Genetics: Early Online, published on June 26, 2017 as 10.1534/genetics.117.203752

Causal genetic variation underlying metabolome differences

Devjanee Swain-Lenz[*,†], Igor Nikolisky[‡], Jiye Cheng[*,§], Priya Sudarsanam[*,†], Darcy Nayler[†], Max V. Staller[*,†], and Barak A. Cohen[*,†,1]

[*]Center for Genome Sciences and Systems Biology, Washington University in St. Louis School of Medicine, St. Louis, Missouri 63110, USA

[†] Department of Genetics, Washington University in St. Louis School of Medicine, St. Louis, Missouri 63110, USA

[‡] Department of Chemistry, Washington University in St. Louis, St. Louis, Missouri, 63130, USA

[§] Center for Gut Microbiome and Nutrition Research, Washington University School of Medicine, St. Louis, MO 63110, USA

[1] Corresponding author

An ongoing challenge in biology is to predict the phenotypes of individuals from their genotypes. Genetic variants that cause disease often change an individual's total metabolite profile, or metabolome. In light of our extensive knowledge of metabolic pathways, genetic variants that alter the metabolome may help predict novel phenotypes. To link genetic variants to changes in the metabolome, we studied natural variation in the yeast *Saccharomyces cerevisiae*. We used an untargeted mass spectrometry method to identify dozens of metabolite Quantitative Trait Loci (mQTL), genomic regions containing genetic variation that control differences in metabolite levels between individuals. We mapped differences in urea cycle

metabolites to genetic variation in specific genes known to regulate amino acid biosynthesis. Our functional assays reveal that genetic variation in two genes, *AUA1* and *ARG81*, cause the differences in the abundance of several urea cycle metabolites. Based on knowledge of the urea cycle we predicted, and then validated a new phenotype, sensitivity to a particular class of amino acid isomers. Our results are a proof of concept that untargeted mass spectrometry can reveal links between natural genetic variants and metabolome diversity. The interpretability of our results demonstrates the promise of using genetic variants underlying natural differences in the metabolome to predict novel phenotypes from genotype.

INTRODUCTION

A fundamental goal in biology is to understand the properties of genetic variants that underlie phenotypic differences between individuals. Because causal genetic variants often change an individual's metabolome (Gauguier 2016; Suhrer and Geiger 2012), metabolomics, the systematic study of metabolites, offers an avenue to identify genetic variants that contribute to phenotypic differences through their effects on metabolism. Understanding how genetic variation in specific genes affects metabolic phenotypes is an important step towards the goal of predicting phenotype from genetic variation. For example, therapeutic outcomes are better when stroke patients receive a dose of warfarin that depends on their genotypes at two metabolic genes rather than a fixed dose (Consortium, T. I. W. P. 2009; Pirmohamed *et al.* 2013). To further explore causal genetic variation in the metabolome, we combined improvements

in untargeted mass spectrometry with a genetically tractable yeast system to uncover gene variants that underlie metabolome differences. Then using knowledge of known metabolic pathways, we predicted novel drug sensitivity phenotypes from genotype.

A large number of metabolite levels can be measured simultaneously either through targeted methods, in which the identities of metabolites are known, or untargeted methods, in which the identities of metabolites are unknown. Targeted methods are typically more quantitative, while untargeted methods can be used to screen a broader range of metabolic phenotypes. Previous metabolomics studies in plants, humans, and lab strains of yeast have either identified genetic variants that effect metabolite levels using targeted methods (Breunig *et al.* 2014; Chen *et al.* 2014; Dong *et al.* 2015), or have identified new metabolic phenotypes between individuals with known causal genetic variation using untargeted methods (Broyart *et al.* 2009; Hu *et al.* 2014, Keurentjes *et al.* 2006). However, few investigators have attempted to map genetic variation using untargeted mass spectrometry, which can lead to discovery of both unknown metabolic and genetic variation (Lewis *et al.* 2014). For instance untargeted methods led to the detection of variation in the chloroquine resistant gene in *Plasmodium* that confers different levels of hemoglobin-derived proteins (Lewis *et al.* 2014). Armed with an extensive knowledge of metabolic pathways, we can further such studies to interpret genotypes to predict novel phenotypes.

Wild strains of the yeast *Saccharomyces cerevisiae* have proven to be useful models of natural genetic variation. Strains of *S. cerevisiae* are

approximately 1% divergent at the nucleotide level, and are phenotypically diverse. For instance, in low nitrogen conditions domesticated wine strains preferentially ferment glucose while natural oak isolates respire and then sporulate (Schacherer *et al.* 2009; Fay and Benavides 2005; Liti *et al.* 2009). Previous studies identified genetic variants that contribute both to the expression and sporulation differences between these natural isolates (Gerke *et al.* 2009), but the genetic variation that causes metabolome differences between wild strains is unknown.

We studied natural variation in the metabolome of *S. cerevisiae*. We used an untargeted mass spectrometry method (Fuhrer *et al.* 2011) to identify dozens of metabolite Quantitative Trait Loci (mQTL), genomic regions containing variation that control differences in levels of unknown metabolites between individuals. We mapped variation in urea cycle metabolites to genetic variation in specific genes known to regulate amino acid biosynthesis (Sophianopoloulou and Diallinas 2005; Dubois and Messenguy 1985). Our functional assays reveal that genetic variation in two genes, *AUA1* and *ARG81*, underlie the differences between individuals' abundance of several urea cycle metabolites. Drawing from knowledge of the urea cycle, we predicted and validated a novel phenotypic difference between strains. The interpretability of our results demonstrates the promise of mapping causal genetic variants underlying complex metabolic phenotypes and further using these variants to predict an individual's phenotype.

Material and Methods

Strains, growth conditions and metabolite extractions

We used 147 genotyped segregants derived from a previously described oak (YPS606) and wine strain (UCD2120) hybrid (Gerke *et al.* 2006). We engineered reciprocal hemizygotes by transforming strains with kanMX4 targeted to the gene of interest. For preparing extracts for mass spectrometry we grew strains overnight in synthetic dextrose media (0.145% yeast nitrogen base minus amino acids/ammonium sulfate, 0.5% ammonium sulfate, 2% dextrose) at 30°C. We diluted overnight cultures into 25 ml of SD media to an OD_{600} of 0.20. Cultures were grown in flasks at 30°C and 300 rpm until mid-log phase. We harvested cells by vacuum filter, and extracted hydrophilic metabolites from 0.2 um filters using 40:40:20 (v/v/v) methanol/acetonitrile/water (Lu *et al.* 2010). We froze and thawed extracts at -80°C and -20°C, respectively, three times. We pelleted cells, and stored the supernatant at -80°C until we performed mass spectrometry. We replicated growth for the parents fifteen times, segregants three times, and reciprocal hemizygotes nine to fifteen times (ARG81-oak::KAN, 9; ARG81-wine::Kan, 12; AUA1-oak::Kan, 15; AUA1-wine::Kan, 15). We randomized samples using a partial block design and extracted biological replicates at different times. To negatively control for non-biological ions specific to the extraction process, we extracted ions from seven samples containing neither media nor cell culture. To test Gap1p activity, parent strains were grown in SD media overnight at 30°C, washed with water, serially diluted and grown at 30°C for three days on SD media plus either 1% L-Proline or 0.1% ammonium sulfate, and with or without 0.16% D-Histidine (Sophianopoloulou and Diallinas 2005; Regenberg and Hansen 2000).

Flow-injection OrbiTrap MS and Data Processing

We directly injected metabolite extracts into a LTQ-Orbitrap Discovery Mass Spectrometer (Thermo Fisher Corporation) without a chromatography phase. The mobile phase for negative mode was isopropanol/water (60:40, v/v) buffered with 5 mM ammonium carbonate at pH 9, and the flow rate was 150 µL/min (fuhrer2011high). We used the R package MALDIquant (Gibb and Strimmer 2012) to process profile mode data. We used a square root transformation on each sample's spectra. For each sample, we removed the baseline and used total ion current to normalize the intensity. We used a signal to noise ratio of 5 and a half-window size of 3 to detect peaks in each sample. To compare peaks across samples, we aligned peaks using a warping function determined by MALDIquant. There were 478 detectable peaks. To eliminate non-biological ions, we filtered out ions that were at least half as abundant in the negative controls as the mean of the segregant samples. Additionally we ensured that our technical replication was within the standard coefficient of variance of 10% by creating a standard of the wine and oak parents mixed at equal amounts. We ran the same standard sample at least once for every fifty samples we ran on the mass spectrometer. We ran all of our samples over the course of four days. To confirm metabolite identity, we compared the candidate peaks of our standard to the profile of known metabolites using high performance liquid chromatography coupled with mass spectrometry.

Statistical and QTL analyses

To identify metabolites that are significantly different in abundance between the two parents, we used a mixed linear model (Bates and Maechler 2010) to describe metabolite abundance with batch as a random effect and genotype as a fixed effect (abundance ~ genotype + batch). To determine the significance of the genotypic effect, we compared our full model to a null model (abundance ~ batch) using a two-way ANOVA with a Benjamini-Hochberg adjustment (FDR = 0.1) (Benajamini and Hochberg1985). We calculated transgression and epistasis for all ions as previously described in Brem and Kruglyak (Gerke *et al.* 2006; Brem and Kruglyak 2005). We chose a conservative cutoff of three standard deviations for transgression to ensure we were not overestimating transgressive effects. We calculated epistasis T as a modified t-test: $T=\Delta/\sigma$, where Δ is the difference between means and σ is the variance. We calculated broad-sense heritability as $H2 = 1 - \sigma_e/\sigma_o$, where is σ_e the expected variance from the parents and is σ_o the observed variance in the segregants.

We used the R package qtl (Broman *et al.* 2003) to map quantitative trait loci to the abundance of metabolites. We permuted the data 1000 times to create a null distribution, and used an automated Haley-Knott regression to identify mQTLs with a 5% significance threshold. For ions with significant QTLs, we again permuted the data 1000 times to create a null distribution, and used composite interval mapping to identify weaker QTL peaks ($P < 0.05$), and used linear models to explain the variance in ion abundance due to candidate QTLs. If an mQTL mapped to multiple metabolites, we took the overlap of the mQTLs ranges for each metabolite, and mapped the overlapping mQTL regions to the *S.*

cerevisiae reference genome to identify candidate genes (Engel *et al.* 2014). We used linear models to calculate the variance due to specific mQTL.

We performed PCA on the segregants and parents using ornithine, glutamine, glutamate, citrulline and arginine as variables using **princomp()** function in R (R Core Team 2014). We used the eigenvectors to calculate broad-sense heritability (Gerke *et al.* 2006). We mapped QTL to the eigenvectors using composite interval mapping as described above. For negative controls we performed the same QTL analyses from all 99 metabolites and the twenty metabolites with individual mQTLs. Additionally we performed the analyses on randomly selected five metabolites from the twenty metabolites with mQTLs, and performed this analysis ten times. We used MANOVA in R to analyze the difference in the urea cycle in reciprocal hemizogytes, and one-way ANOVA in R to analyze the differences in metabolite abundance between reciprocal hemizygotes (R Core Team 2014).

Data Availability

Strains are available upon request. The raw mass spectrometry data obtained in this study will be accessible at the NIH Common Fund's Data Repository and Coordinating Center (supported by NIH grant, U01-DK097430) website, the Metabolomics Workbench, http://www.metabolomicsworkbench.org. Supplemental Table 1 contains a metabolite reporting checklist. Supplemental Table 2 contains processed averages for segregant metabolite abundance. Supplemental Table 3 contains data for transgression, epistasis and heritability. Supplemental Table 4 contains data for linear models of mQTL. Supplemental

Table 5 contains processed averages for parents metabolite abundance. Supplemental Table 6 contains processed data for mixed linear models of parents metabolite abundance. Supplemental Table 7 contains processed averages of reciprocal hemizygotes metabolite abundance.

Results and Discussion

High-throughput measurement of untargeted metabolites

We employed untargeted mass spectrometry to rapidly and systematically quantify abundances of unknown metabolites in natural isolates of the yeast *Saccharomyces cerevisiae.* Previous studies successfully identified causal genetic variation by targeting specific metabolites (Dubois and Messenguy 1985; Chen *et al.* 2014; Dong *et al.* 2015) or by untargeted metabolic analyses in individuals with known genetic variants (Broyart *et al.* 2009; Hu *et al.* 2014). As a complement to these approaches, we instead quantified unknown metabolites in minimally processed extracts by direct injection into a mass spectrometer (Fuhrer *et al.* 2011; Lu *et al.* 2010) (Supplemental Table 1). We chose to sacrifice the resolution of liquid chromatography coupled MS for the speed of the direct injection method, which allowed us to avoid the analytical challenges of mass spectrometer measurement drift over time, and more accurately measure metabolite abundances.

Using a stringent filter for reproducibility, we reliably measured the relative abundance of ninety-nine distinct ions (Supplemental Table 2). To control for non-biological ions, we ensured the ninety-nine ions were more than twice as abundant in the biological samples than the negative controls (Supplemental

Figure 1). To determine the reproducibility of the direct injection approach, we created a reference standard by extracting and pooling metabolites from two independently grown strains. We ran this standard eleven times over the course of the four-day run and determined the median coefficient of variance across biological metabolites was 10%, well within the range of acceptable experimental variation (Supplemental Figure 2) (Lu *et al.* 2010). This conservative analysis revealed that we can use untargeted methods to consistently measure the relative abundance of unknown, biological metabolites (Supplemental Table 2).

Complex genetic architecture underlying natural variation in metabolite differences.

Metabolite abundances are genetically complex traits with alleles that have both small additive and non-additive effects. To define the genetic architecture of metabolite levels we quantified the abundances of metabolites in 147 diploid segregants derived from a cross between a yeast strain isolated from the bark of an oak tree and a yeast strain isolated from a commercial wine barrel (Gerke *et al.* 2006). The continuous distribution of metabolite abundances in the segregants indicates that metabolite levels are controlled by many alleles of small effect (Figure 1A and 1B). We also found statistical evidence from the shape of the phenotype distributions for genetic interactions among alleles that influence metabolite levels, especially for metabolites that displayed transgressive segregation patterns (Figure 1C, Supplemental Table 3) (Brem and Kruglyak 2005). Thus alleles with small additive effects and alleles that display epistatic interactions contribute to natural variation in metabolite levels.

For more than half of all metabolites, abundance in some segregants was more than three standard deviations away from both parents' abundances, which is evidence for pervasive transgression (Brem and Kruglyak 2005) (Figure 1B, Supplemental Table 3). In the most striking examples sixteen metabolites had very low or undetectable levels in both parents, while 75% or more of segregants had high levels of the metabolite (Figure 1B and 1C, e.g. 159.08 m/z). This transgressive segregation pattern is consistent with the hypothesis that the wild parental strains contain compensatory alleles with both positive and negative effects on metabolite levels, which together maintain low levels of certain intermediate metabolites. Recombination of compensatory alleles during meiosis leads to the accumulation of high levels of metabolites in the segregants.

mQTL influence metabolites in the urea cycle

We next identified mQTL by correlating segregating polymorphisms with metabolite levels of unknown metabolites in the panel of segregants (Figure 2). We previously genotyped 225 markers in our 147 segregants (Gerke *et al.* 2009). We detected a genetically complex network of mQTLs with several mQTLs influencing the same metabolite, and several metabolites with multiple mQTL. In total we detected sixteen significant mQTL (Figure 2B) that contribute to the variation of twenty metabolites ($P < 0.05$). Seven mQTL are shared among two or more metabolites. Most metabolites have either one or two detectable mQTL, and four metabolites have either four or five detectable mQTL. To determine the fraction of the variance in metabolite abundance explained by mQTL, we used linear models (Supplemental Table 4). On average individual mQTL explain

11.0% of the variance in metabolite levels, with a range of 6.0% - 22.6%. As expected from our general transgression analysis, we found that of the seven metabolites with multiple mQTLs, four metabolites had mQTLs with effects in opposite directions. This finding further supports the hypothesis that parental strains contain compensatory alleles that maintain optimal metabolite abundances that are similar to each other. To determine whether contributions to mQTL are additive or non-additive, we analyzed the interactions of alleles. Our analysis of epistasis from the phenotype distribution of segregants suggested that interactions between QTL contribute to variation in metabolite levels (Figure 1C). Typically the additive contribution to QTL are larger than non-additive interaction. As the additive contribution of mQTL are small, we expected the effects of interactions to be even smaller. Consistent with this idea, linear models revealed one small but significant interaction term (Supplemental Table 4). Thus, metabolite abundances are largely shaped by many loci with small additive effects, and while interactions do play some role in shaping the distributions of ion abundances, most interaction effects are likely quite small.

We identified several segregating loci that impact urea cycle metabolism. We organized metabolites into pathways by determining the identities of metabolites with the most mQTL. After searching yeast mass spectrometry databases (Jewison *et al.* 2012) for candidate metabolites, we used traditional liquid chromatography coupled with targeted mass spectrometry to compare our samples to standards of these candidate metabolites. In this way we identified glutamine and citrulline as mQTL targets. As citrulline is produced during the

urea cycle, and glutamine biosynthesis is closely connected to the urea cycle, we searched for other possible candidates in the urea cycle (Jewison *et al.* 2012). We confirmed the identity of five metabolites in or adjacent to the urea cycle: citrulline, ornithine, arginine, glutamine and glutamate (Figure 2A). Our results demonstrate that several segregating genetic variants impact urea cycle metabolism and that our rapid untargeted method identified mQTLs that affect an important biochemical pathway.

Because segregating variation in the recombinant progeny influenced metabolites in the urea cycle, we predicted that the parental strains would harbor differences in urea cycle metabolism. Consistent with this prediction we found significantly different levels of citrulline, ornithine, glutamine and glutamate between the parents (Figure 2A, Figure 3A, Supplemental Tables 5 and 6). Notably, humans domesticated wine strains in low nitrogen environments, which may have provided selective pressure on the urea cycle, a nitrogen reclamation pathway (Marsit and Dequin 2015). Our genetic data reveal significant natural variation in urea cycle metabolism between strains from different ecological niches.

Our initial mQTL analysis assumed that each metabolite was independent, but metabolite abundances in the urea cycle are intrinsically linked to one another. Given that multiple metabolites in the urea cycle map to overlapping mQTL, we reasoned that combining metabolite measurements from the urea cycle would improve our power and allow us to narrow the linkage region. In other words, an mQTL could have effects spread across correlated metabolites,

and may have stronger effects on pooled measurements from correlated metabolites. We performed a principal component analysis on the segregants using the five amino acids in the urea cycle, and then remapped mQTL to these principal components (PCs). Using PCs as phenotypes increases the statistical power to detect QTL for correlated and variable data (Chase *et al.* 2002; Mangin *et al.* 1998). Four PCs explain 99.0% of the variance (Table 1). We calculated the broad sense heritabilities (H^2) of each PC, which measures the proportion of phenotypic variability due to genetic variation (Gerke *et al.* 2006; Brem and Kruglyak 2005). The first two PCs have low H^2, which indicates that the majority of phenotypic variability of intracellular metabolites is due to environmental effects. In contrast, PC3 and PC4 have higher H^2, supporting a genetic component to phenotypic variability in the urea cycle. As a negative control, we attempted to map mQTL to PCs derived from all ninety-nine metabolites, the twenty metabolites with mQTL, as well as five random metabolites with mQTL, but found no significant peaks. This suggests that the mQTL with the strongest genetic signal are specific to the urea cycle.

Causal variation in two genes underlies natural variation in urea cycle metabolites

Our pathway level analysis of metabolite abundances narrowed mQTL and revealed promising candidate genes. When we mapped mQTL to PC3 and PC4, we detected multiple QTL peaks (Table 2), two of which overlap with peaks mapped to individual metabolites and contain excellent candidate genes (Engel *et al.* 2014). One peak covers the gene *AUA1*. *AUA1* regulates amino acid

transport in the presence of ammonia, which is removed from the cell via the urea cycle. The wine variant of *AUA1* contains a premature stop codon, which truncates the eighty-four amino acid peptide, to just thirteen amino acids. The mutation rate (dN/dS) between strains is not higher than expected, which suggests the wine strain mutation is relatively new. Another QTL contains *ARG81*, a zinc-finger transcription factor that represses arginine biosynthesis. The number of nonsynonymous mutations between strains is not higher than expected, but when we analyzed our previously published expression data (Gerke *et al.* 2006), we indeed see differential expression of nine out of twenty-six *ARG81* targets, all of which show reduced expression in the wine strain.

We found that *ARG81* and *AUA1* contain causal variants for differences in the urea cycle. We tested our hypothesis that *ARG81* and *AUA1* contain causal genetic variation modulating urea cycle activity using reciprocal hemizygosity assays (Steinmetz *et al.* 2002). We used a multivariate ANOVA (MANOVA) to test whether the genotype of *ARG81* or *AUA1* has an effect across the whole urea cycle. We find that the genotype of *ARG81* has a significant effect on the abundance across all urea cycle metabolites (P=0.03), while the genotype of *AUA1* does not (P=0.23). When we split the MANOVA into separate components, we find the genotypes of *ARG81* and *AUA1* have significant effects on different individual metabolites. We found that the wine *ARG81* allele produces a higher abundance of ornithine than the oak allele, which matches the direction of the effect between the parents but not the QTL model (one-way ANOVA, P=0.03, Figure 4A, Supplemental Table 7). Although our QTL mapping did not detect an

effect of the *ARG81* peak on citrulline, the wine *ARG81* allele also produces a higher abundance of citrulline than the oak allele (one-way ANOVA, P=0.005). Arginine can passively turn into citrulline, which can explain the discrepancies of the mQTL and metabolite data. Additionally, the wine allele of *AUA1* produces a higher abundance of glutamine than the oak allele in the hybrid, which matches the direction of effect between the parents and the mQTL model (one-way ANOVA, P=0.02, Figure 4B, Supplemental Table 7). Both the difference in directionality between the hybrid and parental backgrounds and original mQTL mapping results suggest that there are other alleles that influence glutamine abundance. We conclude that *ARG81* and *AUA1* are novel mQTGs (metabolite Quantitative Trait Genes).

Predicting phenotype from genotype: A novel phenotype deduced from variation in the urea cycle

In principle, genetic variation in metabolism can predict new phenotypes. We hypothesized that both mQTGs control nitrogen metabolism by regulating the gene General Amino acid Permease (*GAP1*) (Figure 5A). *AUA1* post-translationally controls Gap1p by deactivating transport activity in the presence of a strong nitrogen source, such as ammonia (Sophianopoloulou and Diallinas 2005). In poor nitrogen sources such as proline, both *ARG81* and Gap1p are active (Sophianopoloulou and Diallinas 2005). We predict that the small 13 a.a. truncated version of the wine *AUA1* gene is effectively a null allele, which allows the wine strain to upregulate Gap1p to increase amino acid uptake. This would give the wine strain a selective advantage to continue fermenting instead of

sporulating in low nitrogen environments, such as a wine barrel. According to this model, under nitrogen poor conditions, Gap1p should be deactivated in the oak parent relative to the wine strain. To test this prediction we leveraged the fact that stereoisomers of L-amino acids, D-amino acids, are toxic to yeast and only enter the cell through Gap1p. If Gap1p activity is higher in the wine parent than the oak parent, then the wine parent will be more sensitive to the toxic amino acid D-Histidine (Sophianopoloulou and Diallinas 2005; Regenberg and Hansen 2000). Consistent with this prediction we found that the wine strain does not grow as well as the oak strain in the presence of proline and D-Histidine, indicating higher Gap1p activity in the wine parent (Figure 5B). Additionally, in the presence of a strong nitrogen source in which Gap1p is not induced, both parents grow similarly regardless of toxin, indicating that the phenotype is Gap1p dependent. This example demonstrates how linking metabolic pathways to mQTGs can lead to prediction of novel organismal phenotypes.

This work demonstrates the value of identifying genetic variation that underlies natural differences in the metabolome. We have presented a rapid approach for measuring metabolites in an untargeted fashion to systematically identify causal alleles controlling variation in a core metabolic pathway. Most importantly, by leveraging decades of biochemistry to interpret our results, we predicted and then validated a novel cellular phenotype from measured genotypes. The genes we identify as containing causal variation in the urea cycle are coherent with our existing knowledge of natural selection and metabolic pathways. With the current growth in metabolomics and genetics in human

391 studies (Wishart *et al.* 2013; Dharuri *et al.* 2014; Shin *et al.* 2014), similar
392 predictive methods can be used and tested in cell culture to further understand
393 how causal loci of one metabolic phenotype can affect other phenotypes, ranging
394 from metabolic biomarkers to drug sensitivity.

395 **Acknowledgements**

396 We thank Jeffrey Gordon for use of the mass spectrometer, Amy Caudy,
397 Heather Lawson and Gary Patti for advice and discussions, and members of the
398 Cohen lab for valuable feedback. This work was supported by a grant from the
399 National Institutes of Health (R01 GM092910-5).

400 **Author contributions**

401 DSL, PS, and BAC designed experiments. IN and JC developed and performed
402 mass spectrometry protocols. DSL, PS, DN and MVS ran experiments. DSL
403 analyzed data. DSL and BAC wrote the manuscript.

404 **Corresponding author**

405 The authors declare no competing financial interests. Correspondence should be
406 addressed to BAC (cohen@genetics.wustl.edu).

Literature Cited

Bates, D., *and* M. Maechler, 2010 *Package 'lme4.' http://lme4.r-forge.r-project.org/*

Benjamini, Y. and Y. Hochberg, 1995 Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* **57:** 289–300.

Brem, R. B. and L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. *PNAS* **102:** 1572–1577.

Broman, K. W., H. Wu, S. Sen, and G.A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19:** 889–890.

Breunig, J. S., S.R. Hackett, J.D. Rabinowitz, and L. Kruglyak, 2014 Genetic Basis of Metabolome Variation in Yeast. *PLoS Genetics* **10:** 1–15.

Broyart, C., J. Fontaine, R. Moliné, D. Callieu, T. Tercé-Laforgue, *et al*., 2009 Metabolic profiling of maize mutants deficient for two glutamine synthetase isoenzymes using 1H-NMR-based metabolomics. *Phytochem. Anal.* **21:**102–109.

Chase, K., D.R. Carrier, F.R. Adler, T. Jarvik, E.A. Ostrander, T.D Lorentzen, and K.G. Lark, 2002 Genetic basis for systems of skeletal quantitative traits: Principal component analysis of the canid skeleton. *PNAS* **99:** 9930–9935.

Chen, W., Y. Gao, W. Xie, L. Gong, K. Lu, *et al.*, 2014 Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature Genetics* **46:** 714-721.

Consortium, T. I. W. P. 2009 Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. *New England Journal of Medicine* **360:** 753-764.

Dharuri H., A. Demirkan, J.B. van Klinken, D.O. Mook-Kanamori, C.M. van Dujhn, *et al.*, 2014 Genetics of the human metabolome, what is next? *Biochim Biophys Acta* **1842:**1921-1931.

Dong, X., Y. Gao, W. Chen, W. Wang, L. Gong, *et al*., 2015 Spatiotemporal Distribution of Phenolamides and the Genetics of Natural Variation of Hydroxycinnamoyl Spermidine in Rice. *Molecular Plant* **8:** 111-121.

Dubois E. and F. Messenguy, 1985 Isolation and characterization of the yeast ARGRII gene involved in regulating both anabolism and catabolism of arginine. *Mol. Gen. Genet*. **198:** 283-9.

Engel, S. R., F.S. Dietrich, D.G. Fisk, G. Binkley, R. Balakrishnan, *et al.*, 2014 The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3* **4:** 389–398.

Fay J.C. and J.A. Benavides, 2005 Evidence for domesticated and wild populations of Saccharomyces cerevisiae. *PLoS Gen.* **1:** 66-71.Fernie, A.R., A. Aharoni, L. Willmitzer,

Fernie A.R., A. Aharoni, L. Willmitzer, M. Stitt, T. Tohge, *et al*. 2011. Recommendations for Reporting Metabolite Data. *Plant Cell* **23:** 2477-2482.

Fuhrer, T., D. Heer, B. Begemann, and N. Zamboni, N., 2011 High-Throughput, Accurate Mass Metabolome Profiling of Cellular Extracts by Flow Injection–Time-of-Flight Mass Spectrometry. *Anal. Chem.* **83:** 7074–7080.

Gauguier, D., 2016 Application of quantitative metabolomics in systems genetics in rodent models of complex phenotypes. *Archives of Biochemistry and Biophysics* **589:** 158–167.

Gerke, J. P., C.T. L. Chen, and B.A. Cohen, 2006 Natural Isolates of Saccharomyces cerevisiae Display Complex Genetic Variation in Sporulation Efficiency. *Genetics* **174:** 985–997.

Gerke, J. P., K. Lorenz, and B.A. Cohen, 2009 Genetic interactions between transcription factors cause natural variation in yeast. *Science* **323:** 498-501.

Gibb, S. and K. Strimmer, 2012 MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics* **28:** 2270–2271.

Hu, C., J. Shi, S. Quan, B. Cui, S. Kleessen, Z. Nikoloski, *et al*., 2014 Metabolic variation between japonica and indica rice cultivars as revealed by non-targeted metabolomics. Scientific Reports **4**.

Jewison T., V. Neveu, J. Lee, C. Knox, P. Liu, *et al.*, 2012 YMDB: The Yeast Metabolome Database. *Nucleic Acids Res*. **40**.

Keurentjes J.J.B., J. Fu, C.H.R. de Vos, A. Lommen A, R.D.Hall, *et al.*, 2006 The genetics of plant metabolism. *Nat. Genet.* **38:** 842–849.

Lewis I.A., M. Wacker, K.L. Olszewski, S.A. Cobbold, K.S. Baska, *et a*l., 2014 Metabolic QTL analysis links chloroquine resistance in *Plasmodium falciparum* to impaired hemoglobin catabolism. *PLoS Genet*. **10:** e1004085. doi: 10.1371/journal.pgen.1004085 pmid:24391526.

Liti G., D.M. Carter, A.M. Moses, J. Warringer, L. Parts, S.A. James SA, *et al*., 2009 Population genomics of domestic and wild yeasts. *Nature* **458:** 337-341.

Lu, W., M.F. Clasquin, E. Melamund, D. Amador-Noguez, A.A. Caudy, and J.D.

Mangin, B., P. Thoquet and N. Grimsley, 1998 Pleiotropic QTL analysis. *Biometrics* 54**:** 88–99.

Marsit, S. and S. Dequin, 2015 Diversity and adaptive evolution of Saccharomyces wine yeast: a review. *FEMS Yeast Research* **15:** fov067.

Pirmohamed, M., G. Burnside, N. Erikkson, A.L. Jorgensen, C. Hok Toh, *et al*., 2013 A Randomized Trial of Genotype-Guided Dosing of Warfarin. *New England Journal of Medicine* **369:** 2294-2303.

Rabinowitz 2010 Metabolomic Analysis via Reversed-Phase Ion-Pairing Liquid Chromatography Coupled to a Stand Alone Orbitrap Mass Spectrometer. *Anal. Chem.* **82:** 3212–3221.

R Core Team, 2014 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Regenberg, B. and J. Hansen, 2000 GAP1, a novel selection and counter-selection marker for multiple gene disruptions in Saccharomyces cerevisiae. *Yeast* **16:** 1111-9.

Schacherer J., J.A. Shapiro, D.M. Ruderfer, and L. Kruglyak, 2009 Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature* **458**: 342-5.

Shin S.Y., E.B. Fauman, A.K. Petersen, J. Krumsiek, R. Santos, *et al*., 2014 An atlas of genetic influences on human blood metabolites. *Nat Genet* **46:** 543-550.

Sophianopoloulou, V. and G. Diallinas, 2005 AUA1, a gene involved in ammonia regulation of amino acid transport in Saccharomyces cerevisiae. *Molecular Microbiology* **8:** 167–178.

Steinmetz, L. M., H. Sinha, D.R. Richards, J.I. Spiegelman, P.J. Oefner PJ, *et al.,* 2002 Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416:** 326–330.

Suhre, K. and C. Geiger, 2012 Genetic variation in metabolic phenotypes: study designs and applications. *Nature Reviews Genetics* **13:** 759–769.

Wishart D.S., T. Jewison, A.C. Guo, M. Wilson, C. Knox *et al*., 2013 HMDB 3.0 — The Human Metabolome Database in 2013. *Nucleic Acids Res*. **41(D1):**D801-7.

546 Table 1. Principal components for urea cycle

PC	Variance explained	Standard deviation	orn	cit	gln	glu	arg	H^2
1	47.5%	1.5%	-0.572	-0.540	-0.471	-0.385	-0.106	< 0
2	22.0%	1.5%	--	-0.377	--	-0.109	-0.916	0.008
3	19.4%	0.99%	0.430	0.359	-0.427	-0.678	0.212	0.28
4	9.9%	0.71%	-0.160	--	0.769	-.0614	--	0.25

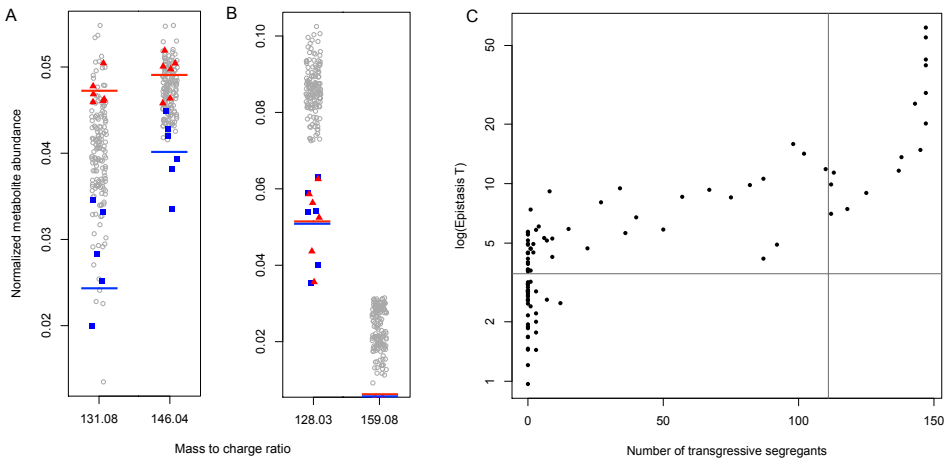
547

548 Table 2. mQTL for urea cycle principal components

PC	Chromosome	cM	Nearest marker	LOD	LOD P	Variance Explained	Additive Effect	Model P
3	6	25	L63	3.26	0.033	3.4%	19.3%	$8.4x10^{-3}$
	10	247	L1016	3.31	0.027	4.6%	-21.9%	$2.1x10^{-3}$
	13	5	L132	3.27	0.031	6.7%	26.2%	$2.5x10^{-4}$
	16	65	L165	5.43	<0.001	9.4%	-31.8%	$1.6x10^{-5}$
4	6	6	L61	6.62	<0.001	19.0%	-31.2%	$1.7x10^{-9}$
	11	11	L113	5.31	<0.001	15.8%	-28.0%	$2.9x10^{-8}$

549

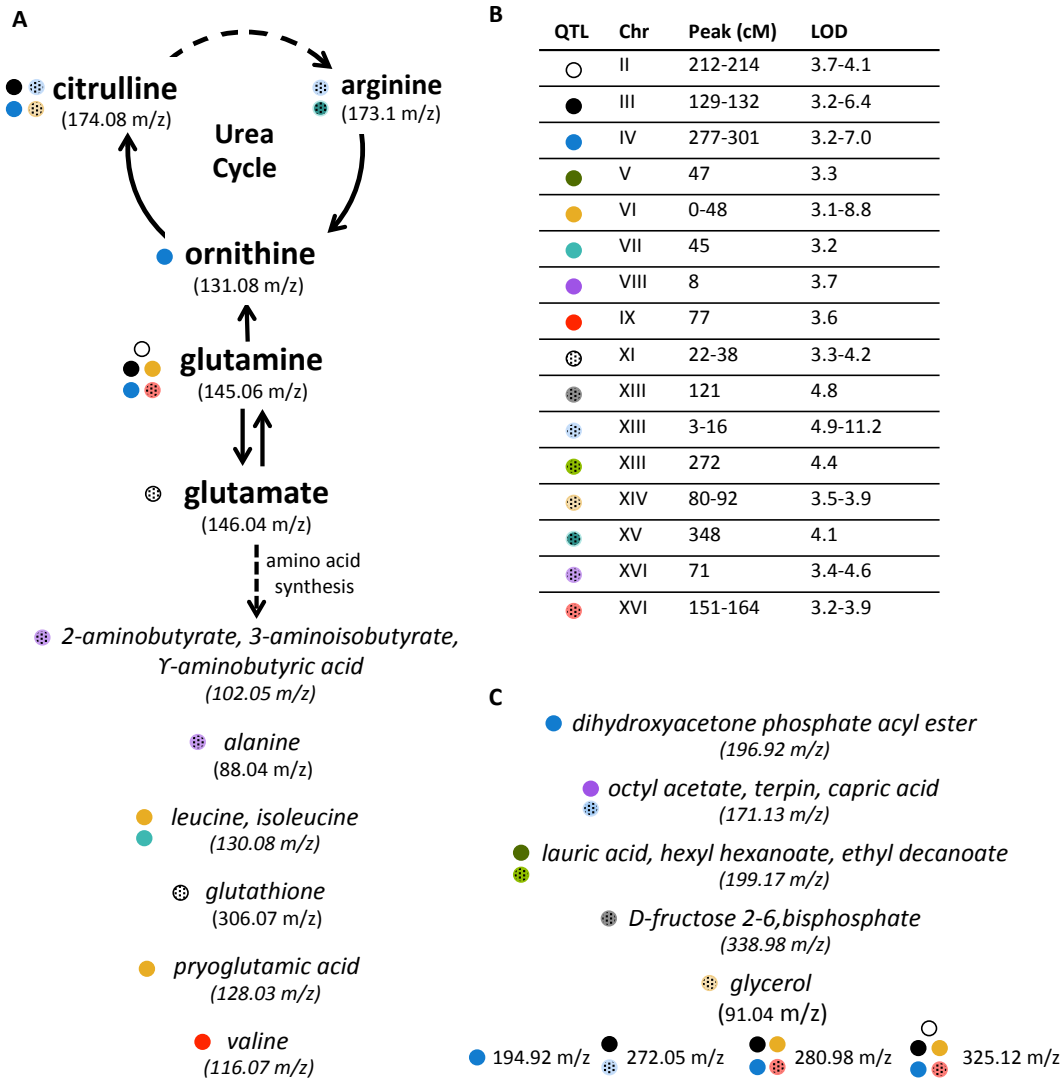
550



551

552 **Figure 1.** Metabolite abundances are complex traits. **(A)** Examples of ions for
553 which the metabolite levels of the segregants (gray) fall between the sample
554 means of the oak (blue) and wine (red) parents (45% of all metabolites) **(B)**
555 Examples of ions for which the metabolites abundance is more than three
556 standard deviations from the parental mean, indicating transgression. **(C)** For
557 each metabolite, the number of transgressive segregants is plotted against T, a
558 score for epistasis (Gerke *et al.* 2006; Brem and Kruglyak 2005). The horizontal
559 grey line indicates a significant T for epistasis. For sixteen metabolites, at least
560 75% of segregants are transgressive (vertical gray line).

561



QTL	Chr	Peak (cM)	LOD
○	II	212-214	3.7-4.1
●	III	129-132	3.2-6.4
●	IV	277-301	3.2-7.0
●	V	47	3.3
●	VI	0-48	3.1-8.8
●	VII	45	3.2
●	VIII	8	3.7
●	IX	77	3.6
●	XI	22-38	3.3-4.2
●	XIII	121	4.8
●	XIII	3-16	4.9-11.2
●	XIII	272	4.4
●	XIV	80-92	3.5-3.9
●	XV	348	4.1
●	XVI	71	3.4-4.6
●	XVI	151-164	3.2-3.9

Figure 2. Mapping of untargeted metabolites reveals 20 metabolites share sixteen QTL. Asterisks represent metabolites significantly different in abundance between parents. **(A)** The confirmed metabolite identity of five amino acids (bold font), which are involved in the urea cycle and nitrogen utilization. Candidate metabolites are italicized. Circles represent individual QTL contributing to metabolite abundance as listed in **(B)**. **(C)** Additional metabolites with at least one QTL.

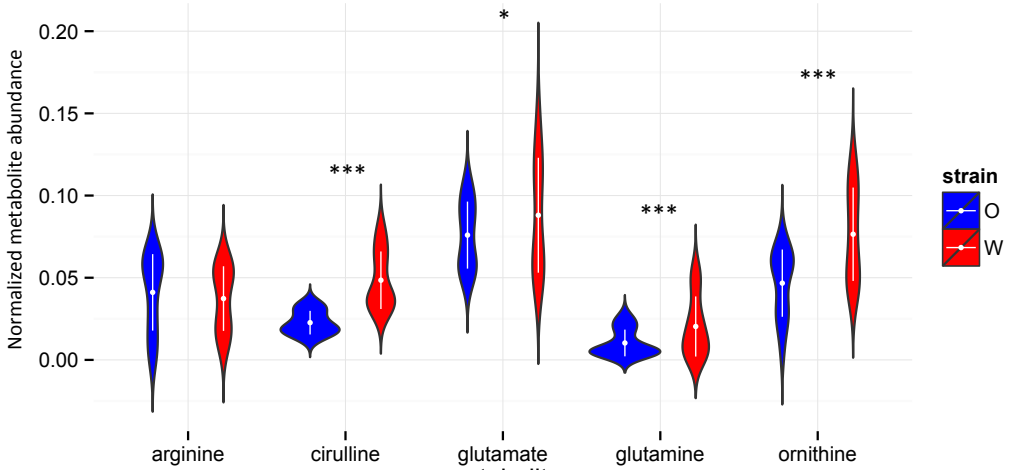
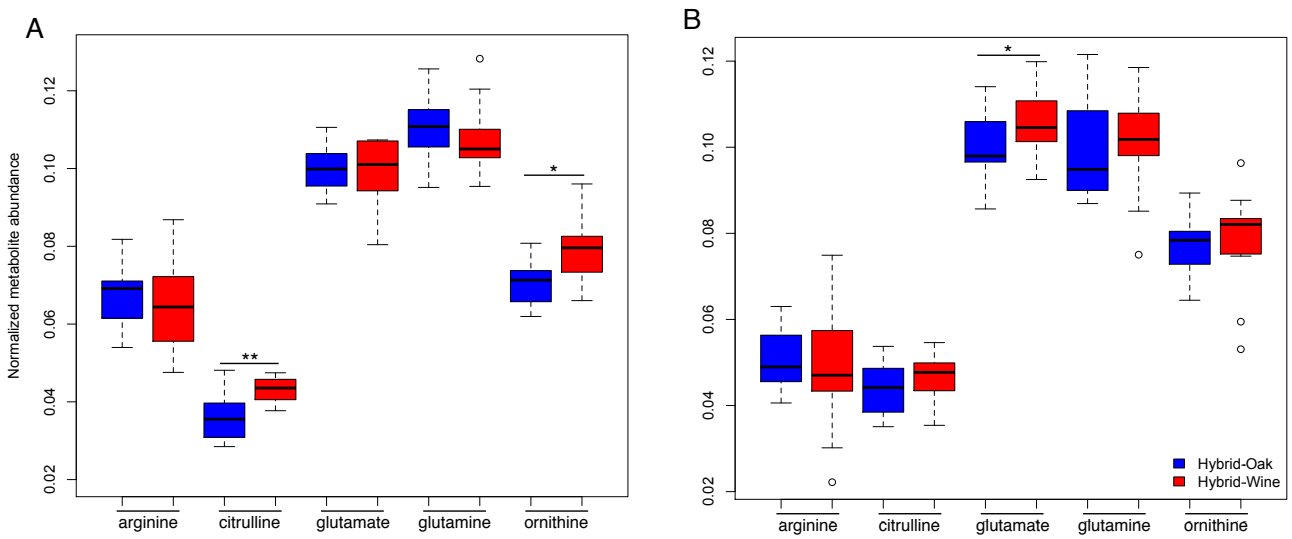


Figure 3. The abundance of urea cycle amino acids differs between parent strains. We measured metabolite abundance in fifteen biological replicates of the oak (blue) and wine (red) parents. White dots and bars represent the mean and standard deviation respectively. We used mixed linear models to measure the variance in abundance due to batch and genotype, and measured significance due to genotype. (*P < 0.05, ***P < 0.005, ANOVA, Benjamini Hochberg correction).



580

581 **Figure 4.** Reciprocal hemizygosity assays reveal causal variation in *ARG81* and

582 *AUA1*. Hybrid strains that contain only the wine (red) and oak (blue) allele of **(A)**

583 *ARG81* and **(B)** *AUA1*. The amino acids in the urea cycle are depicted. * P < 0.05,

584 ** P < 0.005 (one-way ANOVA).

585

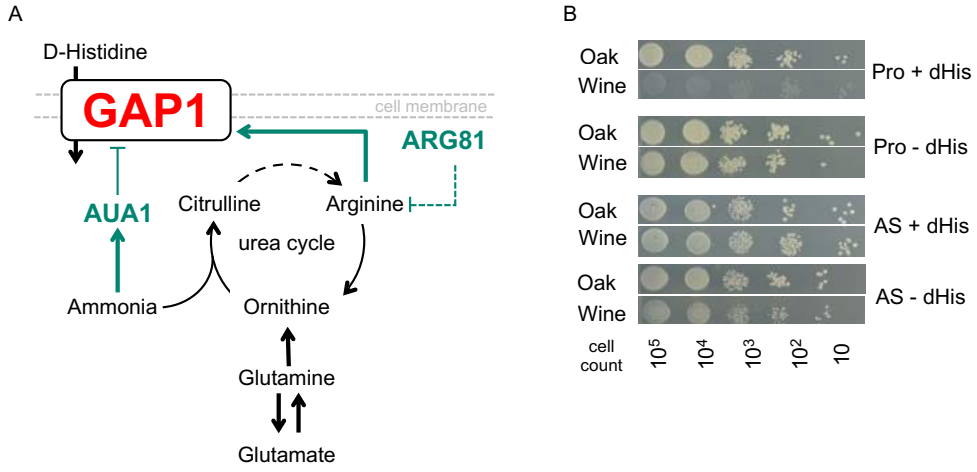
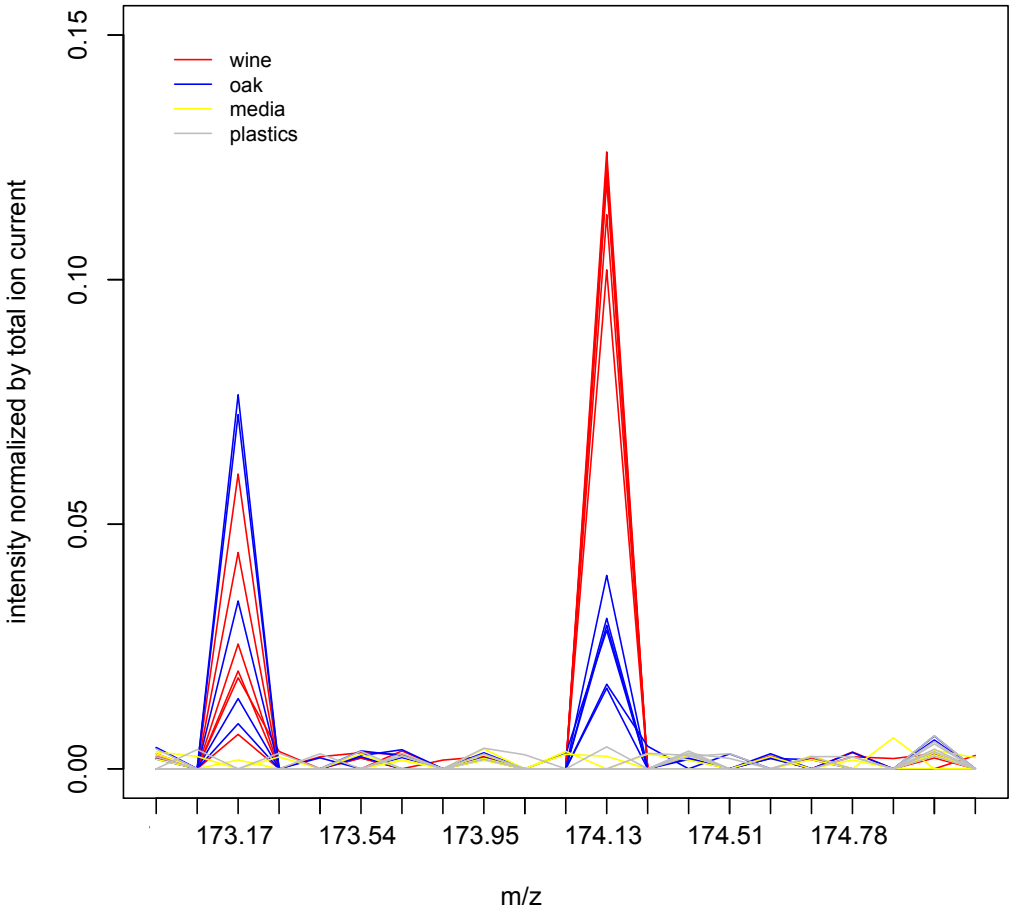
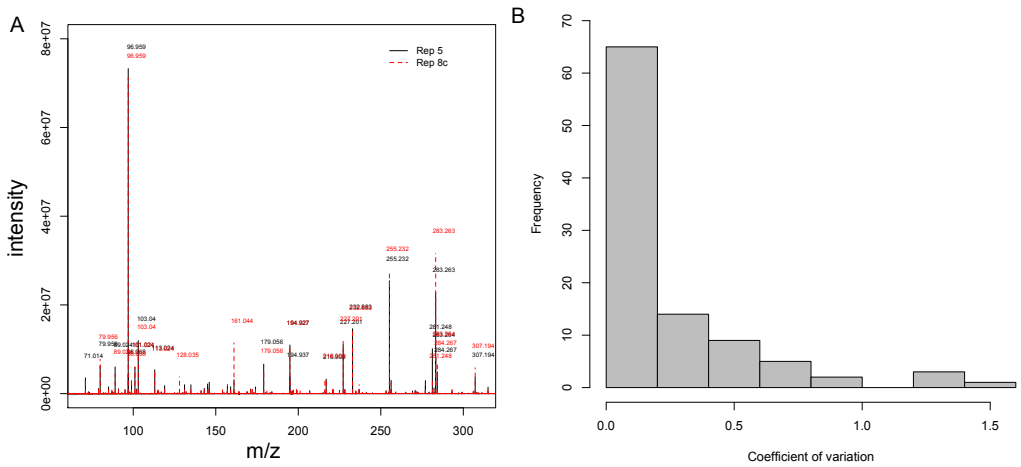


Figure 5. Genetic variation predicts novel drug sensitivity phenotype. **(A)** Model for how genetic variation in *AUA1* and *ARG81* impacts *GAP1* activity. The relative abundances of metabolites (black) induce regulators (green) to modulate Gap1p activity. We hypothesize that the wine alleles of *AUA1* and *ARG81* lead to increased activity of *GAP1*. **(B)** The wine strain does not grow as well as the oak parent in the presence of a poor nitrogen source, proline (Pro) and toxic D-Histidine (dHis), indicating that the wine parent has higher *GAP1* activity than the oak parent. *GAP1* is downregulated in the presence of a strong nitrogen source like ammonium sulfate (AS).



596

597 **Supplemental Figure 1.** Direct injection mass spectrometry captures biological
598 metabolites. We directly injected extractions into the mass spectrometer and
599 normalized reads by total ion current (red and blue lines depict biological
600 samples wine parent and oak parent respectively; yellow and gray depicts
601 negative samples without cells and without either media or cells respectively). To
602 focus on biologically relevant metabolites, we excluded metabolites whose
603 abundance in the negative controls were equal or greater than half of the mean
604 of all biological samples. We excluded the peaks depicted here except the peaks
605 at 173.17 m/z and 174.13 m/z.



606

607 **Supplemental Figure 2.** Direct injection mass spectrometry reproducibility. **(A)**
608 We mixed the oak and wine parents in equal amounts, and extracted the mix to
609 create a standard that was directly injected into the mass spectrometry at least
610 once every 48 runs. We show the raw spectra for two runs of the standard. **(B)**
611 Histogram of the coefficient of variation for each metabolite measured over 11
612 replicates of the standard**.** Frequency is the number of metabolites for a given
613 coefficient of variation.