Chapter 6

Conclusions

6.1 Conclusions and discussion

I have spent the last four years trying to understand the genetic basis of cellular traits, focusing mostly on genetics of gene expression, and trait mapping by selection.

6.1.1 Abundance and importance of eQTLs

We found genetic associations to 30% of the transcript levels, and preliminary results suggest that this number increases to 85% in larger human cohorts. Thus, gene expression level, the most basic cellular phenotype, is certainly influenced by genotype. While it is not surprising that common genetic variation in gene regulatory regions does influence the structure and protein binding affinities of DNA, still only a small amount of variance is explained by genotype.

We (Chapter 3) and others (Foss et al., 2007) have found evidence for little correlation between mRNA and protein levels in the cell. This suggests that small scale gene expression variation is not amplified to protein levels, and rather dampened. This supports the view that many of the eQTLs we find may not have functional consequences in the tissue the gene expression was assayed. However, in other tissues, the effects may be larger.

Strong QTLs, however are candidates for causal regulatory effects, including human GWAS hits (Nica et al., 2010; Nicolae et al., 2010). It could also be

that some of the weak eQTL alleles tag rare variants of large effect, that are not detected with the non-parametric methods commonly used. Furthermore, weak eQTLs in one tissue could be strong in another, where the genes are expressed at higher levels. While a comprehensive resource of human average tissue-specific expression levels has recently been published (Lukk et al., 2010), there is no comprehensive data available yet on the variability of expression levels in all tissues, and their genetic basis. Some projects, e.g. the MuTHER resource (Nica et al., 2011) are beginning to fill this gap.

Some dimensions of eQTL mapping remain understudied. First, there is a question of identifying the causative nucleotide(s). Functional annotation of the loci can help find variants that are in known or predicted protein binding regions, and therefore predicted to be functional. Further correlating the existence of a binding site with expression from the haplotype, which can be possible in mRNA-seq experiments, will give more evidence for the functional impact. Second, from studying cell populations, it is not clear whether the change in expression levels is due to a shift in the mean level in every cell, or a large change in a smaller number of cells. The problem of differentiating between small effects of full penetrance and large effects of low penetrance is fascinating and important, and pertinent to most cellular traits. I hope that large condition-specific effects are at play, as these will be easier to model, once appropriate assays have been undertaken.

6.1.2 Abundance and importance of interactions

We have developed methods to detect genotype-specific effects on cellular traits (Chapter 3). We found that modelling unmeasured cellular phenotypes lowers the dimensionality of the hypothesis space, eases the multiple testing burden, and yields interpretable genetic associations and interactions.

There is a gap between intuition developed in model organisms and findings from studying human cohorts. In models, epistasis has been found almost everywhere, while in humans, it remains elusive. This is a secondary problem. The real cellular mechanism for an epistatic effect is not DNA-DNA interaction, but instead, an interaction of two traits. Thus, the question is not where are epistatic

effects, but rather which traits we need to measure or infer to understand variation in our favourite trait. The genotype then comes into play only as a source of variability for the interacting trait.

We have explored one direction of such interactions, genotype-specific transcription factor and pathway effects. We rely heavily on existing annotations for structuring our model, and have to use inferred phenotypes, as the traits we are really interested in are not measured. Thus, obvious extensions to existing work would include more detailed prior information, as well as modelling other measured traits. I believe that genotype-specific effects are also pervasive in humans, and will be detected using inferred intermediate phenotypes, or assays of further cellular phenotypes.

6.1.3 Trait mapping using artificial selection

We have established a method to map any selectable trait in yeast to narrow intervals, and found many loci to be contributing to heat resistance.

It is not surprising that many loci are responsible for variation in one trait. While simple characteristics can be determined by one specialised protein such as efflux pumps of specific molecules, most cellular traits are determined by the action of entire pathways. Thus, variation in any part of the pathway that also effects its activity will be a source of variability for the phenotype. Any allele that affects a pathway component and has downstream effects will be under selection if pathway activation is selected for.

The lack of fixation of individual alleles is also explained by selection for pathway activation. Once the activation is perturbed enough to produce a fit individual, genotypes of other alleles have no fitness effect, and are under no selective pressure. Alternatively, once enough deleterious alleles are present to abolish the pathway activation and produce an unfit individual, the additional alleles will not influence the fitness of the individual further. Such effects correspond to negative epistatic interactions, and we are validating whether they explain our observations (see Future work below).

It has been reassuring to observe genetic complexity in a simple model organism. Hopefully, much of what we learn about the fine scale structure of complex yeast traits mapped by artificial selection will also translate to higher eukaryotes.

6.2 Future work

The development of the PEER framework (Chapter 2) and the interaction model (Chapter 3) is complete. What remains to be finished is an interface more accessible to the general user. To this end, we are reimplementing both general as well as sparse factor analysis models as an R package. Further work along modelling latent phenotypes will use the MuTHER dataset, and combine information from genotype, gene expression, small RNA expression, methylation, lipid level, metabolite, DEXA scan, and clinical questionnaire data, building towards a generative model of human cellular and molecular physiology, and its relation to disease.

There are many possibilities to extend work on mapping by artificial selection (Chapter 4). There are experiments to be done, analysis to undertake, and models to develop.

Modelling. The first priority is establishing and implementing inference for a correct generative model of allele frequencies and QTLs. Currently, we employ a simplistic smoothing approach. Instead, we have a HMM-like model in mind, where binary indicators designate QTL locations, local recombination rates are modelled as random variables with informed prior distributions, and allele frequencies are inferred taking the above into account. This model would yield a more finescale QTL map, inform of the required sequencing depth for accurate mapping, as well as provide estimates of local recombination rates. The exact QTL locations can be used in designing further genotyping experiments.

Analyses. We have generated a rich dataset, and many questions are not yet answered. We are looking to analyse population genetics and signatures of selection for all QTL regions, run SIFT analyses to detect intolerable alleles, and perform additional computational experiments to assess the effect of candidate causative alleles on mRNA expression levels or protein structure and function.

Finally, we are in a position to assess local recombination rates both from the model proposed above, as well as long insert libraries that we can scan for read pairs with evidence of material from more than one parental strain.

Experiments. We are in the process of extending this work in several directions.

Firstly, we are genotyping 1,000 segregants from the pool after selection at all the QTL loci to find epistatic interactions. Secondly, we have generated a four-way cross of the strains used in the grid of crosses experiment (Chapter 5.2). We are mapping QTLs in this genetically more diverse population, and genotyping and phenotyping 384 of the segregants from the intercross pool. We need to perform a few follow-up experiments for replicating diploid results, and understand how much information is shared between haploid and diploid screens. Finally, it may be interesting to assess the dynamics of the allele frequency change at higher resolution, and in longer term.

Most importantly, we can map any selectable trait to high resolution. This opens a wide range of possible experiments. Most interesting ones pertain to general cell biology that would also transfer to higher eukaryotes, such as DNA damage response, oxidative stress response, ageing, and cell adhesion. We are looking to focus on specific biological questions that can be answered in this model.

We have measured mRNA levels from the pool before and after selection, both steady state, as well as in response to heat shock after 15 minutes. We may be looking to supplement these experiments with protein level measurements of a few key proteins to assess the role of their abundance in the heat resistance trait to trace the phenotypic effect of the alleles. Furthermore, we may attempt to rescue the heat resistance phenotype in some segregants with low fitness by introducing a plasmid that modifies *RAS* activity. Finally, we could generate allele replacement strains for the individual QTLs to assess the effect of the alleles in isolation and specific combinations.