# Chapter 3

# Genetic mapping with inferred traits

### Collaboration note

*This chapter contains work performed in collaboration with Dr. Oliver Stegle and Dr. John Winn for methods development. Oliver developed and implemented the sparse factor analysis model used in this chapter (Stegle et al., 2009), we then expanded on this work jointly (Parts et al., 2011). In particular, I applied the factor analysis model to simulated and real data, and performed the analyses of the results, including all association and interaction mapping. The coauthored manuscript forms the backbone of the chapter.*

Expressing RNA molecules is a highly regulated process that depends on activations of specific pathways and regulatory factors. Such state of the cell is hard to measure (Chapter 1.3.1), making it difficult to understand what drives the changes in the gene expression. To close this gap we apply a statistical model to infer the cell state variables, such as activations of transcription factors and molecular pathways, from gene expression data. We demonstrate how the inferred state helps to explain the effects of variation in the DNA and environment on the expression trait via both direct regulatory effects and interactions with the genetic state. Such analysis, exploiting inferred intermediate phenotypes, will aid

understanding effects of genetic variability on global traits, and help to interpret the data from existing and forthcoming large scale studies.

## 3.1 Expression analysis with cellular traits

Gene expression levels are determined by the state of the cell, as well as genotypes of the gene regulatory regions. A correct model for gene expression should incorporate both effects.

### Context-dependent genetic effects

Locus effects in isolation are not sufficient to account for gene expression variability (see also Chapters 1.2.2 and 1.3.2). Environment and intermediate cellular phenotypes (e.g. transcription factor or pathway activation) can and do have large effects on the measured transcript levels (Brem and Kruglyak, 2005; Gibson, 2008). To understand the genetics of gene expression, we must therefore analyse the consequences of genetic variants in the context of these other factors. Studies in segregating yeast strains have investigated epistatic interactions (Brem and Kruglyak, 2005; Storey et al., 2005), recovering interactions with genotypes of a few major transcriptional regulators. Large scale efforts to map functional epistasis between genes are currently underway with promising initial results (Costanzo et al., 2010). A recent study also searched for genotype-environment effects, and found many gene expression levels affected by an interaction between the environment and the genotype of a major transcriptional regulator (Smith and Kruglyak, 2008). However, much remains to be done in this area. While gene expression has been used as an intermediate phenotype to study the genetics of global traits (Schadt et al. 2005, Emilsson et al. 2008, Chen et al. 2008), genetics of gene expression itself has not been considered jointly with relevant cellular phenotypes such as pathway or transcription factor activations. This is an important gap. It is the state of the cell that determines how genetic variation can affect the gene expression levels, thus a joint analysis with the intermediate phenotypes should inform us about the mechanisms involved – a crucial step for understanding the causes of phenotypic variability.

### Inferring unmeasured cellular traits

Despite their importance, the intermediate phenotypes are usually not measured, thus genetic effects cannot be analysed in their cellular context. Fortunately, statistical approaches have been developed that allow inferring unmeasured factors which influence expression levels from expression data alone. Methods such as principal components analysis (Alter et al., 2000), network components analysis (Liao et al., 2003), surrogate variable analysis (SVA, Leek and Storey, 2007), independent components analysis (Biswas et al., 2008), and the PEER framework (Chapter 2) can be used to determine a set of variables that explain a part of gene expression variability with (usually) a linear model. Their application has been shown to increase power to find expression quantitative trait loci (eQTLs) by explaining away confounding variation (Leek and Storey, 2007; Stegle et al., 2010), and to yield variance components of the expression data that may be interpretable (Stegle et al., 2010).

### Our approach

Here, we perform a thorough joint genetic analysis of a gene expression dataset with intermediate phenotypes inferred from gene expression levels. We revisit the data of Smith and Kruglyak (Smith and Kruglyak, 2008), where the authors looked for gene-environment interactions affecting gene expression levels in a population of segregating yeast strains grown in two different carbon sources. First, we use a variant of a sparse factor analysis model (Rattray et al., 2009; Stegle et al., 2009) to infer intermediate phenotypes from the gene expression levels (Figure 3.1a). Importantly, this method uses prior information to guide the inference of which factors are affecting which genes, as opposed to unsupervised methods (e.g. PEER, SVA, ICA) that learn broad effects. We use Yeastract (Teixeira et al., 2006) transcription factor binding and KEGG (Kanehisa et al., 2002) pathway data as prior information in the model, which allows the inferred phenotypes to be interpreted as transcription factor and pathway activations. We then analyse the variation in the learnt activations, and find that growth condition and segregating locus genotypes have a strong influence (Figure 3.1b). Finally, for the first

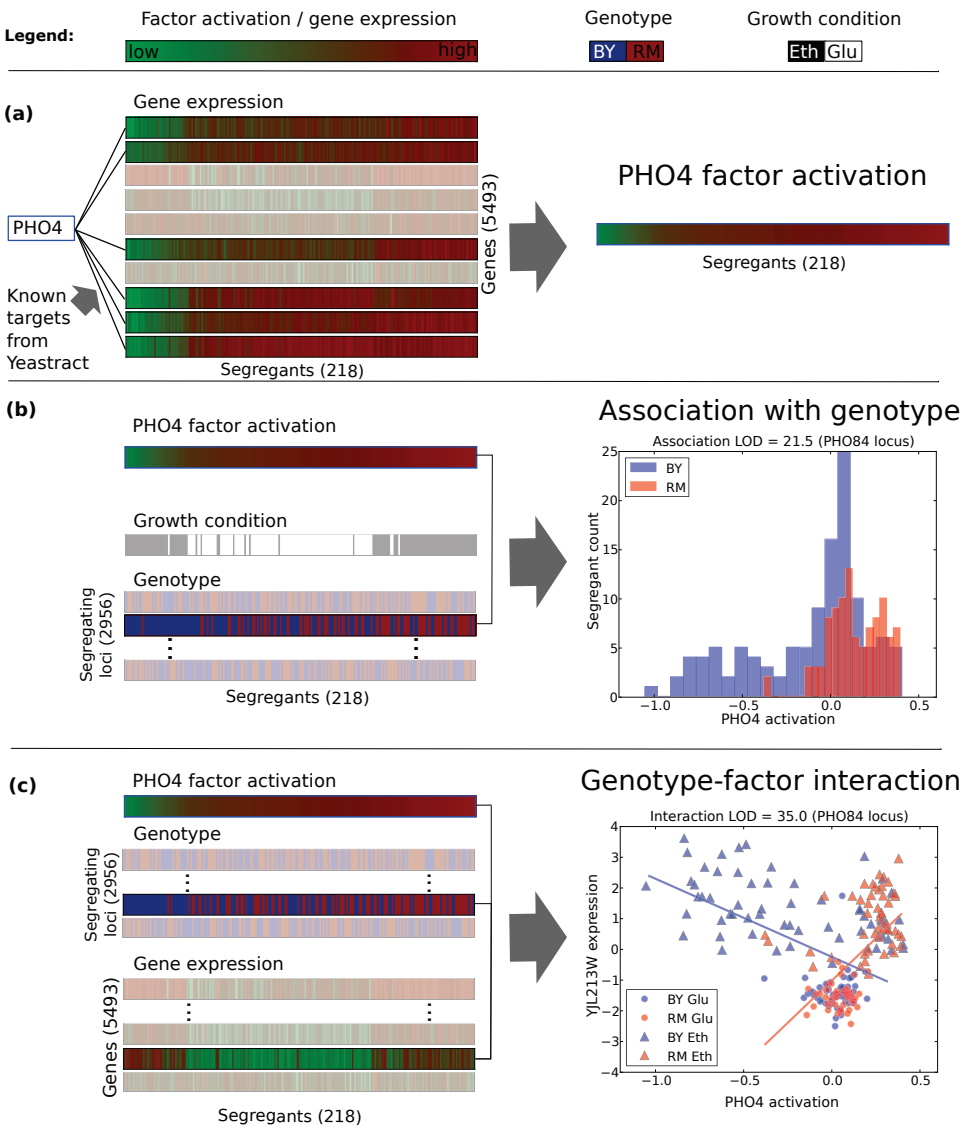


Figure 3.1: Analysing genetic effects in the context of intermediate phenotypes using *PHO4* as an example. **(a)** Intermediate phenotypes are learnt from expression levels using prior information from Yeastract database on the targets of the factor. The highlighted genes are known targets of *PHO4*. These activations are learned jointly for all factors. **(b)** The variation in intermediate phenotypes can be explained by locus genotypes or the growth condition of the segregants. For most loci (greyed out), the genotype is uncorrelated with the factor activation level. For the PHO84 locus at chrIII-46084, not greyed out and indicated by arrow, it is correlated. The plot at right shows the distribution of factor activations stratified by genotype at this locus. **(c)** Some genotypes show a statistical interaction with the inferred intermediate phenotype affecting gene expression levels, in this case YJL213W. See also Figure 3.3.

time, we consider genotype-dependent effects of the inferred intermediate phenotypes. We find genetic interactions with the inferred phenotypes that affect gene expression levels (Figure 3.1c), and identify hotspots in the genome that show an excess of these interactions. We show that many genotype-environment interactions are captured with the estimated intermediate phenotype, helping to interpret the environmental effect, and generate plausible, testable hypotheses for the mechanisms of several determined interactions. We propose that as pathway and transcription factor target annotations improve, our approach will produce even more useful intermediate traits that should be included in analysis and interpretation of high-throughput gene expression data.

## 3.2 Model of expression with unmeasured traits

We used a joint model of genotype and unmeasured trait effects on gene expression data, and used a two-stage inference procedure to estimate the individual effects.

### 3.2.1 Statistical model

The statistical model underlying our analysis assumes that the gene expression levels are influenced by effects of locus genotypes, intermediate factors, and interaction effects between them. These effects jointly influence expression variability in an additive manner, resulting in a generative model for expression $y_{g,j}$ of gene $g$ in individual $j$ of the form:

$$y_{g,j} = \mu_g + \underbrace{\sum_{n=1}^{N} \theta_{g,n} s_{n,j}}_{\text{SNP effect}} + \underbrace{\sum_{k=1}^{K} w_{g,k} x_{k,j}}_{\text{factor effect}} + \underbrace{\sum_{k=1}^{K} \sum_{n=1}^{N} \phi_{g,k,n} \left(s_{n,j} x_{k,j}\right)}_{\text{interaction term}} + \psi_{g,j}. \tag{3.1}$$

Here, $\mu_g$ is the mean expression level, $\psi_{g,j}$ the residual expression, and $\theta_{g,n}$ denote the weights of genotypes of SNPs $s_{n,j}$. The activations $\mathbf{x}_k = \{x_{k,1}, \ldots, x_{k,J}\}$ of $K$ intermediate factors are modelled as unobserved latent variables that linearly influence gene $g$ with weights $w_{g,k}$. Finally, the strength of interaction effects between factor $k$ and SNP $n$ is regulated by the interaction weights $\phi_{g,k,n}$.

On a second level of the model, the latent factor activations $\mathbf{x}_k$ may themselves be associated to the genetic state. Again assuming a linear model, these relations are cast as

$$x_{k,j} = \mu_k + \sum_{n=1}^{N} \underbrace{\beta_{k,n} s_{n,j}}_{\text{SNP effect}} + \epsilon_{k,j}, \tag{3.2}$$

where $\beta_{k,n}$ is the association weight and $\psi_{k,j}$ denotes the observation noise.

While appealing because of its generality, it is hard to perform joint parameter inference in the model implied by Equations (3.1) and (3.2). Here, we follow a two-step approach that yields tractable inferences and allows for statistical significance testing of the relevant factors contributing to the total gene expression variability (Equation (3.1)).

1. **Factor inference:** The latent factors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ and weights $\mathbf{W} = \{w_{g,k}\}$ are inferred from the expression levels alone, not taking the effects of SNPs $s_{n,j}$ via association and interaction into account.

2. **Association and interaction testing:** Significance of associations of factors to SNPs (Equation (3.2)) and SNP-gene-factor interaction terms (Equation (3.1)) are tested conditioned on the state of the inferred factors.

In this scheme, the factor inference is approximated as the contribution of direct SNP effects and interactions is not taken into account while learning. In this context, this approximation is well justified because of the relative effect sizes. The total variance explained by the interactions is small compared to the direct factor effects. If necessary on other datasets, this step-wise procedure could also be iterated, refining the state of the inferred factors given the state of associations and interactions.

### 3.2.2 Trait inference

Factors are inferred using a sparse Bayesian factor analysis model (Rattray et al., 2009; Stegle et al., 2009), presented here for completeness. Starting from the full model in Equation (3.1), the terms for direct genetic associations and interactions are dropped. The remaining factor model explains the expression profile

$\mathbf{y}_j = (y_{1,j}, \ldots, y_{G,j})^{\mathrm{T}}$ of the $G$ genes for segregant $j$ by a product of activations $\mathbf{x}_j = (x_{1,j}, \ldots, x_{K,j})^{\mathrm{T}}$ of the $K$ factors, and the $G$ times $K$ weight matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_G)$ and per-gene Gaussian noise $\boldsymbol{\psi} = (\psi_1, ...\psi_G)^{\mathrm{T}}$

$$y_{g,j} = \mathbf{w}_g \cdot \mathbf{x}_j + \psi_g. \tag{3.3}$$

The expression data $\mathbf{Y}$ is observed, and all other variables are treated as random with corresponding prior probabilities. The indicator variable $z_{g,k}$ encodes whether factor $k$ regulates gene $g$ ($z_{g,k} = 1$) or not ($z_{g,k} = 0$).

$$\begin{aligned} P(w_{g,k}|z_{g,k} = 0) =& \mathcal{N}(w_{g,k}\,;\, 0, \sigma_0) \\ P(w_{g,k}|z_{g,k} = 1) =& \mathcal{N}(w_{g,k}\,;\, 0, 1) \end{aligned} \tag{3.4}$$

The width $\sigma_0$ of the first Gaussian is small, driving the weight to zero. In experiments, we used $\sigma_0 = 10^{-4}$. This existing knowledge about whether a factor affects a gene, extracted from KEGG or Yeastract, is then encoded as a Bernoulli prior on the indicator variables $z_{g,k}$.

$$\pi_{g,k} = P(z_{g,k} = 1) = \begin{cases} \eta_0 & \text{no link} \\ 1 - \eta_1 & \text{link} \end{cases}. \tag{3.5}$$

The variable $\eta_0$ can be thought of as the false negative rate (FNR), the frequency at which prior information is incorrectly set to "no link". Similarly, $\eta_1$ is the false positive rate (FPR) of the observed prior information. We used $\eta_0 = 0.06$ and $\eta_0 = 0.0001$ for Yeastract and KEGG factors, respectively, and $\eta_1 = 0.001$ for both. The ratio of the false positive and false negative rate is motivated by relatively high false positive rates in chromatin immunoprecipitation experiments, and confidence in the KEGG annotations.

Prior probabilities over factors $\mathbf{X}$ are standard Gaussian distributed, $x_{k,j} \sim \mathcal{N}(0, 1)$, and the per-gene noise precisions $\tau_g$, $\psi_g \sim \mathcal{N}(0, \tau_g)$, are a priori Gamma distributed, $\tau_g \sim \mathrm{Gamma}(\tau_g \,|\, a_\tau, b_\tau)$. For the experiments this prior was set to be uninformative with $a_\tau = b_\tau = 0.001$.

Inference in the sparse factor analysis model is achieved using a hybrid of two deterministic approximations, variational learning (VB) (Jordan et al., 1999) and

Expectation Propagation (Minka, 2001), with exact details presented in (Rattray et al., 2009; Stegle et al., 2009).

#### Statistical identifiability of factors and model restarts

In general, factor analysis models are prone to suffering from intrinsic symmetries such as sign flips or factor permutations with impacts on the interpretability of obtained solutions. The informative sparsity prior of the factor analysis model (Equation (3.5)) substantially reduces these ambiguities, as it introduces constraints on possible factor configurations.

As an additional measure, our analysis explicitly takes the variability of factor solutions into account by analysing a set of inference solutions rather than a single point estimate. In the experiments, we performed 20 independent runs of the factor analysis model with parameters randomly initialised from their respective prior distributions, and used this whole ensemble to test for significant association and interaction effects.

### 3.2.3 Association and interaction testing

Following the generative model (Equation 3.1) we use standard association and interaction statistics (Lynch and Walsh, 1998) to test for associations between known variables (genotype of SNP $n$, environment indicator, or mRNA expression level) and the inferred factor activations. For completeness, we first present the model and test statistics used for both associations and interactions, followed by the significance testing approach. The derivation is developed explicitly using the SNP genotype as the known variable and factor activation as the dependent variable; tests for other covariates (or eQTL effects) are performed analogously.

#### Test statistics

We perform independent tests for association between the activation $\mathbf{x}_k$ of individual factor $k$ and genotype $\mathbf{s}_n$ of SNP $n$, fitting a liner model of the form

$$x_{k,j} = \mu_k + \underbrace{\beta_{k,n} s_{n,j}}_{\text{SNP effect}} + \epsilon_{k,j}, \tag{3.6}$$

assuming Gaussian observation noise $\epsilon_{k,j} \sim \mathcal{N}(0, \sigma^2_{k,j})$. For each pair of SNP $n$ and factor $k$, we calculate the association log-odds (LOD) score

$$L^a_{k,n} = \log P(\mathbf{x}_k \,|\, \beta_{k,n}) - \log P(\mathbf{x}_k \,|\, \beta_{k,n} = 0) \tag{3.7}$$

as a test statistic. The weight in the foreground model $\beta_{k,n}$, the mean $\mu_k$ and the noise level $\sigma^2_{k,n}$ are fit by maximum likelihood for every calculation.

Test statistics for the interaction terms are calculated analogously based on an independent interaction model. In short, we calculate the residuals of the factor analysis model and apply a standard interaction model between SNP $n$, factor $k$ and gene $g$. This corresponds to the linear model

$$y_{g,j} = \mu_g + \overbrace{\underbrace{\theta_{g,n} s_{n,j}}_{\text{SNP effect}} + \underbrace{w_{g,k} x_{k,j}}_{\text{factor effect}}}^{\text{direct effects}} + \underbrace{\phi_{g,k,n}\,(s_{n,j} x_{k,j})}_{\text{interaction term}} + \underbrace{\left[\sum_{l \neq k} w_{g,l} x_{l,j}\right]}_{\text{remaining factor effect}} + \psi_{g,j}, \tag{3.8}$$

where the expression level of gene probe $g$ for individual $j$ is described by fitted effects of the tested SNP $s_{n,j}$, learned factor activation $x_{k,j}$ and the interaction term $s_{n,j} x_{k,j}$ with the residuals explained by 0-meaned Gaussian noise $\psi_{g,j}$. The log-odds test statistic for the interaction between factor $k$ and SNP $n$ to influence gene $g$ follows as

$$L^i_{g,k,n} = \log P(\mathbf{y}_g \,|\, \phi_{g,k,n}) - \log P(\mathbf{y}_g \,|\, \phi_{g,k,n} = 0). \tag{3.9}$$

The respective mean variable $\mu_g$, weights $\theta_{g,n}$, $w_{g,k}$ (but not $w_{g,k'}$ where $k' \neq k$), and $\phi_{g,k,n}$, as well as noise variance $\psi_{g,j}$ are fitted independently using maximum likelihood for each factor, gene, SNP triplet. The contribution from all remaining factors is not refit to preserve the sparsity pattern learnt from the factor inference. To reduce the number of effective tests, we used the strongest interaction LOD score $\hat{L}^i_{g,n} = \max_k L^i_{g,k,n}$ across factors, thus performing tests for every SNP and gene pair. This approach corresponds to the assumption that at most a single factor is interacting with a given gene-SNP pair. The consistency of the strongest interacting factor is informative of the identifiability of the interaction effect (see below).

### Incorporating several random initialisations

For all our analysis of intermediate phenotypes, we generated factor inference results from $R = 20$ random initialisations of the model parameters to capture the variability in the model and avoid overfitting of inferred factor activations to local optima (See Statistical identifiability of factors below). Thus, we designed a significance testing scheme based on Q-values (Storey and Tibshirani, 2003) that employs the full set of runs, taking the uncertainty in the factor posterior distributions into account. We present this approach below for associations. Testing interactions is analogous except for the specifics of permutations highlighted in the text. In case of analyses where the multiple restarts are not used (e.g. eQTLs), we calculated Q-values from the single instance. In all cases, the null distribution of LOD scores was obtained by combining all calculated null statistics in the random restart.

**Q-value calculation** For every run $r = 1, \ldots, R$ of the factor analysis model, we evaluated the test statistics of factor associations ($L^a_{k,n}$) for every pair of factor $k$ and SNP $s$. This analysis was then repeated on 20 permuted datasets in each run with the genotypes shuffled with respect to the factor activations, while keeping individual segregants grown in two conditions paired. For interaction LOD scores, the factor activations and gene expression levels were not permuted with respect to each other. From this empirical null distribution of LOD scores in run $r$ (across all SNPs and factors), we calculated Q-values $q^r_{n,r}$ (local FDR) for each candidate association (Storey and Tibshirani, 2003) between SNP $n$ and inferred posterior of factor $k$ in this run.

**Combining Q-values** The Q-values from all runs were then combined into an overall Q-value $q_{k,n} = R^{-1} \sum_{r=1}^{R} q^r_{k,n}$, which was used to assess significance at a given FDR threshold.

From a probabilistic viewpoint, averaging Q-values over multiple restarts of the model can intuitively be thought of as integrating out the uncertainty from the factor inference. For example, for an association test assessing the significance of the weight $\beta_{k,n}$, we are truly interested in the probability of an association being absent (Bayesian Q-value, see for example (Storey, 2003)) given uncertain

inference of the factor activation $P(\mathbf{x}_k \,|\, \mathbf{Y}, \boldsymbol{\pi})$. Conditioned on the observed data $\mathbf{Y}$ and prior $\boldsymbol{\pi}$ this probability follows as

$$P(\beta_{k,n} = 0 \,|\, \mathbf{Y}, \boldsymbol{\pi}, \mathbf{s}_n) = \int_{\mathbf{x}_k} P(\beta_{k,n} = 0 \,|\, \mathbf{x}_k, \mathbf{s}_n) P(\mathbf{x}_k \,|\, \mathbf{Y}, \boldsymbol{\pi}). \tag{3.10}$$

In general this integral is not analytically tractable. Assuming we have instead a number of $R$ samples $\mathbf{x}_k^r$ from the factor posterior, the integral can be approximated by

$$\approx \frac{1}{R} \sum_{r=1}^{R} P(\beta_{k,n} = 0 \,|\, \mathbf{x}_k^r, \mathbf{s}_n) \tag{3.11}$$

in a Monte Carlo fashion. Finally, identifying the null probabilities as Bayesian Q-values we get

$$= \frac{1}{R} \sum_{r=1}^{R} q_{k,n}^r. \tag{3.12}$$

Note that the restarts from the factor analysis model are not exactly samples from its posterior but nevertheless characterise the posterior uncertainty sufficiently well (See also Simulation study below). Full MCMC sampling is computationally infeasible due to the size of the regulatory network; for a comparison of MCMC sampling and deterministic inference as employed here, see Stegle et al. (2009).

## 3.3 Phenotype inference

We inferred intermediate phenotypes on expression levels of 5493 genes from 109 yeast segregants grown in two environmental conditions (Chapter 3.3.2, Smith and Kruglyak (2008)). We performed the inference 20 times with different random initialisations of the parameters.

We considered three alternative types of prior information. First, we hypothesised the factors to be transcription factor activation levels, and used data for 167 transcription factors from Yeastract (Teixeira et al., 2006) to assign a prior probability of a factor affecting a gene expression level. Second, we hypothesised the

factors to be pathway activations, and used KEGG database information (Kanehisa et al., 2002) for 63 pathways for the prior probability of a link between a pathway activation and a gene. Third, for comparison, we employed an uninformative prior, where 30 factors were *a priori* equally likely to affect all genes. The datasets are described in more detail in Chapter 3.3.2 We call the inferred factor activations Yeastract factors, KEGG factors, and freeform factors, respectively.

### 3.3.1 Factor analysis model performance

In-depth comparison of inference approaches for the sparse factor analysis model used is given in other work (Stegle et al., 2009); the model was found to accurately recover factor activations in a setup similar to this study.

One way to further assess the reproducibility of the factor inference is to consider the correlation between the posterior means of individual factor activations. We called the inferred activation of factor $k$ in $u$-th run $\mathbf{x}_k^u = (x_{k,1}^u, ..., x_{k,J}^u)$ reproducible if its Pearson correlation $\rho(\mathbf{x}_k^u, \mathbf{x}_k^v) > 0.7$ for at least 16 of the 20 different $v$. 72 of 167 (31%) Yeastract and 19 of 63 (30%) KEGG factors were reproducible. While we explicitly took the variability between runs into account in further analyses, these numbers are instructive for developing intuition about the model.

### 3.3.2 Datasets used

For completeness, we provide specific details of the datasets used.

Gene expression data from (Smith and Kruglyak, 2008) (GEO accession number GSE9376) was downloaded using PUMAdb (http://puma.princeton.edu). In line with (Smith and Kruglyak, 2008), we considered spots good data if the intensity was well above background and the feature was not a nonuniformity outlier. Transcripts with more than 20% of missing values were discarded. All other missing expression values were replaced with the averages across the corresponding growth condition.

The remaining expression data consisted of 5493 probe measurements for 109 crosses of BY (laboratory) and RM (wild) strains grown in either glucose or ethanol, resulting in a total of 218 individuals. Strain genotypes were kindly

provided by R. Brem. Each of the 109 segregant strains was genotyped at 2956 loci to give a crude map of genetic background.

Transcription factor binding data was downloaded from Yeastract (Teixeira et al., 2006) (Version 1.1438) and contained binary indicators of binding between 174 transcription factors and 5914 genes. We considered the 3000 most variable probes whose corresponding genes were included in the binding matrix, and transcription factors that influenced at least 5 genes. After further discarding probes for which there were no data available, the remaining Yeastract prior dataset consisted of binding data for 167 transcription factors affecting 2941 genes.

Similarly, pathway information was downloaded from the KEGG database (Kanehisa et al., 2002). Only pathways with at least 5 genes were included in the network prior. This filtering procedure retained 63 pathways controlling 1263 genes. The results of Smith and Kruglyak (2008) were not used as a source of information for either of the prior datasets.

## 3.4 Association analysis with inferred phenotypes

First, we looked for the causes and consequences of variability in the inferred intermediate phenotypes.

### 3.4.1 Genotype and environment

Although the factors were inferred jointly from the expression data alone, many factor activations were significantly associated with a locus (SNP) genotype or indicator variable encoding growth in ethanol or glucose as a carbon source ("environment", Tables B.9 to B.11). Thirty two Yeastract factors were associated with a SNP genotype at false discovery rate (FDR) less than 5% and 26 with the environment. Similarly, 7 KEGG factors were associated with a SNP genotype, and one with the environment while 24 freeform factors were significantly associated with a SNP genotype and one with the environment. Some of the genotype associations were due to pleiotropic effects of single loci, while others were private to a locus-factor combination (Tables B.12 to B.14).

Many of these individual associations to Yeastract and KEGG factors can be interpreted by considering the role of the inferred factors and functional annotations of genes at associated loci. We now give some examples to further corroborate the use of factor activations as intermediate phenotypes. All associations are significant at 5% FDR, with corresponding Q-values $q$ (minimal FDR for which the association is significant (Storey and Tibshirani, 2003)) and average log-odds scores $L$ given.

**Yeastract factors.**

Loci associated with Yeastract factor activations encode genes functionally related to the corresponding transcription factor. The *PHO84* (an inorganic phosphate transporter) locus was associated with the *PHO4* (a major regulator of phosphate-responsive genes) transcription factor activation ($q < 0.03, L = 15.5$). The association implicates genetic variation in the transporter as a determinant of the transciptional activation of phosphate-responsive genes through *PHO4* activation. The mechanism of action is likely a switch in transcriptional response when *PHO84*, a high affinity phosphate transporter, is rendered ineffective by a mutation (Wykoff et al., 2007).

The *SUM1* (transcriptional repressor of middle sporulation-specific genes) factor activation was associated with the genotype of the *RFM1* (repression factor of middle sporulation) locus ($q < 10^{-5}, L = 115.2$). This is intriguing since *RFM1* recruits the *HST1* histone deacetylase to some of the promoters regulated by *SUM1* (McCord et al., 2003; SGD project), suggesting that genetic variation in the *RFM1* gene indirectly alters the effect of *SUM1* on individual genes.

There is also a straightforward eQTL that regulates the *HAP1* (heme activation protein) gene expression ($q < 10^{-5}, L = 80.6$), as well as factor activation ($q < 10^{-5}, L = 38.7$). This is a *cis* effect, since the locus is proximal to the gene, and manifests itself as a *trans* eQTL hotspot by affecting expression levels of some of the 170 known *HAP1* targets. Thirty four of the 84 (40%) significant *trans* eQTLs are also known targets of *HAP1*. Our data suggest that the other 50 may either be previously undiscovered targets of *HAP1*, or downstream effects of some of its direct targets.

The *THI2* thiamine metabolism transcription factor activation was associated with the genotype of the *THI5* locus ($q < 10^{-5}, L = 52.2$). This suggests a regulatory role of *THI5* upstream of *THI2* in thiamine biosynthesis for the previously poorly characterised *THI5* gene. This illustrates how our inference allows generating hypotheses for the function for genes that are implicated in a cellular pathway, but not annotated with a specific role.

**KEGG factors.**

Associations to KEGG pathways tend to capture the effect of a pathway component genotype. For example, two amino acid metabolism pathways are associated with locus genotypes of genes in the pathway. The inferred activation of lysine biosynthesis pathway was associated with genetic variation in the *LYS2* locus ($q < 10^{-4}, L = 25.6$), and the activation of arginine and proline metabolism pathway with the *ARG8* locus ($q < 10^{-5}, L = 46.7$), both members of the respective pathways. We thus hypothesise that variants in these genes directly affect the activation of the corresponding pathways. Also, the nitrogen metabolism pathway was associated with the *ASP3* (cell-wall L-asparaginase) gene cluster locus genotype. ($q < 10^{-5}, L = 119.9$). The *ASP3* genes are part of the pathway, and are present in four copies in the reference strain S288c, conferring increased resistance to nitrogen starvation stress. The inferred state of the pathway thus likely corresponds to the *ASP3* copy number via the locus genotype proxy.

Furthermore, the fatty acid metabolism pathway activation was associated with the *OAF1* (oleate-activated transcription factor) locus genotype ($q < 0.01$, $L = 67.1$), which is a known regulator of the pathway (Smith et al., 2007). We thus hypothesise that genetic variants in *OAF1* between the two strains are responsible for differences in fatty acid metabolism in the segregants, as has also been proposed in earlier work (Lee et al., 2009).

Finally, the environment is strongly associated to the very wide metabolic pathways category ($q < 10^{-5}, L = 393.2$). This KEGG entry comprises 619 genes, and captures the effect of the growth condition of the segregants on their metabolic state.

**Freeform factors.**

The freeform factors capture broad variance components in the data, with each factor's activation contributing to very many probe expression levels. Regardless of the unsupervised inference of the activations, they still show strong associations to environment and locus genotypes. However, due to this global nature of the factors, the associations are less straightforwardly amenable to interpretation. The first factor is associated with the environment ($q < 10^{-5}, L = 289.5$), and accounts for any mean shifts in gene expression levels between segregants grown in glucose and ethanol (Table B.11). Several of the other factors are associated with genotypes of "pivotal loci" described before (Brem and Kruglyak, 2005; Smith and Kruglyak, 2008; Yvert et al., 2003). It may be possible to make suggestions about the functionality via methods such as overrepresentation of GO categories within sets of genes with large weights for a factor, such as a recent study that performed a similar association analysis with unsupervised factors (Biswas et al., 2008). Our approach of using existing data for guidance is stronger compared to unsupervised methods as we use evidence of which gene is affected by the factor, thus improving statistical identifiability, and do not rely on an *ad hoc* choice of number of factors. This yields interpretable results that are more useful for generating hypotheses for the consequence of genetic or environmental variation.

Response to small molecule stress has been measured in the same segregants to map drug response loci (Perlstein et al., 2007). This study found eight QTL hotspots, six of which are within 20kb of loci that also show several associations to our inferred intermediate phenotypes (Tables B.12 to B.14), corroborating their pleiotropic effect.

### 3.4.2 mRNA and protein levels

Twenty five of 167 Yeastract factors were associated with the probe expression level measuring the mRNA abundance of the corresponding transcription factor gene (Table B.9, Figure 3.2). Twenty of the 25 (80%) were also significantly associated with a SNP genotype or environment. While statistically significant, these associations do not explain the majority of the factor variability, as only four

Yeastract factors were correlated with their probe expression level with Pearson $r^2 > 0.5$.

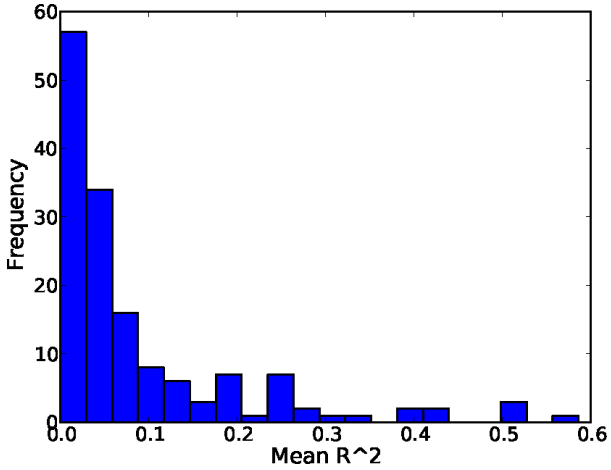


Figure 3.2: Pearson's correlation of Yeastract factors and their corresponding probe expression levels.

The general lack of correlation between factor activation and the corresponding measured expression level for the remaining transcription factors is perhaps not surprising. Presumably what matters for the factor activation is protein activity level, not mRNA abundance. Previous studies have found poor correlation between mRNA and protein expression levels (Foss et al., 2007; Gygi et al., 1999). Also, alternative mechanisms for activation exist. Many Yeastract factors without significant correlation to transcript levels have been shown to be activated not via increase in expression, but other means. For example, *PHO4* is activated by multiple phosphorylation events (Komeili and O'Shea, 1999). Simlarly, nuclear localisation and therefore activation of *ACE2* and *MSN2* are controlled by phosphorylation state (Goerner et al., 1998; O'Conallain et al., 1999). We predict most of the other transcription factors to also be activated by non-transcriptional means.

The protein level of one of the Yeastract factors, *GIS2*, has been assayed quantitatively in a previous study (Foss et al., 2007) for 87 of the 109 segregants we considered in a similar growth condition. For this transcription factor, the

inferred factor activation was better correlated to the protein level than the corresponding probe expression level for 15 of the 20 random initialisations. This example gives further support to treating the inferred factors as meaningful intermediate quantitative traits.

### 3.4.3 eQTL hotspots

As observed before (Brem et al., 2002; Smith and Kruglyak, 2008; Yvert et al., 2003) some segregating loci show significant associations with up to 271 (*IRA2*, regulator of the RAS-cAMP pathway locus) probe expression levels, forming *trans* eQTL hotspots. There are five such hotspots with at least 30 associations each. On average, 32% of the genes associated with a *trans* eQTL hotspot (FDR<5%) are explained by a transcription factor associated with the hotspot locus genotype targeting the gene (Table B.15). In 94% of these cases, the association with the inferred factor activation is stronger than with the locus genotype, and for three of the five hotspots, many additional associations with factor targets are recovered. For example, the *PHO84* locus is associated with the *PHO4* Yeastract factor activation ($q < 0.03, L = 15.5$), as well as 31 probe expression levels in *trans*. Eleven of these are also significantly associated with the *PHO4* factor activation, all showing a stronger association. *PHO4* itself is significantly associated with 454 probes, greatly expanding the range of plausible effects of the *PHO84* locus. This shows that using inferred intermediate phenotypes can reveal additional associations that otherwise would not be statistically significant.

## 3.5 Interaction analysis with inferred phenotypes

Beyond understanding the causes of variability in the inferred traits, we are also interested in their genotype-dependent effects on gene expression levels.

### 3.5.1 Discovering interactions

We scanned the genome for genotype-factor interactions that affect gene expression levels (Figure 3.1c) using a standard linear interaction model, and recovered three broad classes of interactions (Figure 3.3). We tested each locus-gene pair

independently for interaction with any inferred factor using 20 permutations, and information from all the random restarts of the model. If a single factor was observed with the strongest interaction score for a locus-gene pair in at least half the multiple restarts, we interpreted it as the true interacting factor; in other cases, we did not designate a factor to an interaction effect. We give examples of interactions we find below, highlighting how they add to the understanding of the propagation of the genotype effect.

The largest set of interactions was found at the *IRA2* locus. Many Yeastract factors, such as *MIG1*, *HAP4*, *YAP1* and *MSN2* show high interaction LOD scores with this locus (Figure 3.3a). All these corresponding transcription factors act in glucose response, nutrient limitation or stress conditions, which is consistent with the role of *IRA2* in environmental stress response by mediating cAMP levels in the cell. Their factor activations are associated with the environment (Table B.9), and the interactions thus recapitulate gene-environment interactions. While all these factor activations are correlated due to the strong association with the environment, making it hard to identify the true interacting factor, we can narrow the factor down to a few that exhibit strong LOD scores. Identifiability of the interacting factor is hard in general for factors that capture large effects, or have target sets that largely overlap with other factors. The inferred factors do capture the true underlying sources of variability, which is even more useful in settings where not all sources of variability are measured. Also, even having measured the relevant growth condition, we can further interpret the interactions as transcription factor activation having an effect in a specific genetic background in some cases, a more specific claim.

The *PHO4* factor activation is associated with ($q < 0.03, L = 15.5$) and interacts with the *PHO84* locus on chromosome XIII to influence 245 genes (Figure 3.3b). At the same time, the activation also interacts with the environment variable to influence gene expression levels. Notably, the statistical interaction for the *PHO4* expression, *PHO84* genotype and the same gene expression levels also has LOD scores greater than 11. Thus these interactions are not artifactual, but can be traced back to measured quantities for all interacting variables.

We also recovered epistatic interactions that failed the stringent multiple testing criteria on their own, but showed a stronger signal via the intermediate fac-

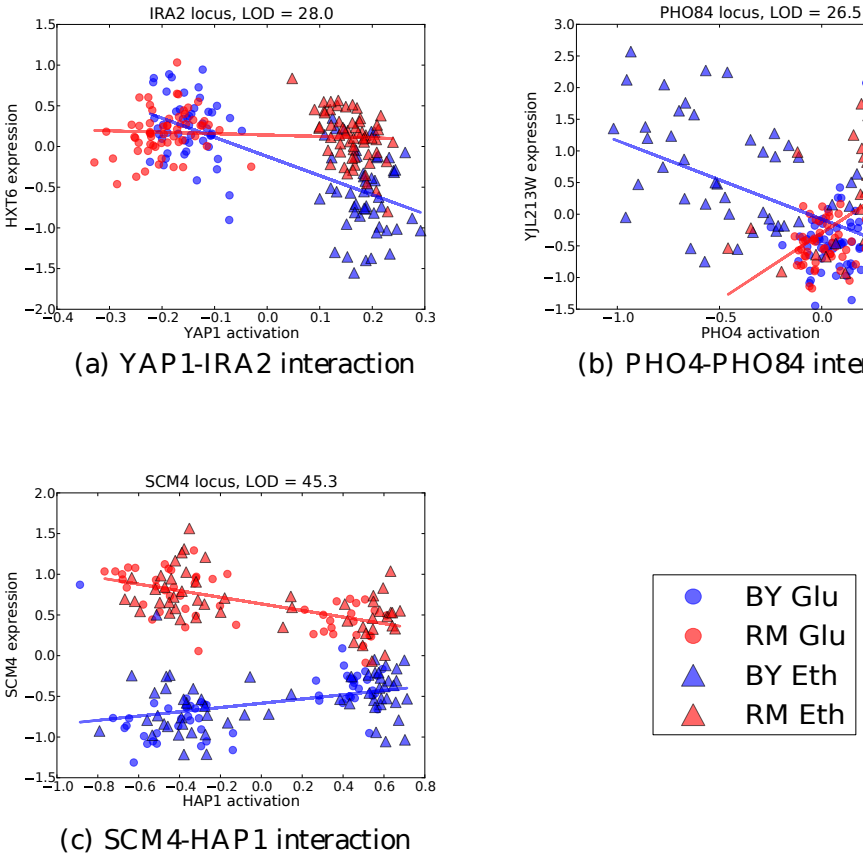


Figure 3.3: Three broad classes of interaction effects between locus genotype and transcription factor activation affecting gene expression (for details see text). Each marker shows the gene expression and factor activation for one individual segregant of either BY (blue) and RM (red) background at the locus, and grown in ethanol (triangles) or glucose (circles) as a carbon source. Maximum likelihood fits for expression data for the BY and RM segregants are plotted as solid lines; an interaction effect corresponds to a difference in slope in the two genetic backgrounds. **(a)** Genotype-environment interaction mediated by the inferred *YAP1* transcription factor activation. **(b)** Interaction between the *PHO84* locus and *PHO4* transcription factor activation, which is associated both with the *PHO84* locus genotype and the *PHO4* probe expression level. **(c)** Epistatic interaction between HAP1 and its target, SCM4, mediated by the HAP1 activation.

tor. For example, *HAP1* factor activation interacts with ($q < 0.01, L = 38.6$) the *SCM4* (suppressor of *CDC4* mutation) locus genotype to influence *SCM4* expression level (Figure 3.3c), while the epistatic interaction LOD score is only 7.9. As *SCM4* has a *HAP1* binding site in its promoter region, it is plausible that genetic variants could directly inhibit *HAP1* binding. This effect would only be observable in case *HAP1* is active, which in turn is controlled by the *HAP1* locus genotype ($q < 10^{-5}, L = 38.7$). This is an example of an epistatic interaction that is mediated by an intermediate phenotype of transcription factor activity.

In total, we found 2,397 genes with a gene-Yeastract factor interaction effect ($q < 0.05$). We also found 2,211 genes that show genetic interactions with KEGG factors and 2,250 with freeform factors. We noted several interaction "peaks" in the genome, such as the *IRA2* locus, where the locus genotype interacts with several genes via one or multiple factors (Figure 3.4). These coincide with *trans* eQTL peaks and gene-environment interaction peaks observed before (Smith and Kruglyak, 2008; Yvert et al., 2003), and have been annotated for potential causal genes.

### 3.5.2 Recovering interactions

We found 10,049 locus-environment interactions affecting 676 gene expression levels (Figure 3.4) using the same model and testing approach as for inferred factor interactions (FDR < 5%). Of these, we recovered 4605 interactions (46%) affecting 505 genes (75%) with the Yeastract factors, 6464 interactions (64%) affecting 572 genes (85%) with the KEGG factors, and 3065 interactions (31%) affecting 420 genes (62%) with the freeform factors. All environment-associated Yeastract factors had a strong interaction LOD scores with the *IRA2* locus, affecting hundreds of genes. These interactions recapitulate the gene-environment interaction reported and validated in the original analysis of the data (Smith and Kruglyak, 2008). It is reassuring that we are able to recover these interactions with the inferred intermediate phenotypes, and to expand their repertoire as well as provide hypotheses for their mechanism.

Preliminary results from an ongoing screen for gene-gene interactions have shown epistatic interactions for 95,445 gene pairs (Costanzo et al., 2010). Three

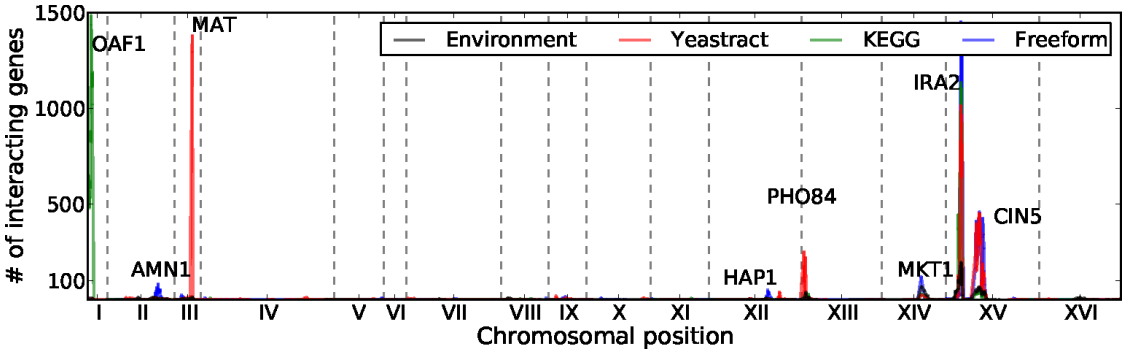


Figure 3.4: Number of genes affected by a genotype-factor interaction for each locus for Yeastract factors (blue), KEGG factors (red), freeform factors (green), and environment (gray).

hundred and sixty eight knockouts of a Yeastract factor gene and an interaction peak gene were tested in this large-scale assay, with 40 epistatic interactions found. We could find interactions for 22 of the tested pairs, and recovered one of the 40 interactions of Costanzo et al. (2010). Our screen is for genetic interactions that are different from the synthetic lethal screen of Costanzo et al. Consistent with this, we find some overlap, but not more or less than expected by chance.

## 3.6 Discussion

Our genetic analysis of the gene expression data from (Smith and Kruglyak, 2008) has shown that inferred intermediate phenotypes are valuable for generating hypotheses about plausible connections between genetic and gene expression variation. Using these inferred cellular phenotypes, we identified loci associated with transcription factor and pathway activations, thus giving the genetic effect a straightforward mechanistic interpretation, and often suggesting a candidate gene responsible for the change. For the first time, we considered and found statistical interaction effects with inferred intermediate phenotypes.

Our work is a step towards interpreting and understanding effects of genetic variants by putting them into cellular context. Conventional analysis, relating genotype and expression levels, is restricted to observed measurements, often

producing only statistical associations instead of a plausible mechanistic view. Going beyond this, our approach yields phenotypic variables at an intermediate level which can be used in the analysis. We showed that these provide additional interpretability and in some settings increase statistical power. Besides standard association and interaction effects between genotype and gene expression, our approach allows more rich hypothesis spaces to be explored, where the dependent variable we model is not a global organism phenotype such as disease label, or a very specific measurement like a single gene expression level. We have shown that this analysis is both feasible, and gives interesting results.

The idea of looking for associations and interactions with inferred intermediate phenotypes will be even more useful in forthcoming studies that include other cellular measurements. The inferred transcription factor or pathway activations allow interpreting the variability in these measured phenotypes as a result of changes in regulator activity or pathway state, bridging the gap between individual molecule measurements, and states of protein complexes, cellular machines, and pathways. We believe that the inferred intermediate phenotypes can be much more informative about the state of the cell and organism than individual locus genotypes and gene expression levels, and will also show stronger associations to downstream cellular and tissue phenotypes.

The intermediate activation phenotype has lower dimensionality compared to the space of genotypes and gene expression levels, which helps against multiple testing issues present in genome-wide scans for epistatic interactions. We were able to infer association and interaction effects, including proxies for epistasis, while finding epistatic interactions by testing all locus pairs is usually hindered by the billions of tests performed (Brem and Kruglyak, 2005; Cordell, 2009; Storey et al., 2005). The incorporation of prior information to infer interpretable factors is a flexible way to reduce the number of tests by capturing relevant parts of the data variation in a few factors, and can also add power if the factor is a better proxy for the true interacting variable.

The inferred transcription factor activations did not mostly correlate with their expression level. This is expected, as the activity of a protein depends on the protein level, localisation, posttranslational modification state, and existence

of binding partners to carry out its function. Expression level alone is often a poor proxy for a measure of protein activity.

A range of prior work has applied linear or generalised linear models to infer unobserved determinants of gene expression levels. For example, broad hidden factors have been inferred from gene expression that are likely to be due to confounding sources and hence can safely be explained away, thereby increasing the power of eQTL studies (Leek and Storey, 2007; Stegle et al., 2010). Although methodologically related, this work has a completely different aim. Also, unsupervised sparse linear models have been applied to infer hidden determinants in gene expression which are subsequently analysed for association to the genetic state (Biswas et al., 2008). This approach is closely related to the "freeform factors" included in this analysis for comparison. Overall, we show that factor learning taking prior knowledge into account adds statistical identifiability of the actual factors thereby providing interpretability. Other interesting approaches perform feature selection to capture relevant properties of the segregating sites in order to pinpoint the causative allele (Lee et al., 2009), or build a predictive (network) model of gene expression, followed by analysing its cliques and subnetworks (Zhu et al., 2008), but neither explicitly model unobserved phenotypes. A very recent paper proposed an integrated Bayesian ANOVA model that explains the gene expression profile by modules (Zhang et al., 2010). These modules in turn are modelled as a function of the genotype, taking direct and epistatic regulation into account. Importantly, both these related approaches infer gene expression determinants in an unsupervised fashion, and hence the interpretation of these association signals can be difficult and remains as a retrospective analysis step. Finally, a methodologically related sparse factor analysis model employing prior information has been applied to a narrower dataset with an aim to explain *trans* eQTL hotspots (Sun et al., 2007). However, the study does not consider the idea of genetic effects in the phenotypic context, or look for interaction effects, which is a primary focus of this work.

There has been speculation that a significant proportion of heritable variability that cannot be attributed to associations with single loci is due to interaction effects. This hypothesis is intuitively appealing, since we expect some genetic variants only to have an effect in a specific context. We have found an abundance

of such statistical interactions, and have shown how some of them help to understand and interpret yeast gene expression regulation. Often, they recapitulate epistatic or gene-environment interactions, but nevertheless add a plausible mechanism of action. It will be especially interesting and important to see how these methods work on large, extensively genotyped and phenotyped human cohorts that are becoming available in the near future.

An open source Python implementation of the statistical models and the analysis pipeline is available from ftp://ftp.sanger.ac.uk/pub/rd/PEER