Chapter 4

Genetic mapping using artificial selection

Collaboration note

This chapter contains work performed in collaboration with many people, most notably Dr. Gianni Liti and Francisco Cubillos. The results have been submitted for publication (Parts et al., 2010), and the manuscript forms the backbone of this chapter. I am including brief validation results from some of their experiments for completeness, detailed acknowledgements are given below.
I conceived and developed the project with Gianni. Gianni and Richard Durbin designed the intercross approach. Gianni, Francisco, and Kanika Jain performed the genotyping, crossing, selection, and validation experiments. Michael Quail prepared the sequencing libraries. Jared Simpson assembled the parental strains. Jonas Warringer performed the phenotyping. I analyzed the sequencing and genotyping data. Amin Zia and Alan Moses performed individual allele analysis.

One approach to understanding the genetic basis of traits is to study their pattern of inheritance among offspring of phenotypically different strains (Mackay et al., 2009; Nordborg and Weigel, 2008; Rockman, 2008). Previously, such analysis has been limited by low mapping resolution, high labour costs, and large

sample size requirements for detecting modest effects. We present a novel approach to map trait loci using artificial selection. We subject a large pool of haploid or diploid twelfth generation progeny between two budding yeast strains to heat stress for extended time. Sequencing total DNA from the pool before and during selection reveals the genetic architecture of heat resistance in this cross. Many regions, some contained within a single gene, change in allele frequency, show evidence of negative epistatic interactions, and exhibit dominant, recessive, and additive effects.

4.1 Trait mapping with natural genetic variation

A central challenge of modern genetics is to identify genes and pathways responsible for variation in quantitative traits. In the last decade, efforts of large international collaborations have revealed numerous loci that influence disease risk in humans by genotyping and phenotyping very large cohorts of individuals (Chapter 1.1.2). However, the effects of single alleles are generally modest, and explain only a small proportion of the heritable variability. Studies in model organisms, where causality can be addressed by reverse genetic tools, can help understand the genetic complexity of such traits (Chapter 1.1.1).

4.1.1 Shortcomings of existing approaches

Mapping the effect of naturally occurring genetic variation on traits is not straightforward even in model organisms. Designed crosses often use manipulated laboratory strains (Ehrenreich et al., 2009), and produce segregants that have to be laboriously genotyped and phenotyped. It is also costly to develop and maintain outbred populations of sufficient size (Valdar et al., 2006). Recently, analysis of a very large pool of recombinant yeast strains has been used to identify quantitative trait loci (QTLs) for multiple traits (Ehrenreich et al., 2010; Segrè et al., 2006; Wenger et al., 2010) without characterizing individual segregants.

4.1.2 Leveraging artificial selection

While Ehrenreich et al. (2010) found many QTL regions, the problem of finding all responsible loci, and localising the trait genes within QTL peaks remains. Furthermore, such analyses in yeast have previously been limited to haploid samples. Here, we present a precise and sensitive approach to QTL mapping, extending the approach of Ehrenreich et al. (2010), and identify trait loci and genes in both haploid and diploid populations. We used a three step process (Figure 4.1). First, we generated intercross lines between two phenotypically different yeast strains. We then applied selective pressure to the pool by growing it in a restrictive condition (40°C heat or 400 μg/ml paraquat) to enrich for individuals with beneficial alleles. Finally, we sequenced the pool before and at multiple timepoints during selection to directly assess the changes in population allele frequencies throughout the genome.

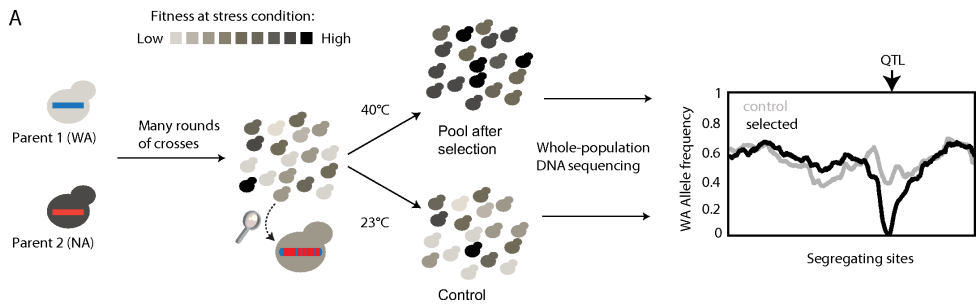

Figure 4.1: A 3-step QTL mapping strategy by crossing two phenotypically different strains for many generations to create a large segregating pool of individuals of various fitness, and growing the pool in a restrictive condition that enriches for beneficial alleles that can be detected via sequencing total DNA from the pool.

Methods used throughout the chapter are outlined in Chapter 4.4. The technical aspects of the experimental approaches designed and performed by collaborators are available elsewhere (Parts et al., 2010).

4.2 Very large segregating yeast population

We generated up to 12 generations of advanced intercross lines (F12 AILs, 12 generations of random mating between YPS128, a North American oak tree bark (NA) strain, and DVGBP6044, a West African palm wine (WA) strain. The resulting haploid or heterozygous diploid intercross pools consisted of 10-100 million random segregants, with a segregating site every 170 bases on average.

We sought to characterise the properties of this mapping pool to assess the increase in number of recombinations, and confirm that the alleles present in the parents are still segregating after many generations of intercross.

4.2.1 Recombination rate

Using many rounds of crosses should expand the genetic map due to reduction of linkage between nearby loci (Figure 4.2, Darvasi and Soller (1995)). To confirm this, we genotyped 30 markers in 96 individual segregants from each of three generations in three regions to assess the change in recombination fraction between adjacent markers.

The genetic distance (measured in 100 times the average number of recombination events) between two chromosome XIII loci separated by 204kb increased from 88 in F1 to 125 in F6 and 180 in F12. We further sequenced two segregants from the F6 pool at low coverage and observed 64 and 68 recombination events, a 125% increase compared to an average of 30 events detected in 96 F1 segregants (Figure 4.2b, Cubillos et al. (2011)).

We observed fewer recombination events than expected if an independent set of crossovers occurred every generation. There are several explanations to this. First, it is known that the recombination rate is not uniform, but accentuated in specific regions (recombination hotspots, Tsai et al. (2010)). Therefore, multiple recombinations can occur at the same site, leading to underdetection of recombination events. Second, it is possible that there is recombination preference in the heterozygous diploids with homozygous regions. We are not able to detect such events. Finally, we have conservatively filtered out very closely spaced events (2kb, Chapter 4.4) as well as subtelomeric events, introducing a further bias.

Some of these issues can be addressed by using a different cross with higher recombination frequency, or mutant strains that exhibit alternative recombination patterns.

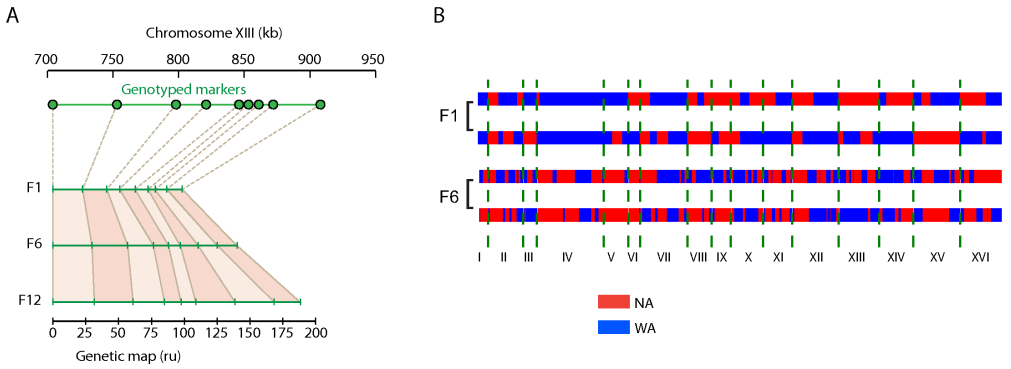

Figure 4.2: Recombination landscape after mulptiple rounds of intercrosses. a) Expansion of the genetic map, measured in recombination units (ru) of 100 times the average number of recombination events from first to twelfth generation (bottom of panel) of a 200-kb chromosome XIII locus genotyped at 9 markers (top of panel). b) Genetic background of two segregants from first (F1) and 6th (F6) generation cross shows a sharp increase in recombination events.

4.2.2 Parental allele frequency

Sequencing total DNA from pools before selection shows that more than 99% of the mappable genome is segregating in the F6 generation with minor allele frequency greater than 10%, and 97% in the F12 generation (Figure 4.3a). A small fraction of the genome is strongly selected for during the intercross rounds, due to alleles favoring sporulation, mating, or resistance to selection steps used in the cross (Chapter 4.4). This allowed us to map 6 regions responsible for these traits as a byproduct of our approach (Table B.17).

4.3 Mapping trait loci using selection

After establishing a large segregating population, we employed it for trait mapping.

4.3.1 QTLs in haploid pool

We sequenced DNA from the F12 haploid pool to an average genome coverage of 25x to 150x (Table B.16) after 0 (T0), 96 (T1), 192 (T2) and 288 (T3, 144 generations) hours of growth at 40°C. There were 19 regions where the inferred allele frequency of the T2 pool changed by at least 10% in each of two biological replicates compared to both the initial pool and the control experiment (Figure 4.3a-c, Table B.18). The NA version of the locus was selected in about two thirds (12/19) of the cases, consistent with it being the more heat resistant strain(Cubillos et al., 2011), however, several antagonistic WA alleles were also selected for. These changes are specific to the heat stress condition, as the same pool exposed to oxidative stress (paraquat, 1.5 mM) yielded a different set of QTLs (not shown). In addition, all the mitochondrial genes were greatly reduced in copy number upon selection (Table B.19). On further testing, 189/189 segregants from F12 selected pool exhibited the petite phenotype when grown in non-fermentable carbon source (glycerol and ethanol), indicating loss of mitochondrial genome.

4.3.2 QTL validation

We validated three of the mapped QTLs. Conventional linkage analysis of 96 F1 segregants also indicated a strong QTL at the right end of chromosome XIII (variance explained 66%). No other strong QTLs were seen in this cross. This region corresponds to the most rapid change in allele frequency with the NA allele fixing early in the selection at T1 (Figure 4.3d). The only other QTL that reached fixation was the GTPase activating protein *IRA1*, a negative regulator of the RAS signalling pathway. Interestingly, three additional genes of the same pathway (*IRA2*, *GPB1* and *GPB2*), as well as some of its targets (*CDC25*, *BCY1*, *CYR1*, mitochondrial genome) were contained in intervals with sharp increase in

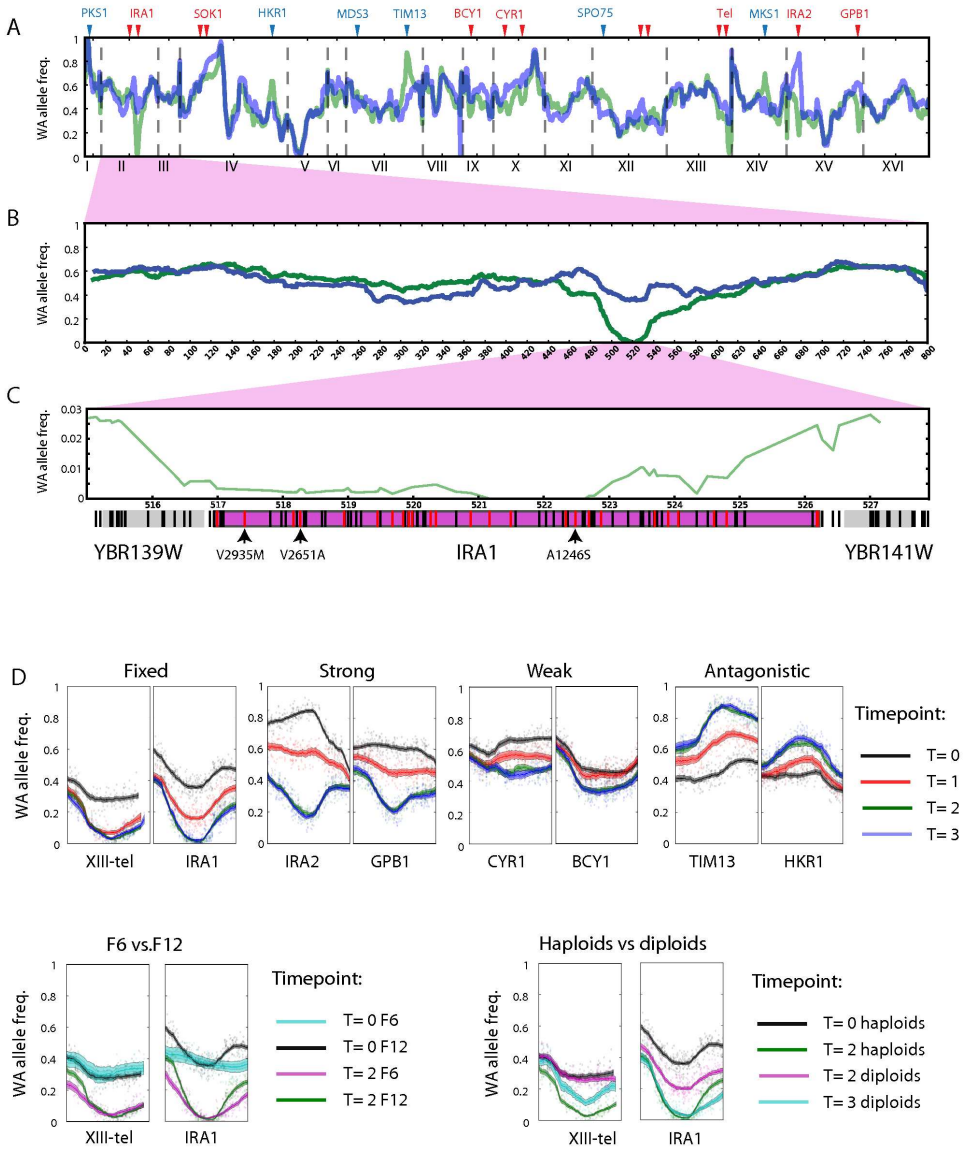

Figure 4.3: Changes in allele frequencies pinpoint QTLs. a-c) WA allele frequency of whole genome (a), chromosome II (b), and *IRA1* region (c) of the F12 pool before (blue) and after selection (green). Lines in gene regions in (c) denote segregating sites (black) and non-synonymous segregating sites (red). The sites with intolerable mutations (SIFT analysis) are highlighted with arrows and designated with the amino acid change. d) Individual examples of mapped QTLs that show differences in strength, beneficial allele, effect of recombination and ploidy. Each window spans 80kb and is centered on the locus with the largest allele frequency change in F12 T2 across two replicas. Shaded regions indicate 90% and 95% confidence intervals.

NA allele frequencies in the F6 or F12 pools, confirming the involvement of the entire RAS pathway in the heat resistance phenotype.

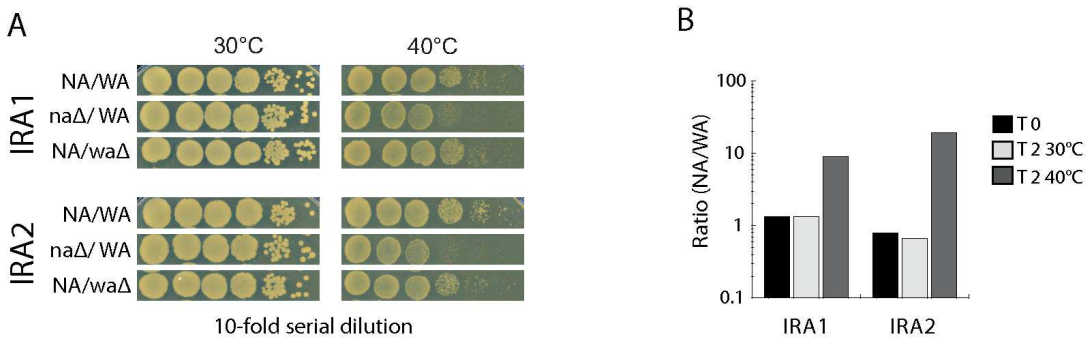

Figure 4.4: *IRA1* and *IRA2* are high temperature growth QTLs. a) Reciprocal hemizygosity confirms that *IRA1* and *IRA2* are high temperature growth QTLs. WA/NA hybrids were individually deleted for the IRA alleles and used to assess their contribution to high temperature growth. Plate spotting assay using 10-fold serial dilution demonstrates better growth of the hybrid when the NA allele is present. b) Competition experiment on hybrids with IRA reciprocal hemizygous deletions (as a) that resembles the selective step applied to the pool. This assay shows that hybrids carrying the NA allele are selected when cells were grown at 40C for 192 hours (T2).

We validated by reciprocal hemizygosity (Steinmetz et al., 2002) that *IRA1* and *IRA2* alleles affect high temperature growth. The effect was evident from a plating assay, growth curves, and competition experiments (Figure 4.4). These genes affect both growth rate (doubling time) and efficiency (final density) with *IRA1* having a stronger effect compared to *IRA2*, consistent with the difference in their final allele frequency. Interestingly, *IRA1* and *IRA2* do not have a pleiotropic effect on growth, even at environmental conditions where RAS activity has a strong influence. The clear identification of the *IRA1* and *IRA2* alleles as a cause of low performance at high temperatures shows that our method can directly map causative genes without any a priori information and without requiring further fine-mapping.

4.3.3 Mapping resolution

The advantage of reduced linkage is evident from narrow mapped intervals, in several cases localising to within single genes (Figure 4.3c-d). For example, in case of *IRA1*, we could map the selected variant down to a small region of the gene (Figure 4.3c), that also harbours the strongest candidate sequence variant between the two strains from SIFT (Ng and Henikoff, 2003) analysis. This resolution is in contrast to that from previous studies based on crosses between strains, including (Ehrenreich et al., 2010), which typically map to large regions containing many genes. An additional advantage of the intercross rounds is the ability to unlink independent QTLs at one locus (Figure 4.3d). There is a risk that long term culturing under stress conditions will select for new adaptive mutations that might rise to high frequencies and dominate the pool. However, as the pool did not become clonal, it is unlikely that haplotypes harbouring strongly adaptive mutations had risen to high frequency during selection (see simulations below).

4.3.4 Lack of fixation

While the alleles with strongest fitness effect, such as at *IRA1* gene (chrII:522kb, Figure 4.3b) and chrXIII subtelomeric region (Figure 4.3d) reached fixation in the pool upon selection, weaker ones required extended selective pressure to rise in frequency (Figure 4.3d), demonstrating the advantage of using extended selection. Three of the 19 QTLs (16%) had reached their T3 allele frequency at T1, but 17 by T2 (89%). Thus, only a minority of two loci were still changing in allele frequency after T2. This indicates that all the remaining haplotypes in the pool have nearly equal fitness in this stress condition, or are so rare even by T3 that change in their frequency does not have a major effect on the average pool genotype. It also suggests that we have saturated for individual loci with independent effects that are present in the founding strains.

4.3.5 Negative epistatic interactions

The fact that for 17 QTLs both alleles remained segregating in the pool after up to 288 hours under selection suggests that these segregating loci cannot have

independent additive effects, as otherwise their beneficial version would continue to rise in frequency after T2. Thus, the pool after selection is a mixture of haplotypes with alternative QTL genotype combinations of similar fitness. This suggests an abundance of negative epistatic interactions, as otherwise the beneficial combination would keep rising in frequency. To test this explanation, we genotyped 192 segregants from the F12 pool after 240 hours of selection (T2.5), and looked for scarcity and abundance of specific allele combinations at the 11 strongest QTLs. None of the two-locus combinations was significantly different from the expectation under independence (lowest one- tailed p=0.09, Fisher's exact test) after correcting for multiple testing. Some evidence for lower than expected deleterious allele combination counts was observed when pooling the counts over all pairs (one-tailed p=0.19, Fisher's exact test). This pattern is consistent with complex control and interactions involving multiple genes.

4.3.6 QTLs in diploid pool

Importantly for drawing comparisons with human studies, we were able to map all the 19 heat resistance QTLs in the pool of heterozygous diploid individuals. The effect of selection was weaker for the diploid pool, as allele frequencies had not reached their equilibrium levels by T2, and continued to change until T3 (Figure 4.3d). Diploid pool allele frequency after selection indicates that the chrXIII QTL is consistent with a dominant effect with the final frequency of the homozygous deleterious genotype being removed from the pool, and the *IRA1* QTL with a recessive effect with the beneficial allele being fixed. While it was expected that we could map the recessive QTLs, it is surprising that the final allele frequencies for the other 17 loci were nearly identical for haploid and diploid segregants, consistent with the selected alleles having additive effects as observed for most human GWAS hits.

4.3.7 Comparison with F1 segregant analysis

The heat growth QTLs we found by linkage analysis of 96 F1 segregants partially overlapped the ones identified using our novel approach (Cubillos et al., 2011). However, the linkage analysis lacked power to detect the weak effects, as only

the strongest chrXIII subtelomere QTL was detected with high confidence, and a chromosome IV QTL with borderline significance. This shows the additional power of our method. Furthermore, data from growth curves suggests that some of the QTL effects may not be detectable by phenotyping the segregants at 40°C, and phenotyping at higher temperatures will be more informative of individual effect sizes.

4.4 Data analysis

4.4.1 Sequencing data handling

Sequencing reads from the intercross pools before and after selection were mapped to the S288c reference genome obtained from the SGRP project website (Liti et al., 2009a) using BWA (Li and Durbin, 2010), with option ' -n 8' to allow mapping of divergent reads from the other strains. Pileup files comprising the genotypes of mapped reads were created for segregating sites inferred from both low-coverage capillary sequencing (Liti et al., 2009a) and the parental strain shotgun sequence mapping to the S288c assembly. For allele frequency inference, sites that were not segregating in the initial population, corresponding to likely false positive variant calls, were filtered out, as well as sites that were noted as heterozygous in either parental strain, indicative of copy number variation. Furthermore, for allele frequency inference, we filtered the variants to have minimum distance of at least 200 bases to ensure that any single read does not contribute disproportionately due to spanning many variants. The mapping pipeline is available upon request.

4.4.2 Segregant analysis

To analyse the genetic background of two individual F6 segregants, we mapped the sequencing reads to the genome as described earlier, and classified every segregating site to stem from one of the two parental strains, or a no-call. A site was called to be from one parent, if it was covered by at least 15 sequencing reads with base and mapping qualities at least 30, and 80% of them had the parental

allele. We conservatively refrained from making a call at low-coverage variants, subtelomeric regions up to 30kb, and variants with ambiguous mapping data. We called a recombination event if a region of at least 2kb from one parent was followed by a region of at least 2kb from the other, and at least 5 calls were made in both regions. This results in a conservative estimate of recombination events, as it discards non-crossovers, and recombination in subtelomeric regions.

4.4.3 Copy number and missing sequence.

We mapped all reads to artificial chromosomes, each containing exactly one gene with 100 flanking bases, and recorded their average sequencing coverage every 100 bases. We used that to infer a copy number for each gene as the average gene coverage normalised by the average sequencing coverage. We also mapped the reads to the assembled contigs from parental sequence data that did not map to the S288c reference; no large allele frequency changes were observed.

4.4.4 Allele frequency inference

Under a simple model, there is an unobserved WA allele frequency f_l at each locus l; we want to infer the posterior distribution of f_l after observing the sequence data. We assume all reads to come from different segregants after filtering segregating sites to be distant, thus every segregant i has one allele a_i observed at some locus l_1 distance $d_{l,i}$ away from l. We take $d_{l,i}$ to be infinity if the loci are on different chromosomes. For that segregant, there is an unobserved allele $b_{l,i}$ at locus l, and the probability that these loci are linked, with no recombination event occurring during the intercross between them, is $q_{l,i} = 1 - \exp(-d_{l,i}\rho)$, where ρ is the recombination rate. We took $\rho = 30(1 + \frac{g-1}{2})$, where g is the number of intercross rounds, as there are on average 30 crossovers per tetrad, and every intercross after the first one has a 50/50 chance of introducing a switch between parental haplotypes. The likelihood of the allele frequency at locus l is

thus $P(D|f_l) = \prod_i P(a_i|f_l)$, where

$$\begin{aligned} P(a_i \,|\, f_l) &= P(a_i, b_{l,i} = \mathrm{WA} \,|\, f_l) + P(a_i, b_{l,i} = \mathrm{NA} \,|\, f_l) = \\ &= P(a_i \,|\, b_{l,i} = \mathrm{WA}) P(b_{l,i} = \mathrm{WA} \,|\, f_l) + P(a_i \,|\, b_{l,i} = \mathrm{NA}) P(b_{l,i} = \mathrm{NA} \,|\, f_l) = \\ &= q_{l,i}^{a_i=\mathrm{WA}} (1 - q_{l,i})^{a_i=\mathrm{NA}} f_l + q_{l,i}^{a_i=\mathrm{NA}} (1 - q_{l,i})^{a_i=\mathrm{WA}} (1 - f_l) \simeq \\ &\simeq q_{l,i} f_l^{a_i=\mathrm{WA}} (1 - f_l)^{a_i=\mathrm{NA}} \end{aligned}$$

Here, we have discarded likelihood terms that require a recombination event, as we will filter $q_{l,i}$ to be large. We approximate the posterior of f_l with a Beta distribution with an uninformative prior, and find the maximum likelihood parameters of the distribution from for segregants for which $q_{l,i} > 0.95$ (0.75 for Fig. 2A-B for smoothness). This inference procedure corresponds to a smoothing approach within a fixed window with the width determined by the recombination rate, and has the effect of discriminating against extreme allele frequencies. The posterior confidence intervals were obtained from the approximated Beta distribution.

4.4.5 QTL inference

We inferred QTLs in the F12 selected pool by comparing the inferred allele frequencies before and after selection. The allele frequencies in the control experiment, propagating the cells without selection, were nearly identical to those before selection. We called QTLs by testing for inequality of the inferred approximate posterior allele frequencies before and after selection. As a simple cutoff, we called a QTL if the inferred allele frequency changed in the same direction by at least 10% in both biological replicas and 25% in total, a change larger than exhibited for the control experiments at permissive temperature of 23 degrees after 192 hours at any locus in either replica. A single QTL was called in any 20kb window, corresponding to the variant with the largest combined allele frequency change over the two replicas.

4.4.6 Linkage analysis in F1 segregants.

We used results from Cubillos et al. (2011) for F1 segregant QTL mapping. In short, we used standard marker regression for the 200 genotyped markers and

3 heat growth phenotypes to map QTLs significant at 5% false discovery rate (FDR) using a standard linear model and 1000 permutations.

4.4.7 Epistatic interaction tests

We used a standard linear model (Chapter 1.4.1) to assess the significance of an epistatic interaction term between two genotyped loci that affects any of the three growth phenotypes assayed for the segregants. No significant interactions were found at 5% false discovery rate (FDR, fraction of expected false positives in all calls), possibly due to the difference of the phenotyping temperature effect in solid and liquid media, or lack of power.

We tested for scarcity and abundance of two-locus genotype combinations for 11 genotyped trait loci in 189 segregants of F12 population after 120 hours in heat stress (Table S9). For each pair of segregating loci, we compiled a contingency table of genotype counts, and applied two-tailed Fishers exact test to calculate the p-value of independence of the loci. We calculated the false discovery rate (FDR, fraction of expected false positive calls) at a range of p-value cutoffs for the set of pairwise tests, and did not find individual interactions at FDR $< 10\%$. As an alternative, we pooled all allele combination counts for beneficial/beneficial (BB), beneficial/deleterious (BD), and deleterious/deleterious (DD) genotype combinations, to test for relative abundance of BB and DD combinations. For the 11 genotyped QTLs, there were $n_{11} = 607$ DD genotypes, $n_{12} = 3959$ BD genotypes, and $n_{22} = 5739$ BB genotypes for the $N = n_{11} + n_{12} + n_{22}$ genotypes. We then compiled a contingency table with the observed and expected combination counts calculated from the fraction of genotyped beneficial alleles, and calculated the chi-squared test p-value. This test is appropriate as the sample size is large. We also repeated the test for QTLs found within individual pathways; no p-values were significant at 10% cutoff.

4.5 Simulation experiments

We now expand on the argument for lack of adaptive mutations dominating the pool, as well as allele frequency changes under simplifying conditions in a

simulation scenario. We simulated data from a simple generative model to explore the potential of adaptive mutations to dominate a haploid pool, as well as effects of more than one allele in haploid and diploid pools.

4.5.1 Adaptive mutations.

First, we provide three lines of computational evidence for lack of new adaptive mutations with large effect on intercross pool allele frequency during selection. Second, we demonstrate how interaction effects can account for lack of fixation, and ploidy can be responsible for the difference in response time to selection.

Firstly, the fitness requirement of adaptive mutations to dominate the pool is too high. A single adaptive mutation begins at very low initial frequency, $f_1 = \frac{1}{N}$, where we take N, the total number of segregants in the pool to be 10^7. The doubling times for the segregants range from 1.5 hours in permissive condition (or for fit segregants in restrictive condition) to 2 hours for unfit segregants in restrictive condition. Let us assume an adaptive mutation rises to the same frequency as the total frequency of haplotypes with beneficial alleles at the two loci that reach fixation - the *IRA1* and chrXIII subtelomeric loci (initial frequency $f_0 = 0.25$) all of which have doubling times $t_0 \sim 1.5$ hours. Over $T = 288$ hours of selection, the following identity must then hold for the doubling time t_1 of the adaptive mutation: $f_1 2^{\frac{T}{t_1}} \geq f_0 2^{\frac{T}{t_0}}$, or $t_1 \leq \frac{T}{\log_2 f_0 - \log_2 f_1 + \frac{T}{t_0}}$. Plugging in numbers for f_1, f_0, T, this gives $t_1 \leq 1.34 = 0.9 t_0$ Thus, in order to rise to appreciable frequencies in the very large pool, the haplotype with the adaptive mutation must grow 10% faster in restrictive condition than the segregants do in the permissive condition. If such mutations were possible, they would be more likely to rise during the many months of intercross rounds, not during the span of four days. However, in this case, the allele, not the haplotype, will be selected for, as further intercross rounds separate the adaptive mutation from the haplotype on which it arose.

Dominating adaptive mutations would drive the pool allele frequencies to extremes. In the very long run, the haplotype with the adaptive mutation will be the only one left in the pool, as no recombination happens during selection. As the frequency of the adaptive mutation rises in the pool, the pool loses heterozygosity

and genetic complexity, and the frequency of the NA allele at all segregating loci will be driven to 0 or 1. If a haplotype with an adaptive mutation is present at high frequency in the pool, we would expect to see an allele frequency change from the initial pool at all loci towards the genotype of that haplotype, which we do not observe.

Adaptive mutations would continue to rise in frequency after 192 hours. We do not observe global allele frequency changes after 192 hours. However, as outlined above, haplotypes with adaptive mutations should continue to rise in frequency in the pool. These three lines of evidence point to little contribution from adaptive mutations to the final segregant pool allele frequency makeup. Adaptive mutations for sporulation, mating, or growth can arise during intercross, and could be traced. However, for QTL mapping, we are conditioning our analysis on all the segregating sites present in the pool at the beginning of selection, regardless of whether they were present in the parental strains.

4.5.2 Effects of selection on allele frequency.

We simulated allele frequency changes under simple assumptions for various scenarios. While standard (e.g. Hartl and Clark (2006)), the results give intuition for allele frequency changes observed.

Haploid individuals. We fixed the initial allele frequency of any locus to be 0.5 for simplicity, and calculated its change over generations in a deterministic way. For a one locus trait, the individuals with genotype '1' were assumed to have a fitness advantage s, which changed the rate at which they survived to the next generation, with the frequency $f_{l,t}$ of locus l at generation t was taken to be $\frac{(1+s)f_{l,t-1}}{(1+s)f_{l,t-1}+(1-f_{l,t-1})} = \frac{1+s}{1+sf_{l,t-1}}f_{l,t-1}$ If $s > 0$, f_l increases, and if $s < 0$, it decreases in a near-geometric manner. For these one locus haploid pools, the beneficial allele asymptotically approaches fixation, with the speed depending on the magnitude of the selection coefficient (Figure 4.5).

In case two loci are contributing, the calculation remains almost unchanged, but now the effect of selection is assumed to act only on the '11' genotype. In this case, if $s > 0$, the haplotypes with '11' genotype are fitter than the

others, and again are driven to fixation. However, if $s < 0$, the '11' genotype is selected against, and will be purged from the pool in the long run. Both alleles will still be present at each locus (Figure 4.5A). We hypothesize such interactions to be responsible for the lack of fixation upon over 100 generations of selection. The usual intuition behind this is that fitness depends on functioning of a specific pathway. While any single mutation does not alter the functionality of the pathway, there are many possible combinations of genotypes that render it defective. These combinations are selected against, producing a change in allele frequency, but not fixation of any allele.

Diploid individuals. As the diploid individuals propagate clonally just like haploids, we have to trace the frequency of the genotypes, not alleles, since there is no further mixing of the haplotypes between individuals. We can therefore treat a one locus trait in diploids, identically to a two-locus trait in haploids, and find that for traits where the beneficial allele behaves in an additive or recessive way, selection drives the frequency of beneficial allele to fixation, and for dominant beneficial alleles, the homozygous non-beneficial allele combination is selected against (Figure 4.5). We observed QTLs with final allele frequencies as well as their speed of change consistent with both recessive (*IRA1*) and dominant (chrXIII subtelomere) beneficial alleles (main text). However, when the QTL acts in an additive manner, the allele frequency change is identical to that of the haploid pool.

If interaction effects are responsible for the allele frequency change, the effect can again be dominant, additive, or recessive. The differences to a one-locus model are slower effect of selection, as the fittest haplotype has lower initial frequency, and less extreme final allele frequency in case the interaction effect is dominant, as there are fewer genotype combinations selected against (Figure 4.5B).

4.6 Discussion

We have presented an accurate, sensitive, quick approach for QTL mapping in yeast. It is straightforward to apply our method to any selectable trait. We

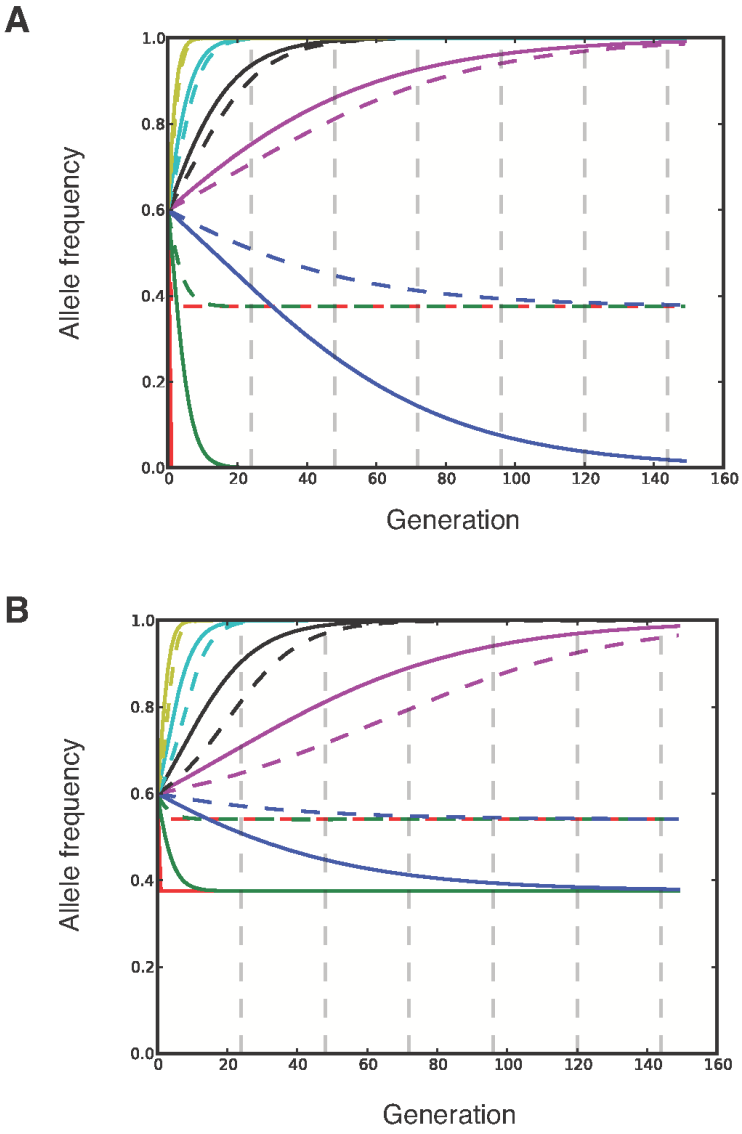

Figure 4.5: Haploid (solid lines) and diploid (dashed lines) pool allele frequency changes for 1-locus (a) and 2-locus effects (b). Initial allele frequency of a locus is 0.6. Individual lines correspond to different fitness modifiers, from top to bottom: $+1, +0.3, +0.1, +0.03, -0.03, -0.3, -1$.

expect to be able to extend these mapping populations to include more of the genetic diversity in the species by crossing a larger number of parental strains. As we were also able to map the trait loci in the diploid pool, there is a potential to establish an outbred yeast population that can be used as a model for natural diploid genome-wide association studies as carried out in humans.