch correction using the *sva* approach (Leek et al., 2012). Supervised surrogate variable analysis (with 1 variable) was applied without specifying experimental factors (Nygaard et al., 2016) using 50% of least varying peptides as controls. Estimated surrogate effects were regressed out from the peptide signal. Finally, for each protein, the signals of all peptide groups were geometrically averaged. External standard quality control (QC) samples were prepared as a mixture of all proteomes and were measured every 8-12 injections. After applying batch correction, QC samples are clustered together around 0 on a scaled PCA plot, showing that the batch correction strategy has removed most of the confounding effects Figure S2.

Diploid strains were analysed with a slightly modified proteomics workflow (Demichev et al., 2018). Briefly, proteins were extracted in 6M urea/ 0.1M ammonium bicarbonate using a bead beater (Spex Geno/Grinder). After reduction and alkylation with dithiothreitol (5mM) and iodoacetamide (10mM), respectively, proteins were digested overnight with trypsin. The resulting peptides were cleaned-up using 96-well MACROSpin plates (Nest Group). Samples were measured on a Waters nanoAcquity coupled to a SCIEX TripleTOF 6600. The peptides were separated with a 20min gradient on a Waters HSS T3 column (300um x 150mm, 1.8um) using a flow rate of 5ul/min. SWATH MS/MS acquisition scheme with 40 variable size windows and 35ms accumulation time was used. Raw data were processed with DIA-NN (version 1.2) using the default settings and mass accuracy set to 20 ppm and 12 ppm at the MS2 and MS1 level, respectively.

Enzyme Expression Analysis

After correction for batch effects, differential protein expression analysis was performed using *limma* (Smyth, 2005). Geometrically averaged fold-changes as outputted by *limma* were omitted, instead fold-change ratio of mean signals between mutant and parental strain were used throughout the manuscript. The Benjamini-Hochberg (BH) false discovery rate (FDR) control procedure (Benjamini and Hochberg, 1995) was applied after performing all comparisons using *p.adjust* as implemented in R-core *stats*

package. Additionally, we used a cut-off of 40% change noted as $\log_2$(fold-change) of ±0.485 which we refer over the manuscript as $\log_2$(1.4/0.714) up-/down regulated empirically determined cut-off to account for any potentially unaccounted batch-to-batch variation left to further eliminate any potential false discoveries. For this, we calculated protein expression fold-changes for each protein in the QC sample (a mixture of all samples) and identified a tiny fraction of proteins that was significantly (adj P-value < 0.01) differentially expressed between different batches. The fold-change exceeding the median obtained from the distribution of these extreme cases was used as the cut-off throughout the analysis.

To estimate enzyme copy numbers, we compared two datasets recorded by fluorescence microscopy (Ghaemmaghami et al., 2003) and one by mass spectrometry (Kulak et al., 2014), showing overall agreement of enzyme copy number fraction of the total yeast proteome (~35% and ~36% respectively). To calibrate the relative abundance changes (Figure 1E) for each mutant, every enzyme's copy number (Kulak et al., 2014) was multiplied by the fold-change if it was differentially expressed (BH adjusted p-value <0.01) compared to a WT strain. We then calculated the percentage change of total enzyme copy numbers in mutant comparing to the parental strain.

Flux Coupling Analysis

To identify active metabolic reactions, the reactions that are important for cellular growth, we performed flux coupling analysis (FCA) (Burgard et al., 2004) under several growth conditions, i.e. minimal media conditions, synthetic complete with and without oxygen. Physiological data for constraints were obtained from Van Hoek et al. (1998). Simulations were performed using an improved iMM904 model (Zomorrodi and Maranas, 2010). Reactions that were fully or partially coupled to biomass growth were considered as active. The flux coupling analysis procedure was implemented in C++ and solved using the IBM ILOG CPLEX Optimization Studio 12.7.1.

Metabolic Control Analysis

Metabolic control analysis was performed on the basis of the *S. cerevisiae* kinetic model (Smallbone et al., 2013), which was downloaded from the BioModels database (http://www.ebi.ac.uk/biomodels-main/) under ID MODEL1303260018. Enzyme abundances in the model were adjusted by multiplying original model's enzyme values by kinase mutants enzyme fold-changes, considering only significantly changed enzymes (BH adjusted p-value < 0.01), resulting in a model for every mutant. The steady-state simulations and calculations of control coefficients were performed with libRoadRunner (Somogyi et al., 2015) using the Python API. Metabolites and fluxes were considered to be in steady-state if there was less than a 10^{-6} increment in the solution. The overall flux control coefficients were calculated as described in Millard et al. (2017), i.e. taking for every enzyme the second norm over all its concentrations/flux control coefficients that were parameterised on it, e.g. for flux $CJ_E^{overall} = \sqrt{\sum_j (C_E^J)^2}$ analogously, overall concentrations control coefficients were calculated using $CC_E^{overall} = \sqrt{\sum_s (C_E^S)^2}$.

Metabolic regulation clusters were identified by first transforming all flux control coefficients across samples into PCA space, then calculating Euclidean distance using the first 10 components (retaining over 90% of variance) and performing Ward's hierarchical agglomerative clustering. The number of cluster was then identified using two independent graphical methods "dindex" and "huber" as implemented in NbClust R package (Charrad et al., 2014).

Statistical Modelling

Data Preparation

Depending on the metabolomics experiment (Figure S13), we assigned a proteome measurement of matching genotype to each metabolite sample. For Dataset 1, replicates of proteomics experiments were averaged per genotype. For Datasets 2 and 3, where multiple biological replicates were available, we assigned random proteome measurements to the corresponding genotype. Then we only kept metabolic enzymes, as annotated in genome-scale yeast metabolic network reconstruction (Herrgård et al., 2008). Next, the metabolic network was converted to a bipartite metabolite-enzyme graph. Based on metabolic network topology, we selected enzyme neighbours at various metabolite network neighbourhood radii (Figures 4 and 5) for each measured metabolite. In total, for each modelled metabolite we created 3 response-predictor data matrices corresponding to different metabolic network radii. These were then used as basis for modelling of metabolite concentration data. For this, we used the batch corrected label-free protein quantifications and metabolite concentration measurements. Network manipulations were performed by calling routines from *igraph* library R package (Csardi and Nepusz, 2006).

Data Transformation

To each of the response-predictor matrices, a combination of data transformation methods were applied, specifically quantile normalization (Smyth, 2005), log-transformation, Box-Cox (Sakia, 1992) for predictors (enzyme levels), and log and Box-Cox transformations for responses (metabolite levels/concentrations). Both predictors and responses were standardized to have zero mean and unit variance. To reduce the dimensionality within the predictor space, predictors were transformed onto principal component space (PCA) for metabolite concentration modelling by machine learning. The choice of of number of principal components to retain was based on cumulative coverage of 99% of predictors variation. All data transformations were performed either using R base functions or the *preProcess* function as implemented in *caret* R package (Kuhn, 2008).

Metabolite Concentration Regression Modelling

The computational analysis is divided in the use of multiple linear regression (MLR), 'explanatory part', Figure 4), and using machine learning regression algorithms ('predictive part', Figures 4 and 5).

MLR modelling with exhaustive feature selection mainly was applied for exploratory purposes to identify readily-interpretable biologically meaningful associations while more advanced regression algorithms were used for metabolite concentration predictions. Model selection for MLR case was performed using the following procedures:

For each metabolite sample (without replacement) we created 1000 random subsets, each of them having 90% of the original data.

1) For each subset, we exhaustively evaluated all possible multiple regression models and chose the one with minimum Akaike information criterion (AIC) (as implemented by *regsubsets* function from the *leaps* R package) (Lumley and Miller, 2004)
2) We kept the top 5 most frequent models among all subsets
3) and remove outlier points based on Studentized residuals, exceeding Bonferroni adjusted p-value < 0.05 as implemented in *outlier Test* function in the car R package (Fox and Weisberg, 2011).
4) We removed influence points if any exceeded Cook's distance thresholds, as calculated based on 4/(N−k−1), where N is the number of observations and k is the number of explanatory variables.
5) We tested for the presence autocorrelation using Breusch-Godfrey test (Bonferroni adjusted p-value < 0.05) (as implemented in *bgtest* function in *lmtest* package (Zeileis and Hothorn, 2002);
6) and determined how the obtained models explain the data by calculating the adjusted R^2 value to assess the model fit.
7) To account for finite sample size, AIC was calculated according to the formula N*log(RSS/N) + 2*k (Faraway, 2016), where N is the number of observations, k is the number of explanatory variables, RSS is the residual sum of squares of linear model. Such ranking is not based on hypothesis testing, wherefore does not require FDR correction (Faraway, 2016). Models displaying the highest adjusted R^2 are presented in the main text (Figure 4), the rest of the candidate models are present in Figure S15.

For MLR in the explanatory part, we used only the expression of the first enzyme neighbours of metabolites as features for the metabolite concentration modelling. Scaling was applied to predictors and responses without applying PCA transformation with exception of ATP where the number of predictors exceeded the number of samples. For machine learning (ML) regression (predictive part), we tested 12 algorithms. These are a generalised linear model with stepwise AIC feature selection, ridge regression with foba sparse learning algorithm, partial least squares regression, elastic net regression, lasso, multivariate adaptive regression spline, support vector machine regression, model averaged neural network for regression, recursive partitioning tree, bagged recursive partition tree, conditional inference tree and tree with stochastic gradient boosting. These ML methods were combined with the all possible data transformation strategies as described above. To identify the best predictive model, for each metabolite and data transformation, we optimised each model's hyperparameters by retaining the model having the minimal average 100 times repeated 10-fold cross-validated root-mean-square error between the prediction and metabolite concentration measurement. The model's hyperparameters were optimised using the unified *caret* interface (Kuhn, 2008). Then, the algorithm and the combination of data transformations demonstrating the best predictive performance was expressed as cross-validated R^2 and used to compute metabolite concentrations with the all the proteomic data as input. Each metabolite was finally assigned the algorithm demonstrating the best predictive performance (Figure S17). Source code with grid ranges of hyperparameters for each algorithm are available through GitHub (https://github.com/alzel/regression_models).

The importance of variables was estimated by calling *varImp* function as implemented in *caret* R package (Kuhn, 2008). Variable importance is dependent on the particular algorithm (Kuhn, 2008). Coefficients are scaled to 100% based on the most important variable. In the present analysis, variables are principal components of the enzyme abundance matrix and we considered the variable to be important if it had an importance coefficient >50% and up to 10 enzymes with the highest absolute loading from each of the the components were chosen. In Figure 6C, for visualization purpose we displayed only enzymes with importance coefficient >90% and with up to 5 enzymes with highest absolute loading per component.

Enzyme Saturation

KM values of enzymes for *S. cerevisiae* were obtained from the BRENDA database (Chang et al., 2015; Schomburg et al., 2000) accessed via its Python API on 1.10.2015. Metabolite names were manually mapped to substrate names of BRENDA records. Since the database contains the records of multiple enzymes, including recombinant and modified proteins, only the enzymes that did not match "mutant|recombinant" pattern in the comment section were used for the analysis. For absolute concentration determination we used calibration as described previously (Mülleder et al., 2016a), i.e. by adjusting for dilution used in metabolite extraction protocols and normalising by cell volume (Figures 4F–4H). For analysis we collected 5ml cultures at OD595 1.5, the extraction volume (100 μl for Dataset 2 and 400 μl for Dataset 3) and used the values for cells/OD595 (3.2*107) and cell volume (45.54 fL) for the strain BY4741 in synthetic minimal medium. Cell volume estimates were obtained from Petrezselyova et al. (2010).

Note on the Relationship between Enzyme Expression Changes in the High and Low Abundant Fraction of the Proteome

As microLC-SWATH-MS captures preferentially the high abundant fraction of the the proteome (Vowinckel et al., 2018) we made use of transcriptional profiles as previously recorded for the kinase deletion strains in exponentially growing cells (van Wageningen et al.,

2010), in order to assess enzyme expression also within the genes of lower expression level. Enzyme encoding transcripts on average account for 15% of the total transcriptomic impact of kinase deletion (using the thresholds for differential expression as defined in van Wageningen et al. [2010]) (Figure S5).

In the subset of transcripts that directly correspond to the proteins as quantified by microLC-SWATH-MS, this value is 27%; in the subset of transcripts where no protein values are available (largely representing the low abundant fraction of the proteome), this value is 11%. Kinase-dependent enzyme level changes hence dominate to 1/3rd the highly abundant fraction of the proteome, while they are also significant, but less dominating factor, in differential gene expression in the low abundant transcript fraction.The comparison of our proteomes with this transcriptional data needs to be seen in the context that the transcriptional profiles were recorded from yeast grown in amino acid supplemented media. This fact yielded some interesting observations from the comparison on its own. Indeed, the difference between amino-acid supplemented and minimal media was reflected in a lower correlation between transcriptional and proteomic data as it is typically reported in exponentially growing yeast. This confirms our recent study revealing the importance of biosynthetic metabolism as global factor in cellular gene expression (Alam et al., 2016). Although transcriptome and proteome fold-changes correlated significantly in many of the kinase knock-outs (Pearson $r > 0.25$, p-value < 0.01, Figure S5), none of the Pearson correlation coefficients (PCCs) exceeded a value of >0.5 (Figure 1G); the median value, 0.12, was much lower, and in several strains the correlation was insignificant (Figures 1G and S5). Furthermore, we find that the proximity of kinases to transcription factors in protein-protein interaction networks (Szklarczyk et al., 2015) is a negative indicator of enzyme level changes (Wilcoxon rank sum test, p-value < 0.05). Hence, the more upstream a kinase is compared to a transcription factor, the more enzymes are affected (Figure S6). In contrast, the number of protein-protein interactions reported for each kinase, and the betweenness in the protein-protein interaction networks, did not show significant correlations with number of affected enzymes (Figure S6).

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were done in R (R Core Team, 2015) with specific packages as indicated in each methods section. For the basic data manipulation and visualization we used the R tidyverse package compilation (Wickham and Grolemund, 2016). Hypothesis testing to assess means of population differences were mainly done using non-parametric Wilcoxon Rank Sum test, unless indicated otherwise in specific cases. Sample size estimation were not performed in any of the experiments. For growth experiments at least n=3 biological replicates were analysed unless stated otherwise.

DATA AND SOFTWARE AVAILABILITY

The raw proteomics mass spectrometry data have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaíno et al., 2016) partner repository with the dataset identifier PRIDE: PXD010529. All code used to generate figures in the manuscript are available through Github repository: https://github.com/zelezniak-lab/kinase_metabolism. All data from this manuscript is deposited at: https://doi.org/10.5281/zenodo.1320288.

Cell Systems, Volume *7*

Supplemental Information

Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts

Aleksej Zelezniak, Jakob Vowinckel, Floriana Capuano, Christoph B. Messner, Vadim Demichev, Nicole Polowsky, Michael Mülleder, Stephan Kamrad, Bernd Klaus, Markus A. Keller, and Markus Ralser

Supplementary Information

Machine learning predicts the yeast metabolome from the quantitative proteome of kinase knock-outs

Aleksej Zelezniak[1,2,3,4], Jakob Vowinckel[2,5], Floriana Capuano[2], Christoph Messner[1], Vadim Demichev[1,2], Nicole Polowsky[2], Michael Mülleder[1,2], Stephan Kamrad[1,7], Bernd Klaus[6], Markus Keller[2,8] and Markus Ralser[1,2,9*]

[1]The Francis Crick Institute, Molecular Biology of Metabolism laboratory, London, United Kingdom

[2]Department of Biochemistry and Cambridge Systems Biology Centre, University of Cambridge, Cambridge, United Kingdom

[3]Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

[4]Science for Life Laboratory, KTH – Royal Institute of Technology, Stockholm, Sweden

[5]Biognosys AG, Schlieren, Switzerland

[6]Centre for Statistical Data Analysis, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

[7]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

[8]Medical University of Innsbruck, Innsbruck, Austria

[9]Department of Biochemistry, Charité Universitaetsmedizin Berlin, Berlin, Germany

*Lead author. +44 1223 761346, markus.ralser@crick.ac.uk

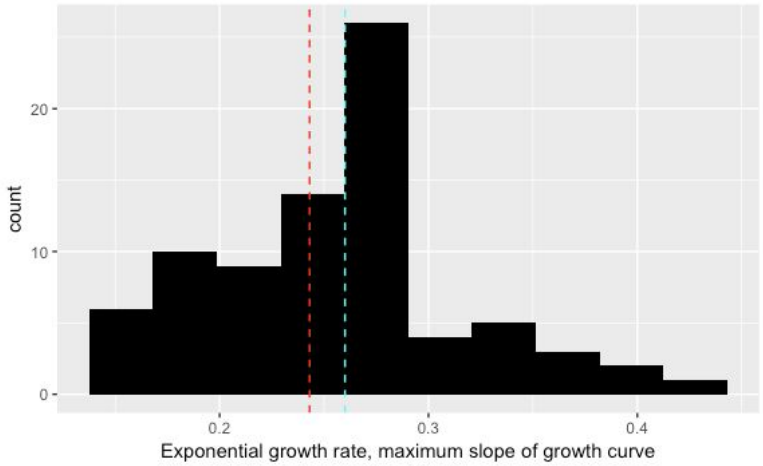


Figure S1. Related to Figure 1; Growth rate in kinase mutants. Many kinase mutants (median growth rate of all kinases, cyan dotted line) exhibit growth rates similar to WT (red dotted line). Data is non-normally distributed with mass center close to WT-strain. The growth curves were fitted using non-parametric (without growth law assumption) spline model as implemented in R growFit package (Kahm et al., 2010).

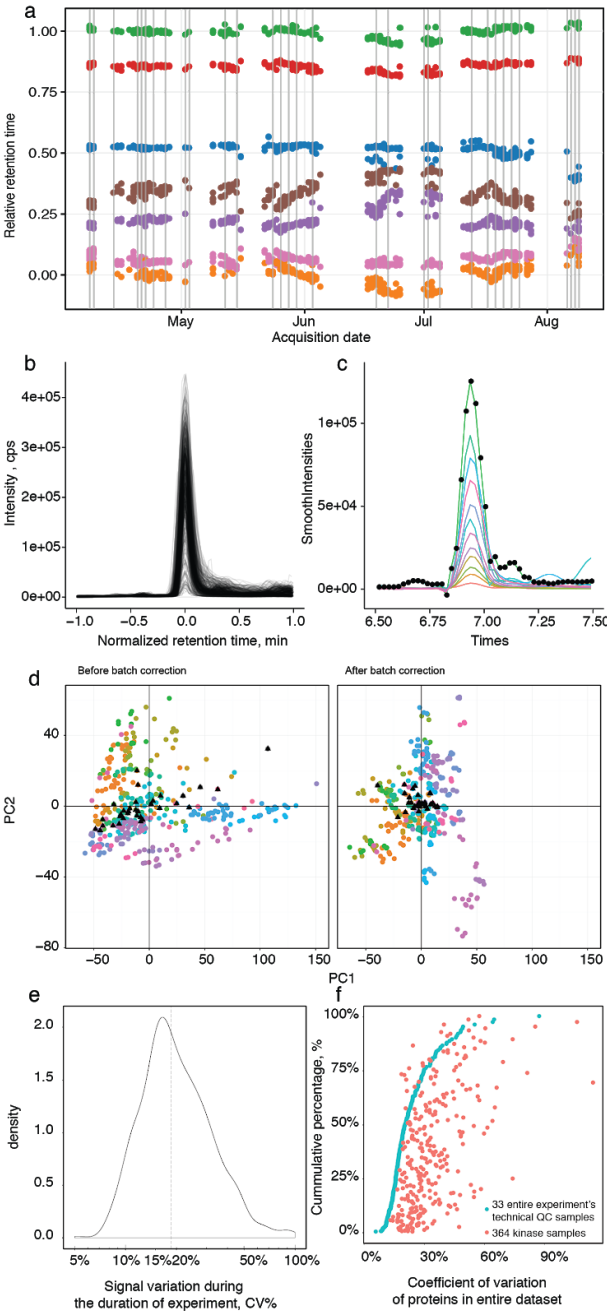


Figure S2. Related to Figure 1; Quality of large proteomics experiment a)microLC-SWATH-MS (Vowinckel et al., 2018) was applied to systematically record the proteomes of *Saccharomyces cerevisiae kinase gene* deletion strains. Shown are retention time stabilities during the measurement of 397 yeast full-proteome tryptic digests by microLC-SWATH-MS over a four-month acquisition period. The median retention time drift was as low as +/- 5.7%, as illustrated by the retention of standard peptides (iRT, coloured points). The rightmost coloured dots represent average peptide retention time with standard deviation (in % of iRT retention) of total chromatographic runtime. Grey lines indicate the processing of a standardised proteome digest (quality control (QC) sample) to monitor instrument performance, to normalise for batch effects, as well as to determine adequate cut-off values for determining differential protein expression. **b)** Overlay of 397 extracted ion chromatograms representing a typical iRT peptide (IGSEVYHNLK) illustrates chromatographic robustness. **c)** our microLC-SWATH-MS implementation covered the typical chromatographic peak with a 1.31s scan cycle so that the illustrated example peptide IGSEVYHNLK (left) is covered by 9 MS^2 and 3 MS^1 ions (different colours in the chromatogram), each by 10 measurements (black dots) in the average sample. This high coverage helps to obtain precise quantification. **d**) Batch correction of microLC-SWATH-MS proteomic data. Before batch correction signal is technically confounded by the acquisition date as demonstrated by variation of external QC control samples (black triangles), colours represent different experimental batches acquired in the period of 4 months. Batch correction reduces variation associated with acquisition date as demonstrated by grouping of QC samples (right panel). **e)** Technical variation of label-free protein quantification as determined by calculating coefficient of variation of combined fragment signal batch corrected intensities in all quality control samples. x - axis is log-scaled, dotted line is the median of the CV% values (19%). **f)** The coefficient of variation (CV) at whole-process technical and biological levels. The CV of technical replicates in 93% of measured metabolic enzymes were lower than in kinase samples, resolving biological signal from technical noise. Kinase samples are sorted as QC replicates.

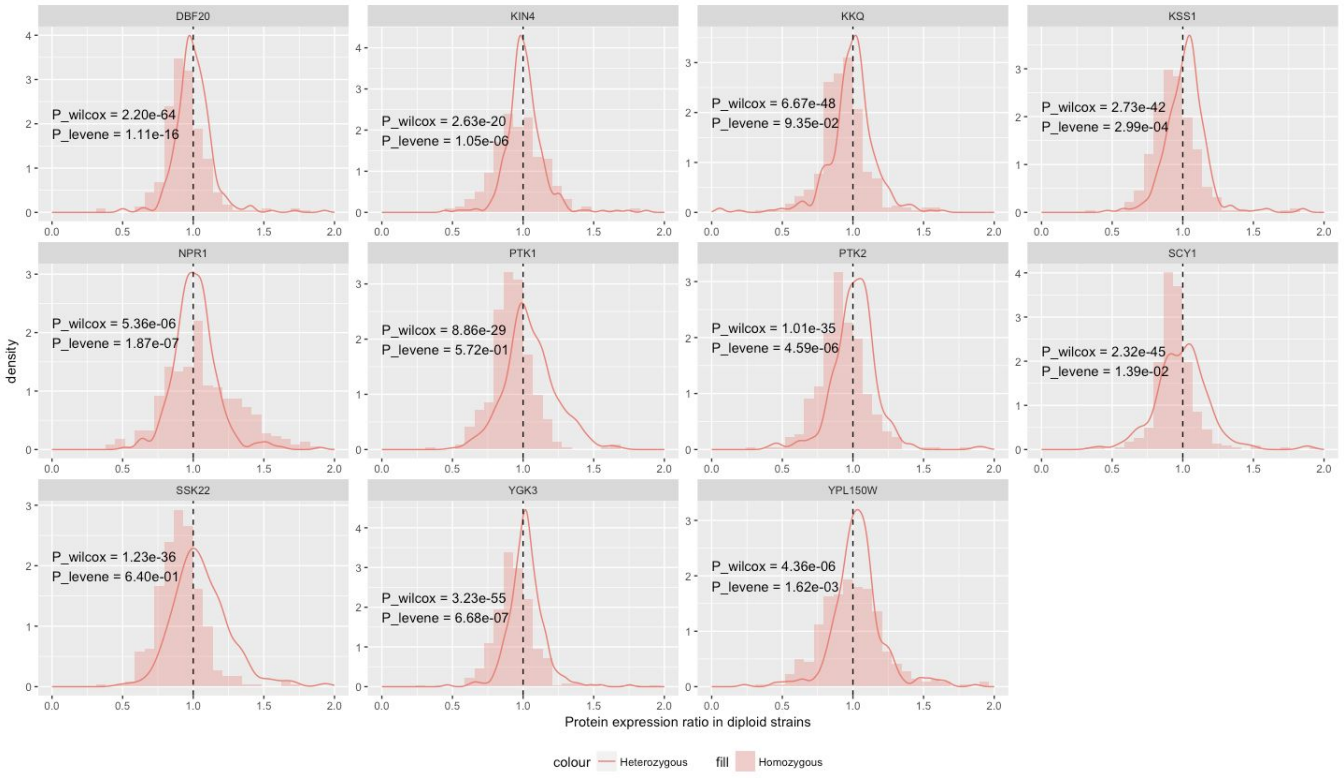


Figure S3. **Related to Figure 1; Kinase deletions in diploids.** Enzyme expression in ten heterozygous vs homozygous kinase mutants, generated by mating the *MATa* strains as used in our study (Mülleder et al., 2012) with a wild-type strain (BY4742) or a complementary kinase knock-out in the MATalpha background. Homozygous diploid kinase mutants have much stronger gene expression changes compared to the wild-type, relative to the corresponding heterozygous strains to which one kinase copy was re-introduced by mating with the MATalpha kinase-wild-type strain. Histogram represents a ratio between kinase homozygous diploid mutant and diploid BY4741-*Δhis* parental strain. Density plot shows ratio between heterozygous mutants normalized by their respective MATa/*MAT*α-*Δkinase* vs *MAT*a/*MAT*α-WT diploids. The dotted line corresponds to no change respective to the wild-type control proteome.

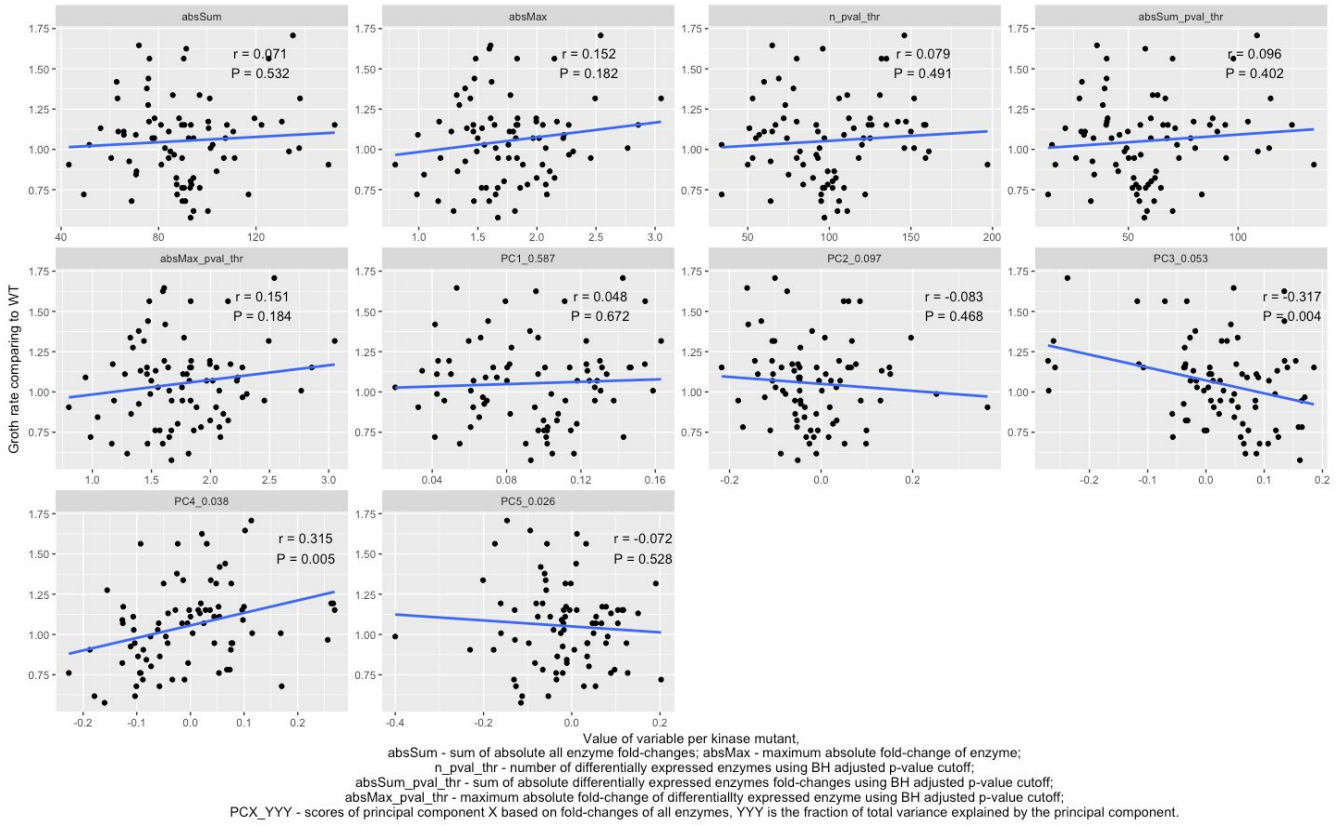


Figure S4. Related to Figure 1; Changes in growth rate are not the main cause of differential enzyme expression in kinase knock-outs, as the main principal components of enzyme expression show no correlation with growth rate changes. Such correlation is however obtained between principal components 3 and 4, that capture 5.3% and 3.8% respectively, of total enzyme expression. More than 90% of enzyme expression changes in kinase knock-outs are not associated to growth rate changes.

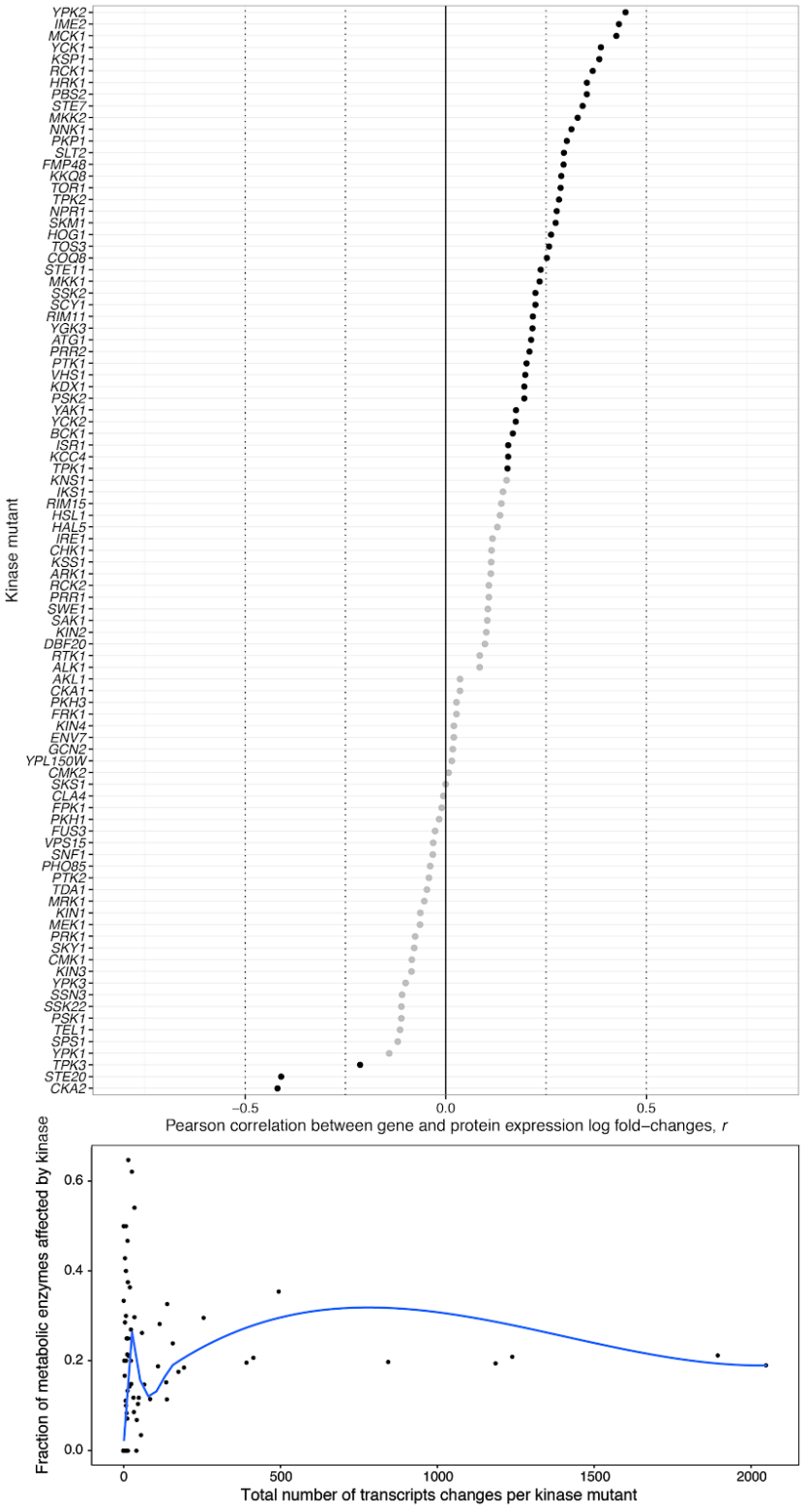


Figure S5. **Related to Figure 1; Correlation between mRNA and protein expression in kinase mutants.** Top: Correlation between mRNA and protein expression $\log_2$ fold-changes in kinase knock-outs. Grey points represent correlation coefficients where p-value exceeded significance cutoff >0.01. Bottom: Fraction of differentially expressed enzymes-coding genes, in comparison to all

differentially expressed genes, at the transcriptional level in kinase deletion strains (van Wageningen et al., 2010). Multiple kinase transcriptomes are characterized by a high number of differentially expressed enzymes, 16% on average. As on the proteome, the total effect size does not determine the relative occurrence of enzyme-encoding transcripts. Fold-change and p-value cutoffs for differential gene expression were obtained from the the original publication (van Wageningen et al., 2010).

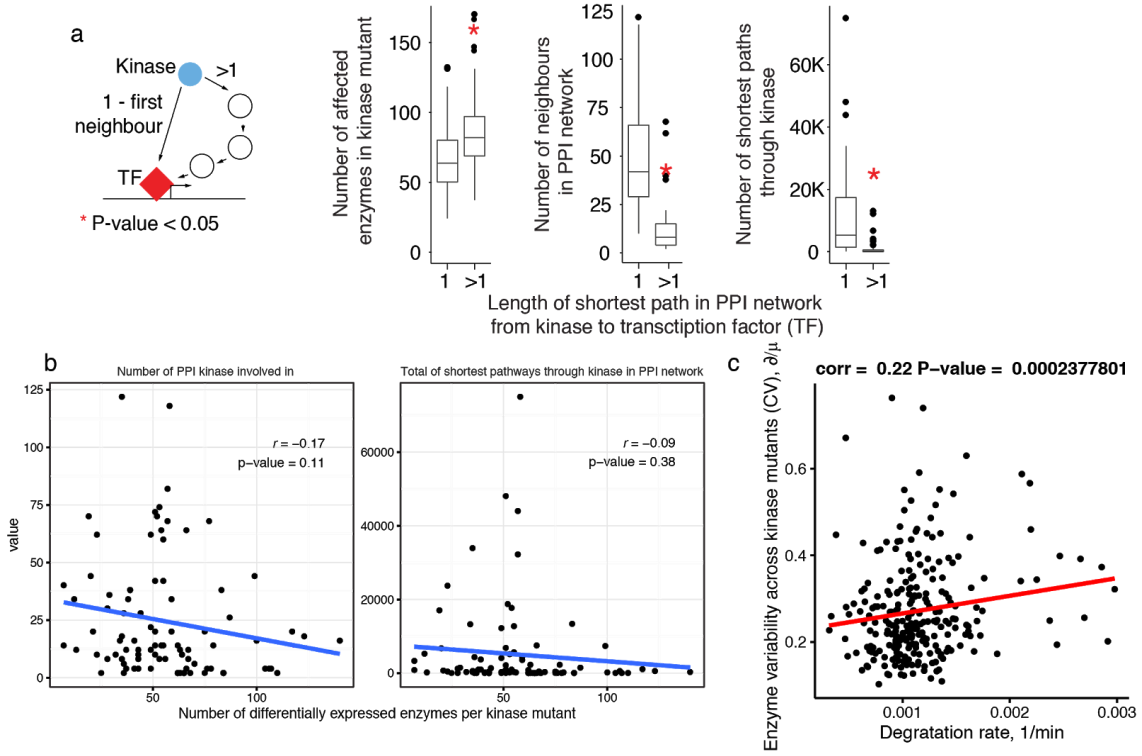


Figure S6. Related to Figure 1; a) The more distant a kinase is to a transcription factor (TF) in a protein-protein interaction network, the more enzyme levels it affects (Wilcoxon rank sum test, p-value < 0.05) (left). Conversely, kinases which directly interact with a transcription factor, have a higher network centrality (middle) and an increased betweenness (right). However, their importance in PPI network had no influence on the number of differentially expressed enzymes (b). **b)** Perturbation size, expressed as a number of differentially expressed enzymes in contrast to wild type strain, is not correlated with number of protein-protein interactions (PPI) kinase involved in (left panel) or number of shortest paths going through kinase in PPI networks kinase (right panel). **c)** Protein degradation rate (x-axis) and the likelihood of an enzyme to be regulated by a kinase, expressed as coefficient of variation across all kinases mutants (y-axis) are weakly correlated. Enzyme degradation rates were obtained from (Christiano et al., 2014)**.** For network analysis a collection of yeast protein-protein interactions (PPI) was obtained from the STRING database (Szklarczyk et al., 2015) (version 10, downloaded on 2015-06-03). We constructed a high confidence (STRING score > 900) PPI network based only on experimentally validated interactions. Transcription factor annotation were obtained based on GO slim (www.yeastgenome.org) categories by selecting terms matching the “nucleic acid binding transcription factor activity” pattern. Graph manipulations and network analysis were performed using the *igraph* library as implemented in R package (Csardi and Nepusz, 2006).

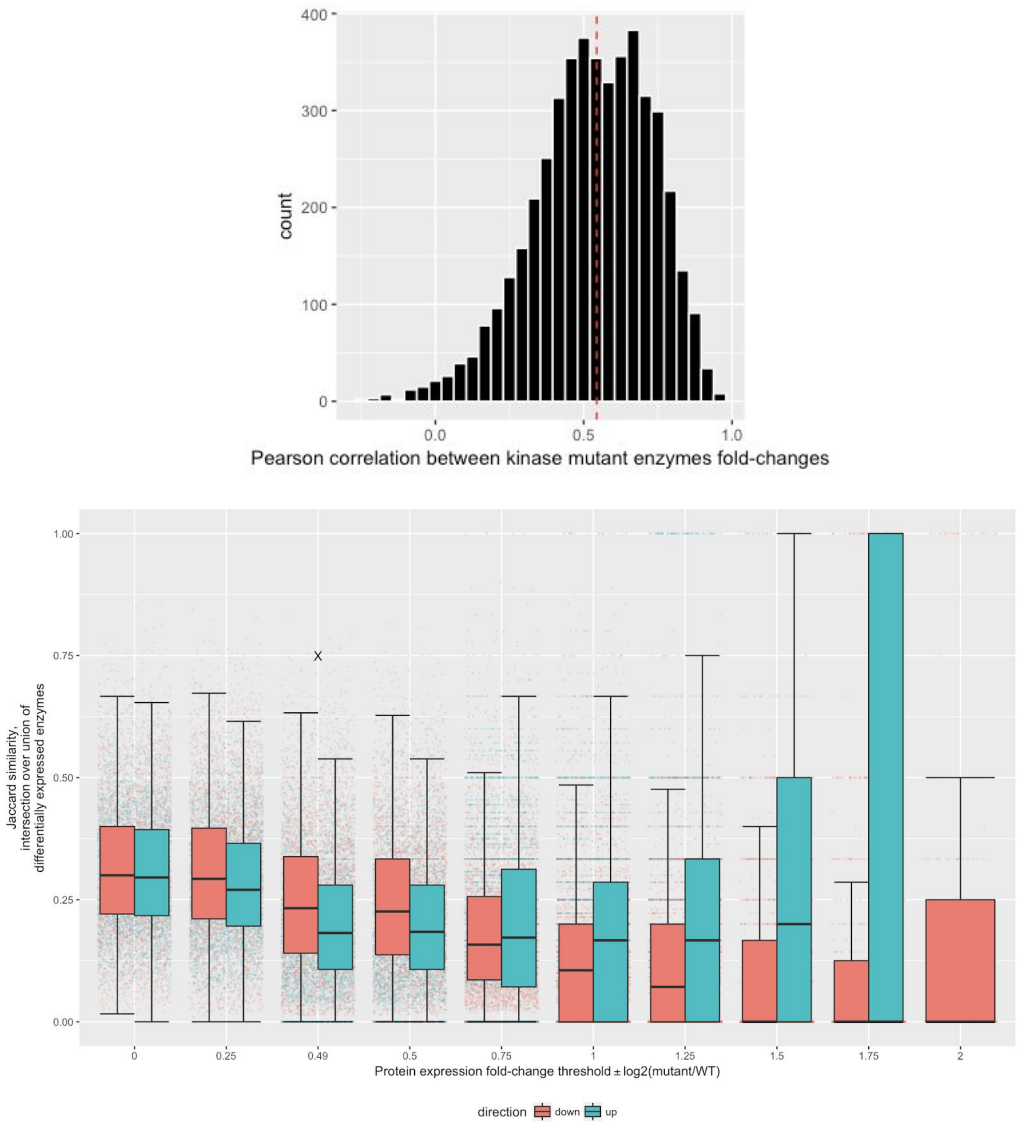


Figure S7. Related to Figure 2; Sensitivity analysis of differential expression. Top panel: Distribution of correlations of kinase mutant enzyme fold-changes. The median correlation of between kinase signatures is ~0.5. A simple linear model build on this basis shows that only 25% of expression changes of one kinase can explain changes of the other, leaving ¾ of the proteome changes being specific to the typical kinase deletion. Hence, also with this metric, the conclusions holds: each kinase deletion leaves a highly specific signature in the enzyme expression proteome. Bottom panel: Sensitivity analysis applied to protein differential expression cutoffs in kinase knock-out strains. Symbol "X" denotes the threshold applied in our study. As one can observe similarity of differentially expressed genes is low between kinase knock-out proteomes even when consider no, or conservative, fold-change cutoff. Dots on the background represent Jaccard similarity of all pairwise kinase comparisons of differentially expressed enzymes. Please note, that as typical for enzyme expression experiments, with there are very few genes that have very strong fold-change concentration changes, thus dots are not anymore gradually scattered once large thresholds are applied.

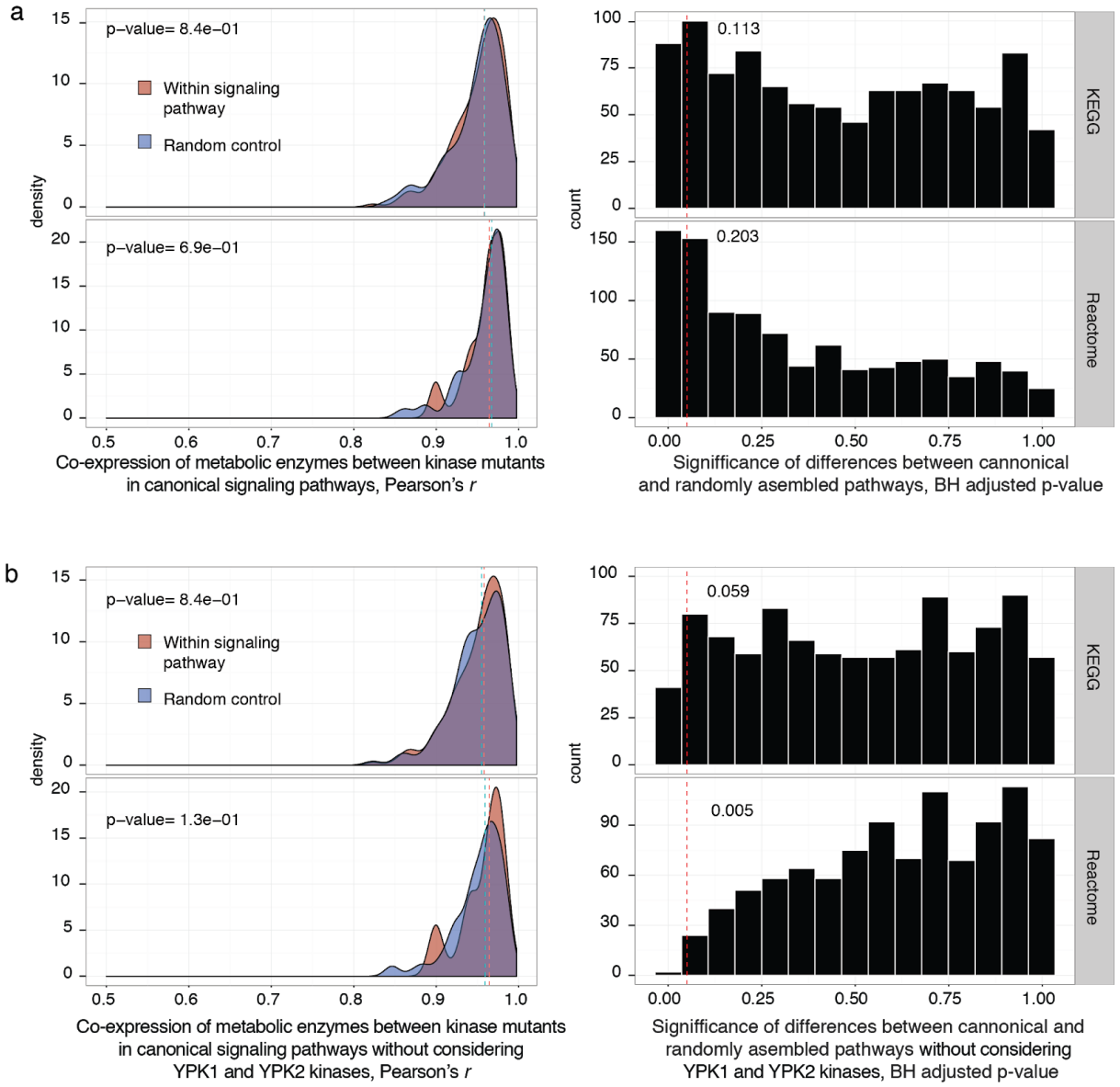


Figure S8. Related to Figure 2; Kinase mutants expression mapping to signaling pathways (part 1). Kinase signaling pathways as assembled in KEGG and REACTOME, and the kinase associations within them, fail to explain enzyme co-expression upon kinase deletion. a) The distribution of the correlation coefficients between enzyme expression levels in kinases mutants that are annotated to the same signaling pathway from KEGG or Reactome databases (left panel). The distribution corresponding to random assignment (of kinases to signaling pathways of the same size as the annotated signaling pathways) is shown for comparison. Random pathways and signaling pathways predict enzyme expression changes not statistically different. Right panel, distribution of p-values from tests (Wilcoxon rank sum) comparing co-expression of canonical signalling versus 1000 times randomly assigned pathways of the same size. Dotted red vertical line denotes a fraction of significantly detected differences (BH adjusted p-value <0.05) between coexpression in canonical pathways and random background. b) same as in (a), but removing *YPK1/ YPK2* - the kinase pair that is most frequently annotated to signaling kinases. In total 49 kinases were assigned to signaling pathways in both databases.

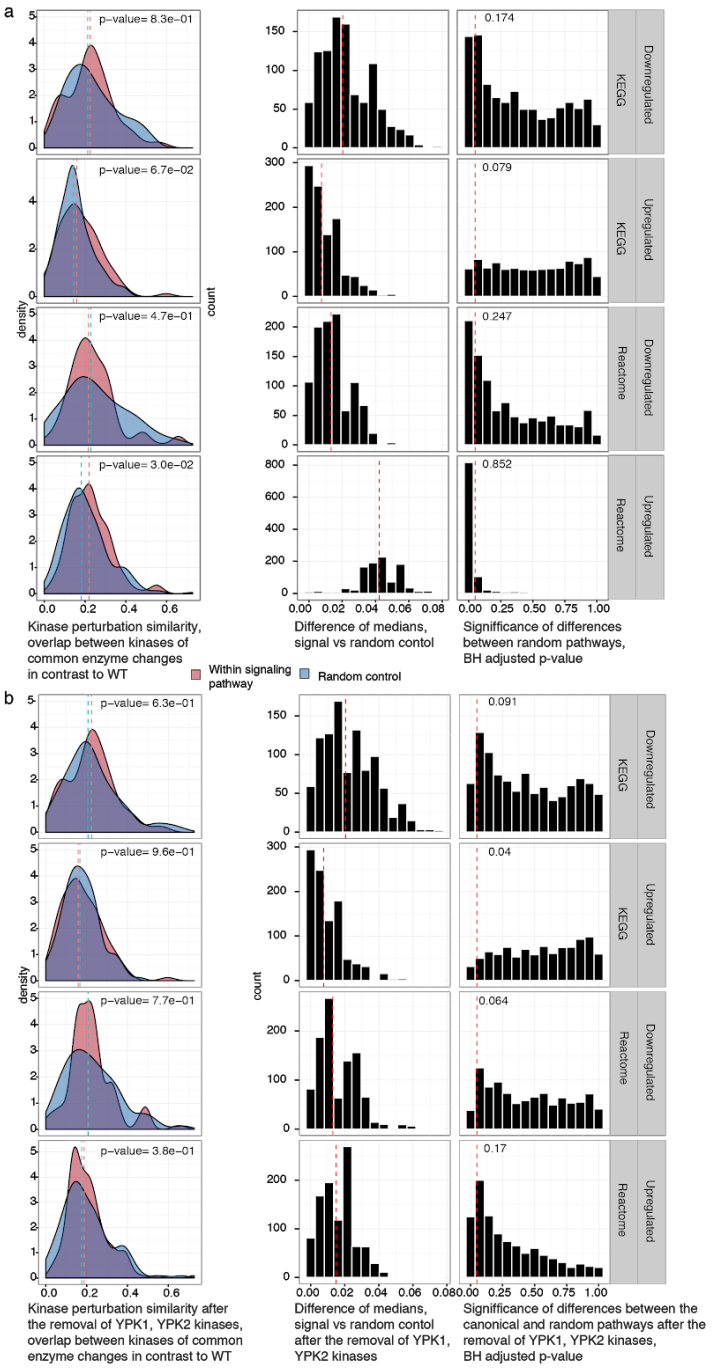


Figure S9. Related to Figure 2; Kinase mutants expression mapping to signaling pathways (part 2). The conventional topology of signaling pathways, and the kinase associations within them, fail to explain enzyme expression upon kinase deletion. The distribution of the overlaps in up-/downregulated metabolic enzymes levels (>|$\log_2$(fold-change)|, BH adj. p-value < 0.01 in contrast to WT strains, see STAR Methods) in kinases mutants that are annotated to the same signaling pathway from KEGG or Reactome databases (left panel). The distribution corresponding to random assignment of kinases to signaling pathways of the same size is shown for comparison. [Middle panel] distribution of median differences of overlaps between canonical kinase and randomly assembled pathways of the same size. Right panel, distribution of p-values from tests (Wilcoxon rank sum) comparing overlaps in canonical signalling versus 1000 times randomly assigned pathways of the same size. Dotted red vertical line

denotes a fraction of significantly detected differences (BH adjusted p-value <0.05) between overlaps in canonical pathways and random background. b) same as in (a), but removing *YPK1, YPK2* - the kinase pair that is most frequently annotated to signaling kinases. In total 49 kinases were assigned to signaling pathways in both databases.

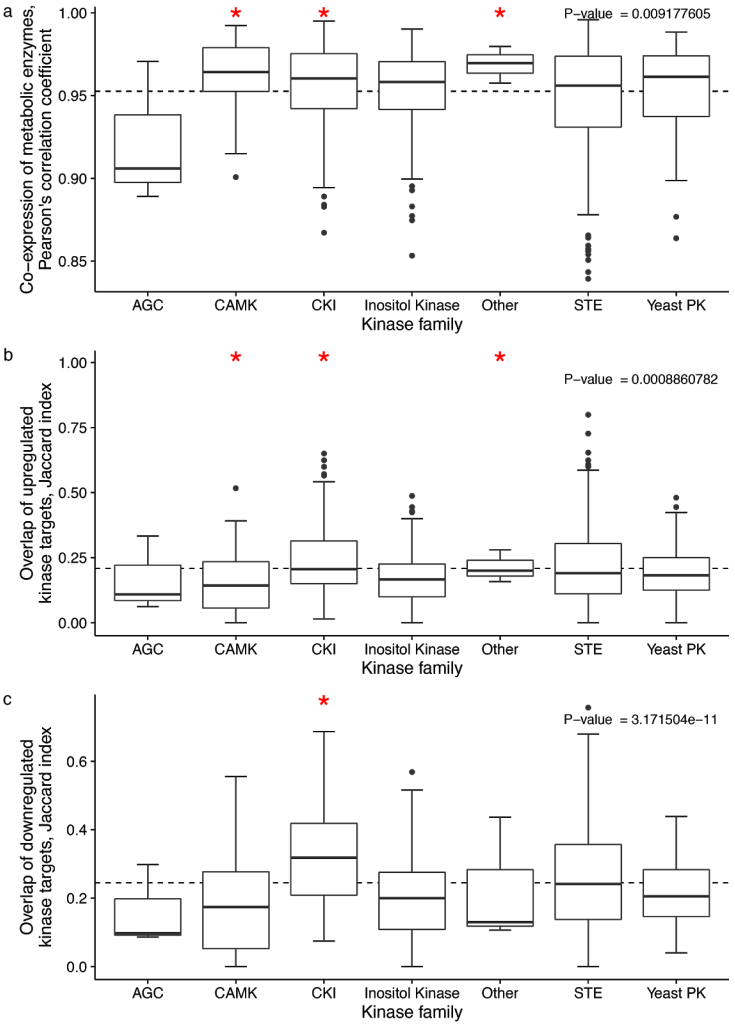


Figure S10. **Related to Figure 2; Co-expression of enzymes between kinase families.** Definition of kinase classes was taken from (Hunter and Plowman, 1997). a) Co-expression of metabolic enzymes between kinase within the class of kinases, expressed as Pearson's correlation coefficient. b,c) overlaps of up-/downregulated metabolic enzymes in kinase mutants. P-values denote significance of one-way ANOVA test using kinase family (classes) as categorical variable, stars depict variables that are significantly different from mean response levels (dashed line).

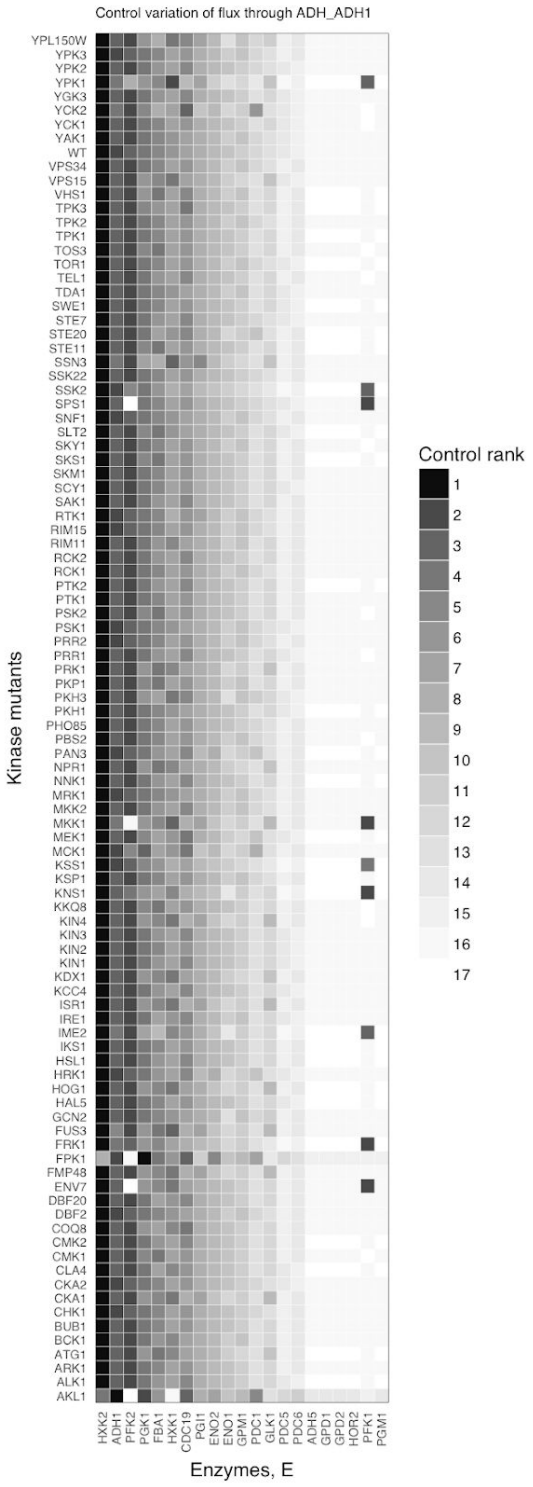


Figure S11. **Related to Figure 3; Flux control variation over alcohol dehydrogenase (*ADH1)* in kinase knockouts.** Control of flux through alcohol dehydrogenase (*ADH1)* reaction shifts to other enzymes depending on the enzymes expression in each kinase mutant. Control coefficients are ranked with the highest as rank 1.

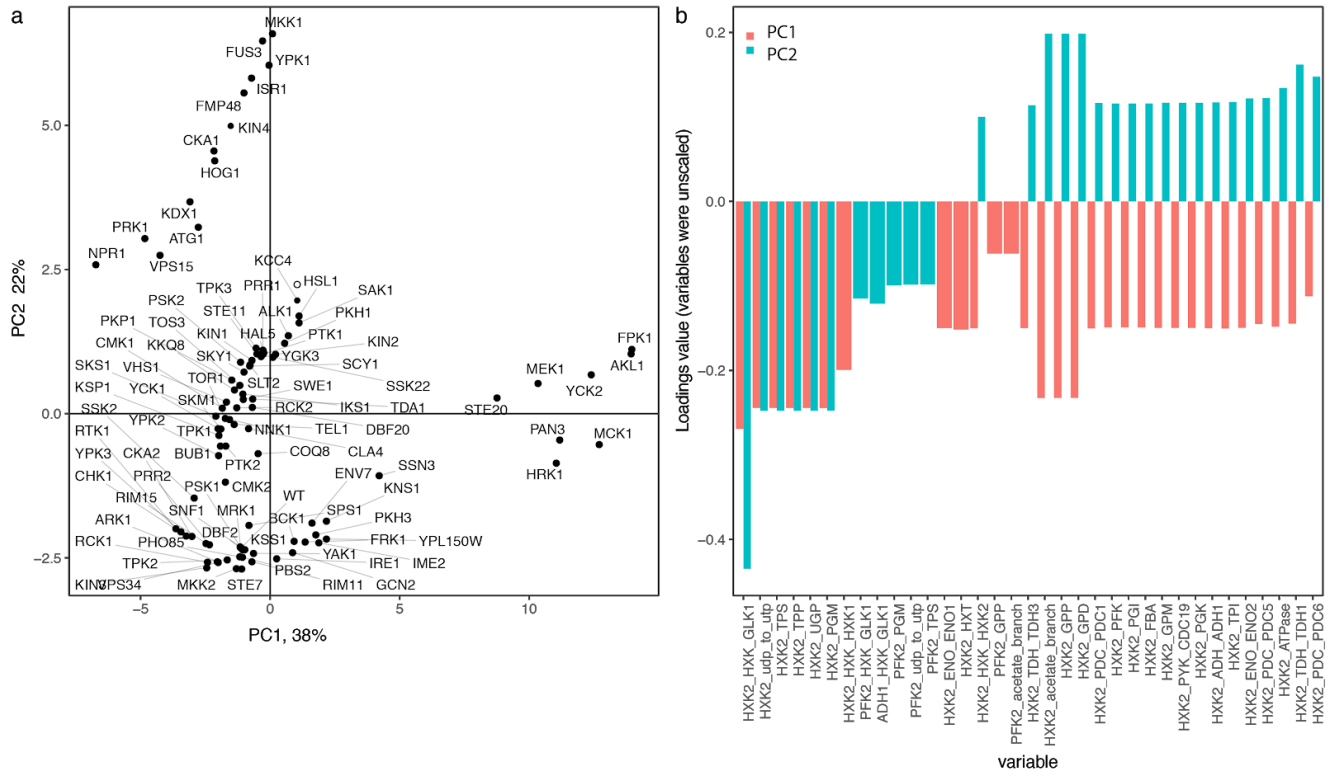


Figure S12. **Related to Figure 3; Flux control variation in kinase mutants using principal component analysis.** a) Principal component plot of flux control coefficients (FCC) for every kinase mutant. FCC were not scaled. Values on axes labels represent percentage of total variance explained by each of the component. b) Loadings for 2 principal components, for each component top 30 (absolute values) FCC are plotted colored according to the component variable loads on.

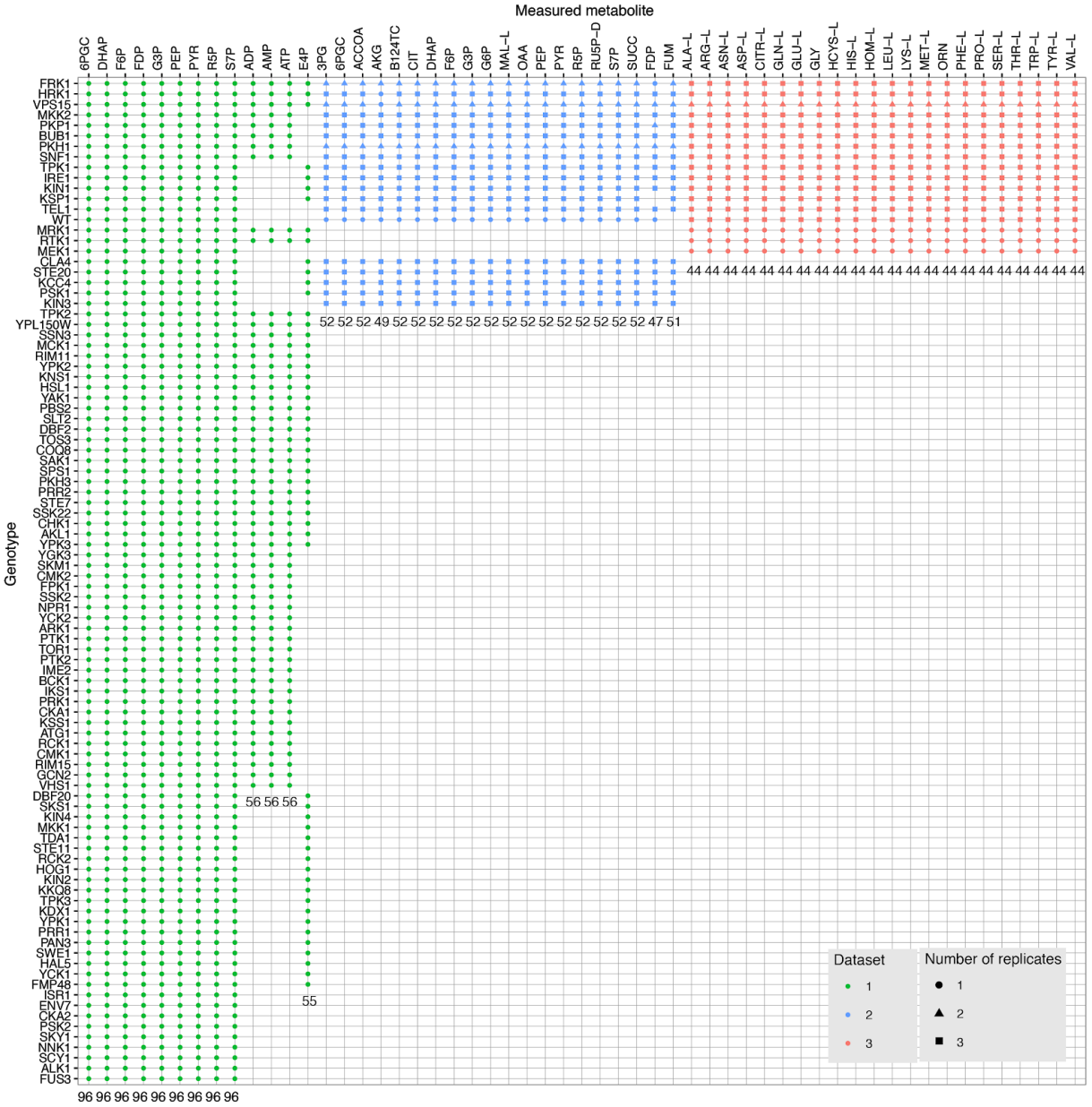


Figure S13. Related to Figure 3, Figure 4, Figure 5; Overview over the metabolite datasets used in the study. Numbers displayed below of each metabolite indicate the total of samples where metabolite was measured.

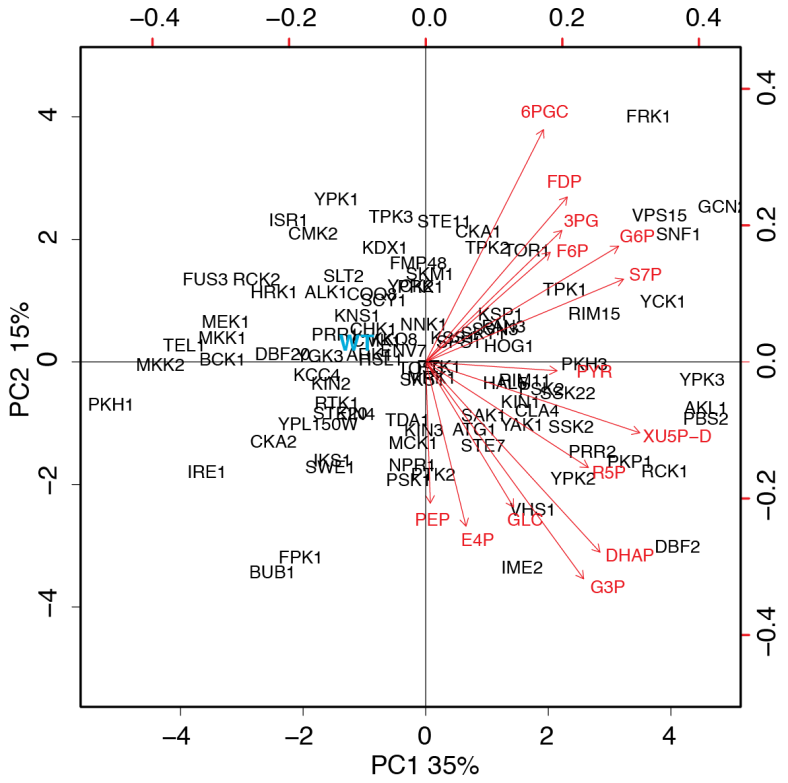


Figure S14. **Related to Figure 3, Figure 4, Figure 5; Principal component analysis of the metabolic profiles caused by kinase deletion.** Percentages on the left and bottom axes denote the proportion of the total metabolite concentration variance captured by the first two principal components. The arrows denote the contribution of each metabolite towards the principal component (loading plot). In blue is highlighted wild-type strain. For visualization purposes only, missing concentration values were imputed as described in (Honaker et al., 2011). Metabolite abbreviations: FDP Fructose 1,6 bisphosphate, 6PGC: 6 -phosphogluconate, FDP: fructose 1,6-bisphosphate, 3PG: 3-phosphoglycerate, F6P: fructose 6-phosphate, G6P: glucose 6-phosphate, S7P: sedoheptulose 7-phosphate, PYR: pyruvate, XU5P-D: xylulose 5-phosphate, R5P: ribose 5-phosphate, DHAP: Dihydroxyacetone phosphate, G3P: Glyceraldehyde 3-phosphate, GLC: glucose, E4P: erythrose 4-phosphate, PEP: phosphoenolpyruvate.

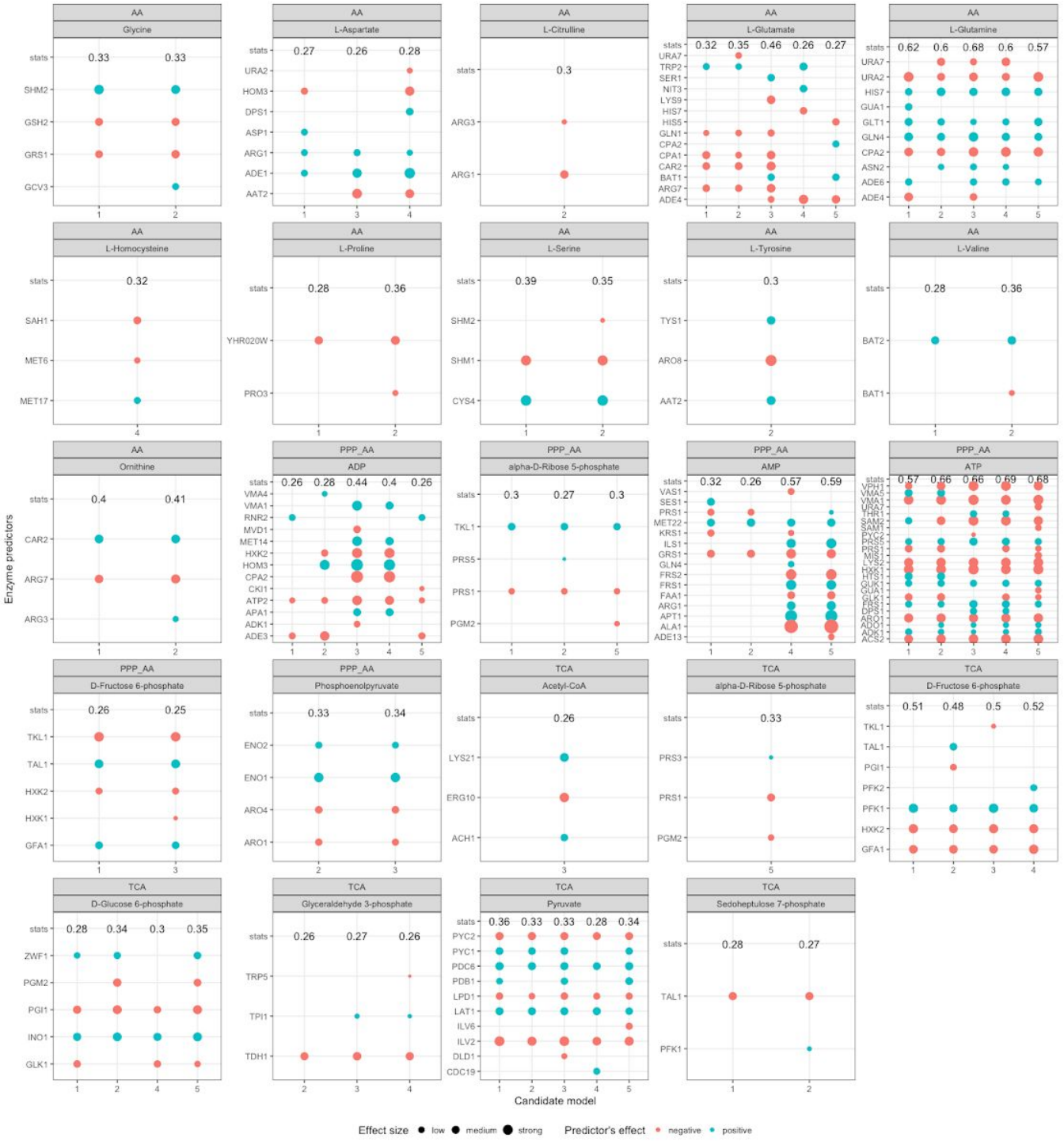


Figure S15. Related to Figure 4; Metabolite concencentration models formulated using multiple linear regression model with exhaustive feature selection. Stats - represents adjusted R^2, all models were diagnosed for the presence of autocorrelation, outliers and influential points (Methods). Presented models have adjusted R^2 value >0.25 and p.value < 0.01. In the main text the models with highest adj. R^2 are presented. For ATP metabolite, due to its connectivity in metabolic network the number of explanatory variables was exceeding the number measured samples, thus before feature selection the explanatory

variables were transformed onto principal components to reduce the dimensionality. From each component we chose 2 highest absolute loadings and assigned corresponding regression coefficient from selected feature.

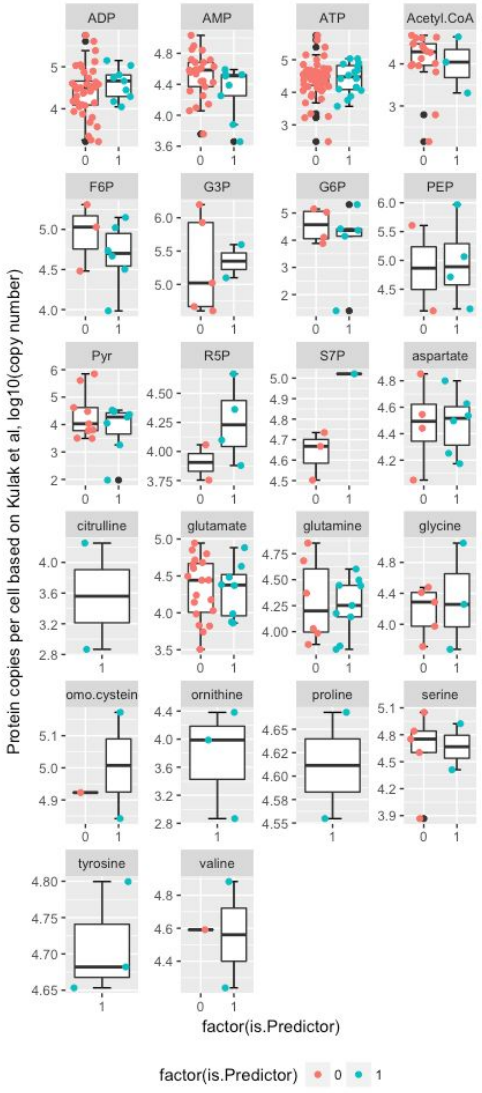


Figure S16. **Related to Figure 4; No difference in absolute copy numbers between the best explanatory variables of metabolite concentrations and the rest enzymes.** The best predictors were selected by exhaustive feature selection using multiple linear regression. For tyrosine, homo-cysteine and ornithine the only measured directly metabolizing enzymes are the ones which are displayed and therefore solely identified as predictors.

	Algorithm	RMSE	RsquaredCV	Data transformation	Dataset	Metabolite	Pathway	Network radius
1	enetModel	0.72	0.58	Box–Cox	AA	L–Arginine	Glutamate family	2
2	plsModel	0.61	0.67	Box–Cox	AA	L–Aspartate	Aspartate family	2
3	plsModel	0.60	0.66	Box–Cox	AA	Glycine	Serine family	3
4	enetModel	0.63	0.59	Box–Cox	AA	L–Histidine	His&nucleotide	2
5	plsModel	0.73	0.57	Box–Cox	AA	L–Homoserine	Other	2
6	fobaModel	0.66	0.65	log	AA	L–Alanine	Pyruvate family	3
7	gbmModel	0.73	0.51	log	AA	L–Asparagine	Aspartate family	2
8	fobaModel	0.77	0.55	log	AA	L–Glutamate	Glutamate family	3
9	plsModel	0.56	0.72	log	AA	L–Phenylalanine	Aromatic family	3
10	plsModel	0.62	0.73	log Quantile	AA	L–Citrulline	Other	3
11	fobaModel	0.75	0.55	log Quantile	AA	L–Glutamine	Glutamate family	1
12	earthModel	0.81	0.47	log Quantile	AA	L–Homocysteine	Other	2
13	fobaModel	0.69	0.58	log Quantile	AA	L–Leucine	Pyruvate family	2
14	rpartModel	0.87	0.44	log Quantile	AA	L–Methionine	Serine family	1
15	gbmModel	0.59	0.75	log Quantile	AA	Ornithine	Other	3
16	plsModel	0.71	0.60	log Quantile	AA	L–Valine	Pyruvate family	3
17	earthModel	0.75	0.54	log	AA	L–Serine	Serine family	3
18	fobaModel	0.54	0.72	Box–Cox	AA	L–Lysine	Glutamate family	2
19	rpartModel	0.80	0.51	Box–Cox	AA	L–Proline	Glutamate family	2
20	plsModel	0.65	0.60	Box–Cox	AA	L–Threonine	Aspartate family	3
21	fobaModel	0.52	0.75	Box–Cox	AA	L–Tryptophan	Aromatic family	2
22	svmRModel	0.77	0.52	Box–Cox	AA	L–Tyrosine	Aromatic family	1
23	enetModel	0.82	0.41	Box–Cox	PPP_AA	ADP	Energy metabolism	3
24	fobaModel	0.78	0.46	Box–Cox	PPP_AA	AMP	Energy metabolism	3
25	fobaModel	0.74	0.55	Box–Cox	PPP_AA	Erythrose 4–phosphate	PPP	2
26	enetModel	0.90	0.31	log Quantile	PPP_AA	ATP	Energy metabolism	1
27	svmRModel	0.88	0.34	Box–Cox	TCA	cis–Aconitate	TCA	3
28	fobaModel	0.79	0.50	Box–Cox	TCA	D–Fructose 6–phosphate	Glycolysis	2
29	plsModel	0.78	0.54	Box–Cox	TCA	Fumarate	TCA	3
30	svmRModel	0.78	0.52	log	TCA	Acetyl–CoA	TCA	2
31	fobaModel	0.86	0.39	log	TCA	Dihydroxyacetone phosphate	Glycolysis	2
32	rpartModel	0.80	0.54	log	TCA	D–Glucose 6–phosphate	Glycolysis	2
33	svmRModel	0.75	0.48	log Quantile	TCA	2–Oxoglutarate	TCA	1
34	gbmModel	0.63	0.53	log Quantile	TCA	Citrate	TCA	2
35	plsModel	0.72	0.56	log Quantile	TCA	Glyceraldehyde 3–phosphate	Glycolysis	3
36	fobaModel	0.78	0.46	log Quantile	TCA	Oxaloacetate	TCA	2
37	fobaModel	0.72	0.58	log Quantile	TCA	D–Ribulose 5–phosphate	PPP	3
38	fobaModel	0.87	0.43	log Quantile	TCA	Succinate	TCA	2
39	rpartModel	0.81	0.57	log Quantile	TCA	3–Phospho–D–glycerate	Glycolysis	3
40	earthModel	0.67	0.65	log	TCA	alpha–D–Ribose 5–phosphate	PPP	3
41	gbmModel	0.72	0.55	log	TCA	Sedoheptulose 7–phosphate	PPP	2
42	earthModel	0.81	0.40	Box–Cox	TCA	L–Malate	TCA	3
43	fobaModel	0.58	0.69	Box–Cox	TCA	Phosphoenolpyruvate	Glycolysis	3
44	gbmModel	0.80	0.48	Box–Cox	TCA	Pyruvate	Glycolysis	2
45	plsModel	0.73	0.53	Box–Cox	TCA	D–Fructose 1,6–bisphosphate	Glycolysis	3
46	enetModel	0.70	0.56	Box–Cox	TCA	6–Phospho–D–gluconate	PPP	3

Figure S17. **Related to Figure 5; Best performing ML algorithms for metabolite concentration predictions.** Metabolite data was Box-Cox transformed using the maximum log-likelihood method for parameter estimation. RsquaredCV - 100 times repeated 10-fold cross-validated R^2. Algorithms abbreviations: enetModel - Elastic net regression, plsModel - partial least squares regression, fobaModel - ridge regression with variable selection, earthmodel - multivariate adaptive regression splines, svmRModel - support vector machine regression, Hyperparameter tuning grid ranges for each algorithm are deposited at github https://github.com/alzel/regression_models/blob/master/regression_models.R

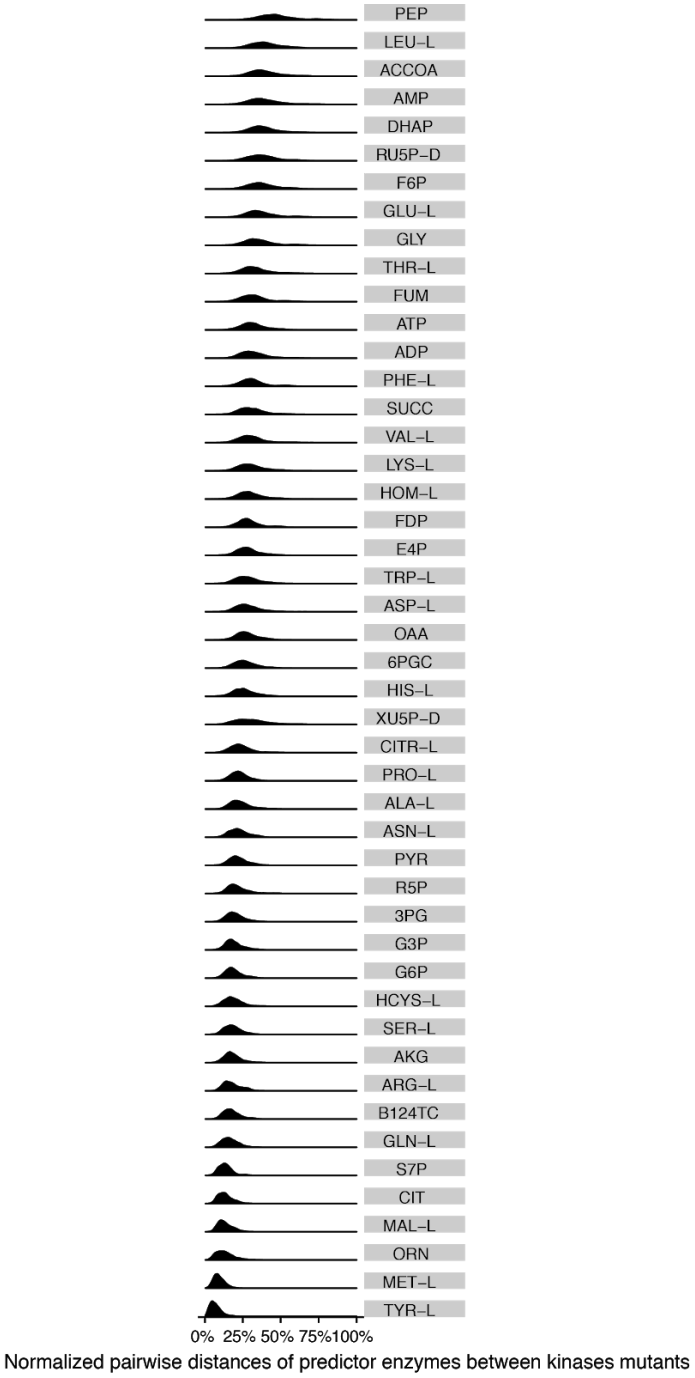


Figure S18. **Related to Figure 5; Regulatory specificity of enzyme predictors.** Kinases interact with metabolite concentrations with different degree of specificity, illustrated as response similarity distribution for each metabolite. More distant values (upper density plots) imply specific response in metabolite predictors. To compare predictor responses between metabolites, predictors were standardized (to mean zero and unit variance), and then the Euclidean distance of standardized enzyme expression was computed pairwise between each kinase mutant and normalized to 100% by the most distant kinase pair.

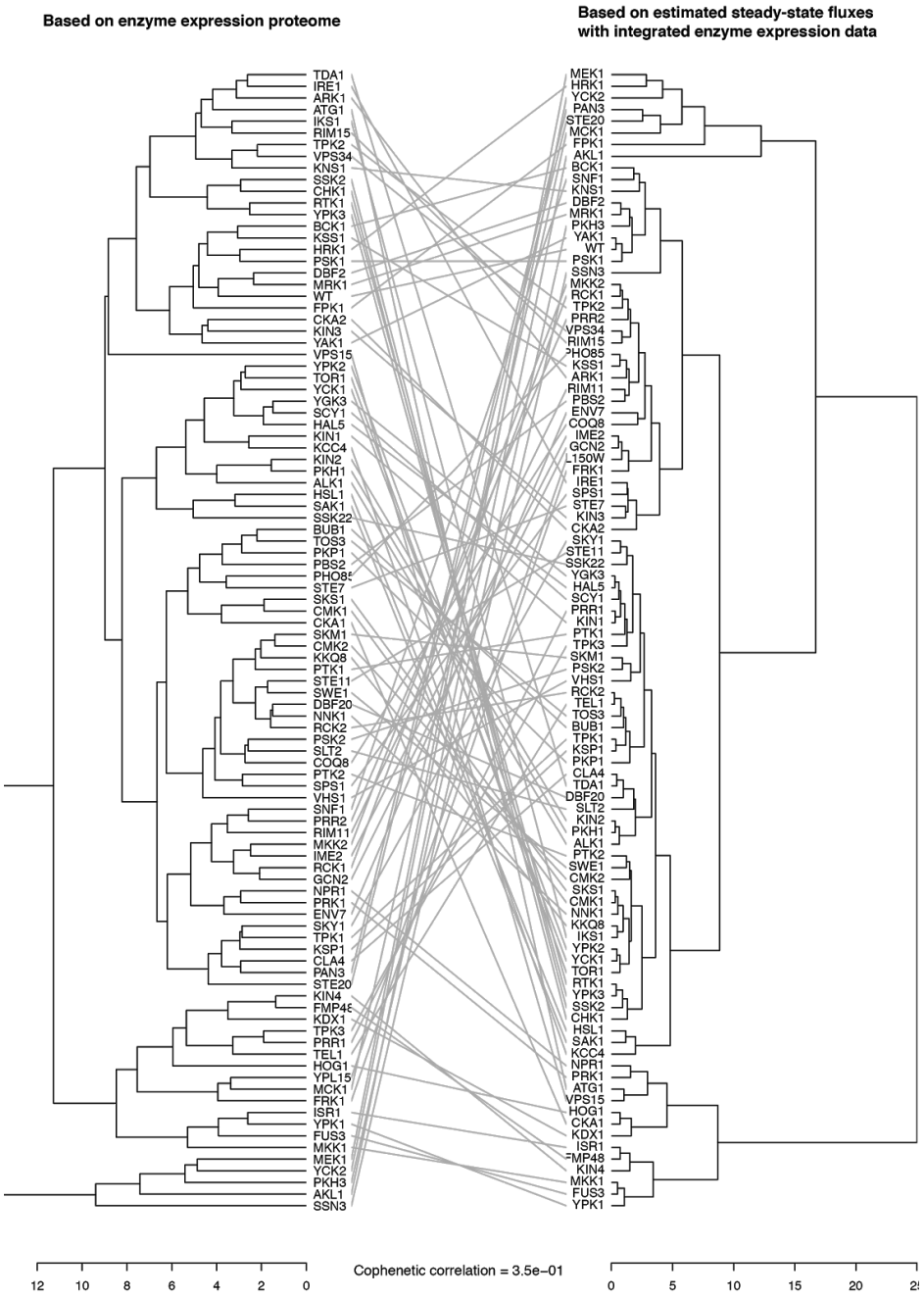


Figure S19. **Related to Figure 6; Correlation between enzyme expression and metabolic fluxes.** The nonlinear nature of metabolism regulation as highlighted by a low correlation of enzyme expression and fluxes - that were estimated by upon introducing experimentally measured enzyme abundances change into a quantitative model of glycolysis (Smallbone et al., 2013). Hierarchical clustering of kinase mutants on the basis the enzyme expression levels (left panel) and mutant fluxes calculated using same enzyme abundance for modelling (right panel). Each variable, either flux and proteins MS signal levels, were standardized by subtracting mean of the value and dividing by its standard deviation among all mutants. Using Euclidean distance between strain profiles both matrices then were hierarchically clustered with complete linkage agglomeration. Replicates of proteomics measurements were averaged per genotype and were used for both analyses.

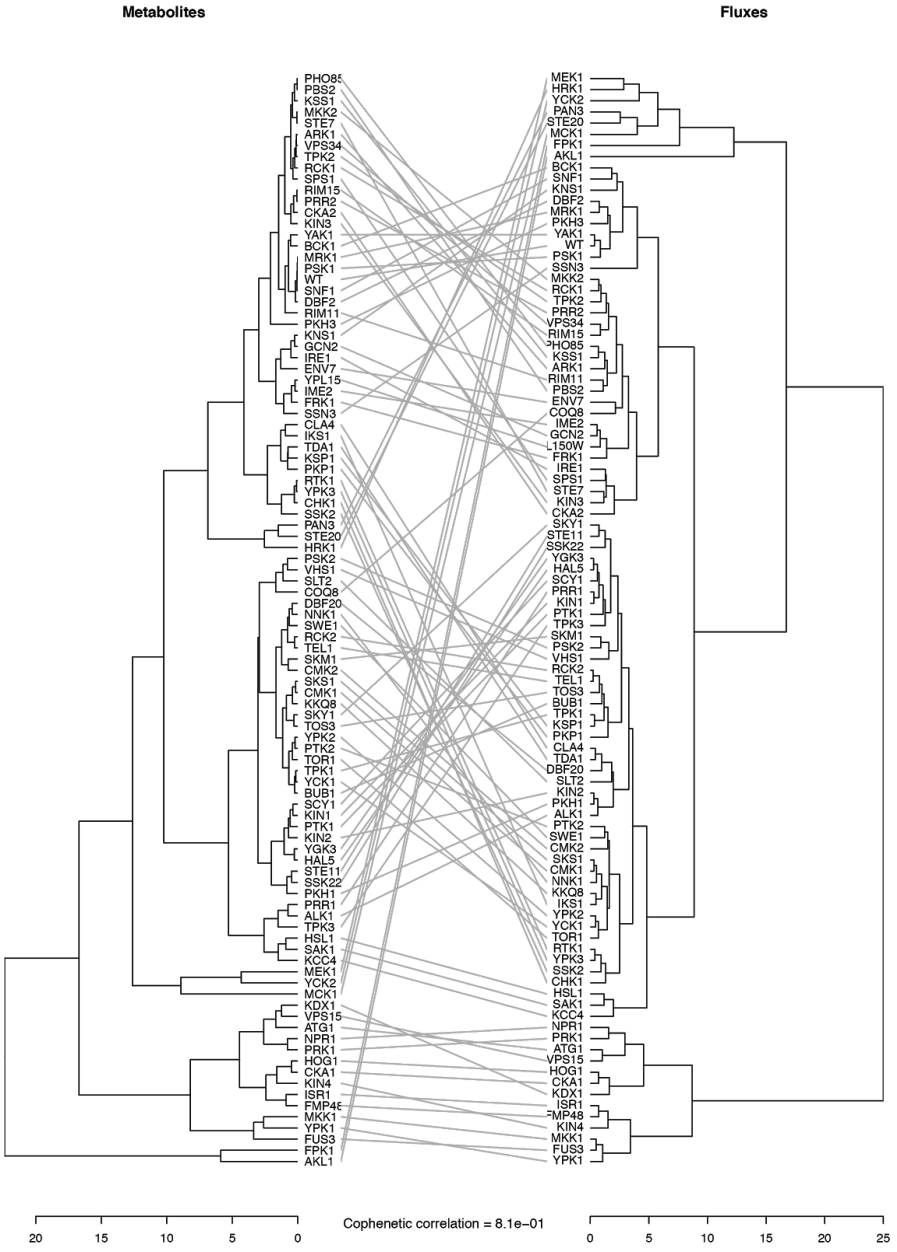


Figure S20. Related to Figure 6; Correlation between metabolite concentrations and metabolic fluxes. Metabolite concentrations are highly correlated with metabolic fluxes highlighting stronger dependency of the flux on metabolite levels rather than enzymes in kinetic model of central carbon metabolism. Analogous results were reported in several recent recent studies (Hackett et al., 2016; Millard et al., 2017). Hierarchical clustering of kinase mutants on the basis of modeled metabolite concentrations with incorporated enzyme expression levels (left panel) and mutant fluxes calculated using same enzyme abundance for modelling (right panel). Each variable, either flux and proteins MS signal levels, were standardized by subtracting mean of the value and dividing by its standard deviation among all mutants. Using Euclidean distance between strain profiles both matrices then were hierarchically clustered with complete linkage agglomeration. Replicates of proteomics measurements were averaged per genotype and were used for both analyses.

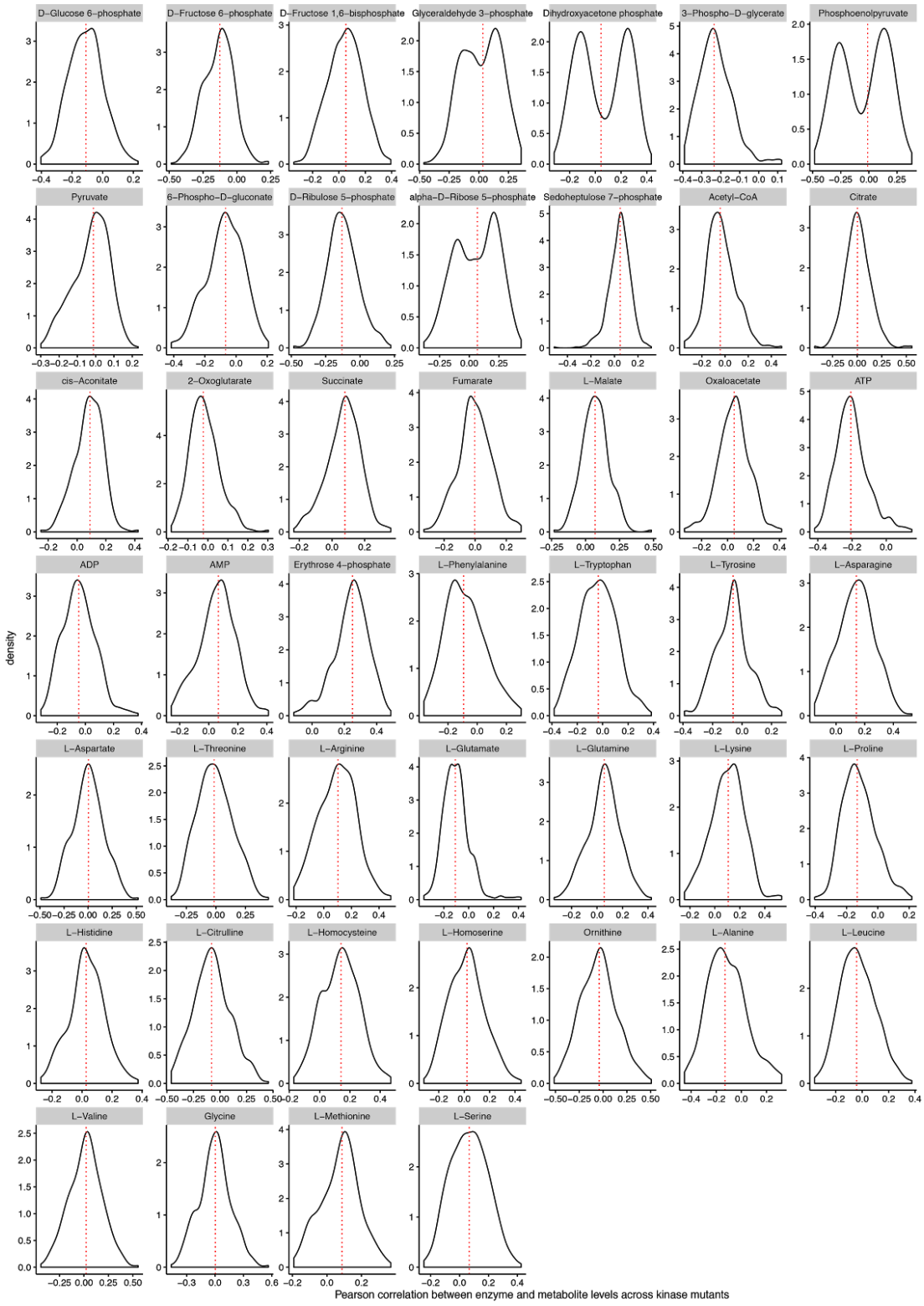


Figure S21. Related to Figure 6; Correlation between metabolite levels and enzyme abundances. Distribution of Pearson's correlation coefficients between metabolite levels and all measured enzyme abundances across all kinase mutants.

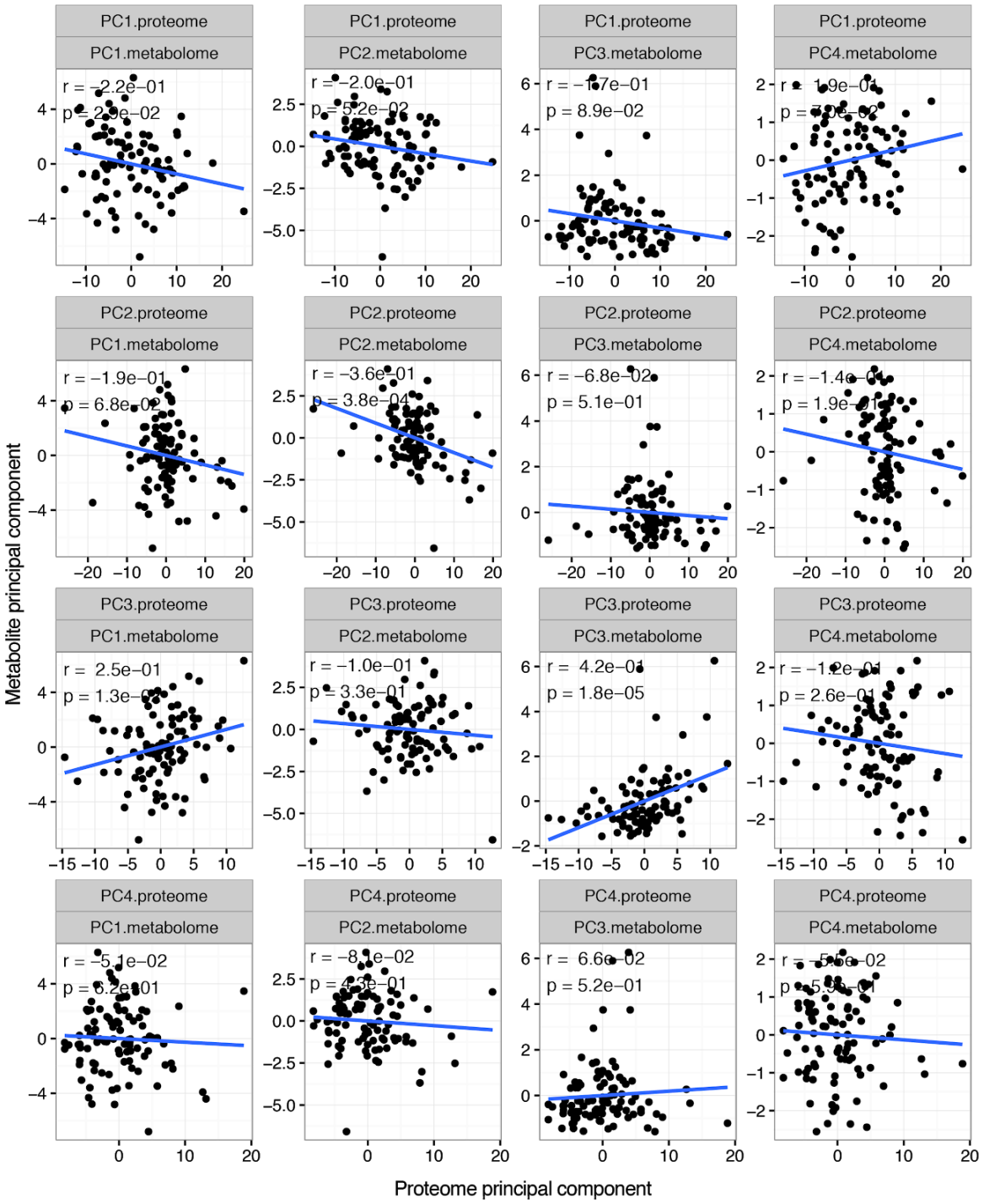


Figure S22. **Related to Figure 6; Correlation of first four principal components of metabolite and proteome data.** Before performing principal component analysis, each dataset was scaled, mean centered and normalized to unit variance. Metabolite data from dataset 1 (Supplementary Figure 20) was used for analysis. Missing metabolite measurements were imputed using *amelia* (Honaker et al., 2011). Replicates of proteomics measurements were averaged per genotype. Data was matched by genotype.