1 Case study: multitrait GWAS of 41 growth traits in *Saccharomyces cerevisiae*

Things to elaborate on further:

- missing data and imputation: MAR, NMAR; One crucial parameter in the imputation strategy are the number of predictor variables one choses for the imputation of a given target variable. As a general rule, using every bit of available information yields multiple imputations that have minimal bias and maximal certainty (Meng 1995; Collins et al. 2001). This principle implies that the number of predictors should be chosen as large as possible. Including as many predictors as possible tends to make the MAR assumption more plausible, thus reducing the need to make special adjustments for NMAR mechanisms (Schafer 1997).

- ’multivariate imputations by chained equations (MICE) [**?**], which generates imputations for incomplete multivariate data by Gibbs sampling.

- The imputation itself was done by predictive mean matching (PMM) [**?**]. PMM is a semi-parametric imputation method which can preserve non-linear relations in the data.

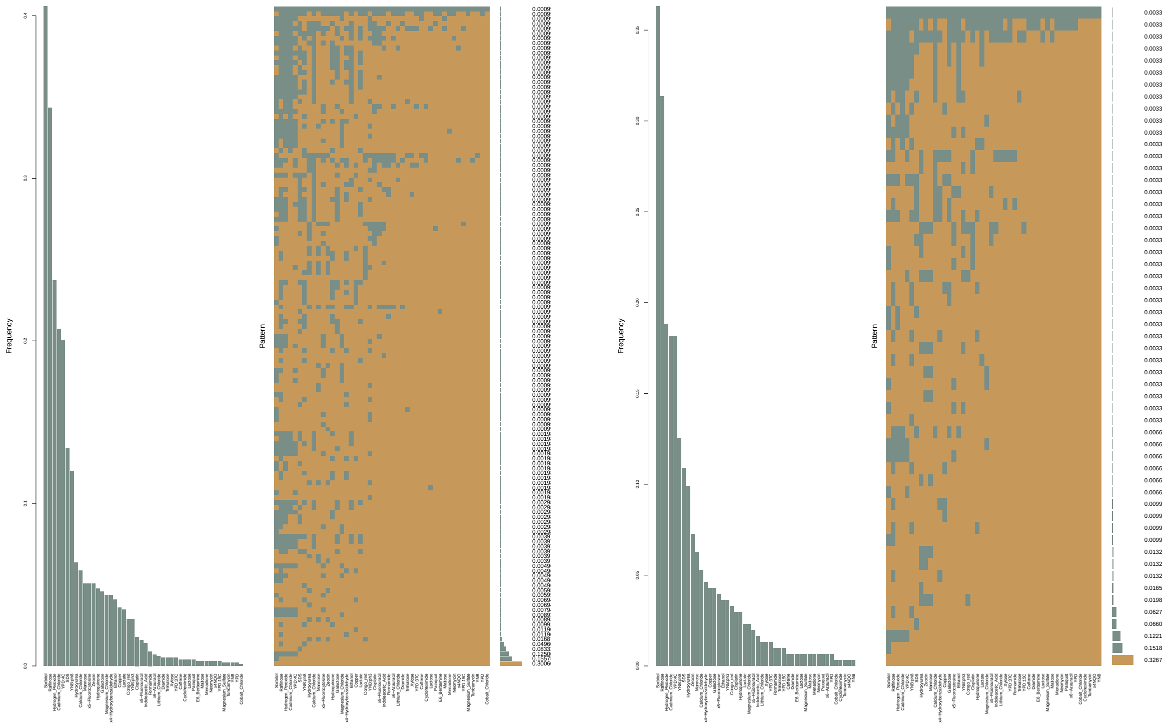
- multiple testing correction: Bonferroni, Benjamini and Hochberg, effective number of tests

The yeast dataset from a study by Bloom and colleagues [**?**] is used as case study dataset to show the feasibility of LiMMBo.

1.1 Data

Public dataset The dataset consists of phenotype and genotype data of 1,008 prototrophic haploid *Saccharomyces cerevisiae* segregants derived from a cross between a laboratory strain and a wine strain strains. It contains 11,623 unique genotypic markers obtained via short-read sequencing for all 1,008 segregants (no missing genotypes). For phenotyping, they grew segregants on agar plates under 46 different conditions, including different temperatures, pH and nutrient addition (see labels in Figure 2). The phenotypes were definded as end-point colony size normalized relative to growth on control medium. For the remainder of this chapter, a trait is defined as the normalised growth size in one conditions.

Phenotype imputation. Out of the 1,008 segregants, 303 were phenotyped for all 46 traits. Missing phenotypes are not evenly distributed with some traits such as cobalt chloride being present for almost all samples and others such as sorbitol or raffinose are lacking in more than a third of the samples. The distribution of trait missingness across all samples is depicted in Figure 1a.



(a) Full dataset **(b) Simulated dataset**

Figure 1: Frequencies and distributions of missing values in the yeast phenotype data. In both panels, the histogram (left) shows the frequency of missing values for a particular trait. The aggregation plot (middle) depicts all existing combinations of missing and non-missing values in the traits. The bar chart and numbers (right) show the frequencies of occurrence of the different combinations (R Package: *VIM* [**?**]).**(a) Full dataset.** The full dataset contains normalised colony sizes for growth in 46 different conditions of 1,008 genotyped yeast segregants as derived from [**?**]. 306 segregants are fully genotyped (bar chart, orange bar). **(b) Simulated dataset.** Fully-phenotyped dataset with simulated missing values.

The LMM framework relies on all samples being fully genotyped and phenotyped and does not accept missing values. In order to use the largest possible subset of the data, I investigated imputation strategies for the missing phenotypes. I used the subset of 303 fully phenotyped samples to determine traits suitable for imputation. I simulated data with a similar pattern of missingness as observed in the original dataset by subsampling the full dataset to the subset size and overlaying the observed missingness pattern onto the subset of 303 fully phenotyped samples. The resulting pattern is depicted in Figure 1b. Similar results in frequencies of fully phenotyped samples and combination of missing/non-missing traits can be observed when comparing it to the original frequencies and patterns (Figure 1a). I chose the MICE framework [**?**] with PMM as the imputation method to determine the most suitable imputation parameter settings in the simulated dataset which would then be applied to impute the real missing values in the full dataset. The predictor variables for each trait were determined based on its pair-wise Spearman's rank correlation coefficient ρ with all other traits in the dataset (Figure 2). In addition, only predictor traits that had been measured in at least 20% of the samples in the dataset were considered. Different predictor variable set-ups were examined based on increasing thresholds for the Spearman's Rank correlation coefficient: $\rho = \{0, 0.1, 0.2, 0.3\}$. Further parameters for MICE are the number of multiple iterations m (set to $m = 20$) and the number of iterations $maxit$ (set to $maxit = 30$). For each predictor set-up, MICE was initiated with the same seed for the random number generator to ensure comparability. The goodness of the imputation was evaluated by computing the correlation of the imputed values (averaged across iterations m) to the experimentally observed ones. Traits where the imputed values correlated to the original ones by more then 95% in at least one of the predictor set-ups were retained in the analysis. For five traits (cadmium chloride, hydrogen peroxide, raffinose, YNB:ph8, YPD:4C), no suitable predictors could be determined and these were excluded from further analyses (Figure 3, red labels). For each trait, the predictor scheme that yielded the highest correlation between the imputed and observed data was chosen for the imputation of missing values in the full dataset. Missing values were imputed in segregants that were phenotyped for at least 80% of the traits. The final dataset contained 981 segregants phenotyped for 41 traits.

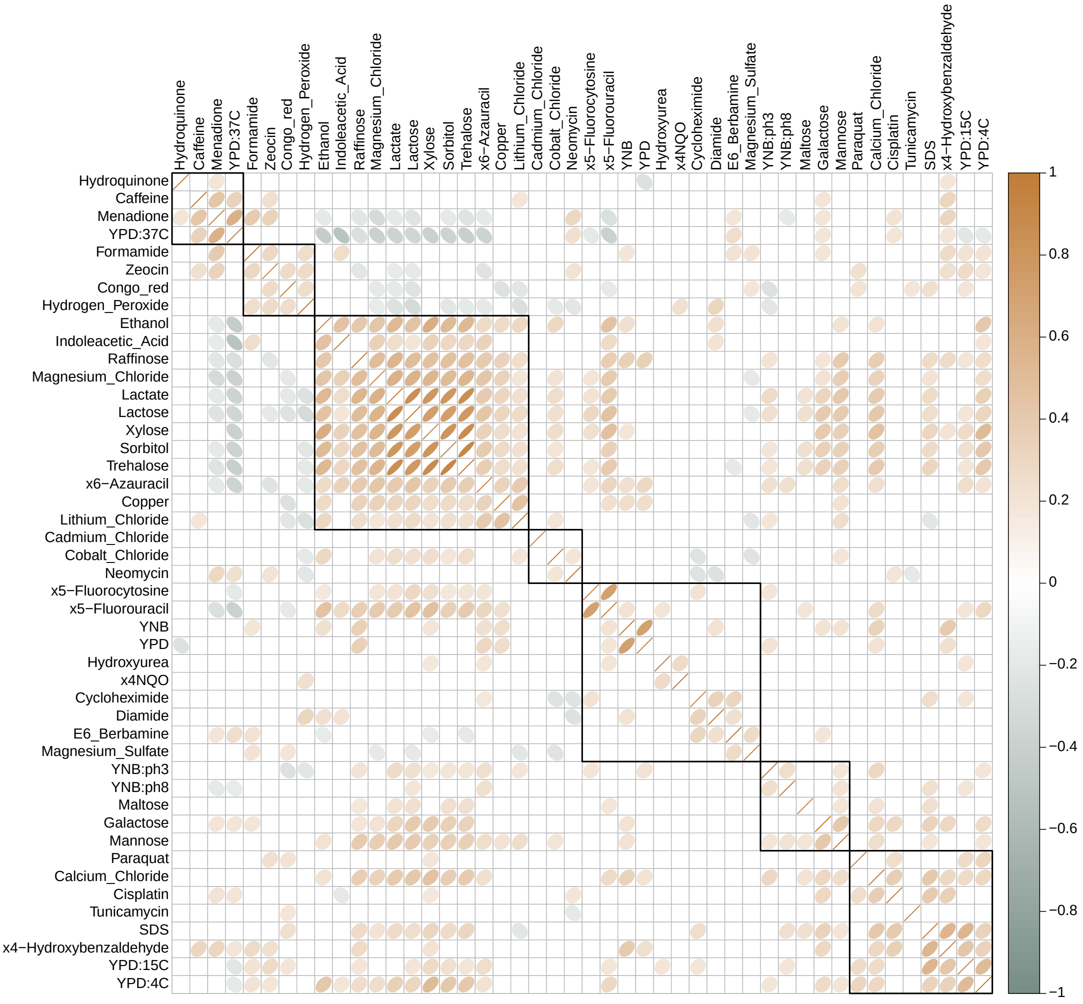


Figure 2: Pair-wise correlations of 46 growth traits in *Saccharomyces cerevisiae*. For each trait pair, Spearman's rank correlation coefficient ρ and the p-values of the correlation were computed. The p-values were adjusted for multiple testing according to Benjamini and Hochberg's method [**?**]. The strength and the direction of significant correlations ($p < 0.05$) are depicted above. Unsignificant correlations are left blank. The traits are clustered based on complete-linkage clustering of $(1 - \rho)$ as distance measurement (R Package: *corrplot*).

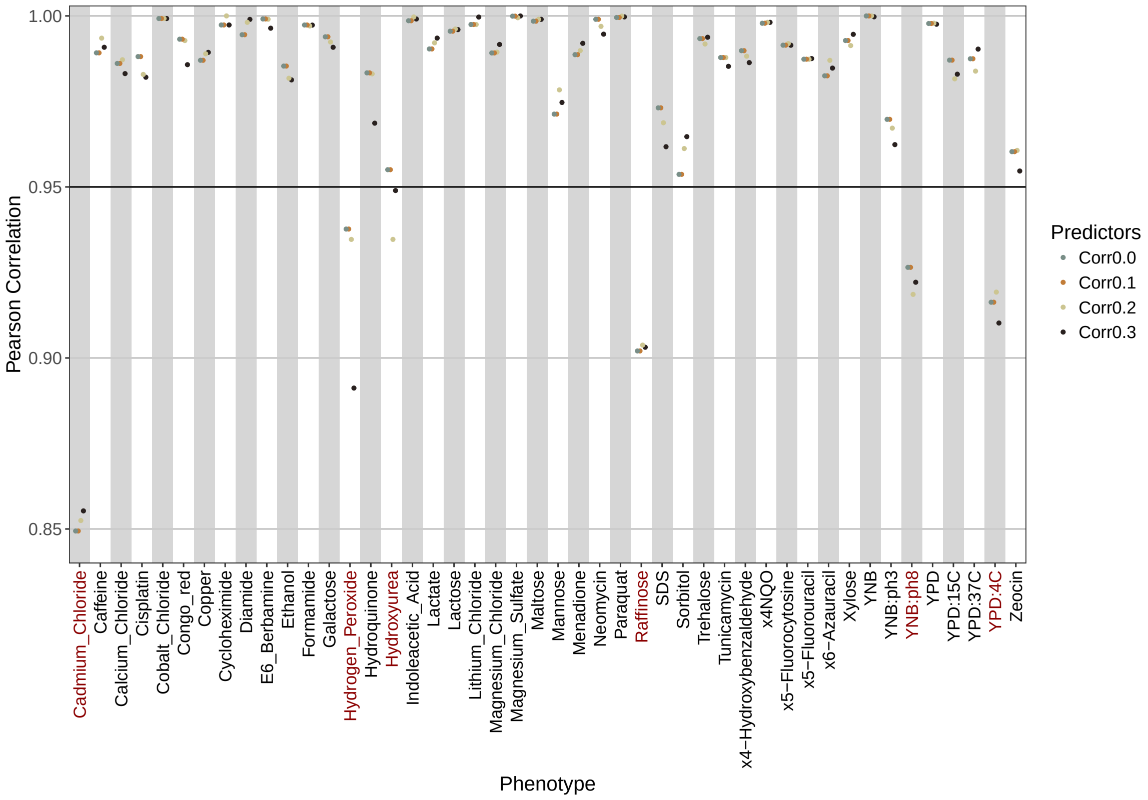


Figure 3: Correlation between imputed and experimentally observed trait values. In the subset of 306 fully phenotyped samples, missing values were introduced and subsequently imputed via MICE. Different predictor sets were tested, differing in the predictors traits included. Sets were constructed based on different Spearman's rank correlation coefficient: traits were considered predictors if their correlation with the target trait was greater than a given threshold. For each predictor setup ($\rho = \{0, 0.1, 0.2, 0.3\}$, $m = 20$ multiple imputations and $maxit = 30$ iterations of MICE were conducted. The goodness of the imputation was evaluated by computing the correlation of the imputed values (averaged across iterations m) to the experimentally observed ones. Traits with at least one correlation greater than the 0.95 (black vertical line) were retained in the dataset. For traits labeled in red, the imputation was considered to be unreliable and the traits were excluded from further analyses (R Package: *mice* [**?**]).

1.2 LiMMBo increases power in detecting genetic association in yeast

- Assessing significance of association in yeast crosses via permutations [??]

In order to test the performance of LiMMBo as a means for mtGWAS in a suitable set-up, i.e. in a cohort with related individuals, I conducted and compared an any effect mt-LMM-GWAS of 41 quantitative yeast traits to the results obtained in independent st-LMM-GWAS of the same traits. Figure 4 depicts the manhattan plot of both the st-LMM-GWAS p-values (adjusted for multiple testing by the effective number of tests [?]) and the mt-LMM-GWAS p-values. On several chromsomomes, mt-LMM-GWAS peaks (blue) are observed whereas no st-LMM-GWAS peaks (green) can be detected, demonstrating the increase in power by jointly modeling the traits. The heatmap below the mahattan plot shows the effect size estimates for each SNP for all 41 traits. Effect sizes were clustered based on their inner product across all SNPs.

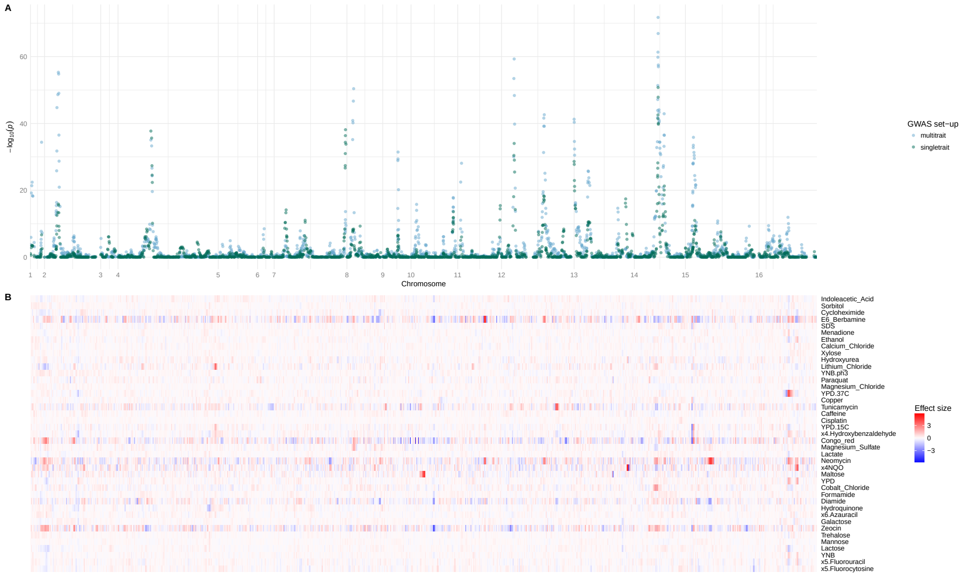


Figure 4: st-LMM-GWAS and any effect mt-LMM-GWAS of 41 quantitative traits in yeast. (a) Manhattan plot of p-values from mt-LMM-GWAS (blue) and st-LMM-GWAS (green). Single-trait p-values are the minimum p-value per SNP for any of the 41 singletrait GWAS, adjusted for multiple testing by the effective number of test ($T_{eff} = 33$). (b) Effect size estimates across the 41 jointly tested traits in the multitrait GWAS. Effect size estimate positions in the heatmap correspond the SNP positions in the manhattan plot.

References

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. & McVean, G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65. doi:10.1038/nature11632

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. & Zondervan, K.T. (2010) Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–73. doi:10.1038/nprot.2010.116

Delaneau, O., Marchini, J. & Zagury, J.F. (2012) A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–81. doi:10.1038/nmeth.1785

Delaneau, O., Zagury, J.F. & Marchini, J. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, 10(1):5–6. doi:10.1038/nmeth.2307

Howie, B., Marchini, J. & Stephens, M. (2011) Genotype imputation with thousands of genomes. *G3 (Bethesda, Md.)*, 1(6):457–70. doi:10.1534/g3.111.001198

Howie, B.N., Donnelly, P. & Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529. doi:10.1371/journal.pgen.1000529

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–13. doi:10.1038/ng2088

Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. (2012) Improved heritability estimation from genome-wide SNPs. *American journal of human genetics*, 91(6):1011–21. doi:10.1016/j.ajhg.2012.10.010

Teo, Y.Y., Inouye, M., Small, K.S., Gwilliam, R., Deloukas, P., Kwiatkowski, D.P. & Clark, T.G. (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics (Oxford, England)*, 23(20):2741–6. doi:10.1093/bioinformatics/btm443

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, 437(7063):1299–320. doi:10.1038/nature04226

UK10K Consortium (2014) UK10K Project. *http://www.uk10k.org*