Information-sharing models for computational genetics

by

Matthew Douglas Edwards

B.S., Computer Science, Duke University (2008)
B.A., Mathematics, Duke University (2008)
S.M., Electrical Engineering and Computer Science,
Massachusetts Institute of Technology (2011)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author .....
Department of Electrical Engineering and Computer Science
May 20, 2016

Certified by .....
David K. Gifford
Professor
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Information-sharing models for computational genetics

by

Matthew Douglas Edwards

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Modern genetics has been transformed by a dramatic explosion of data. As sample sizes and the number of measured data types grow, the need for computational methods tailored to deal with these noisy and complex datasets increases. In this thesis, we develop and apply integrated computational and biological approaches for two genetic problems.

First, we build a statistical model for genetic mapping using pooled sequencing, a powerful and efficient technique for rapidly unraveling the genetic basis of complex traits. Our approach explicitly models the pooling process and genetic parameters underlying the noisy observed data, and we use it to calculate accurate intervals that contain the targeted regions of interest. We show that our model outperforms simpler alternatives that do not use all available marker data in a principled way. We apply this model to study several phenotypes in yeast, including the genetic basis of the surprising phenomenon of strain-specific essential genes. We demonstrate the complex genetic basis of many of these strain-specific viability phenotypes and uncover the influence of an inherited virus in modifying their effects.

Second, we design a statistical model that uses additional functional information describing large sets of genetic variants in order to predict which variants are likely to cause phenotypic changes. Our technique is able to learn complicated relationships between candidate features and can accommodate the additional noise introduced by training on groups of candidate variants, instead of single labeled variants. We apply this model to a large genetic mapping study in yeast by collecting multiple genome-wide functional measurements. By using our model, we demonstrate the importance of several molecular phenotypes in predicting genetic impact.

The common themes in this thesis are the development of computational models that accurately reflect the underlying biological processes and the integration of carefully controlled biological experiments to test and utilize our new models.

Thesis Supervisor: David K. Gifford
Title: Professor

Acknowledgments

I am grateful to the many people that enriched my graduate school experience, as well as those that set me on this path and contributed to my personal and intellectual development.

First, I thank my advisor David Gifford for his unwavering support and encouragement throughout my graduate career. He always has his students' best interests at heart, and I appreciate his astute scientific guidance coupled with giving me the freedom to explore, learn, and discover. I thank my thesis committee member and genetic mentor Gerry Fink for his generosity with his time and wisdom and for exposing me to the complicated world of genetics and molecular biology. I thank my final committee member Tommi Jaakkola for his insightful comments and advice throughout this work.

I am grateful to past and present members of the Gifford lab, who have made the lab a great place to work and contributed both to my scientific work and my social life. I thank Alex Rolfe for early encouragement and sage system administration advice, Shaun Mahony for his profound warmth and generosity, Charlie O'Donnell for wise suggestions and insightful conversations, Tatsu Hashimoto for his collaborative spirit and inspiring productivity, Chris Reeder for his encouragement and positive example, and Haoyang Zeng for his dedication and good humor. I thank Yuchun Guo, Rich Sherwood, Nisha Rajagopal, Tahin Syed, Jonas Mueller, Rujian Chen, Daniel Kang, Grace Yeo, Saber Liu, and Logan Engstrom for their friendship, camaraderie, and encouragement over the years. I also am grateful to Jeanne Darling and Patrice Macaluso for their guidance and support on many occasions.

I thank the Fink lab members that taught me many things about yeast genetics through lab meetings and informal conversations and generously shared their time and skills during my brief foray into the messy world of wet lab biology. I am especially grateful to Anna Symbor-Nagrabska and Lindsey Dollard for their dedication on our collaborative projects.

I thank my previous teachers and mentors, and unfortunately I do not have space to list all the positive influences I've been fortunate enough to have. In particular I thank John Noland for his positive example in high school, Owen Astrachan for his early encouragement when I began as an undergraduate, and Alex Hartemink for his guidance and advice in research and life during and after my undergraduate years. I have him to thank for starting me on the computational biology path, and it is no accident that many of the principles explored here and elsewhere in my graduate work are inspired by and reflect his intellectual

influence.

I would also like to thank my friends near and far for their welcome distractions that contributed to my happiness throughout my graduate school years. I have tremendously enjoyed our ultimate games, bike rides, ski trips, game nights, and cookouts, and hope to continue these far into the future.

Finally, I want to thank my family. While I could not always adequately explain what I do on a day-to-day basis, you all made sure I never doubted your love and support. I thank my extended family and new Lee family, as well as my sister Lauren for her cheerfulness and tolerance to my shenanigans. I thank my parents Mark and Melissa for their support and encouragement in my educational pursuits, and for exposing me to the apparent family business of computer science while still giving me the freedom to choose my own path. Last but not least, I want to thank my wife Amanda. She supported me before, during, and hopefully after my graduate work, and helped me get across the finish line despite the times when I doubted my ability to do so. I could not have done this without you.

To my parents

Contents

1 Introduction **17**

1.1 Background 18

1.2 Summary of contributions 19

1.3 Thesis outline 19

1.4 Previously published work 20

1.5 Collaborators 20

I Efficient genetic mapping with pooled sequencing **21**

2 Statistical models for genetic mapping using pooled sequencing **23**

2.1 Introduction 23

2.1.1 Targeted experiments 24

2.1.2 Challenges 25

2.1.3 Previous statistical methods 26

2.1.4 Approach 26

2.2 Methods for MULTIPOOL 27

2.2.1 Obtaining allele frequency measurements 27

2.2.2 Multi-locus model 28

2.2.3 Model specification 29

2.2.4 Inference: discrete model 31

2.2.5 Inference: continuous approximation 32

2.2.6 Statistical tests for a single experiment 35

2.2.7 Statistical tests for multiple experiments 36

2.3 Methods for MULTIPOOL2 37

3 Application of statistical models for pooled sequencing 41

3.1 Simulation results . . . 41

3.2 Mapping results with previously validated causal loci . . . 42

3.2.1 Single-locus comparisons . . . 42

3.2.2 Large pool results . . . 43

3.3 Mapping the genetic basis of strain-specific essential genes . . . 45

3.3.1 Methods . . . 46

3.3.2 Mapping results . . . 48

3.3.3 Validation of a killer virus link . . . 49

3.3.4 Complex genetic basis of strain-specific essential genes . . . 52

3.4 Conclusions . . . 53

3.4.1 Future computational work . . . 55

3.4.2 Biological insight from pooled sequencing studies . . . 55

II Leveraging functional annotations to improve genetic mapping 58

4 Statistical models for integrating functional annotations with genetic mapping 59

4.1 Introduction . . . 59

4.1.1 Challenges . . . 60

4.1.2 Related work . . . 61

4.2 Statistical models . . . 62

4.2.1 Converting instance probabilities to bag probabilities . . . 62

4.2.2 Optimization and model fitting . . . 64

4.2.3 More complex instance-level probability models . . . 64

4.2.4 Detailed comparison to related models . . . 65

4.3 Reweighting association studies . . . 66

4.4 Simulation results for multi-instance classification . . . 67

4.4.1 Dataset . . . 67

4.4.2 Results . . . 68

4.5 Simulation results for a genetic mapping study . . . 69

4.5.1 Dataset . . . 70

4.5.2 Results . . . 70

4.6 Conclusions 72

5 Genetic mapping using functional annotations applied to yeast genetics 73

5.1 Introduction 73

5.2 Functional annotations 74

5.2.1 Methods and data sources 75

5.3 Functional annotation results 79

5.4 Genetic mapping results using functional annotations 82

5.5 Conclusions 90

6 Conclusions 93

6.1 Future computational work 93

6.2 Future biological work 94

6.3 Final thoughts 95

List of Figures

2-1 Pooled sequencing experimental design example 39
2-2 Graphical model showing multi-locus dependencies 40

3-1 Mapping accuracy in simulated datasets 45
3-2 Localization of *RAD5* using 4-NQO selected replicate 2 46
3-3 Localization of *IRA1* using heat tolerant replicates 1 and 2 47
3-4 Mapping results for *ski7*Δ and *ret2*Δ viability modifiers 49
3-5 Fine mapping of suppressors interacting with *ski7*Δ on chromosomes 4 and 12 50
3-6 Fine mapping of suppressors interacting with *ret2*Δ on chromosomes 14 and 15 51
3-7 Mapping results for *bem1*Δ and *ski7*Δ viability modifiers 52
3-8 Fine mapping of a suppressor interacting with *bem1*Δ on chromosome 15 . 53
3-9 Experimental design for assessing nonchromosomal and chromosomal genetic interactions 54
3-10 *ski7*Δ viability is dependent on killer virus in a Σ1278b background 55
3-11 The lethality of multiple gene deletions is dependent on killer virus in a Σ1278b background 56
3-12 Chromosomal variants exhibit a minimal dependence on dsRNA presence in an S288c background 57

4-1 Multi-instance classification model performance on simulated data 69
4-2 Genetic mapping performance using a simulated yeast cross 71

5-1 Example training labels along the yeast genome 74
5-2 Example features for yeast genetic mapping 75
5-3 Yeast functional annotation correlations 80
5-4 Example learned prior probability model for yeast genetic mapping 83

5-5 Comparison of bag-level prediction probabilities to control probabilities . . . 88
5-6 Overlap between high- and low-scoring variants and reported association regions . . . 89
5-7 Predictions around multiple validated loci . . . 91

List of Tables

2.1 Parameters used in the MULTIPOOL model 29

3.1 Analyzed experiments for published large pool datasets 44
3.2 Localization of known associated genes in large drug-selected pools 44
3.3 Localization of known associated genes in large heat-selected pools 44
3.4 Analyzed experiments for strain-specific essential genes 48
3.5 Associated suppressors identified by comparing opposite phenotypic extremes 50
3.6 Associated suppressors identified with single pools 50
3.7 Overlap between strain-specific essential genes and killer-associated genes. . 52

4.1 Multi-instance classification model variants 67
4.2 Summarized performance of multi-instance classification models on simulated data 68
4.3 Summarized performance of multi-instance classification models on simulated yeast genetics data 72

5.1 Annotations used in yeast genetic mapping experiments 76
5.2 Analyzed conditions for yeast genetic mapping study 76
5.3 Yeast functional annotation counts 81
5.4 Yeast proteomics differential analysis results 81
5.5 Yeast RNA-seq differential analysis results 82
5.6 High-magnitude model terms in a yeast functional annotation experiment . 84
5.7 Annotation feature significance in multi-instance logistic regression models . 86
5.8 Grouped annotation feature significance in multi-instance logistic regression models 86
5.9 Negative control grouped annotation feature significance tests 87

Chapter 1

Introduction

Modern genetics has been transformed by an explosion of data. Led by technological advances in collecting genotypes and high-throughput sequencing data, medical and population genetic studies are growing larger and larger. These increases in data sizes also come with increased noise and analytical complexity, particularly when coupled with complicated population-based mapping schemes.

In parallel, functional surveys such as the NIH ENCODE [Consortium et al., 2012] and the Roadmap Epigenomics [Bernstein et al., 2010] projects are generating a wide breadth of measurements describing the human genome in numerous cell types and genetic backgrounds. However, designing approaches that can efficiently combine and use these rich and complex datasets is a difficult task.

This thesis aims to attack two related problems in this area, focusing on the joint development of computational models and biological datasets. First, we consider pooled genetic studies, which are efficient study designs for model organisms in particular. We design and then use a genetic mapping approach to learn new insights in yeast genetics. Second, we propose an algorithm that can link functional annotations genome-wide, such as ENCODE or Roadmap data, to observed genetic associations. We collect a wide variety of functional information in yeast and apply our models to learn which annotations are most predictive of genetic associations.

1.1 Background

We first consider the concepts underlying genetic mapping in general, which plays a central role in both projects discussed in this thesis. For a more thorough introduction to these topics, see previous reviews and the references therein [Altshuler et al., 2008, Civelek and Lusis, 2014, Lander and Schork, 1994, Lehner, 2013].

Genetic mapping relies on the comparison of the inheritance pattern of a trait with the inheritance pattern of a particular genetic segment. Broadly speaking, and in well-controlled contexts where confounding factors can be avoided or corrected, a genetic segment that tends to be inherited or appear along with a trait is said to be associated with the trait. The task of genetic mapping then is two-fold: collecting or tracking genetic variation in a set of individuals and measuring or obtaining trait values. The key prerequisite is determining a particular group of individuals to study, which determines and informs the particular statistical tests used to assess how nonrandom inheritance of genetic variants will be manifested. For instance, linkage studies in extended families will have a different pattern of inheritance than association studies that consider large groups of unrelated individuals.

For tracking genetic variations in individuals, technological advances over the past decades have broadened the particular types of variations that can be identified. The earliest work used known marker regions with observable phenotypes, and later work extended genetic mapping to medium-scale molecular variations including short length polymorphisms. With advances in microarray and sequencing technology, single-base changes can now be surveyed with relative ease. Determining the status of genetic variants, referred to as genotyping, will play a role in both parts of this thesis. The second task of genetic mapping, determining phenotypes, varies determining on the particular studied trait. For human diseases, phenotypic labels are determined by clinical analysis or individual patient reports. For the model organism phenotypes considered in this thesis, phenotypes are often observed growth rates measured by proxy using colony size. Phenotypes may be assigned individually or may be used in specific experimental scenarios to identify large batches of individuals that have a particular property. In all cases, the details of the experimental design inform the statistical tests used to perform genetic mapping and the downstream interpretation of

the results. In later chapters of this thesis we will derive and use specific techniques, and we will go into greater detail on their application and interpretation.

1.2 Summary of contributions

In this thesis, we propose and implement multiple machine learning models that are coupled to specific biological tasks. In each case, we use the properties of the biological problem to design algorithms that are flexible and efficient while appropriately handling the uncertainty present in the targeted datasets. Specifically, we offer the following contributions:

- The development of an efficient algorithm for analyzing genetic mapping studies conducted by pooled sequencing. Our model considers noisy sequencing data from multiple locations in the genome to make accurate predictions of where causal variants may reside.
- The application of pooled sequencing genetic studies to unravel the basis of strain-specific essential genes in yeast, where we confirm the complex genetic basis of this phenomenon and demonstrate the link of an inherited cytoplasmic virus.
- The design and implementation of machine learning algorithms designed to learn classification rules from examples where possibly noisy labels are assigned to groups of examples, instead of single examples.
- The collection of a large set of functional information describing a yeast genetic mapping study and the application of our machine learning algorithms to understand which functional measurements can improve the genetic mapping task.

1.3 Thesis outline

In Chapter 2 we describe our computational model of genetic mapping using pooled sequencing. In Chapter 3, we apply the computational framework to several biological datasets. We analyze the genetic basis of strain-specific essential genes and observe the complex genetic basis of this phenomenon, along with the involvement of an inherited cytoplasmic

virus. Chapter 4 contains our machine learning models that learn from groups of labeled training examples. In Chapter 5 we collect a large set of information describing a yeast cross and apply our statistical models to determine which data types are most helpful in genetic mapping.

1.4 Previously published work

Portions of the algorithmic material in Chapter 2 were published in [Edwards and Gifford, 2012]. Some of the biological results in Chapter 3 were discussed in [Edwards et al., 2014]. The work in Part II is being prepared for publication.

1.5 Collaborators

The biological results in Chapter 3 were obtained as part of a close collaboration with members of Gerald Fink's laboratory at the Whitehead Institute, including Anna Symbor-Nagrabska, Lindsey Dollard, Brian Chin, Ifat Rubin-Bejerano, and Doug Bernstein. Other results in the same chapter arose from a collaboration with Pierre Côte in Charlie Boone's laboratory at the University of Toronto. The biological results in Part II were a collaboration with members of Gerald Fink's laboratory and Amanda Edwards in the laboratory of Wilhelm Haas at the MGH Cancer Center. The high-throughput sequencing data throughout this thesis depended on the help and insight of members of the MIT BioMicro Center and the Whitehead Genome Technology Core.

Part I

Efficient genetic mapping with pooled sequencing

Chapter 2

Statistical models for genetic mapping using pooled sequencing

In this chapter, we present the computational models and statistical tests underlying the MULTIPOOL and MULTIPOOL2 methods. These two models are computational methods for genetic mapping in model organism crosses that are analyzed by pooled genotyping. Unlike other methods for the analysis of pooled sequence data, we simultaneously consider information from all linked chromosomal markers when estimating the location of a causal variant.

2.1 Introduction

Advances in high-throughput DNA sequencing have created new avenues of attack for classical genetics problems. A robust method for determining the genetic elements that underlie a phenotype is to gather and group individuals of different phenotypes, interrogate the genome sequences of each group, and identify elements that are present in different proportions between the groups. However, the uncertainty from pooling and the challenge of noisy sequencing data demand advanced computational methods. We describe MULTIPOOL and MULTIPOOL2, multi-locus methods for analyzing high-throughput DNA sequencing reads obtained from large pools of phenotypically-extreme individuals.

The MULTIPOOL model analyzes informative sequencing reads with a discrete dynamic

Bayesian network, which we extend with a continuous approximation that allows for rapid inference without a dependence on the pool size. In MULTIPOOL2, we extend the model to handle multiple linked loci using a Gaussian process regression framework to describe the unobserved allele frequency in each pool. Both MULTIPOOL methods generalize to include biological replicates and case-only or case-control designs for binary and quantitative traits.

Our increased information sharing and principled inclusion of relevant error sources improve resolution and accuracy when compared to existing methods, localizing associations to single genes in several cases. MULTIPOOL is freely available at `http://cgs.csail.mit.edu/multipool/` and MULTIPOOL2 is available at `http://cgs.csail.mit.edu/multipool2/`.

2.1.1 Targeted experiments

We focus on model organism experiments where two strains are crossed and the progeny are grouped and pooled according to phenotype. We describe and model experiments for haploid organisms that are hybrids between two strains, but we note that the models we develop should generalize to more sophisticated crosses or diploid organisms. When two strains vary in a phenotype, analyzing progeny with extreme phenotypes should elucidate the genetic basis of the trait. The main idea is that polymorphic loci that do not affect the phenotype will segregate with approximately equal frequency in the progeny (regardless of phenotype), while loci that influence the trait will be enriched in opposite directions in the extreme individuals, according to the effect size of each locus. This approach assumes that the causal loci have sufficiently strong main effects to be detectable via any type of pooled analysis. This pooled study design is also referred to as "bulk segregant analysis" [Michelmore et al., 1991] in model system genetics. Selection and pooling based on a quantitative phenotype can identify quantitative trait loci (QTLs), so this procedure can also be viewed as a type of pooled QTL mapping. Figure 2-1 illustrates the experimental design at a broad level, though there are many ways to design crosses and experimental selections to produce pools that may be analyzed by MULTIPOOL and MULTIPOOL2.

Bulk segregant analysis with high-throughput sequencing has been applied in yeast to study drug resistance in [Ehrenreich et al., 2010], high temperature growth in [Parts et al.,

2011], and viability on alternate carbon sources in [Wenger et al., 2010]. Related pooled sequencing experiments used fly [Andolfatto et al., 2011] and *Arabidopsis* [Schneeberger et al., 2009] model systems. In human, analogous pooled sequencing studies currently require target capture methods and a preselected set of candidate loci [Calvo et al., 2010].

Pools may be selected from a single phenotypic extreme, opposite extremes, or one extreme and a control sample. Pools may also be obtained by grouping based on binary traits rather than quantitative phenotype extremes. Early studies used microarrays for pooled genotyping [Borevitz et al., 2003, Hazen et al., 2005, Brauer et al., 2006], but recent developments in high-throughput sequencing allow its use as a direct genotyping platform. While genotyping or sequencing individuals is an alternate choice, the appeal of pooled analysis is the dramatic reduction in cost while maintaining the statistical power of large sample sizes. Specifically, pooled genotyping allows for sequencing costs that scale with the number of pools, rather than the number of individuals. When collecting large pools of individuals is relatively easy, as with certain model organism designs, pooled sequencing can lead to lower experiment costs by several orders of magnitude. See [Sham et al., 2002, Jawaid et al., 2002, Macgregor et al., 2008] for a discussion on pooled association studies and experiment design considerations.

2.1.2 Challenges

Pooled genetic mapping studies using high-throughput sequencing present a number of unique difficulties. The core statistical quantity of interest, the allele frequency in each pool, is observed only indirectly. The strain-specific read counts that are used to estimate the allele frequencies are corrupted by sampling noise at most reasonable sequencing depths, read mapping errors [Degner et al., 2009], reference genome inaccuracies, and biological bias during sample preparation. In addition, the allele frequency measurements are nonuniformly spaced along the genome, depending on the polymorphism structure between the strains of interest. As an illustration, we refer to the bottom two plots in Figure 2-1 which show simulated 50X average sequencing coverage using polymorphisms from two yeast strains. Linkage implicates a wide region along the shown chromosome, and the allele frequencies

estimated from read counts are noisy and not necessarily highest at the exact location of the causal allele.

However, the unbiased nature of genotyping via high-throughput sequencing results in nearly saturated marker coverage where almost all polymorphisms are queried. This avoids the laborious process of marker discovery and assay design required by earlier genotyping technologies. The dense marker coverage also allows for a high degree of information sharing, which motivates the methods underlying MULTIPOOL and MULTIPOOL2.

2.1.3 Previous statistical methods

Previous statistical approaches to analyzing pooled genotyping data have focused on alternate regimes where genetic markers are relatively sparse and measurements are relatively accurate. Often, only single loci are tested for association, necessarily ignoring data from nearby markers. Additionally, single-locus methods encounter difficulties with missing data, such as regions that are difficult to sequence or map or have very few polymorphisms.

Earlier work applied hidden Markov models (HMMs) to fine mapping within small regions with fewer number of markers [McPeek and Strahs, 1999, Morris et al., 2000], and was extended to pooled genotype measurements in similar scenarios [Johnson, 2007]. However, these methods relied on computationally intensive sampling methods and were applied to datasets with only a few dozen markers. Conceptually similar methods have been explored for human studies, focusing on utilizing haplotype structure in the analysis of pooled experiments [Homer et al., 2008]. In more recent pooled sequencing experiments, a sliding-window method was applied on p-values from local tests in [Ehrenreich et al., 2010], while a local weighted method motivated by a probabilistic model was given in [Parts et al., 2011]. However, these models do not explicitly model the location of the causal locus while considering all relevant marker data.

2.1.4 Approach

MULTIPOOL and MULTIPOOL2 are designed for experimental crosses and dense noisy genotyping, as obtained by sequencing, and handles datasets with tens or hundreds of thousands

of markers. We develop statistical models that can combine information across many nearby markers while accounting for the nonuniform noise levels introduced by varying sequencing depth and marker spacing. The specific advances we present with MULTIPOOL and MULTIPOOL2 include:

- A model-based framework that allows for information sharing across genomic loci and incorporation of experiment-specific noise sources. These methods improve on previous approaches that rely on heuristic techniques to select sliding window sizes, which may sacrifice resolution.
- Statistical tests using an information-sharing dynamic Bayesian network (DBN) or Gaussian process model (GP) that report robust location estimates and confidence intervals. The multi-locus methods allow for principled inference even in regions without strain-specific markers and reduce experimental noise when many markers are available.
- Extensions of our method to any number of replicates and multiple experimental designs, within the same principled statistical framework.

2.2 Methods for MULTIPOOL

We develop inference methods for the pool allele frequency at a particular genome position, given the pooled read samples. First, we propose generative models which describe the experimental process. Next, these models are used to construct likelihood-based statistics to assess the significance of associations in multiple experimental designs.

2.2.1 Obtaining allele frequency measurements

All sequencing reads from a particular pooling experiment are aligned to one strain's reference genome using the short read aligner `bwa` [Li and Durbin, 2010a]. To increase specificity, only uniquely-mapping reads are considered. In practice, any short read aligner that can produce or export its output to the standard SAM format is compatible with this workflow. Next, a whole-genome pileup is generated using `samtools` [Li et al., 2009a]. A genome

pileup lists the particular base calls at each genomic position, using the set of mapped sequencing reads. The genome pileup produces reference and non-reference allele counts at each base. Using single-strain sequencing data, strain-specific bases can be determined and identified in the pileup of the pooled experiments. The result is a list of allele-specific read counts at many polymorphic sites across the genome. The coverage of the marker sites will vary according to local sequencing depth and mappability [Degner et al., 2009], and the density will vary according to the local polymorphism level. A similar approach was applied to generate allele counts in [Ehrenreich et al., 2010].

2.2.2 Multi-locus model

We first present the motivation and details for the MULTIPOOL model, which we later will extend to the full MULTIPOOL2 formulation. MULTIPOOL uses a probabilistic model that considers one chromosome at a time and explicitly models the effect that recombination and pool size have on neighboring allele frequencies. The model is a dynamic Bayesian network that describes the changing allele frequencies in the pool along a chromosome. The chromosome of interest is segmented into discrete blocks of equal size. A hidden state corresponding to each block reflects the pool allele frequency in the pool at that locus, varying along the genome as recombination causes random fluctuations. Each locus may emit sequencing reads according to its local pool allele frequency (hidden state). These reads may originate from multiple markers falling within the same region or a single marker. When there are no polymorphisms or mappable reads available in a region, the locus has no emissions and therefore the observed data do not directly constrain the hidden state at that locus. Finally, a particular locus may include the causal gene and therefore be directly associated with the phenotype. We assume there is only one causal locus in an analyzed region. For the genetic mapping problem, the causal locus is unknown and the key inference task is identifying its location and degree of association with the phenotype.

Name	Description
N	Number of individuals in the pool
L	Length of the analyzed portion of the genome, in segments
p	Allele frequency at the causal locus
x_i	Unobserved pool allele frequency at segment i in the genome
x_c	Unobserved pool allele frequency at the causal locus in the region
y_i	Observed (noisy) pool allele frequency at segment i in the genome
d_i	Observed read depth at segment i in the genome
r	Recombination frequency between adjacent genome segments

Table 2.1: **Parameters used in the MULTIPOOL model.** We list the parameters used in the MULTIPOOL model, as introduced throughout the text. We highlight that the model operates over chunks or segments of the genome, so the indices in each vector (presented as scalars in the table) refer to segments and not bases.

2.2.3 Model specification

The pool is composed of N individuals. An unknown causal locus is linked to the phenotype and displays association with allele frequency $p \neq \frac{1}{2}$ in the population. Loci that are not associated with the phenotype and are not linked to the causal locus segregate at frequency $p = \frac{1}{2}$ in the population. The pool allele frequencies are unobserved and are given for each genome segment i by x_i, $i = \{1, .., L\}$. The observed allele frequency measurements y_i are obtained from the mapped sequencing reads. We also define d_i, the total informative reads at each locus. This quantity is determined by the local sequencing depth and the number of mappable polymorphisms. The recombination frequency r gives the probability of an odd number of crossovers between adjacent genome segments in one individual in the pool. We do not model crossover interference, and therefore assume that recombination events are independent along the genome. The dependencies encoded in this model can be expressed as a graphical model, shown in Figure 2-2. While the example figure shows a particular choice of the causal locus, the inference task consists of selecting among all possible choices (model structures) for the causal locus and the population allele frequency p. The population allele frequency is the allele frequency of the causal locus that would be observed in an infinitely-large pool (the population), and depends on the strength of the locus's association. Subsequent sections develop efficient methods for calculating likelihoods for all relevant model structures by reusing intermediate computations. We provide the variables used in the model descriptions in Table 2.1.

Emission probabilities

The probability of observing a set of sequencing reads conditioned on the pool fraction y_i at locus i and a total informative read count d_i can be calculated using the binomial distribution:

$$y_i \cdot d_i \sim \text{Bin}(d_i, x_i). \tag{2.1}$$

This formulation models the read count proportion exactly with a discrete model. An approximation, applicable to high read counts, can be obtained with a Gaussian distribution:

$$y_i \sim \mathcal{N}(x_i, \frac{x_i(1-x_i)}{d_i}). \tag{2.2}$$

Technical pooling variance that increases the local measurement noise, such as allele-specific PCR amplification bias, could be assumed to act in locus-independent manner and be modeled with increased variance in this expression.

Transition probabilities

In practice, the genome segments are chosen to be small enough so that r is effectively the probability of a single recombination event occurring. We can determine the transition probabilities from x_i to x_{i+1} by considering the k individuals that switch from the first strain to the second and the j individuals of the reverse case. We know $k \sim \text{Bin}(Nx_i, r)$ since each of the Nx_i individuals with the first strain's ancestry at locus i will switch strain type when a recombination event occurs, with probability r. Similarly, $j \sim \text{Bin}(N(1-x_i), r)$. Thus:

$$x_{i+1} = x_i - \tfrac{k}{N} + \tfrac{j}{N}. \tag{2.3}$$

Employing normal approximations for the binomial distributions and dividing by N, we obtain an approximation for the transition probabilities:

$$x_{i+1} \sim x_i - \mathcal{N}(x_i r, \frac{x_i r(1-r)}{N}) + \mathcal{N}((1-x_i)r, \frac{(1-x_i)r(1-r)}{N})$$

$$= \mathcal{N}(x_i(1-2r)+r, \frac{(1-r)r}{N}). \tag{2.4}$$

This formulation shows that the latent allele frequencies form a first-order autoregressive Gaussian process with mean $\frac{r}{1-(1-2r)} = \frac{1}{2}$ and variance $\frac{(1-r)r/N}{1-(1-2r)^2} = \frac{1}{4N}$, which can be verified with a single-locus analysis. This equivalence will be developed further in MULTIPOOL2, presented in Section 2.3.

Initial probabilities

The causal locus node induces a particular distribution over hidden states, depending on the selected population allele frequency p:

$$x_i \cdot N \sim \text{Bin}(N, p). \tag{2.5}$$

The normal approximation is:

$$x_i \sim \mathcal{N}(p, \frac{p(1-p)}{N}). \tag{2.6}$$

2.2.4 Inference: discrete model

Inference of the hidden state values can proceed outwards from the causal locus, using the conditional independence structure of the model. We describe the algorithms in terms of standard HMM techniques, but note that a more general treatment in terms of message passing is also possible.

The observed data likelihood $\Pr(\mathbf{y})$, conditioned on a particular causal allele at x_c (model structure) and population allele frequency p, is obtained by conditioning on the values of the causal locus:

$$\Pr(\mathbf{y}|p) = \sum_{j=0}^{N} \Pr(\mathbf{y}_{\{c+1,..,L\}}|x_c = \tfrac{j}{N}) \Pr(\mathbf{y}_{\{1,..,c-1\}}|x_c = \tfrac{j}{N}) \Pr(y_c|x_c = \tfrac{j}{N}) \Pr(x_c = \tfrac{j}{N}|p). \tag{2.7}$$

The first term in the sum operates on an HMM with rightwards arrows in its graph, while the second term operates on an HMM with leftwards arrows (see Figure 2-1). However, the latent states form a reversible Markov chain, allowing us to reverse the arrows in the left graphical model fragment. After this transformation, the likelihood computations for all choices of the causal node x_c use the same graphical structure over the latent states $\mathbf{x}$ when conditioned on the causal node x_c: two chains with all rightwards arrows, separated by the conditioned node x_c. Using this fact, we can compute the desired likelihoods with intermediate computations from a single graphical model.

We compute the product of the first three terms in the sum, $\Pr(\mathbf{y}|x_c)$, using the posterior distribution of x_c computed using an HMM with no causal locus ($\Pr(x_c|\mathbf{y})$). The posterior distributions are calculated using the forward-backward algorithm [Bishop, 2007, Murphy, 1999], using the transition and emission distributions given previously. The unconditional marginal distribution $\Pr(x_c)$ is computed using the stationary distribution of the latent allele frequencies in the noncausal model.

$$\Pr(\mathbf{y}|p) = \sum_{j=0}^{N} \Pr(\mathbf{y}|x_c = \tfrac{j}{N}) \Pr(x_c = \tfrac{j}{N}|p) = \Pr(\mathbf{y}) \sum_{j=0}^{N} \frac{\Pr(x_c = \frac{j}{N}|\mathbf{y})}{\Pr(x_c = \frac{j}{N})} \Pr(x_c = \tfrac{j}{N}|p) \quad (2.8)$$

Running the forward-backward algorithm requires considering all transitions in each chromosome block, leading to a runtime quadratic in the size of the pool: $O(N^2L)$. This dominates the cost for the final step of computing $\Pr(\mathbf{y}|p)$ for all causal locus locations and a fixed p, which is $O(NL)$. The quadratic dependence on the pool size renders the exact modeling of large pools prohibitive, motivating the continuous approximation in the next section.

2.2.5 Inference: continuous approximation

The previous inference procedure applied to discrete hidden states where the pool composition is modeled exactly, but yielded inference algorithms that require time quadratic in the size of the pool. For large pools, we can relax this requirement and avoid the quadratic burden by modeling the allele frequency as a continuous value. The graphical model is

linear-Gaussian since the transitions and observations are linear functions of the latent variables, subject to Gaussian noise. In a linear dynamical systems formulation, the model is:

$$x_{i+1} = x_i - x_i r + (1 - x_i)r + w = (1 - 2r)x_i + r + w, \tag{2.9a}$$

$$y_i = x_i + v_i. \tag{2.9b}$$

Where:

$$w \sim \mathcal{N}(0, \frac{(1-r)r}{N}), \tag{2.10a}$$

$$v_i \sim \mathcal{N}(0, \tfrac{1}{4d_i}). \tag{2.10b}$$

The per-locus observation noise v_i can be approximated with the sample variance from the observed y_i, depending on y_i and d_i, or upper bounded by $\frac{1}{4d_i}$. The posterior probabilities over the continuous latent states can be calculated with the Kalman filtering and smoothing equations, analogous to the two recursive functions used to calculate the posterior probabilities for HMMs [Bishop, 2007, Ghahramani and Hinton, 1996, Murphy, 1999]. The Kalman filtering equations yield the conditional distribution of the latent state given the preceding observations with a recursive estimate:

$$\Pr(x_i|\mathbf{y}_{\{1,..,i\}}) = \mathcal{N}(x_i; \mu_i, \sigma_i^2), \tag{2.11a}$$

$$\mu_i = (1 - 2r)\mu_{i-1} + r + K_i(y_i - (1 - 2r)\mu_{i-1} - r), \tag{2.11b}$$

$$\sigma_i^2 = (1 - K_i)P_{i-1}. \tag{2.11c}$$

Where:

$$P_{i-1} = (1-2r)^2\sigma_{i-1}^2 + \frac{r(1-r)}{N}, \tag{2.12a}$$

$$K_i = \frac{P_{i-1}}{P_{i-1} + \frac{1}{4d_i}}. \tag{2.12b}$$

The recursions begin with the stationary distribution parameters:

$$\mu_0 = \tfrac{1}{2}, \tag{2.13a}$$

$$\sigma_0^2 = \tfrac{1}{4N}, \tag{2.13b}$$

$$P_0 = \tfrac{1}{4N}. \tag{2.13c}$$

The Kalman smoothing equations use the filtered results (forward estimates) to create estimates of the hidden state using the entire observation sequence, recursing backwards:

$$\Pr(x_i|\mathbf{y}) = \mathcal{N}(x_i; \hat{\mu}_i, \hat{\sigma}_i^2), \tag{2.14a}$$

$$\hat{\mu}_i = \mu_i + J_i(\hat{\mu}_{i+1} - (1-2r)\mu_i - p), \tag{2.14b}$$

$$\hat{\sigma}_i^2 = \sigma_i^2 + J_i^2(\hat{\sigma}_{i+1}^2 - P_i). \tag{2.14c}$$

Where:

$$J_i = \frac{\sigma_i^2(1-2r)}{P_i}, \tag{2.15a}$$

$$\hat{\mu}_L = \mu_L, \tag{2.15b}$$

$$\hat{\sigma}_L^2 = \sigma_L^2. \tag{2.15c}$$

As in the discrete section, the posterior distributions of the latent states under a null model can be used to compute the desired data likelihoods for all possible causal models. Required integrals are computed numerically using a fixed number of points. Specifically:

$$\Pr(\mathbf{y}|p) = \int_0^1 \Pr(\mathbf{y}|x_c = j) \Pr(x_c = j)\, \mathrm{d}j = \Pr(\mathbf{y}) \int_0^1 \frac{\Pr(x_c = j|\mathbf{y})}{\Pr(x_c = j)} \Pr(x_c = j|p)\, \mathrm{d}j. \tag{2.16}$$

Since the probability distributions during inference are represented with a constant number of parameters instead of a full vector (as in the discrete case), inference is more efficient. Specifically, computing the required quantities $\Pr(x_c = j|\mathbf{y})$ for all c requires $O(L)$ time. This removes the dependence on the size of the pool that was present in the discrete method, allowing MULTIPOOL to perform accurate inference in very large pools.

2.2.6 Statistical tests for a single experiment

With these computations in place, we can compare all values of the causal locus and the trait association, measured by p. For each locus, we construct a likelihood ratio statistic comparing the hypotheses of association and no association:

$$LR(c) = \frac{\max_{p'} \Pr(\mathbf{y}|p = p')}{\Pr(\mathbf{y}|p = \frac{1}{2})} = \max_{p'} \sum_{j=0}^{N} \frac{\Pr(x_c = \frac{j}{N}|\mathbf{y})}{\Pr(x_c = \frac{j}{N})} \Pr(x_c = \tfrac{j}{N}|p = p'). \tag{2.17}$$

The simplification occurs because the likelihood under the noncausal hypothesis at any locus is the same, namely $\Pr(\mathbf{y})$ from the noncausal HMM. A similar likelihood is obtained with the continuous model:

$$LR(c) = \max_{p'} \int_0^1 \frac{\Pr(x_c = j|\mathbf{y})}{\Pr(x_c = j)} \Pr(x_c = j|p = p')\, \mathrm{d}j. \tag{2.18}$$

We perform the maximization over p' numerically and calculate the likelihood ratio for

all positions of the causal locus by reweighting the posterior probabilities.

2.2.7 Statistical tests for multiple experiments

We can analyze replicate experiments by forming a coupled dynamic Bayesian network. This analysis present two replicates, but the methods generalize to any number of coupled experiments. In this situation, the same sampling distribution is induced at the shared causal locus in two coupled chains. The joint data likelihood factors since the chains are conditionally independent given the selection node p:

$$LR(c) = \max_{p'} \frac{\Pr(\mathbf{y}_1, \mathbf{y}_2 | p = p')}{\Pr(\mathbf{y}_1, \mathbf{y}_2 | p = \frac{1}{2})} = \frac{\max_{p'} \Pr(\mathbf{y}_1 | p = p') \Pr(\mathbf{y}_2 | p = p')}{\Pr(\mathbf{y}_1) \Pr(\mathbf{y}_2)}. \tag{2.19}$$

The maximization over p' must consider the product of the data likelihoods in the replicates. For designs where paired experiments are expected to show opposite effects, each experiment selects an optimal population allele frequency p. In this case, the null hypothesis is the coupled model where the two experiments share the same population allele frequency. The likelihood ratio is:

$$LR(c) = \frac{\max_{p_1, p_2} \Pr(\mathbf{y}_1 | p = p_1) \Pr(\mathbf{y}_2 | p = p_2)}{\max_{p_3} \Pr(\mathbf{y}_1, \mathbf{y}_2 | p = p_3)} = \frac{\max_{p_1} \Pr(\mathbf{y}_1 | p = p_1) \max_{p_2} \Pr(\mathbf{y}_2 | p = p_2)}{\max_{p_3} \Pr(\mathbf{y}_1 | p = p_3) \Pr(\mathbf{y}_2 | p = p_3)}. \tag{2.20}$$

The numerator is the product of two single-experiment maximizations, while the denominator is the coupled model likelihood that was presented for replicate analysis.

Using these results, MULTIPOOL reports $\log_{10}$ likelihood ratios (LOD scores in the genetics community), maximum-likelihood estimates (MLE) of the causal locus location, and approximate credible intervals for the location of the causal locus. Assuming a uniform prior over causal locus locations, $\Pr(x_c | \mathbf{y}) \propto \Pr(\mathbf{y} | x_c)$ for a particular set of observations $\mathbf{y}$. In each case we fix p at its MLE, but could alternately integrate it out. Therefore, we can compute multi-locus statistics that include information from the entire dataset in experiments where multiple pools are available.

2.3 Methods for MULTIPOOL2

The formulation for MULTIPOOL2 is based on and depends on the modeling approach for MULTIPOOL, but generalizes it in a way that allows for more flexibility in hypothesis testing and considering nearby linked causal loci. We will present it briefly here, based on the modeling decisions made in Section 2.2.

We start with Equation 2.4 from the MULTIPOOL model, which describes the unobserved allele frequency in the pool under the null model of no causal locus. As noted, the unobserved allele frequencies form an autoregressive Gaussian process, which is completely defined with a mean function and covariance function. This Gaussian process is equivalent to a Ornstein-Uhlenbeck process, which has been previously described in the QTL literature for mapping populations [Lander and Botstein, 1989]. The corresponding kernel in this case is the absolute exponential function [Williams and Rasmussen, 2006].

We note that the kernel-based smoothing interpretation of the Gaussian process in this case is similar to the approach taken in [Magwene et al., 2011], who instead apply a tri-cube kernel as an empirical choice based on similarity to existing LOESS methods [Cleveland, 1979].

Inference in this model proceeds by using standard Gaussian process regression techniques. Each marker with observed sequencing reads from the two possible alleles is used as a noisy frequency estimate, with a observation noise dependent on the sequencing depth (as in Equation 2.2). Imputation of the unobserved allele frequencies in the pool produces posterior estimates of the allele frequencies, with the key added benefit of covariance terms between any pair (or larger group) of candidate markers for hypothesis testing [Murphy, 2012]. These terms allow for the accurate comparison of data likelihoods while fixing one or both alleles to a specific value, permitting the detection of association in nearby linked genes. These covariance terms could not be calculated using the Kalman filtering approach given for the original version of MULTIPOOL.

The downside of this new approach is that inference, at least with standard Gaussian process regression algorithms, requires $O(L^3)$ computation (assuming densely-spaced markers throughout the region of length L). When a user does not wish to analyze a region for

multiple linked associations, the simpler computational demands of MULTIPOOL (linear in L) result in more practical time requirements.

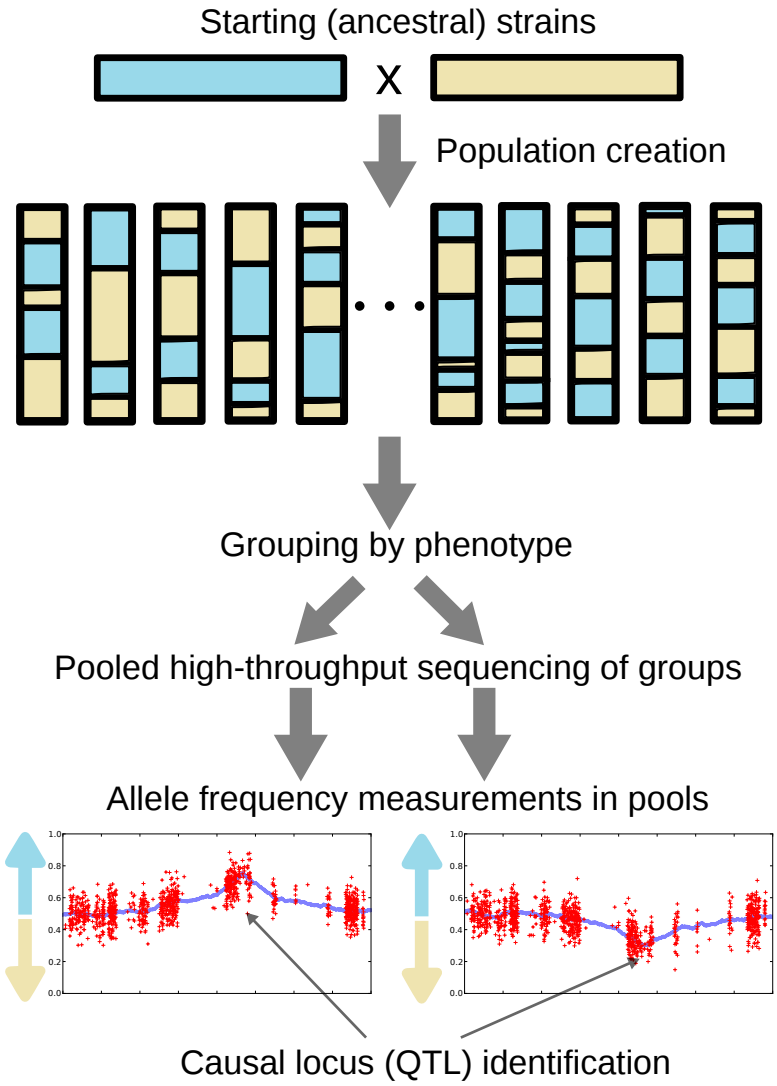

Figure 2-1: **Pooled sequencing experimental design example.** Strains are crossed and hybrid progeny are collected. The progeny are grouped by phenotype and the pooled DNA of each group is subjected to high-throughput DNA sequencing. Loci that affect the phenotype show an enrichment for one strain in each pool, while other unlinked loci segregate evenly. The bottom two plots show simulated (unobserved) allele frequencies in the pool with blue lines and (observed) allele frequencies computed from simulated 50X sequencing coverage in red.

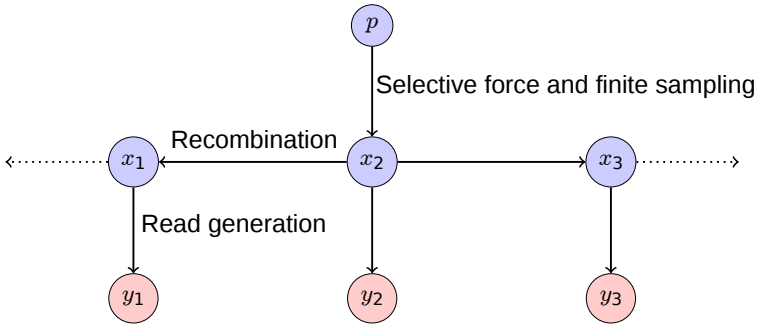

Figure 2-2: **Graphical model showing multi-locus dependencies.** Dynamic Bayesian network used by MULTIPOOL to capture the dependence between nearby loci in a pooled sequencing experiment. Allele frequencies in the pool influence the mix of observed sequencing reads at each locus. Here, the causal allele is x_2 and its value is determined by sampling N individuals to create the pool from the population allele frequency p (where $p \neq \frac{1}{2}$ indicates association).

Chapter 3

Application of statistical models for pooled sequencing

In this chapter, we present a range of results using the computational models of MULTIPOOL and MULTIPOOL2. We begin with a set of simulations, where the causal locus location is known exactly. We then proceed to mapping experiments where the causal loci has been verified in previous studies or is otherwise already known. Finally, we present novel data applying the MULTIPOOL family of models to problems in yeast genetics. We include supporting data for several hypotheses generated by the mapping results, and we discuss how they complement or support published findings in the literature.

We note that since the publication of the first version MULTIPOOL [Edwards and Gifford, 2012], the method and implementation has been used by other groups to advance genetic mapping studies in a broad range of traits [Albert et al., 2014b, Clowers et al., 2015, Treusch et al., 2015]. For a detailed comparison of pooled sequencing studies and computational methods, including a broader range of experimental designs than contemplated here, see [Schlötterer et al., 2014] or [Schneeberger, 2014].

3.1 Simulation results

In order to understand the benefit of MULTIPOOL versus single-locus tests on deeply-sequenced pools, we conducted a series of simulations. A causal locus was chosen with

population allele frequency $p = 0.75$ and many pools of sizes $N = 100$, 1000, and 10000 were created. SNP locations and relative per-SNP sequencing depths were calculated from experimental datasets in yeast. Average read coverage (sequencing depth) was varied from 10X to 150X, and 100 datasets of each type were simulated. The MLE causal allele location was calculated with the single-pool DBN model. The single-locus test analyzed allele frequencies computed with 1kb sliding windows, based on the method in [Jawaid et al., 2002]. The root mean square errors for each pool size and sequencing depth is shown in Figure 3-1. In the $N = 100$ cases, the mapping accuracy is predominantly controlled by the small pool size. This leads to little improvement with increased sequencing depth. The larger pools show higher accuracy with increased sequencing depth, but MULTIPOOL is always more accurate with a lower sequencing requirement. These simulations were conducted without additional read mapping noise or other noise sources, and so the absolute results should be interpreted conservatively.

3.2 Mapping results with previously validated causal loci

We also analyze pooled sequencing data recently generated by two groups [Ehrenreich et al., 2010, Parts et al., 2011]. The groups generated haploid yeast individuals with hybrid backgrounds from two strains and performed various phenotypic selections. Table 3.1 lists the datasets and their sequencing depths. While each experiment generated many statistically significant novel results, we limit ourselves to mapping comparisons involving target genes that have been validated using targeted follow-up experiments. We note that even though a target gene may be verified as affecting the trait, an untested nearby gene may affect the localization results.

3.2.1 Single-locus comparisons

In cases where the associated region is localized to a single gene, we compare the LOD scores from MULTIPOOL to a likelihood ratio computed using allele frequencies calculated by summing allele read counts in sliding windows. The data likelihoods under the causal and noncausal models are calculated according to the model in [Jawaid et al., 2002], with the

genotyping noise calculated from the local informative read depth. We use 50-bp genome segments in the dynamic Bayesian network (DBN) model and set the recombination rate in the model to the empirical average in yeast [Mancera et al., 2008].

3.2.2 Large pool results

The first set of large pools was used to characterize the genetic basis of resistance to the DNA-damaging agent 4-NQO. The genes *RAD5* and *MKT1* were validated as affecting 4-NQO resistance with follow-up experiments, so we use them as test cases for our model. The control pools showed no association around the validated loci, so we applied MULTIPOOL's one-pool test for association using the continuous model. Table 3.2 shows the distances from the MLE peak estimate to the middle of the target gene from MULTIPOOL and sliding-window tests.

MULTIPOOL localizes *RAD5* to within the gene body, without a dependence on choosing an appropriate sliding window size. The 90% credible interval of the location contains six genes, centered on *RAD5*. A localization example using one replicate is shown in Figure 3-2. *MKT1* is localized to within 3 kb of the coding sequence, with the 90% credible interval covering *MKT1* and eight other genes.

The second set of large pools was constructed to study the genetics of heat tolerance, using repeated crosses to reduce linkage disequilibrium (increased r in our model). In this study, the genes *IRA1* and *IRA2* were verified as affecting heat sensitivity with direct assays. Table 3.3 reports the distance from the MLE estimate to the center of the target gene using MULTIPOOL and sliding window methods. *IRA2* is localized to within the coding sequence, but the predictions for *IRA1* are consistently upstream of the gene's location. Upon further investigation, the peak around *IRA1* appears to contain another (untested) associated locus. Figure 3-3 plots the estimated allele frequencies and LOD scores in the surrounding region. The relevant 90% credible intervals for the causal locus location include *IRA2* in all datasets, but do not include *IRA1*. This may support the hypothesis of another linked gene in the associated region.

Name	Read length	Pool size	Cov. (rep. 1)	Cov. (rep. 2)	Source (ref.)
4-NQO viable	76	≈10000	67.7	85.0	[Ehrenreich et al., 2010]
Control	76	≈10000	36.1	79.5	[Ehrenreich et al., 2010]
Heat tolerant	76 (paired)	≈10000	152.4	84.8	[Parts et al., 2011]
Control	76 (paired)	≈10000	79.0	75.2	[Parts et al., 2011]

Table 3.1: **Analyzed experiments for published large pool datasets.** Each condition was assayed with two replicates. Coverage is the average reads per marker. Due to the protocols used, precise quantification of the pool size is difficult. We used the listed values as conservative choices since the reported ranges are larger in most cases.

Dataset	Target	DBN dist.	1kb window dist.	10kb window dist.
4-NQO viable rep. 1	*RAD5*	5305	18355	14605
4-NQO viable rep. 2	*RAD5*	745	6195	3145
Combined	*RAD5*	805	755	3145
4-NQO viable rep. 1	*MKT1*	3223	15127	1673
4-NQO viable rep. 2	*MKT1*	5223	15127	5423
Combined	*MKT1*	4323	15127	5423

Table 3.2: **Localization of known associated genes in large drug-selected pools.** Distances are reported in bases from the MLE to the center of the target gene.

Dataset	Target	DBN dist.	1kb window dist.	10kb window dist.
Heat tol. rep. 1A	*IRA1*	10589	16739	14739
Heat tol. rep. 1B	*IRA1*	10889	20689	6389
Heat tol. rep. 2	*IRA1*	8889	2589	17289
Heat tol. rep. 1A	*IRA2*	311	3240	511
Heat tol. rep. 1B	*IRA2*	961	17670	1661
Heat tol. rep. 2	*IRA2*	340	4190	2390

Table 3.3: **Localization of known associated genes in large heat-selected pools.** Distances are reported in bases from the MLE to the center of the target gene. The results for *IRA1* suggest an additional associated gene; see the text and Figure 3-3.

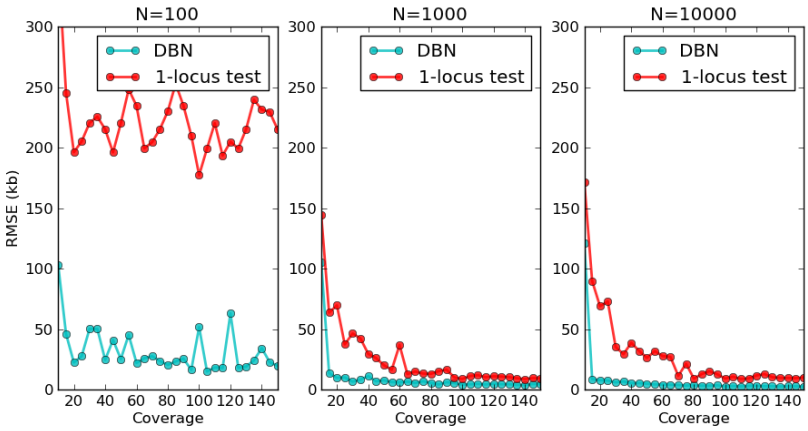

Figure 3-1: **Mapping accuracy in simulated datasets.** Mapping accuracy is shown as root mean square error (RMSE) in kilobases (kb) from the known location. The coverage reports the average sequencing depth (reads per marker) in the experiment. Each point is calculated using 100 simulated experiments. The DBN points show the accuracy of the MLE using the MULTIPOOL's one-pool test, while the 1-locus test shows the accuracy of directly testing allele frequencies calculated in 1-kb sliding windows.

3.3 Mapping the genetic basis of strain-specific essential genes

A recent yeast genetics study uncovered a surprising phenomenon where two closely-related yeast strains had different sets of essential genes [Dowell et al., 2010]. That is, for dozens of genes, a deletion in one strain was lethal but the same gene deletion in the other strain was viable. This intricate background-dependent response to a genetic perturbation emphasizes the complexity of understanding genotype-to-phenotype relationships. Extending the analogy to human disease genetics, this finding demonstrates the difficulty of interpreting and predicting the impact of potential disease-causing variants in individual patients.

The original study conducted a comprehensive gene deletion survey of the *Saccharomyces cerevisiae* strain Σ1278b and compared the set of discovered essential genes to the set of essential genes in the reference strain S288c, which has already been defined [Giaever et al., 2002]. After validation experiments to confirm the high-throughput measurements, 44 genes were discovered to only be essential in Σ1278b, while 13 were only essential in S288c. Genetic analysis of Σ1278b-S288c hybrids (Table S3 in [Dowell et al., 2010]) suggested that the genetic basis of several strain-specific essential genes was complex and depended on the inheritance of more than a single suppressor, but no mapping experiments or other

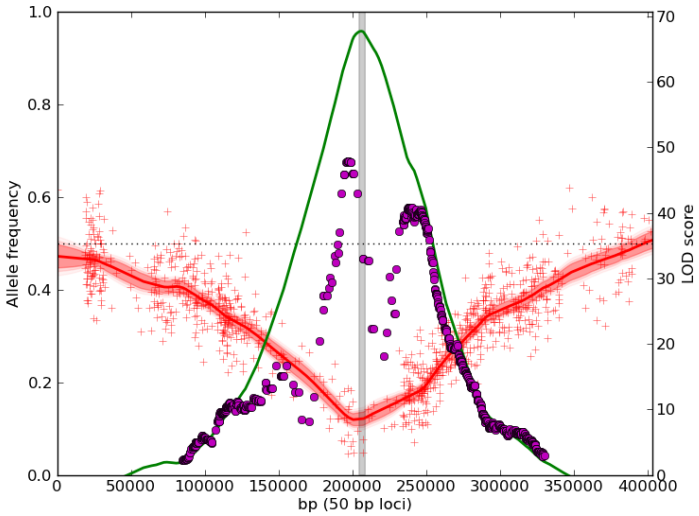

Figure 3-2: **Localization of *RAD5* using 4-NQO selected replicate 2.** The red line and shaded region show the inferred allele frequencies in the pool using MULTIPOOL, and the red pluses plot the observed allele frequencies from the sequencing data (scale on left axis). Regions without pluses do not have polymorphisms or mappable reads. The magenta dots show LOD scores computed using tests of allele frequencies calculated using 10kb sliding windows, while the green line shows the LOD scores calculated using MULTIPOOL (scale on right axis). The gray box shows the position of *RAD5*, the verified causal gene in the region.

direct assays were performed to support this conclusion. Here, we apply pooled sequencing mapping and the MULTIPOOL methods to obtain a greater understanding of the genetic basis of strain-specific essential genes in yeast.

3.3.1 Methods

The goal of the mapping experiments is, for a given strain-specific essential gene, to identify the genetic elements that allow the strain to survive in one background. That is, we are searching for suppressor elements from the non-essential strain that buffer or otherwise interact with the given gene deletion and result in the strain-specific phenotype.

We followed the same general strategy for all studied strain-specific essential genes, with minor technical differences in obtaining knockouts or identifying phenotypic extremes. Diploid hybrids of Σ1278b and S288c were constructed, with a single strain-specific gene of interest missing (deleted) from both strains or only one strain but marked with an antibiotic marker for later recovery. These diploids underwent meiosis and haploid F_1 hybrids with

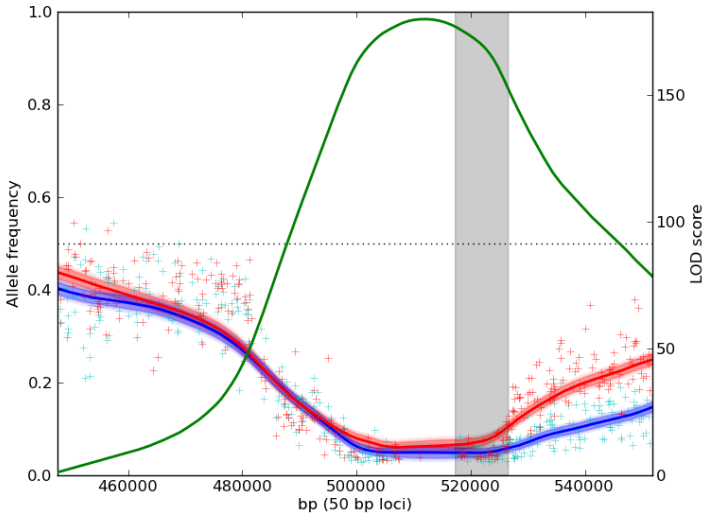

Figure 3-3: **Localization of *IRA1* using heat tolerant replicates 1 and 2.** The red and blue lines and shaded regions show the inferred allele frequencies in the two replicates using the DBN method, and the pluses plot the observed allele frequencies. The green line shows the LOD scores calculated using the DBN two-pool method. The gray box shows the position of *IRA1*, the reported and verified association in this region. However, an uncharacterized association upstream of *IRA1* may be the cause of the extended range of low allele frequencies and the shifted estimate of the peak location.

mosaic genomes consisting of ancestry from both parents were obtained. If one active copy of the gene of interest was present in the cross, the antibiotic resistance marker was used to recover only haploids without the gene. Haploids present at this stage were viable without the gene of interest, meaning they possessed a suppressor from one strain (or equivalently, lacked an interacting "lethality" allele from the strain where the deletion was lethal). These hybrids were pooled and genomic DNA was extracted and sequenced. Table 3.4 lists the studied genes and generated pools, along with basic sequencing statistics from the analyzed pools.

For the smaller pools (fewer than 300 individuals), the haploid members were obtained by manual tetrad dissection. For the larger pools, haploid members were obtained by collecting random spores and haploid status was guaranteed by counterselection against diploids (canavanine or thialysine resistance via *CAN1* and *LYP1* alleles) [Yan Tong and Boone, 2006]. For Pool 4 (*ski7*Δ) and Pool 6 (*ret2*Δ), individual haploids that could not survive the gene deletion were identified by replica plating and then testing via 5-FOA

Name	Read length	Pool size	Coverage
Pool 1 (ski7 viable, kil-k)	76	70	150.9
Pool 2 (ski7 viable, kil-k)	76	70	138.6
Pool 3 (ski7 viable, kil-k)	39	≈ 600	141.8
Pool 4 (ski7 inviable, kil-k)	39	≈ 300	154.7
Pool 5 (ret2 viable, kil-0)	39	288	89.9
Pool 6 (ret2 inviable, kil-0)	39	186	95.6
Pool 8a (ret2 viable, kil-0)	40	960	90.2
Pool 8b (ret2 viable, kil-0)	40	960	84.7
Pool 8c (ret2 viable, kil-0)	40	960	47.0
Pool 8d (ret2 viable, kil-0)	40	960	48.1
Pool 9a (ski7 viable, kil-0)	40	960	42.6
Pool 9b (ski7 viable, kil-0)	40	960	70.9
Pool 9c (ski7 viable, kil-0)	40	960	58.4
Pool 9d (ski7 viable, kil-0)	40	960	66.0
Pool 10a (bem1 viable, kil-0)	40	960	7.6
Pool 10b (bem1 viable, kil-0)	40	960	7.2
Pool 10c (bem1 viable, kil-0)	40	960	17.6
Pool 10d (bem1 viable, kil-0)	40	960	7.8

Table 3.4: **Analyzed experiments for strain-specific essential genes.** Individual yeast spores that were either viable or inviable after a gene deletion were pooled and analyzed. Coverage is the average reads per marker. For details on the strain construction and pooling methods between each pool, see the text.

counterselection to remove a *SKI7* plasmid [Boeke et al., 1987] or temperature sensitivity with a *RET2*-ts allele [Li et al., 2011].

The haploid individuals were combined in pools of varying sizes and their genomic DNA was sequenced using Illumina instruments. Table 3.4 shows the pooling strategy and read coverage for each experiment. Markers were obtained by analyzing single-strain sequencing data [Dowell et al., 2010] and identifying segregating variants [Li et al., 2009b, Li and Durbin, 2010a], as described in Section 2.2.1.

3.3.2 Mapping results

Here we discuss several of the key findings of the mapping experiments here. First, we use the contrast testing mode of MULTIPOOL to identify differences between pools of opposite phenotypic extremes. In the context of the strain-specific essential gene suppressor mapping study, these are F_1 individuals that were either viable or inviable after deletion of a strain-specific essential gene. Figure 3-4 shows the LOD scores across the genome for *ski7*Δ and

*ret2*Δ pools (pools 3 vs. 4 and 5 vs. 6, using the labels in Table 3.4). After excluding known positive controls in the mapping results, we observe that each strain-specific gene has two strong (LOD>10) suppressor signals. The positions and strengths of each candidate suppressor are given in Table 3.5.

We also study larger pools obtained by random spore analysis, with results given in Table 3.6. Interestingly, in these experiments the *ski7*Δ pools show no associated peaks beyond the positive controls arising from the strain construction designs. This is in sharp contrast to the earlier *ski7*Δ results, shown in Figure 3-4 and Table 3.5, where two strong suppressors were observed. Further examination of the strains used in these studies showed that the pools where suppressors were observed possessed a dsRNA that encodes the yeast K1 killer toxin [Magliani et al., 1997, Schmitt and Breinig, 2006] (labeled as kil-k in Table 3.4), while the pools where no suppressors were observed did not (labeled as kil-0). Thus, we conclude that the *SKI7* strain-specific essential trait is killer-dependent, while the *BEM1* and *RET2* traits are not. More details on the killer virus interaction effect, including single-strain validation experiments, are given in the next section.

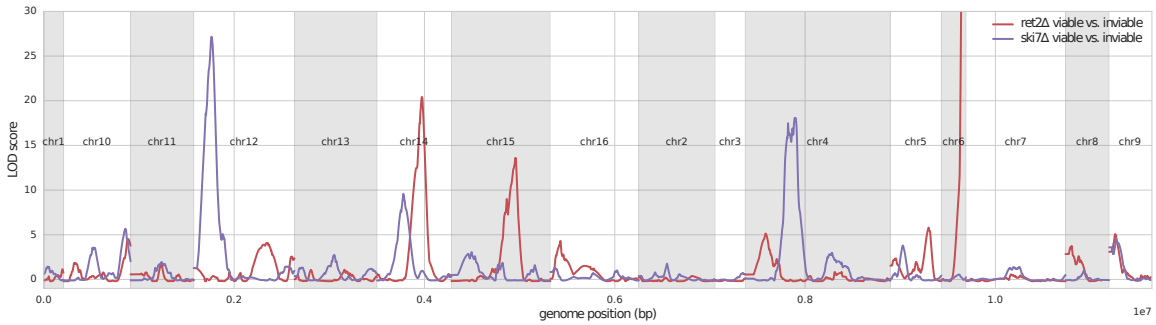

Figure 3-4: **Mapping results for *ski7*Δ and *ret2*Δ viability modifiers.** Genome-wide results using MULTIPOOL in contrast mapping mode to find significant differences between pooled allele frequencies (pools 3 vs. 4 and 5 vs. 6). The peak on chr6 for the *ret2*Δ pool is the location of the *RET2* locus, which had a fixed strain ancestry due to the strain construction pattern in this experiment. Two strong (LOD>10) suppressors are observed for each strain-specific essential gene.

3.3.3 Validation of a killer virus link

To determine the prevalence of chromosomal and killer virus [Magliani et al., 1997, Schmitt and Breinig, 2006] interactions, we analyzed 17 single gene deletions with a growth defect

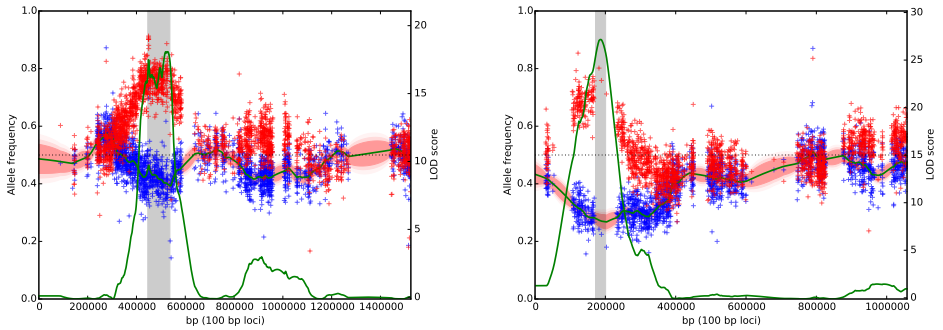

Figure 3-5: **Fine mapping of suppressors interacting with *ski7*Δ on chromosomes 4 and 12.** The red line and shaded region show the inferred allele frequencies in the pool using MULTIPOOL, and the red and blue pluses plot the observed allele frequencies from the sequencing data (scale on left axis). Regions without pluses do not have polymorphisms or mappable reads. The green line shows the LOD scores calculated using MULTIPOOL in contrast mode, comparing the two pools (scale on right axis). The gray box shows the 90% credible interval reported by MULTIPOOL.

Strain-specific gene	Chromosome	Region	Max LOD score
RET2	chr14	454800-483200	20.4
RET2	chr15	634100-682700	13.5
SKI7	chr4	446400-537100	18.1
SKI7	chr12	172100-200800	27.2

Table 3.5: **Associated suppressors identified by comparing opposite phenotypic extremes.** The reported region is the 90% confidence interval reported by MULTIPOOL.

in Σ1278b for interactions with the killer virus, with full results presented in [Edwards et al., 2014]. Although these constructed deletion variants are not observed in natural populations, we note that hundreds of putative natural loss-of-function variants are observed across multiple wild and laboratory yeast strains [Liti et al., 2009, Schacherer et al., 2009, Strope et al., 2015].

We generated all four possible combinations of alleles to study the effects of a gene deletion interacting with the killer virus, as shown in Figure 3-9. We found that 6 of the

Strain-specific gene	Chromosome	Region	Max LOD score	Genes
BEM1	chr15	672800-682200	125.2	*DCI1*, *LAS17*, *RPS30B*, *FYV12*, *SER1*, *GSP2*
SKI7	(none)	(none)	n/a	n/a

Table 3.6: **Associated suppressors identified with single pools.** The reported region is the 90% confidence interval reported by MULTIPOOL.

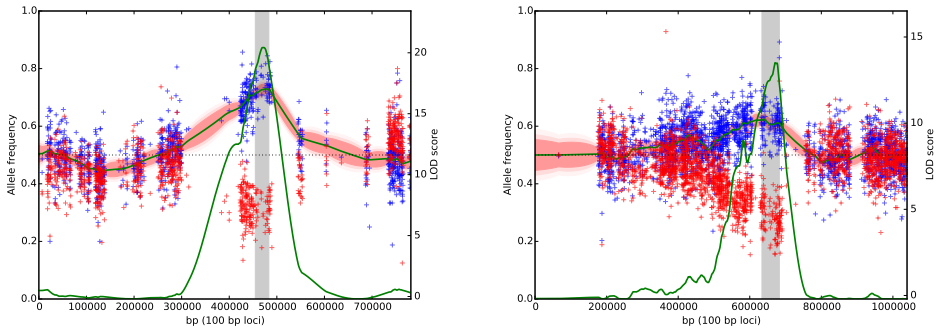

Figure 3-6: **Fine mapping of suppressors interacting with *ret2*Δ on chromosomes 14 and 15.** The red line and shaded region show the inferred allele frequencies in the pool using MULTIPOOL, and the red and blue pluses plot the observed allele frequencies from the sequencing data (scale on left axis). Regions without pluses do not have polymorphisms or mappable reads. The green line shows the LOD scores calculated using MULTIPOOL in contrast mode, comparing the two pools (scale on right axis). The gray box shows the 90% credible interval reported by MULTIPOOL.

17 gene deletions (*pep7*Δ, *pep12*Δ, *pho88*Δ, *ski8*Δ, *vps16*Δ, and *ski7*Δ) grew more slowly when the strain contained the dsRNA virus. The inhibitory effect of the dsRNA virus varied from total (*ski7*Δ) to slight (*vps16*Δ). The colonies are shown in Figures 3-10 and 3-11. The dsRNA encodes a toxin that is secreted and kills strains lacking the dsRNA virus, but it has not been known to kill cells that carry it under our conditions. Strains carrying the dsRNA virus are resistant to killing by the toxin [Pagé et al., 2003, Wickner, 1992].

Supporting the mapping experiments presented here, the most extreme chromosomal mutation-dsRNA interaction was with the *ski7*Δ allele, which is lethal in the presence of the dsRNA (the heterozygous diploid +/*ski7*Δ gives rise to two viable and two dead *ski7*Δ haploid progeny; Figure 3-10) and viable in the absence of the nonkiller dsRNA. Thus, the lethality or viability of the *ski7*Δ deletion in the Σ1278b background is completely dependent on the nonchromosomal information despite having the same chromosomal DNA sequences. In contrast, the S288c background tolerates the deletion with and without the killer toxin, as shown in Figure 3-12.

We next sought to assess the connection of the whole set of strain-specific essential genes from [Dowell et al., 2010] to the killer virus, even though only a selected subset were studied with mapping experiments reported here and single-strain experiments in [Edwards et al., 2014]. We compared the list of strain-specific essential genes to the results from a

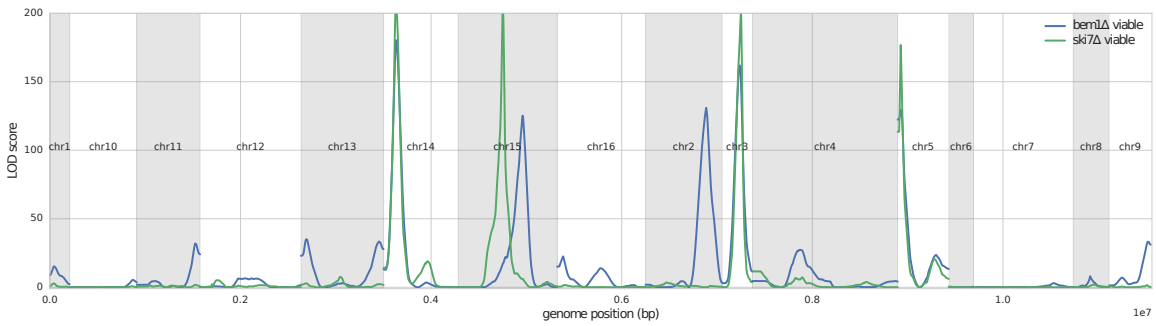

Figure 3-7: **Mapping results for *bem1*Δ and *ski7*Δ viability modifiers.** Genome-wide results using MULTIPOOL in single-pool mapping mode to find significant differences between pooled allele frequencies in large pools (pools 9 and 10, with replicates combined). The peaks on chr14, chr3, and chr5 are positive control peaks arising from the haploid-specific markers used in the cross (*CAN1*, *LYP1*, and *MAT*). The peak on chr15 in the *ski7*Δ pool is the *SKI7* locus and the peak on chr2 in the *bem1*Δ pool is the *BEM1* locus. Therefore we observe that *BEM1* has one strong suppressor and *SKI7* has none, in this cross.

Name	Sigma-specific essential	S288c-specific essential	Not strain-specific essential
Killer sensitive genes	5 *	0	150
Killer resistance genes	5 **	0	66
Not identified in study	34	13	≈ 6000

Table 3.7: **Overlap between strain-specific essential genes and killer-associated genes.** The gene sets reported in [Dowell et al., 2010] and [Pagé et al., 2003] are compared, and a statistically significant enrichment between sigma-specific essential genes and genes whose knockouts are associated with killer toxin phenotypes is observed. In the table, * denotes significance with $p < 0.0012$ and ** denotes significance with $p < 0.000015$, both using a hypergeometric test.

genome-wide survey that used the yeast knockout collection to find genes that conferred increased or decreased sensitivity to the K1 killer toxin [Pagé et al., 2003]. The results, given in Table 3.7, demonstrate a statistically significant overlap between the Σ1278b-specific essential genes as a class and the killer virus, but not the S288c-specific essential genes. This may be due to the fact that Σ1278b typically possesses the killer virus in the "wild", whereas S288c does not [Fink and Styles, 1972].

3.3.4 Complex genetic basis of strain-specific essential genes

Considering the mapping results in total, we observe that for two of the three studied strain-specific essential genes, at least two strong interacting genes (candidate suppressors) are detected (Figures 3-5 and 3-6). And qualitatively, there may be a longer tail of weaker

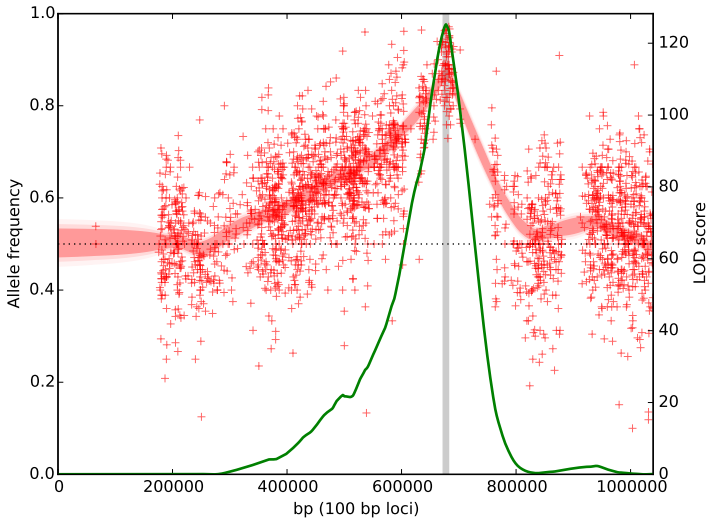

Figure 3-8: **Fine mapping of a suppressor interacting with *bem1*Δ on chromosome 15.** The red line and shaded region show the inferred allele frequencies in the pool using MULTIPOOL, and the red pluses plot the observed allele frequencies from the sequencing data (scale on left axis). Regions without pluses do not have polymorphisms or mappable reads. The green line shows the LOD scores calculated using MULTIPOOL (scale on right axis). The gray box shows the 90% credible interval reported by MULTIPOOL.

suppressors that are also present but with lower significance scores in the mapping results (Figure 3-4). The strain-specific gene with the simplest apparent genetic model of one interacting gene, *BEM1*, was also reported as the simplest genetic architecture in indirect experiments in [Dowell et al., 2010] (Table S3).

We note that this genetic complexity, arising completely from wild variants segregating in the studied pairs of yeast strains, is in addition to the interaction effects with the cytoplasmic yeast killer virus discussed in the previous section. This multilayered complexity and interaction pattern is another demonstration of the difficulty of predicting phenotype from genotype without sophisticated models.

3.4 Conclusions

We presented MULTIPOOL and MULTIPOOL2, computational methods to map genetic elements from pooled sequencing studies. Taking advantage of recent increases in throughput, these experimental designs use sequencing to provide unbiased and labor-efficient genotyping. As

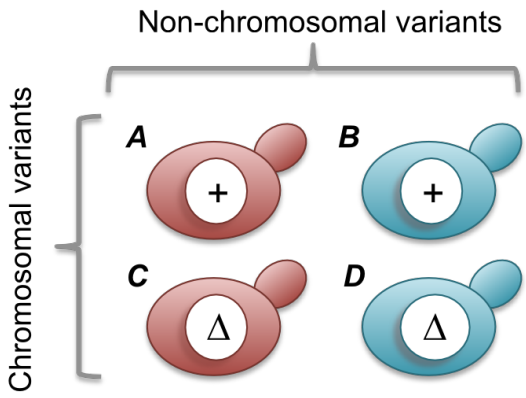

Figure 3-9: **Experimental design for assessing nonchromosomal and chromosomal genetic interactions.** To determine whether particular nonchromosomal factors (red and blue represent different cytoplasmic information, e.g., the yeast killer virus) interact with a gene deletion variant (denoted by + and Δ symbols in the nucleus), we construct all four possible haploid strains combining these two nuclear and nonchromosomal factors. Phenotypic measurements of the four controlled genotypes are then compared to understand the effect the nonchromosomal element has on the phenotype. A genetic mechanism controlled only by the chromosomal gene deletion will result in strains *C* and *D* showing a similar growth defect relative to strains *A* and *B*, whereas an interaction between chromosomal and nonchromosomal genotype could yield a growth defect confined to strain *D*.

throughput continues to increase, similar studies will be extended to larger and more complex genomes. By including all relevant data in a unified framework, the MULTIPOOL methods improve the analysis of these experiments with increased accuracy and the principled estimation of association intervals. The statistical framework is most beneficial for the case where there are many noisy markers, as observed in genotyping via sequencing. In these cases, combining information across the genome is critical in reducing noise and increasing statistical power. More generally, the methods developed and applied in this work support the application of selection and pooled genotyping for experimental organisms. When experimental procedures can create medium or large allele frequency differences, the responsible genes can be mapped with great precision. These methods do not require the step of explicit polymorphism discovery or genotyping array design, yielding large time and cost savings.

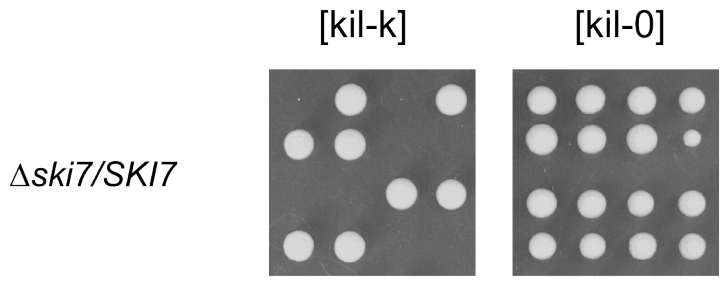

Figure 3-10: ***ski7*Δ viability is dependent on killer virus in a Σ1278b background.** Each row is the result of a dissection of meiotic products from a diploid. The four spores from a single meiosis were placed from left to right in each row. In all tetrads, the larger two colonies are those with the wild-type chromosomal allele. The *ski7*Δ mutation is viable in the absence of dsRNA virus [kil-0] (Right) and lethal in the presence of dsRNA virus [kil-k] (Left). All of the colonies in the right panel lack dsRNA virus ([kil-0]). In the panel on the left the two spore clones that failed to grow in each quartet contain the deletion. The two viable clones are wild type (+) and contain the dsRNA virus ([kil-k]).

3.4.1 Future computational work

Future work could replace our uniform prior over possible causal locus locations with an informative prior that uses conservation data, functional information, or other relevant data types (as in [Lee et al., 2009]). More ideas in this direction, though applied to unpooled study designs, are explored in Part II of this thesis. Other extensions include a more subtle handling of read mapping ambiguities and SNP calling uncertainty. One possibility is to use expected (average) counts under an error-aware probabilistic model instead of hard assignments, which should scale gracefully as certainty lowers. This could reduce our reliance on a particular aligner and SNP calling strategy.

3.4.2 Biological insight from pooled sequencing studies

Our mapping studies analyzing the basis of strain-specific essential genes shows that the pattern of genetic interactions is complex and the phenotype of a mutation may be modified by inherited viral state. The mapping results extend and complement the indirect genetic data presented in [Dowell et al., 2010] and provide a small candidate set of potential suppressors to test with single-gene validation experiments.

Our confirmation of a killer virus effect shows that the nonchromosomal contribution to heritability can be large and, in some cases, can completely mask the effect of a chromosomal

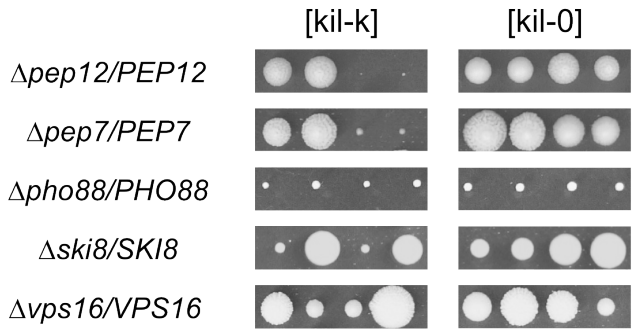

Figure 3-11: **The lethality of multiple gene deletions is dependent on killer virus in a Σ1278b background.** Each row is the result of a dissection of meiotic products from a diploid. The four spores from a single meiosis were placed from left to right in each row. In all tetrads, the larger two colonies are those with the wild-type chromosomal allele. The presence of the dsRNA virus inhibits the growth of several mutations in the Sigma background. Some of the mutants with the dsRNA virus grew extremely slowly and were visible only after 10 days of incubation. For each mutation the meiotic spores with the dsRNA virus and the one without it were dissected on the same plate.

mutation (a gene deletion in our study). Nonchromosomal elements may have affected previous yeast studies [Sinha et al., 2008, Steinmetz et al., 2002, Ben-Ari et al., 2006, Deutschbauer and Davis, 2005, Kim and Fay, 2009] that crossed a strain carrying a dsRNA virus, as many feral yeast strains do [Drinnenberg et al., 2011], with a virus-free strain such as the reference strain S288c [Fink and Styles, 1972].

Previous yeast studies analyzing the basis of quantitative traits (quantitative trait locus mapping) have either not carefully controlled nonchromosomal modifiers or have fixed them so that their influence is eliminated. Our results complement one such study that recovered the chromosomal determinants accounting for almost all of the additive portion of heritability of several traits by dramatically expanding study sizes [Bloom et al., 2013]. In this previous study, potentially confounding nonchromosomal effects were mitigated by standardizing on a single mitochondrial background and by using only dsRNA virus-free strains [Edwards et al., 2014]. The control of nonchromosomal factors in model organism experiments and the inability to do so in "wild" human populations could account for part of the recent success gap between model- and human-focused genetic studies.

Our findings on the relative ubiquity of nonchromosomal genetic effects have profound implications for the association between disease susceptibility and genetic variation in hu-

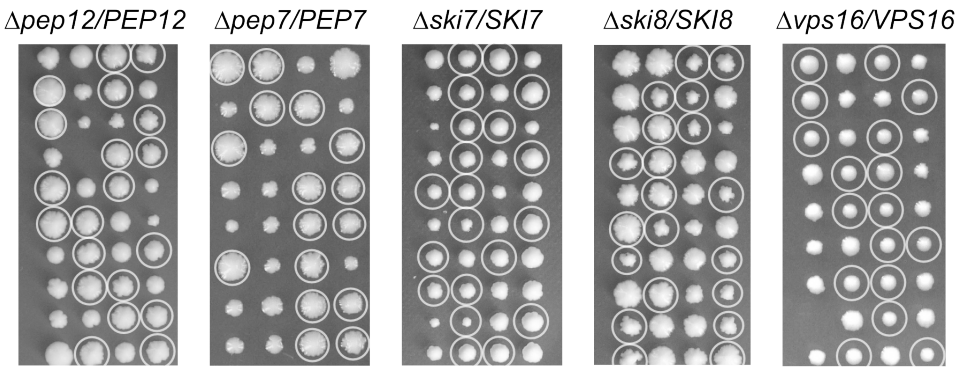

Figure 3-12: **Chromosomal variants exhibit a minimal dependence on dsRNA presence in an S288c background.** Selected heterozygous gene deletion strains (+/Δ) were analyzed in an S288c [kil-k] background, as opposed to the Σ1278b [kil-k] strains discussed earlier or the wild type S288c [kil-0] background. Each row shows four products from a single meiosis, with the wild type (+) spores circled and the two spores with the deletion allele (Δ) remaining unmarked. The growth advantage of the wild type spores is not significantly different from 1.0 (equal growth) for *SKI7*, *SKI8*, and *VPS16* (p=0.12, 0.04, and 0.35 respectively, Mann-Whitney *U* test). For *PEP7* and *PEP12*, the wild type spores have a small growth advantage (2.41 and 1.78 respectively) that is more similar to the advantage observed in Σ1278b [kil-0] strains (1.47 and 1.82) than the Σ1278b [kil-k] strains (89.35 and 64.67). The growth advantage is calculated as the average over all tetrads of the averaged wild type colony sizes divided by the averaged knockout colony sizes. This quantity is used as a robust unitless measurement that allows for comparison across multiple plates and imaging batches.

mans. For the viral interaction case, these elements are not currently captured by genotyping assays, and therefore current studies cannot measure their impact. However, they may be inherited or manifest as a shared environmental factor. Both cases could contribute to the complexity of modeling disease heritability. The inclusion of nonchromosomal interactions adds another dimension to the estimation of heritability in wild populations and susceptibility to common diseases in humans.

Part II

Leveraging functional annotations to improve genetic mapping

Chapter 4

Statistical models for integrating functional annotations with genetic mapping

4.1 Introduction

Standard approaches to genetic mapping use the correlation between genotype and phenotype in a mapping population to implicate certain genetic markers as associated with a phenotype of interest. However, these statistical approaches treat all candidate causal loci as identical, and discard any additional or external information available to distinguish variants. In this work, we propose a statistical model that integrates external information about genetic markers to improve predictions about what types or patterns of annotations are likely to be suggestive of causal regions. We propose and demonstrate models that combine genetic mapping results and sets of functional annotations to discover common annotation patterns that may underlie genetic associations.

The intuition behind our approach here is to take advantage of recent sets of high-throughput measurements describing the genome in order understand the basis of genetic associations. For instance, large sets of cell-type specific expression measurements or chromatin accessibility data may shed light on which particular regions or variants are

linked to a disease phenotype. Here we propose to use a set of annotations, including molecular measurements or categorical descriptions, to learn flexible probability rules that are able to distinguish associated from non-associated genetic regions in a mapping study. Multiple groups have reported the enrichment of genetic associations in certain classes of genomic regions identified by molecular measurements. The methods we discuss in Section 4.1.2 present many representative examples of these results, including [Hu et al., 2011, Karczewski et al., 2013, Maurano et al., 2012, Parker et al., 2013, Paul et al., 2011, Paul et al., 2013, Schaub et al., 2012, Trynka et al., 2013].

Specifically, we use labeled groups of genetic variants and annotation or feature vectors describing them to build probabilistic models that enable one to predict the group-level labels of associated blocks. The spatial correlation pattern of genetic variants in a population (linkage disequilibrium) causes associated regions to be typically reported as blocks that contain multiple candidate causal variants. These blocks of variants typically contain only a small number of truly causal variants in relation to the size of the block, so using the whole block as positive training data is inaccurate. This motivates the use of "multi-instance" machine learning [Foulds and Frank, 2010], a supervised learning setting where training labels are given over batches (termed bags) over training examples (termed instances, hence the name multi-instance) where some of the instances in a bag may not be associated with the bag label. In this chapter, we develop and implement several new variants of multi-instance learning models and apply them to simulated classification and genetic datasets. In Chapter 5 we apply these models to a yeast genetic mapping study.

4.1.1 Challenges

We encounter several modeling challenges when tackling this problem:

- Positive labels are available only for groups of variants, making informative patterns that distinguish positive and negative examples harder to learn. As discussed, this is due to inheritance patterns in the mapping populations where nearby variants are likely to be inherited from the same parent or ancestors.

- The annotation model (underlying classification model) may be nonlinear or contain

interaction terms between separate components. The candidate annotations we include in our models have different sources, may be correlated in certain regions, and may reflect complex conditional relationships.

- We may have many training instances, separated in labeled bags of varying sizes. A learning algorithm should handle different types and scales of input and be able to run efficiently on large datasets.

The algorithms we develop, implement, and test in the next sections aim to address these challenges.

4.1.2 Related work

In the genetics literature, two recent models attack the same problem of using external information to improve genetic mapping studies [Pickrell, 2014, Kichaev et al., 2014]. The models are called `fgwas` and `PAINTOR`, respectively. They both use observed genetic association strength, in the form of Bayes factors, and binary associations to form probabilistic models for identifying causal variants. Some differences in our approach from these existing techniques include:

- We do not need to explicitly model the number or presence of causal variants within each associated region. `fgwas` models one causal locus per fine mapping region, while `PAINTOR` can in practice handle up to three causal loci. Our multi-instance formulation is agnostic to the number of causal variants per region, and does not require their particular locations to be estimated.

- In contrast to `fgwas` and `PAINTOR`, we do not use the observed association strength in each positive bag. We effectively binarize these measurements and treat all associated regions equally. This may be a strength when analyzing multiple studies or when summary statistic data per-locus is not available, but it discards useful information when it is available. On the other hand, both `fgwas` and `PAINTOR` assume that the annotation model correlates with genetic association, where the real relationship could

be closer to a threshold behavior where the magnitude of the effect at a causal locus depends on other factors.

- One of our proposed models has a richer instance-level probability model [Hornik, 1991, Collobert and Bengio, 2004], which can be extended further in future work to include more layers [Bengio, 2009]. The two existing models only use linear combinations of the features.

Other existing genetics approaches focus on identifying relevant annotations or molecular measurements underlying reported regions from human genome-wide association studies [Farh et al., 2015, Finucane et al., 2015, Maurano et al., 2012, Trynka et al., 2013], along with reweighting approaches for dissecting associated regions in a fine-mapping context [Kichaev et al., 2014, Trynka et al., 2015].

In the machine learning literature there are several existing approaches that convert logistic regression into the multi-instance setting by combining instance-level probabilities [Ray and Craven, 2005, Raykar et al., 2008, Zhang et al., 2005]. All of these models use linear models to form the logistic regression log-odds decision function.

4.2 Statistical models

We start with a set of labeled training instances $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ...\}$ where the binary label is $y_i \in \{-1, 1\}$ and the feature vector for training example i is $\mathbf{x}_i$. We use a logistic regression model where the conditional distribution of the label is Bernoulli with the probability of being a positive instance ($y = 1$) defined by a sigmoid function of a linear combination of weights on the feature vector:

$$\Pr(y_i = 1|\mathbf{x}_i) = \sigma(\mathbf{w} \cdot \mathbf{x}_i + b) = \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{x}_i - b)}. \tag{4.1}$$

4.2.1 Converting instance probabilities to bag probabilities

In the multi-instance setting training labels are given over bags or groups of instances rather than single instances. A positive bag may not necessarily consist of all positive members, and

typically we are interested in challenging problems where only a small number of instances in each positive bag are positive. We thus require a way to transform instance-level probabilities to bag-level probabilities that reflects this property. We first examine the "noisy-OR" model, where the probability of having a positive bag is calculated using the probability that not all bags are negative. This approach, or related variants of it, has been applied in several earlier multi-instance learning approaches [Ray and Craven, 2005, Raykar et al., 2008, Zhang et al., 2005].

$$\Pr(y_j^{\text{bag}} = 1|\mathbf{X}) = 1 - \prod_{i \in \text{bag}_j} [1 - \Pr(y_i = 1|\mathbf{x}_i)] \tag{4.2}$$

To keep the notation simple, we let bag_j be the set of instance indices for all instances belonging to bag j. We assume, following the standard multi-instance learning framework, that negative bags are accurately labeled. That is, when a bag is labeled negative, all instances in it are truly negative. Therefore, we convert all negative bags into singleton bags with negative labels, so that the bag mapping probabilities present above avoid requiring different logic to handle negative bags.

Alternate choices for bag probability models

Another choice is to simply assign the probability of the bag as the instance probability with the highest probability of being positive. While this approach lacks the probabilistic justification of the noisy-OR calculation, it has appealing empirical properties in that it does not depend on the size of a bag or the addition of additional negative instances to the bag. The bag-level mapping function for this case is:

$$\Pr(y_j^{\text{bag}} = 1|\mathbf{X}) = \max_{i \in \text{bag}_j} [\Pr(y_i = 1|\mathbf{x}_i)]\,. \tag{4.3}$$

In our implementations we use the following "smooth max" approximation for computational simplicity, with $p_i \equiv \Pr(y_i = 1|\mathbf{x}_i)$ for notational simplicity and with α as a tuneable

smoothing parameter:

$$\Pr(y_j^{\text{bag}} = 1|\mathbf{X}) = \frac{\sum_{i \in \text{bag}_j} p_i \exp(\alpha p_i)}{\sum_{i \in \text{bag}_j} \exp(\alpha p_i)}. \tag{4.4}$$

We observe that setting α to 0 results in an arithmetic mean of the instance probabilities and in the limit as α increases towards infinity, the function becomes a maximum of the instance probabilities. For all our experiments we set α to 10.

4.2.2 Optimization and model fitting

Under the logistic regression variants of this multi-instance classification approach, we identify optimal weights by maximizing the observed data likelihood over all bags:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\text{argmax}} \prod_j \Pr(y_j^{\text{bag}}|\mathbf{X}). \tag{4.5}$$

The weight vector $\mathbf{w}$ is used in the underlying instance probability models (e.g. Equation 4.1) which are combined to make the bag-level probability terms (via Equations 4.2 or 4.3). In all model variants presented here, we optimize the weight vectors by gradient descent using a symbolic computation engine [Bergstra et al., 2010, Bastien et al., 2012]. We ran experiments to determine optimal learning parameters, and we used these to set defaults of a learning rate of 0.1 and the addition of a Nesterov momentum term in the gradient descent procedure [Sutskever et al., 2013].

4.2.3 More complex instance-level probability models

We can extend this logistic regression approach by using a richer instance-level probability model. In order to capture nonlinear effects and possible interactions among features, we employ a multilayer perceptron (MLP) [Rosenblatt, 1958]. The instance-level probability is modeled as

$$\Pr(y_i = 1|\mathbf{x}_i) = \sigma(\mathbf{w} \cdot \mathbf{z}(\mathbf{x}_i)), \tag{4.6}$$

where $\mathbf{z}$ is a hidden layer with size H so that the final layer with weight vector $\mathbf{w}$ also has size H, instead of the size of the feature vectors $\mathbf{x}_i$ as in the original logistic regression model.

The hidden layer, using the notation of [Murphy, 2012], is:

$$\mathbf{z}(\mathbf{x}_i) = g(\mathbf{V}\mathbf{x}_i) = [g(\mathbf{v}_1 \cdot \mathbf{x}_i), ..., g(\mathbf{v}_H \cdot \mathbf{x}_i)]. \tag{4.7}$$

Each unit in the hidden layer has its own weight vector $\mathbf{v}_k$ and applies a nonlinearity function $g(\cdot)$ to its particular weighting of the input features. Here we used a sigmoid function for $g(\cdot)$, though another choice arising in more modern applications would be rectified linear units (ReLU) [Glorot et al., 2011]. We have omitted the bias terms in the MLP equations above for simplicity, but in our implementations we include them in both layers and for each hidden unit (for a total of $H+1$ bias terms). In all experiments reported here, we employ $H = 5$ hidden layers in our "deep" multi-instance logistic regression models.

4.2.4 Detailed comparison to related models

An earlier model, `fgwas` [Pickrell, 2014], proposes a related model to learn empirical priors using external annotations for genetic associations. The `fgwas` model uses a combined region-level and variant-level modeling approach to incorporate external weighting to explain a set of genetic associations. First, we ignore the variant-level modeling or assume it produces a uniform prior on each variant, and investigate the region-level scoring. Following the approach of [Maller et al., 2012] (Supplementary Note, Section 6.3.2), we can combine the Bayes factors for all variants in a given region i by summing them and obtaining a single BF_i^{reg}. This relies on a uniform prior over the possible causal loci in this region, which we have assumed here. This results in a simplified version of the data log likelihood function given in Equation 11 of [Pickrell, 2014]:

$$LL(\mathbf{w}) = \sum_i \log[1 + \sigma(\mathbf{w} \cdot \mathbf{x}_i^{\text{reg}})(BF_i^{\text{reg}} - 1)]. \tag{4.8}$$

The dependence on genotypes and phenotypes from a particular study is encapsulated in the Bayes factors BF_i^{reg}, which can be computed using only summary statistics. Here we use the notation of this work where the (regional) annotations are in the feature vectors $\mathbf{x}_i^{\text{reg}}$ and the modeling task is to learn the optimal weight vector $\mathbf{w}$. We contrast that to the

logistic loss observed in a plain logistic regression approach, where instead of a Bayes factor for each region we have a binary label y_i:

$$LL(\mathbf{w}) = \sum_i \log[1 + \exp(-y_i(\mathbf{w} \cdot \mathbf{x}_i))]. \tag{4.9}$$

We see that at a regional level, the `fgwas` formulation encourages strong associations (large Bayes factors) to have strong prediction probabilities by the unbounded inclusion of the Bayes factor terms. As the `fgwas` and `PAINTOR` models are tailored for human genetic applications with dramatically different linkage disequilibrium structure than our yeast case studies here, we defer a full comparison and competitive benchmarks to future work.

4.3 Reweighting association studies

We follow a simple approach to include the evidence we obtain with our annotation-based model as informative empirical prior. We consider the posterior odds ratio of association by expressing it as the product of the data likelihood ratio and a prior odds term:

$$\frac{\Pr(\text{association}|\text{data})}{\Pr(\text{no association}|\text{data})} = \frac{\Pr(\text{data}|\text{association})\Pr(\text{association})}{\Pr(\text{data}|\text{no association})\Pr(\text{no association})}. \tag{4.10}$$

In typical association studies, the prior odds term is the same for all variants, so it does not affect the overall distribution of association scores. Here, however, we employ a model that yields varying prior probabilities of association, based on external functional annotations. Therefore we use the log prior odds from our annotation-based model to reweight the log likelihood ratios observed in the original genetic mapping study:

$$\log \text{posterior odds} = \log \text{odds} + \log \text{prior odds}. \tag{4.11}$$

This schematic equation is general, and can apply with the log odds term on the right side of the equation being either a LOD score or a Bayes factor, depending on the particular statistical assumptions and testing procedures used in the original genetic mapping study.

4.4 Simulation results for multi-instance classification

In this section, we evaluate the performance of our multi-instance learning models on a simulated classification dataset. Here we temporarily leave the genetic application aside and focus on a general classification task. We evaluate four variants of our MI-LR models, described in Table 4.1. We compare them to existing classification model choices, logistic regression and support vector machines [Boser et al., 1992, Cortes and Vapnik, 1995]. For the existing models, there is no way to directly model the multi-instance aspect of the problem. Therefore, we apply the bag training label to each member instance of the bag. As the size of positive bags increases, this training label is increasingly inaccurate for the non-positive members of each positive bag.

	Noisy or (Eqn. 4.2)	Smooth max (Eqn. 4.4)
Logistic regression (Eqn. 4.1) Deep logistic regression (Eqn. 4.6)	MI-LR (noisy or) Deep MI-LR (noisy or)	MI-LR (smoothmax) Deep MI-LR (smoothmax)

Table 4.1: **Multi-instance classification model variants.** The rows show the instance-level probability models and the columns enumerate the bag-level probability models that combine the instance-level probabilities to form bag likelihoods. A complete probabilistic model is defined by choosing a variant at both modeling layers and optimizing the observed data likelihood according to that model.

4.4.1 Dataset

We analyze samples constructed following the "Madelon" dataset procedure from the NIPS 2003 Feature Selection Challenge [Guyon et al., 2004]. Each experiment consists of 2500 instances, equally split into positive and negative bags. Only one instance in each positive bag is truly positive, while the rest are negative. There are 40 candidate features, with 10 relevant to the classification label [Guyon et al., 2004]. We examined bag sizes of 1 (where the multi-instance models reduce to plain classification), 5, 10, and 20. As the bag sizes increase, the learning problem is more difficult both because the ratio of negative to positive instances in each positive bag increases and because, since we fixed the number of negative instances and total instances, the total number of (true) positive instances in the training dataset decreases.

4.4.2 Results

Each model is trained on bag-labeled data and tested using 4000 held-out instances generated with the same distribution as the training data. The instance-level labels for the test instances are used to evaluate the performance of each classifier. Figure 4-1 shows the receiver-operating characteristic curves for each model for the four tested bag sizes. The summarized areas under the ROC curve (AUCs) are given in Table 4.2.

We observe that with bags of size 1, the multi-instance logistic regression has the same performance as plain logistic regression, as expected. The smoothmax model performs slightly worse, which may reflect numerical or optimization issues in the implementation of the algorithm, since the objective function for bags of size 1 should be equivalent between the two multi-instance variants.

With smaller bag sizes, the deep multi-instance model variants outperform the linear models (MI-LR, logistic regression, and SVM). However, as the bag sizes grow, the richer models are more difficult to learn from the noisy bag labels and the deep models lose their advantage. When the bag size reaches 20, the deep models are outperformed by the simpler linear multi-instance models.

As the bag size increases, the non-multi-instance models (logistic regression and SVM) show degraded performance. The multi-instance models also have lower performance, but they are able to maintain improved performance compared to the other models.

Model	Bag size 1	Bag size 5	Bag size 10	Bag size 20
Logistic regression	0.913	0.851	0.775	0.631
SVM	0.913	0.841	0.765	0.645
MI-LR	0.913	0.869	**0.856**	**0.810**
MI-LR (smoothmax)	0.901	0.788	0.808	0.795
Deep MI-LR	**0.950**	**0.927**	**0.858**	0.785
Deep MI-LR (smoothmax)	0.946	0.802	0.817	0.794

Table 4.2: **Summarized performance of multi-instance classification models on simulated data.** The area under the receiver operating characteristic curve (AUC) for each model and a range of bag sizes is given. The curves underlying this data are shown in Figure 4-1. For small bag sizes, the deep multi-instance models outperform the other model choices. As bag sizes increase, the multi-instance models maintain greater performance than the non-multi-instance models and the deep multi-instance models lose their advantage.

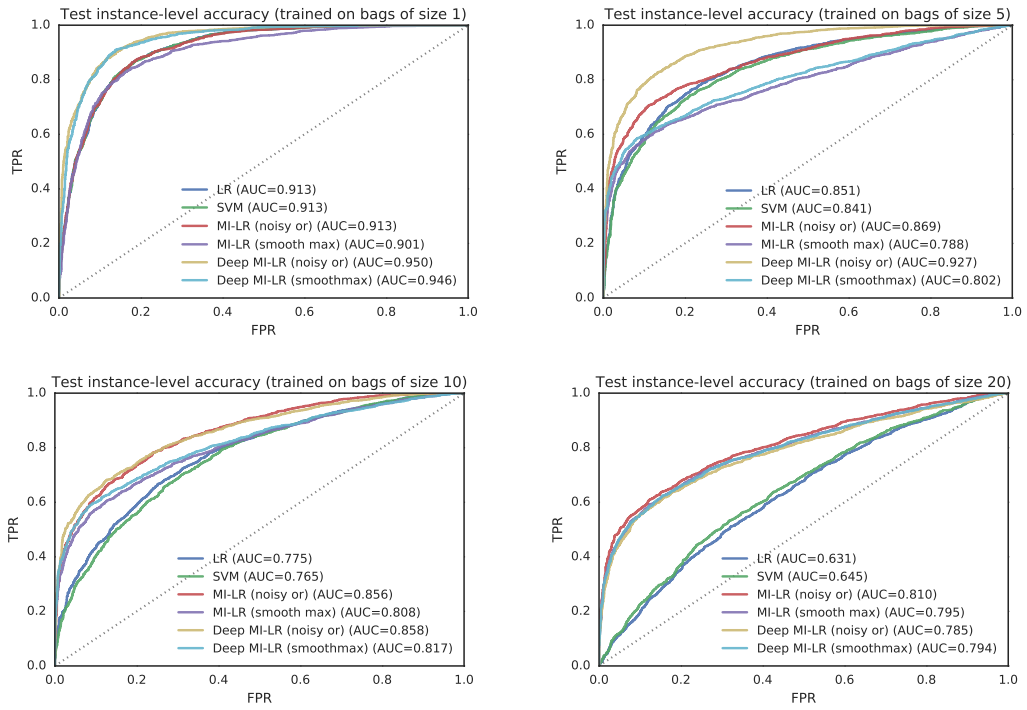

Figure 4-1: **Multi-instance classification model performance on simulated data.** Each model was trained on 2500 examples (instances), split evenly into negative and positive bags. The bag size ranged from 1 to 20 and separate results for each selected size are shown in each panel. The instance-level true positive rate is plotted against the instance-level false positive rate based on an evaluation on held-out test data.

4.5 Simulation results for a genetic mapping study

We next proceed to apply our multi-instance classification algorithms to a simulated genetics dataset. We will simulate a yeast cross with individual genotypes taken from a previous study [Bloom et al., 2013]. This is the same study that we will analyze using true functional annotation data in Chapter 5. Here, we simulate phenotypes and annotations so that we retain complete control and knowledge over which loci are causal. This allows us to evaluate the performance of each model as we test multiple modeling assumptions.

We use Equation 4.11 to combine the LOD scores from the simulated genotypes and the log prior odds from the trained annotation models we compare.

4.5.1 Dataset

The mapping panel we study are F_1 progeny of a cross between the yeast strains BY and RM11, so each individual has ancestry from roughly half of each parental strain on average. All the individuals are haploid. We obtain genotypes from 1008 individuals, using the authors' imputation procedure that assigns genotypes to all 11623 variants in this cross [Bloom et al., 2013].

We randomly select 50 causal variants, each with equal effect sizes with randomized directions, and set the total heritability of the trait to 50%. These parameters are in the range of reported biological values and are partially influenced by existing simulation techniques used in other work [Kichaev et al., 2014]. In yeast the true effect size spectrum depends on the trait of interest, but the distribution is non-uniform and the heritabilities are typically higher than we use here [Bloom et al., 2013, Ehrenreich et al., 2010]. We justify the uniformity assumption since it makes evaluation simpler and less variant-specific. In addition, it makes the overall simulation procedure more difficult for our algorithms, along with the lower heritability simulated here.

We use two synthetic functional annotations, with a baseline presence of 20% in all variants, and enriched four-fold in causal variants. These parameters are similar to those studied in previous genetic fine mapping evaluations [Kichaev et al., 2014]. To test the multi-instance learning procedure, we train our models on simulated associated regions of fixed size (the bag size) around each causal variant. We take 5% of the remaining variants to use as negative examples in the training phase. In contrast to previous approaches for this problem [Pickrell, 2014, Kichaev et al., 2014], we do not use the association strength at each locus when training our annotation model.

4.5.2 Results

After training the multi-instance and non-multi-instance models on the simulated associated regions of the genome together with the annotations, we evaluated the predictions of causal variant status genome-wide. We calculate a reweighted genetic association test using the log prior odds of each annotation model. In Figure 4-2 we show the receiver operating

characteristic curves for all variants, evaluating the prediction probabilities. Table 4.3 shows the summarized area under the receiver operating character curve statistics, comparing each method across a range of bag sizes. We see that the multi-instance logistic regression model outperforms all other models, except with bags of size 1 where the logistic regression model attains equal performance, as expected. We observe that the deep MI-LR model does not obtain good performance in this simulation, perhaps because the increased complexity of its models are too difficult to fit with the limited noisy training data present in these simulations.

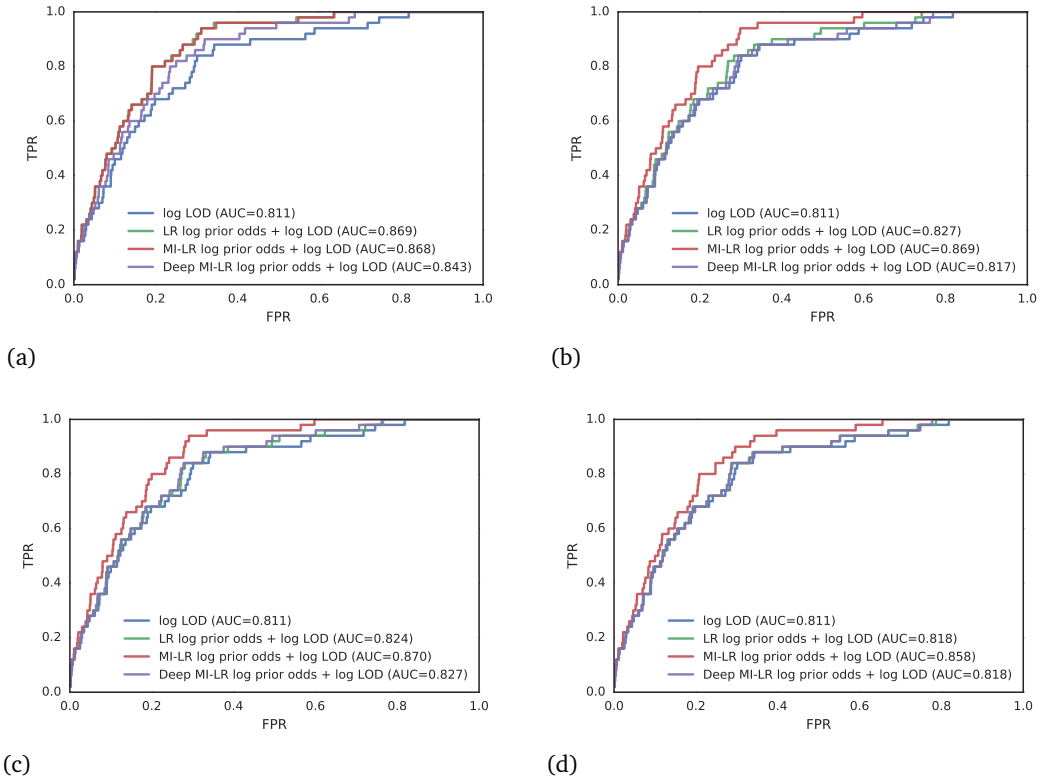

Figure 4-2: **Genetic mapping performance using a simulated yeast cross.** Each panel shows results from different bag sizes: (a) 1, (b) 5, (c) 10, and (d) 20. In all cases, the multi-instance logistic regression has the best performance. With bags of size 1, it reduces to logistic regression, which has the same performance.

Model	Bag size 1	Bag size 5	Bag size 10	Bag size 20
No prior	0.811	0.811	0.811	0.811
Logistic regression prior	**0.869**	0.827	0.824	0.818
MI-LR prior	**0.868**	**0.869**	**0.870**	**0.858**
Deep MI-LR prior	0.843	0.817	0.827	0.818

Table 4.3: **Summarized performance of multi-instance classification models on simulated yeast genetics data.** The area under the receiver operating characteristic curve (AUC) for each model and a range of bag sizes is given. The curves underlying this data are shown in Figure 4-2. With singleton bags, the logistic regression model is equivalent to the multi-instance logistic regression model. As bag sizes grow, the multi-instance model shows improved performance relative to the other approaches. In these simulations, the deep MI-LR model does not outperform the standard logistic regression model.

4.6 Conclusions

We have developed several multi-instance classification models and shown their success on multiple simulated tasks. In future work, we can consider iteratively mapping and estimating covariate models, instead of the single-pass approach we currently implement. We can also extend this work in the direction of previous fine mapping approaches [Kichaev et al., 2014, Pickrell, 2014] and incorporate the strength of association within each targeted region, or across the genome. Further extensions in this area may rely on multiple studies considered jointly, where the annotation model may also have some latent structure reflecting the similarity measures between related phenotypes.

Chapter 5

Genetic mapping using functional annotations applied to yeast genetics

5.1 Introduction

In this chapter, we apply the computational models developed in Chapter 4 to a compendium of data collected describing an existing yeast cross. The focus will be on collecting a large set of molecular measurements as well as sequence features, and we will use our statistical models to determine which functional covariates have the most influence in determining the probability of genetic association. Our goal here is twofold: (1) determining which classes of annotations have the most value in the yeast genetics context and (2) showing that our reweighting strategy has significant signal in identifying genetic blocks that are likely to include causal variants.

Example training data

We analyze a large yeast cross consisting of 1008 individuals, where complete genotype and phenotype information is available across 46 conditions [Bloom et al., 2013]. We consider reported associated regions across all conditions, as given by the original authors of the study. Figure 5-1 shows a small sample of annotation levels across a region of the yeast genome. In cases where associated regions overlap from multiple conditions, we only use

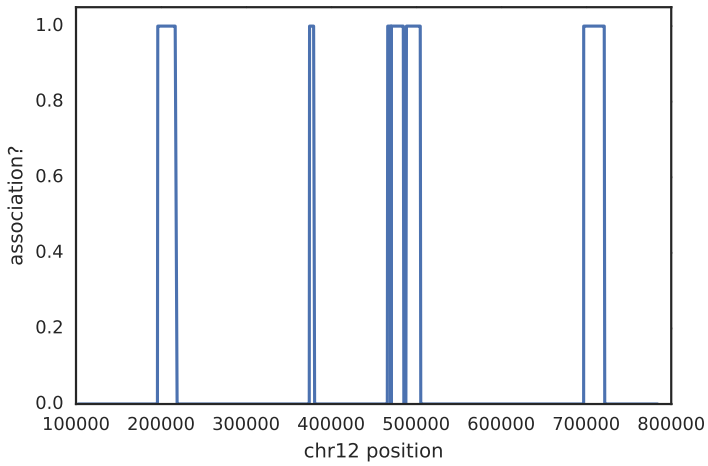

Figure 5-1: **Example training labels along the yeast genome.** Reported genetic associations from a yeast mapping study are used to identify and assign positive labels to contiguous blocks of the genome. Each block is used a positive bag, following the multi-instance learning framework. The variants that are not reported as associated with any condition are used as negative instances by our learning algorithms.

the one with highest reported association strength. Each contiguous block is used a positive bag in our multi-instance classification framework. We discuss how we treat the multiple conditions more fully in Section 5.4.

For candidate predictor features, we collect a large set of functional measurements describing yeast genes and individual variants. The details of the annotations we obtain and collect will be discussed in later sections, but a schematic example is shown in Figure 5-2.

5.2 Functional annotations

We list the functional annotations we collected and used in Table 5.1. We classify them into three broad categories: sequence-based, population genetic, or molecular. Sequence-based means the annotation class or label depends on direct examination of the variant in the context of known biology, such as if a variant is inside a gene or not. Population genetic means we classified variants based on their properties or distribution in related strains or species, and includes existing conservation metrics. Molecular is broader, and refers to the use of high-throughput datasets to identify strain- or condition-specific genes or genomic regions. In this study we pursue RNA-level measurements through assays like

Figure 5-2: **Example features for yeast genetic mapping.** A collection of binary and continuous features are collected and plotted across the yeast genome. The correlation scale locally depends on the particular type of feature and if it represents information about a region or gene or about a single variant. See Table 5.1 for a complete description of the annotations we used and their regional scales.

RNA-seq, protein-level measurements through ribosome profiling and shotgun proteomics, and chromatin accessibility through FAIRE-seq and DNase-seq.

5.2.1 Methods and data sources

Here we describe the sources of the functional annotations we used. For datasets constructed as part of this work, we go into greater detail on the assay designs and parameters. For datasets obtained from previous work, we describe the processing pipeline we employed and refer to the original publications for the details on the primary data collection.

Yeast strain details

Since the original study [Bloom et al., 2013] studied a cross between the yeast strains BY and RM11, we analyzed these two parental strains in multiple conditions. The strains were obtained from the Fink laboratory collection. The specific conditions are listed in Table 5.2, matching a subset of the conditions given in [Bloom et al., 2013].

Saccharomyces cerevisiae strains RM11 and S288c were cultured overnight at 30° in 5 mL liquid yeast peptone dextrose (YPD) starting from frozen stocks or single colonies scraped

Feature	Category	Source	Type
Coding or non-coding	Sequence	this work*	Per variant
Synonymous or non-synonymous	Sequence	this work*	Per variant
Indel or SNP	Sequence	this work*	Per variant
GERV score	Sequence	this work**	Per variant
Allele frequency across other strains	Population genetic	this work*	Per variant
PhastCons (conservation score)	Population genetic	UCSC	Per variant
SIFT (conservation score)	Population genetic	UCSC	Per variant
Strain-specific mRNA (RNA-seq)	Molecular	this work	Per gene or region
Strain-specific protein (proteomics)	Molecular	this work	Per gene or region
Allele-specific mRNA (RNA-seq)	Molecular	[Albert et al., 2014a]	Per gene or region
Strain-specific translation (ribo-seq)	Molecular	[Albert et al., 2014a]	Per gene or region
Allele-specific translation (ribo-seq)	Molecular	[Albert et al., 2014a]	Per gene or region
Open chromatin (FAIRE-seq)	Molecular	[Lee et al., 2013]	Per gene or region
Strain-specific open chromatin	Molecular	[Lee et al., 2013]	Per gene or region

Table 5.1: **Annotations used in yeast genetic mapping experiments.** * denotes the use of data from the Saccharomyces Genome Database [Cherry et al., 2012] and compiled or processed for this work. ** denotes the use of the method in [Zeng et al., 2015] applied to the data from [Hesselberth et al., 2009, Zelin et al., 2012].

off solid plates. After reaching saturation overnight, samples were inoculated into 7.5 mL liquid YPD at an initial OD_{600} of 0.1. These samples, in 10 mL aliquots, were diluted to the target condition concentrations given in Table 5.2 and grown for several hours until they reached a final OD_{600} of 0.8 to 0.9, approximately at mid-log growth phase. After reaching mid-log growth phase, cells were collected from the media on ice, spun down at 3,000xg at 4° for 4 minutes, resuspended and washed once in chilled deionized water, split into aliquots for parallel replicate processing, and spun down at 13,000 rpm at 4° for 5 minutes. The resulting cell pellet was flash-frozen and stored for downstream genomic processing.

Condition	Concentration
YPD	n/a
Cycloheximide	50 ng/mL
Diamide	1.5 mM
Ethanol	2%
Hydrogen peroxide	$375\,\mu$M

Table 5.2: **Analyzed conditions for yeast genetic mapping study.** We collected a set of molecular measurements in the two parental strains in the listed conditions, which are a subset of those studied in earlier work [Bloom et al., 2013].

Transcriptomic datasets

To obtain transcriptomic measurements, we extracted and sequenced total RNA from yeast cultures in specific conditions. Frozen cell pellets were resuspended and total RNA was extracted with Trizol. Quality was verified by Bioanalyzer analysis. The purified RNA was used to make Illumina RNA-seq libraries using Illumina TruSeq kits. Libraries were sequenced on an Illumina HiSeq 2000 instrument, with 100+100 paired end reads. Coverage for the analyzed samples is given in Table 5.5.

The raw reads were mapped to the yeast genome using the software package `bwa` [Li and Durbin, 2010b] and read counts were extracted for each gene using published yeast gene annotations. To identify condition- and strain-specific genes, we employed a generalized linear model (GLM) within the `edgeR` framework [Robinson et al., 2010]. The results of this analysis are presented in Table 5.5.

Proteomic datasets

To collect proteomic measurements, we performed shotgun proteomics to obtain quantitative measurements of all yeast peptides. Frozen cell pellets were resuspended in yeast lysis buffer and suspensions were lysed on a mini bead beater. Protein was isolated with a methanol/chloroform precipitation as described previously [Wessel and Flügge, 1984]. Peptides were labeled with TMT-10plex reagents from Thermo Scientific for quantitative analysis. TMT-labeled peptides were identified using tandem/triple-stage mass spectrometry on an Orbitrap Fusion mass spectrometer (Thermo Scientific).

Proteomic data analysis was performed on an in-house, SEQUEST-based [Eng et al., 1994] software platform [Huttlin et al., 2010]. MS2 spectra were searched against a protein sequence database containing all protein sequences in the *S. cerevisiae* ORF database and the human UniProt database, as well as that of known contaminants. Ambiguous peptides with sequences in more than one protein were assigned to the protein with the most matching peptides [Huttlin et al., 2010].

The normalized log intensities at the protein level were analyzed using a multivariate regression approach to identify strain- and condition-specific proteins. The results of this

analysis are presented in Table 5.4.

Chromatin openness datasets

We also used genome-wide open chromatin measurements obtained via sequencing to identify transcriptionally active regions, including active promoters and coding sequences. An earlier study [Lee et al., 2013] performed FAIRE-seq [Giresi et al., 2007] on the two parental strains to study genetic influences on chromatin openness genome-wide. We remapped the raw sequencing data obtained from the original study and computed genome-wide coverage using `bedTools` [Quinlan and Hall, 2010]. We smoothed the coverage by 500-bp sliding windows and normalized them using whole-genome sequencing data from the same strains, in order to avoid copy-number artifacts. We took regions that had normalized smoothed coverage above the 90th percentile in the experiment as open chromatin in each strain. We took regions that differed by more than two-fold in normalized smoothed coverage between the two strains as strain-specific.

Allele-specific translation and expression

To obtain allele-specific translation and transcription measurements, we analyzed data from a recent study that collected molecular measurements from a BY-RM hybrid diploid strain [Albert et al., 2014a]. By analyzing sequencing data at locations where mutations distinguish the two strains, strain-specific (allele-specific) *cis* effects can be identified. We analyzed processed RNA-seq and ribo-seq data at the gene level from [Albert et al., 2014a], reported in their supplementary materials. We took strain-specific genes based on an FDR threshold of 0.0001 and allele-specific genes determined by analyzing a BY-RM hybrid diploid with an FDR threshold of 0.05.

Sequence-based features

To collect sequence-based annotations, we downloaded the yeast reference genome from the *Saccharomyces* Genome Database (SGD) [Cherry et al., 2012] and identified variants using RM11 whole-genome sequencing data from unpublished data and previous studies [Bloom

et al., 2013]. We used the `snpEff` tool [Cingolani et al., 2012] to identify and categorize which variants affected protein-coding genes.

Allele frequencies

To determine variant allele frequencies across multiple *S. cerevisiae* strains, we collected a large set of single-strain yeast sequencing datasets and identified variants across all the experiments. We downloaded multiple whole-genome yeast sequencing datasets available from references linked on [Cherry et al., 2012] and elsewhere. We remapped all the raw reads using a uniform pipeline and performed joint variant calling using the `samtools` package [Li et al., 2009b]. We ended up with a library of 167 strains. We assigned variants as common ("HighAF1") if the allele frequency in the compendium was between 65% and 95%, to account for reference genome errors. We assigned variants as rare ("LowAF1") if they were present in fewer than 25% of the samples.

5.3 Functional annotation results

We first show the total counts of the binary annotations, reported in Table 5.3. We can also visualize the pairwise correlation between each annotation, as shown in Figure 5-3. The annotation categories span a qualitative range of rare to common, with "HIGH" impact amino acid changes (for example, early or lost stop codons or frameshift mutations) as very rare, and coding variants as relatively common, reflecting the high gene density in yeast.

Proteomics results

The condition-specific analysis results from our shotgun proteomics data is presented in Table 5.4. There were very few condition-specific proteins, which correlates with the results from RNA-seq reported in the next section. Interestingly, the pleiotropic gene *MKT1* is one of the 34 ethanol-specific proteins. These results may reflect noise in the measurement assay, overly conservative statistical modeling for differential analysis, or the fact that the condition-specific perturbations were not long or harsh enough to induce stable protein-level changes in the cells. However, our tests do identify 971 strain-specific proteins.

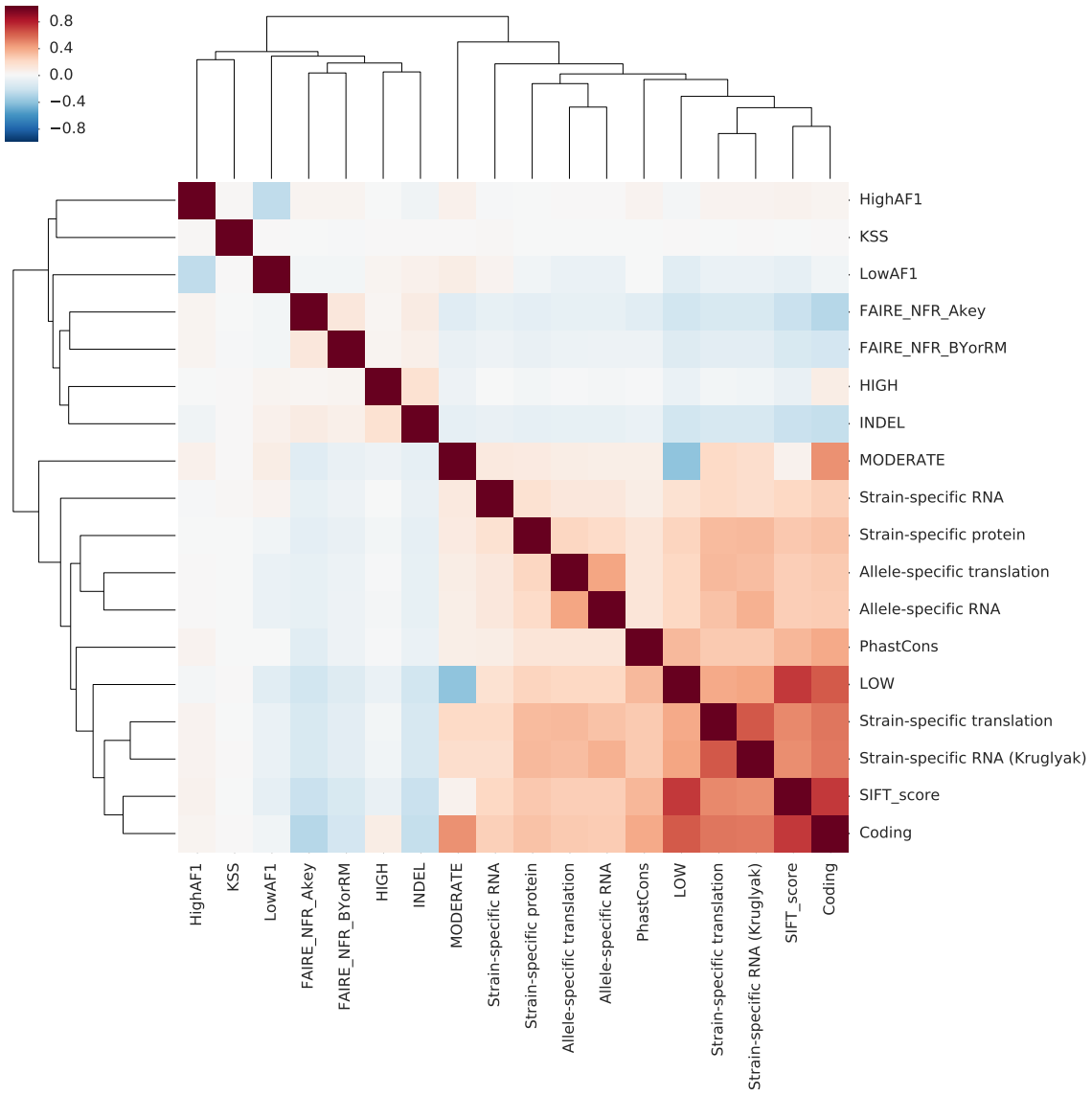

Figure 5-3: **Yeast functional annotation correlations.** The pairwise correlation coefficient between each functional annotation is computed across all variants. Negative correlations are observed for pairs of exclusive annotations, like “MODERATE” versus “LOW” amino acid change impact.

	Count	Fraction
Coding	29744	0.591
LOW	17658	0.351
Strain-specific translation	14714	0.292
Strain-specific RNA (Kruglyak)	14355	0.285
MODERATE	11647	0.231
LowAF1	10088	0.200
HighAF1	10073	0.200
Strain-specific protein	5501	0.109
FAIRE_NFR_BYorRM	5163	0.103
Allele-specific translation	4555	0.090
FAIRE_NFR_Akey	4455	0.088
Allele-specific RNA	4352	0.086
Strain-specific RNA	3798	0.075
INDEL	3277	0.065
HIGH	439	0.009

Table 5.3: **Yeast functional annotation counts.** The counts for all binary annotations used in our models are reported above. For gene- or region-based features, the feature applies to all variants within the region.

Condition	Condition-specific	Interactions
YPD	n/a	n/a
Cycloheximide	0	0
Diamide	0	0
Ethanol	34	0
Hydrogen peroxide	0	0

Table 5.4: **Yeast proteomics differential analysis results.** All numbers are at an FDR threshold of 0.1. Condition-specific analysis was performed versus the baseline YPD condition, so no results are reported in the YPD row. Interaction means a gene that had a significant strain-condition interaction term. 971 proteins were strain-specific without a condition dependence.

Transcription results

The condition-specific transcriptional analysis, along with read count data, is presented in Table 5.5. Similar to the proteomics results, we see a small number of condition-specific genes, with the exception of ethanol response. We observe a greater transcriptional response to ethanol stress, reflecting the greater speed with which the cells respond in this condition or the greater importance of forming a proper cellular environment in rising ethanol concentrations. With this dataset, we did have statistical power to detect interaction effects between strain and condition, that is, condition-specific effects that manifest themselves in

only one strain.

Condition	BY1 reads	BY2 reads	RM1 reads	RM2 reads	Condition-specific	Interaction
YPD	14.3M	16.3M	15.2M	10.3M	n/a	n/a
Cycloheximide	13.0M	14.9M	11.2M	9.5M	6	2
Diamide	12.2M	14.0M	12.5M	14.2M	45	35
Ethanol	14.3M	15.1M	15.8M	13.6M	1020	197
Hydrogen peroxide	14.3M	12.7M	12.9M	10.4M	8	10

Table 5.5: **Yeast RNA-seq differential analysis results.** All numbers are at an FDR threshold of 0.1. Condition-specific analysis was performed versus the baseline YPD condition, so no results are reported in the YPD row. Interaction means a gene that had a significant strain-condition interaction term. 1279 genes were strain-specific without a condition dependence.

5.4 Genetic mapping results using functional annotations

With the assembled functional annotations, we are ready to apply our models from Chapter 4.

Training data

We use 121 associated regions from 46 conditions, as reported in [Bloom et al., 2013]. In our current work, we pool associated regions from multiple conditions together. This combining step is due to practical limitations in our training data, on two axes. First, training a condition-specific model would employ only a small number of associated regions per condition. In many cases, attempting to train a condition-specific model in this way would have more candidate features (annotations) than associated regions (targeted positive bags). Second, a condition-specific model would work best or have the greatest potential predictive power with condition-specific features. Our condition-specific molecular measurements, summarized in Tables 5.4 and 5.5, were not successful at identifying large groups of condition-specific genes in conditions besides ethanol tolerance. The remainder of our annotation features are constant across conditions, and therefore may be expected to provide little condition-specific signal to our models. Therefore, we combined all conditions for the subsequent analyses, in an effort to identify a core or shared genetic association logic for these yeast strains.

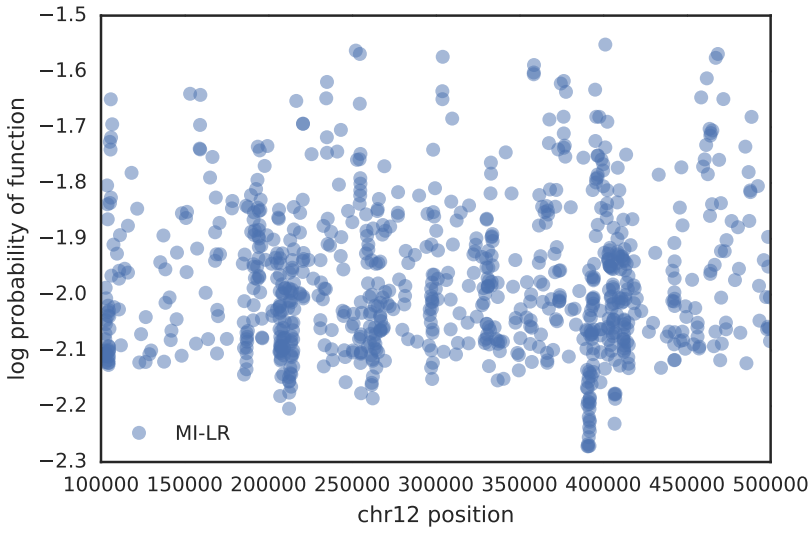

Figure 5-4: **Example learned prior probability model for yeast genetic mapping.** We plot the model predictions for a multi-instance logistic regression model trained on our set of yeast functional annotations.

We merge overlapping associated regions by taking the one with highest reported association strength. We train the multi-instance logistic regression and deep multi-instance logistic regression models described in Chapter 4. Figure 5-4 shows the example probability distribution predicted by the multi-instance logistic regression plotted across a small region of the genome, where the predicted scores vary at a fine spatial scale across the region.

Important model terms

To assess the relative importance of the model coefficients in our linear multi-instance logistic regression model, we fit multi-instance logistic regression models to 20 datasets where the causal peak locations had been randomly shifted through the genome. This allowed us to observe an empirical null distribution of coefficient values. We shifted the regions instead of permuting them in order to maintain the same local correlation structure in the candidate annotation features. We computed an empirical threshold by choosing the 99th percentile of the observed coefficient magnitude distribution in these control experiments. In Table 5.6 we show the functional annotations that surpassed this threshold and their coefficient values. We observe that strain-specific protein has the highest predictive value, followed by another

	MI-LR coefficients
Strain-specific protein	0.131
Strain-specific RNA (Kruglyak)	0.103
PhastCons	0.080
FAIRE_NFR_Akey	0.076
MODERATE	0.074
Allele-specific RNA	0.058
LOW	-0.040

Table 5.6: **High-magnitude model terms in a yeast functional annotation experiment.** Coefficients whose magnitude was larger than the 99th percentile observed in an empirical null distribution are shown here. The majority of features are in the molecular category described above, along with two amino acid change features and a conservation score feature.

molecular feature: strain-specific RNA levels. Of these seven terms, four are molecular features, two are sequence-based describing the impact of amino acid changes, and the final is a general conservation score. All of the selected features are positive, meaning that they contribute to a greater likelihood of a region containing a causal variant, except the "LOW" impact amino acid feature. The "LOW" impact feature direction reflects the assumption that synonymous amino acid changes are less likely to lead to genetic associations.

Significance analysis

To judge the relative importance of single features and classes of features in this dataset, we trained sets of pruned models where single features or groups of features were excluded from the model. We optimized a multi-instance logistic regression model in each case and conducted a likelihood ratio test by comparing the training data log likelihoods from the full model that used all possible annotations and the candidate reduced model.

Table 5.7 shows the raw p-values and corrected false discovery rates for each annotation used in our model. Only strain-specific protein levels are significant, and only when considered alone. This lack of importance for any single feature may reflect the correlation structure present in the annotations, where other features can substitute or compensate for any single excluded feature.

To partially address the correlation structure present in the annotations, we conducted stricter model comparisons that exclude groups of features at a time. In this way, we can

assess the importance of classes of annotation features. Table 5.8 shows the results of this analysis. First, we test the importance of each of the three classes of annotations: sequence, molecular, and genetic (as listed in Table 5.1). We observe that all three classes contribute significantly to the model, with multiple-testing corrected FDRs below 0.05. We next compare the relative contributions of three types of molecular annotations: protein, RNA, and chromatin. At this resolution, only protein measurements produce a significant contribution when considered alone. We note that this analysis is partially confounded by the correlation patterns present in the annotation data. That is, the value or significance of a particular annotation should be interpreted not as its true core importance, but as its ability to contribute additional signal not captured by alternate features that remain in the model in our testing procedure. Additionally, our annotation importance conflates the value of the quantity being measured and the success of our particular assay or study in capturing that quantity. It is possible that annotations that are not judged significant in this analysis may prove significant with improved assay techniques or data processing.

Finally, as a negative control we conduct the same group-wise annotation significance tests using a null multi-instance logistic regression model trained on shuffled associated regions (positive bags). The results are shown in Table 5.9 and confirm our expectation that no class of annotations is significant in this setting.

Bag probability analysis

For this yeast dataset, we do not know the location of the truly causal variants in the population. Many of the reported associations are novel due to the increased size and scope of this genetic mapping study, and even many previously known association regions have not been refined to the single-gene or single-mutation level by validation experiments. Therefore we are forced to use indirect means to assess the predictions given by the model. First, we look at the distribution of predicted bag-level probabilities for causal regions compared to random genomic locations. The bag-level probabilities (computed as in Equation 4.2) combine the per-variant (instance-level) probabilities and allow for probabilistic comparisons between different groups of variants. We take each of the predicted bag-level log probabilities

	p-value	FDR
Strain-specific protein	0.013	0.241
MODERATE	0.056	0.343
PhastCons	0.059	0.343
HighAF1	0.076	0.343
Strain-specific RNA (Kruglyak)	0.155	0.558
LOW	0.287	0.770
FAIRE_NFR_Akey	0.299	0.770
KSS	0.428	0.811
Strain-specific RNA	0.464	0.811
Coding	0.496	0.811
FAIRE_NFR_BYorRM	0.496	0.811
Allele-specific RNA	0.724	1.000
Strain-specific translation	0.833	1.000
HIGH	0.833	1.000
SIFT_score	0.894	1.000
LowAF1	1.000	1.000
Allele-specific translation	1.000	1.000
INDEL	1.000	1.000

Table 5.7: **Annotation feature significance in multi-instance logistic regression models.** We computed the importance of each annotation feature by excluding it from the model and comparing the optimized log likelihood to the log likelihood obtained using a full model. A likelihood ratio test computed using the chi-square distribution gives the p-values shown in the first column, which were corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure to give the false discovery rates (FDR) in the second column. We observe that only strain-specific protein level is significant when considered alone.

	p-value	FDR
molecular	0.0003	0.0016
genetic	0.0113	0.0284
protein	0.0149	0.0284
sequence	0.0189	0.0284
RNA	0.0729	0.0874
chromatin	0.1339	0.1339

Table 5.8: **Grouped annotation feature significance in multi-instance logistic regression models.** We computed the importance of groups of annotation feature by excluding them from the model and comparing the optimized log likelihood to the log likelihood obtained using a full model. A likelihood ratio test computed using the chi-square distribution gives the p-values shown in the first column, which were corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure to give the false discovery rates (FDR) in the second column. We observe that each main class of annotations, molecular, genetic, or sequence, significantly improve the model. Of the molecular features, we see that protein-related annotations outperform RNA and chromatin measurements in terms of improving model fit.

	p-value	FDR
sequence	0.277	1.000
genetic	1.000	1.000
molecular	1.000	1.000
protein	1.000	1.000
RNA	1.000	1.000
chromatin	1.000	1.000

Table 5.9: **Negative control grouped annotation feature significance tests.** Using a negative control multi-instance logistic regression model trained on shuffled associated regions, we computed the importance of groups of annotation feature by excluding them from the model and comparing the optimized log likelihood to the log likelihood obtained using a full model. A likelihood ratio test computed using the chi-square distribution gives the p-values shown in the first column, which were corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure to give the false discovery rates (FDR) in the second column. We observe that no annotation class significantly improves the model fit.

of the positive bags and subtract the mean of 100 random bag-level log probabilities, where the random bags are bags of the same size assigned to random locations in the genome. The random bag locations still use contiguous chunks of variants, in order to maintain the spatial correlation structure present in the annotation features. We visualize the distribution of these log differences in Figure 5-5. Model predictions generated using the true training dataset show an enrichment in the positive direction (median of 0.048), whereas predictions generated using a shuffled control training dataset are centered at 0 (median of -0.008).

Enrichment of genetic associations with high- and low-scoring variants

Another approach to lend support to our predictions is to examine the variants with the highest and lowest predicted functional score from our trained annotation model, according to predicted instance-level probabilities (Equation 4.1). We expect that variants with high predicted instance-level association probabilities are more likely to reside in positive regions, that is, observed associated regions in the original cross. Conversely, we expect that variants with low predicted instance-level association probabilities are less likely to reside in positive regions. This comparison is noisy for at least two reasons, however. First, a truly noncausal variant may be near a causal variant by chance. Second, our predictions are general and not condition-specific. The fact that a given variant is not reported inside a genetic association in the current study may mean that the variant in question has no phenotypic link, or it may

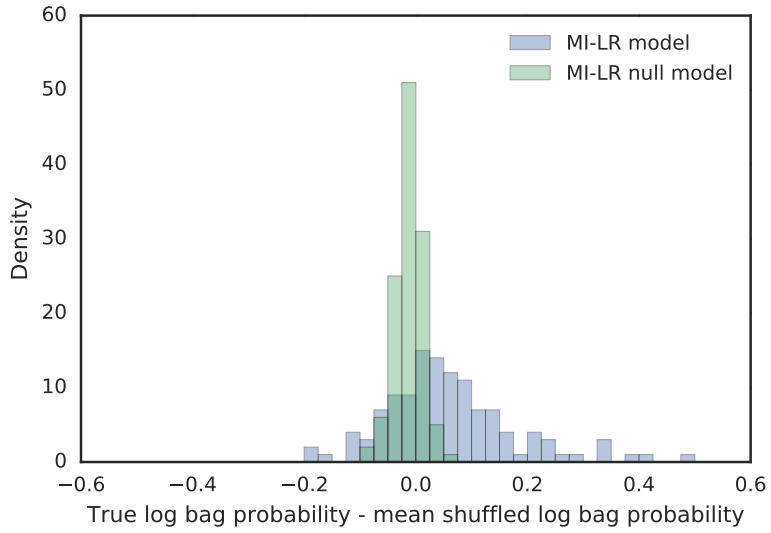

Figure 5-5: **Comparison of bag-level prediction probabilities to control probabilities.** Each positive bag was scored using a multi-instance logistic regression model, and the predicted log bag probability was compared to the average log probability of equally-sized bags placed in 100 random genomic locations. With the true model, these log probability differences are mostly positive, whereas the control model places the log probability differences around zero.

mean the variant has an effect in a condition that has not been studied. We cannot separate these confounding factors, but we hope that an aggregate signal exists despite these noise sources. In Figure 5-6, we show the observed overlaps between the top and bottom sets of predicted association scores and reported associated regions. We observe that variants with the very lowest predicted scores are almost never in associated regions, whereas variants with the highest predicted scores have a high overlap with associated regions. The difference between high- and low-scoring variants persists through the top and bottom 5000 variants that we consider.

Case studies

As an additional validation task, we consider a small set of genes where previous genetic mapping studies have performed targeted validation experiments to identify the precise mutations driving the phenotypic association. Figure 5-7 shows three genes and the predicted log probabilities in windows around each gene.

We first consider *MKT1*, which has been shown to associate with multiple phenotypes in

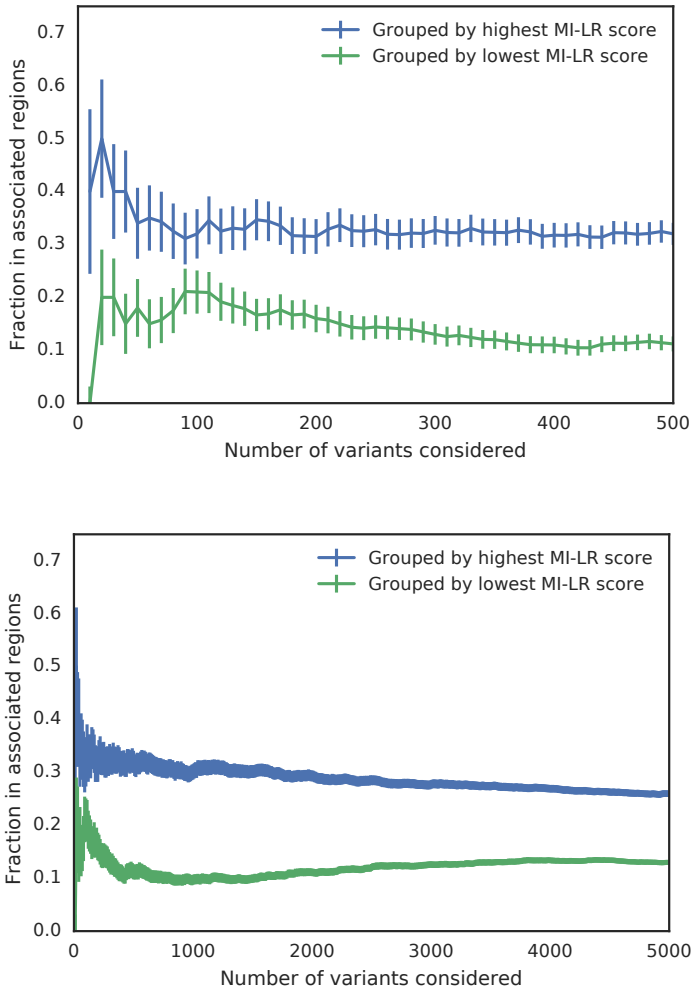

Figure 5-6: **Overlap between high- and low-scoring variants and reported association regions.** The variants with highest and lowest predicted association probability from a trained multi-instance logistic regression model are compared for their overlap with reported association regions. The top panel is a zoomed version of the bottom panel where only the top and bottom 500 variants are shown. Variants are considered in blocks of increasing size, starting at 10 variants and increasing in bins of 10 variants. The plotted error bars are standard deviations calculated using a normal approximation to the estimated binomial proportion.

yeast [Ben-Ari et al., 2006, Demogines et al., 2008, Deutschbauer and Davis, 2005, Dimitrov et al., 2009, Ehrenreich et al., 2010, Sinha et al., 2008, Steinmetz et al., 2002]. Validation experiments have confirmed the D30G mutation as driving the association, with a small effect reported by some authors for the downstream K453R mutation. In this fine mapping region of 52 variants, these two causal variants have the highest predicted probability from our model.

We next consider *RAD5*, where the I791S and G535R mutations have been shown to underlie drug sensitivity phenotypes [Demogines et al., 2008, Fan et al., 1996]. Here these two causal variants are in the top 5 of 51 variants reported in this region.

Finally, we consider *PMR1*. Recent work [Sadhu et al., 2016] highlighted the impact of the F548L mutation by testing the five coding mutations shown in Figure 5-7. However, all variants in *PMR1* are in the middle to low end of the distribution of predicted association probabilities in this region. Therefore we conclude that our model does not capture the factors leading to the genetic association observed in *PMR1*, though it does succeed for the other validated variants we considered.

5.5 Conclusions

In this chapter, we have applied our multi-instance learning algorithms to build predictive models of which genetic variants influence phenotypes in yeast. We collected and generated a large dataset of functional measurements tailored to a specific yeast cross, and demonstrated that several annotations are valuable in a final prediction model. We observed statistical signals showing the efficacy of our models and verified their success in several single-gene case studies, while also observing genes where the model predictions did not completely reflect known biology. We conclude that these types of models have promise and that their relative success in the future depends on increased sample sizes, both in the number of reported genetic associations to use as training examples and in the size and variety of candidate functional associations to use as features. As both dimensions of the problem grow, we should be able to build condition-specific models where more nuanced models of the biological processes involved in each specific phenotype lead to greater prediction accuracy.

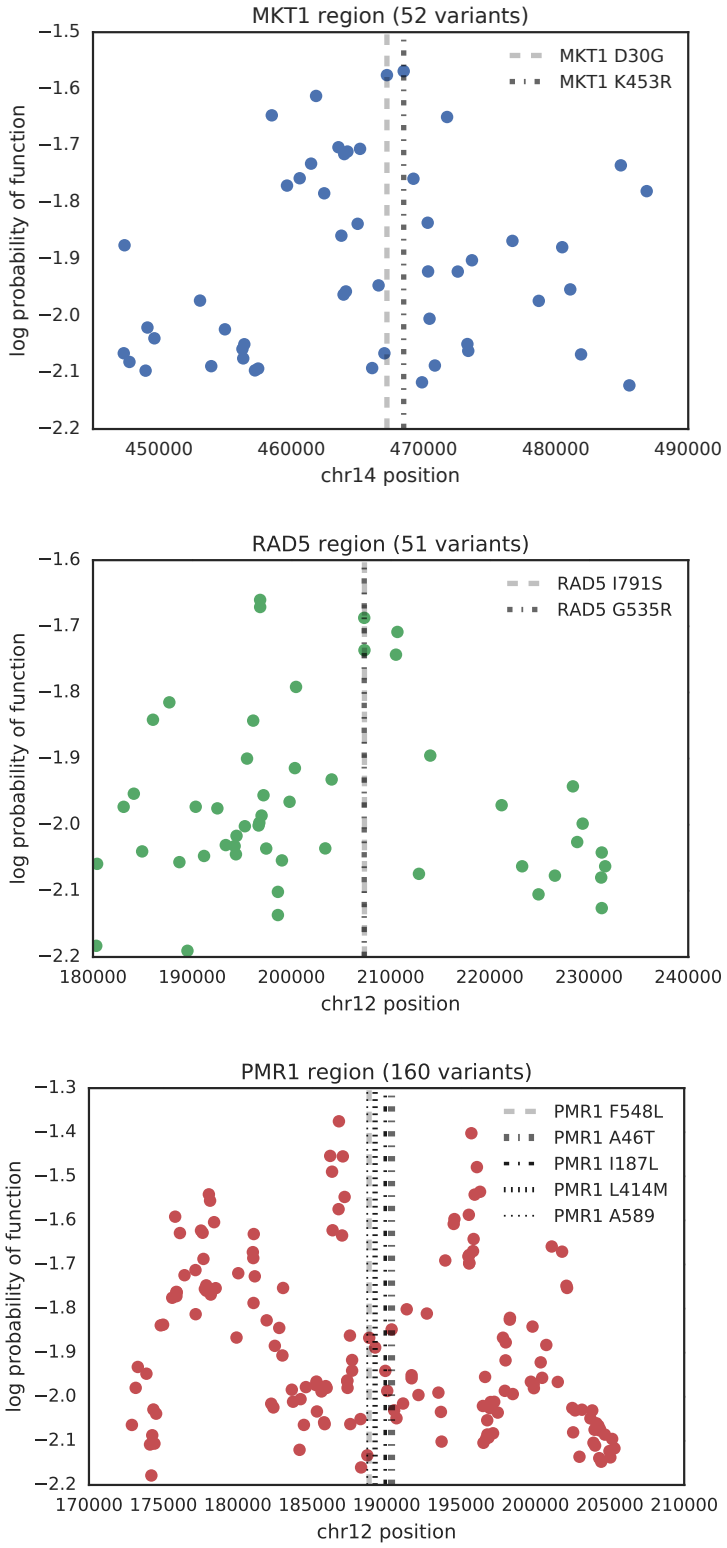

Figure 5-7: **Predictions around multiple validated loci.** Predicted log association probabilities from a trained multi-instance logistic regression model are shown around genes where causal variants have been validated to single-base precision. See the text for more details on each gene and which variants are known to be causal.

Chapter 6

Conclusions

In this thesis, we presented two algorithmic approaches to problems in modern genetics. First, we proposed a computational model of pooled sequencing that integrates information from multiple markers, handles noisy data in a principled manner, and yields accurate location estimates for genetic mapping. Second, we described a statistical model that uses functional annotation information to predict which genomic regions are likely to cause phenotypic changes. For both problems, we used supporting biological experiments to test and gain insights from our computational models. For the pooled sequencing models, we analyzed the basis of strain-specific essential genes in yeast and demonstrated its complex genetic basis. We also verified the involvement of an inherited cytoplasmic virus in certain strain-specific essential genes. For the multi-instance classification models, we trained a classifier using yeast functional annotations and showed that it scored causal genomic regions higher than noncausal genomic regions. We demonstrated that certain functional annotations were useful in the predictive model, most notably including strain-specific protein level.

Future work can proceed along two axes: computational and biological. We discuss each area in turn in the next sections.

6.1 Future computational work

On the computational side, richer models can be developed to require fewer assumptions or handle greater variability in the data. For the pooled sequencing model, it may be useful to

model a varying recombination rate genome-wide. This rate could be given as input from external data sources or learned from the pooled genotype data itself. If proven accurate, this type of model could even be used to derive strain-specific recombination maps directly from pooled sequencing data. The model could be generalized to more complicated settings, such as multi-parent crosses [Cubillos et al., 2013, Ehrenreich et al., 2012]. To our knowledge, pooled genetic mapping has not yet been applied to multi-parent crosses.

The differential testing model for comparing pools could be extended using a full generalized linear model (GLM) framework that would allow for replicates of opposite conditions, testing of batch effects, or the consideration of multiple nested experimental conditions. This would be similar to existing work in the RNA-seq analysis literature [Robinson et al., 2010]. From user feedback on the first version of MULTIPOOL, we have realized that these complicated experimental designs are becoming more popular.

For the multi-instance model, a multivariate output or latent structure model in the predicted classes could yield improvements. Instead of learning from a single set of associated regions, the regions could be separated by phenotype or grouped by related sets of phenotypes. As discussed previously, including the association strength in the classification model may allow for performance gains if the strength of a genetic effect is related to the strength of the annotation model prediction. This extension would also allow our model to consider a wider spectrum of associated regions, instead of only narrower high-confidence regions. Avoiding an exclusive reliance on genome-wide significant hits has been discussed and successfully applied in recent work [Finucane et al., 2015].

6.2 Future biological work

On the biological side, these algorithms could be applied to larger and richer datasets. Pooled genetic mapping studies continue to be applied in model organisms and humans [Schlötterer et al., 2014]. As the number of studied phenotypes grows, the discovered genetic associations can be compared and in cases of possible pleiotropy, studies may be combined to yield greater resolution of the identified causal loci.

For the multi-instance classification model, human data is a clear application. There are

more known associated regions in human, as well as larger and richer functional annotations. In this case, however, it will be important to develop and use models that are able to handle the extra complexity of these data, including the richness of multiple phenotypes and more classes of molecular measurements. The link between certain functional annotations, obtained in particular tissue types or genetic backgrounds, will most likely vary depending on the selected phenotype.

6.3 Final thoughts

This thesis explored two problems in computational genetics, motivated by the increasing size and complexity of modern molecular and genetic data. We derived efficient algorithms that are able to combine information within and across datasets in ways that are faithful to the underlying biological processes and the noise sources arising from experimental equipment. We used our computational techniques to study novel datasets from yeast genetics and obtained specific insights illuminating the complexity of going from genotype to phenotype. We expect and hope that the strategic combination of computational models and large biological datasets will continue to yield important discoveries in the future.

Bibliography

[Albert et al., 2014a] Albert, F. W., Muzzey, D., Weissman, J. S., and Kruglyak, L. (2014a). Genetic influences on translation in yeast. *PLoS Genetics*, 10(10):e1004692.

[Albert et al., 2014b] Albert, F. W., Treusch, S., Shockley, A. H., Bloom, J. S., and Kruglyak, L. (2014b). Genetics of single-cell protein abundance variation in large yeast populations. *Nature*, 506(7489):494.

[Altshuler et al., 2008] Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322(5903):881–888.

[Andolfatto et al., 2011] Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., and Stern, D. L. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, 21(4):610–617.

[Bastien et al., 2012] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., and Bengio, Y. (2012). Theano: new features and speed improvements. In *Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop*.

[Ben-Ari et al., 2006] Ben-Ari, G., Zenvirth, D., Sherman, A., David, L., Klutstein, M., Lavi, U., Hillel, J., and Simchen, G. (2006). Four linked genes participate in controlling sporulation efficiency in budding yeast. *PLoS Genetics*, 2(11):e195.

[Bengio, 2009] Bengio, Y. (2009). Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2(1):1–127.

[Bergstra et al., 2010] Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*.

[Bernstein et al., 2010] Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., et al. (2010). The nih roadmap epigenomics mapping consortium. *Nature Biotechnology*, 28(10):1045–1048.

[Bishop, 2007] Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer.

[Bloom et al., 2013] Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–237.

[Boeke et al., 1987] Boeke, J. D., Trueheart, J., Natsoulis, G., and Fink, G. R. (1987). [10] 5-fluoroorotic acid as a selective agent in yeast molecular genetics. *Methods in Enzymology*, 154:164–175.

[Borevitz et al., 2003] Borevitz, J. O., Liang, D., Plouffe, D., Chang, H., Zhu, T., Weigel, D., Berry, C. C., Winzeler, E., and Chory, J. (2003). Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Research*, 13(3):513–523.

[Boser et al., 1992] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM.

[Brauer et al., 2006] Brauer, M. J., Christianson, C. M., Pai, D. A., and Dunham, M. J. (2006). Mapping novel traits by array-assisted bulk segregant analysis in saccharomyces cerevisiae. *Genetics*, 173(3):1813–1816.

[Calvo et al., 2010] Calvo, S. E., Tucker, E. J., Compton, A. G., Kirby, D. M., Crawford, G., Burtt, N. P., Rivas, M., Guiducci, C., Bruno, D. L., Goldberger, O. A., Redman, M. C., Wiltshire, E., Wilson, C. J., Altshuler, D., Gabriel, S. B., Daly, M. J., Thorburn, D. R., and Mootha, V. K. (2010). High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex i deficiency. *Nature Genetics*, 42(10):851–858.

[Cherry et al., 2012] Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., et al. (2012). Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1):D700–D705.

[Cingolani et al., 2012] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92.

[Civelek and Lusis, 2014] Civelek, M. and Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, 15(1):34–48.

[Cleveland, 1979] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

[Clowers et al., 2015] Clowers, K. J., Heilberger, J., Piotrowski, J. S., Will, J. L., and Gasch, A. P. (2015). Ecological and genetic barriers differentiate natural populations of saccharomyces cerevisiae. *Molecular Biology and Evolution*, 32(9):2317–2327.

[Collobert and Bengio, 2004] Collobert, R. and Bengio, S. (2004). Links between perceptrons, mlps and svms. In *Proceedings of the Twenty-first International Conference on Machine Learning*, page 23. ACM.

[Consortium et al., 2012] Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[Cubillos et al., 2013] Cubillos, F. A., Parts, L., Salinas, F., Bergström, A., Scovacricchi, E., Zia, A., Illingworth, C. J., Mustonen, V., Ibstedt, S., Warringer, J., et al. (2013). High-resolution mapping of complex traits with a four-parent advanced intercross yeast population. *Genetics*, 195(3):1141–1155.

[Degner et al., 2009] Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212.

[Demogines et al., 2008] Demogines, A., Smith, E., Kruglyak, L., and Alani, E. (2008). Identification and dissection of a complex dna repair sensitivity phenotype in baker's yeast. *PLoS Genetics*, 4(7):e1000123.

[Deutschbauer and Davis, 2005] Deutschbauer, A. M. and Davis, R. W. (2005). Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nature Genetics*, 37(12):1333–1340.

[Dimitrov et al., 2009] Dimitrov, L. N., Brem, R. B., Kruglyak, L., and Gottschling, D. E. (2009). Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of saccharomyces cerevisiae s288c strains. *Genetics*, 183(1):365–383.

[Dowell et al., 2010] Dowell, R. D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D. A., Rolfe, P. A., Heisler, L. E., Chin, B., et al. (2010). Genotype to phenotype: a complex problem. *Science*, 328(5977):469–469.

[Drinnenberg et al., 2011] Drinnenberg, I. A., Fink, G. R., and Bartel, D. P. (2011). Compatibility with killer explains the rise of rnai-deficient fungi. *Science*, 333(6049):1592–1592.

[Edwards and Gifford, 2012] Edwards, M. D. and Gifford, D. K. (2012). High-resolution genetic mapping with pooled sequencing. *BMC Bioinformatics*, 13(Suppl 6):S8.

[Edwards et al., 2014] Edwards, M. D., Symbor-Nagrabska, A., Dollard, L., Gifford, D. K., and Fink, G. R. (2014). Interactions between chromosomal and nonchromosomal elements reveal missing heritability. *Proceedings of the National Academy of Sciences*, 111(21):7719–7722.

[Ehrenreich et al., 2012] Ehrenreich, I. M., Bloom, J., Torabi, N., Wang, X., Jia, Y., and Kruglyak, L. (2012). Genetic architecture of highly complex chemical resistance traits across four yeast strains. *PLoS Genetics*, 8(3):e1002570.

[Ehrenreich et al., 2010] Ehrenreich, I. M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J. A., Gresham, D., Caudy, A. A., and Kruglyak, L. (2010). Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature*, 464(7291):1039–1042.

[Eng et al., 1994] Eng, J. K., McCormack, A. L., and Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989.

[Fan et al., 1996] Fan, H.-Y., Cheng, K. K., and Klein, H. L. (1996). Mutations in the rna polymerase ii transcription machinery suppress the hyperrecombination mutant hpr1 δ of saccharomyces cerevisiae. *Genetics*, 142(3):749–759.

[Farh et al., 2015] Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J., Shishkin, A. A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343.

[Fink and Styles, 1972] Fink, G. R. and Styles, C. A. (1972). Curing of a killer factor in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 69(10):2846–2849.

[Finucane et al., 2015] Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235.

[Foulds and Frank, 2010] Foulds, J. and Frank, E. (2010). A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(01):1–25.

[Ghahramani and Hinton, 1996] Ghahramani and Hinton, G. E. (1996). Parameter estimation for linear dynamical systems. *University of Toronto Technical Report*, 6(CRG-TR-96-2):1–6.

[Giaever et al., 2002] Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., et al. (2002). Functional profiling of the saccharomyces cerevisiae genome. *Nature*, 418(6896):387–391.

[Giresi et al., 2007] Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., and Lieb, J. D. (2007). Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17(6):877–885.

[Glorot et al., 2011] Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 315–323.

[Guyon et al., 2004] Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. (2004). Result analysis of the nips 2003 feature selection challenge. In *Advances in Neural Information Processing Systems*, pages 545–552.

[Hazen et al., 2005] Hazen, S. P., Borevitz, J. O., Harmon, F. G., Pruneda-Paz, J. L., Schultz, T. F., Yanovsky, M. J., Liljegren, S. J., Ecker, J. R., and Kay, S. A. (2005). Rapid array mapping of circadian clock and developmental mutations in arabidopsis. *Plant Physiology*, 138(2):990–997.

[Hesselberth et al., 2009] Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., et al. (2009). Global mapping of protein-dna interactions in vivo by digital genomic footprinting. *Nature Methods*, 6(4):283–289.

[Homer et al., 2008] Homer, N., Tembe, W. D., Szelinger, S., Redman, M., Stephan, D. A., Pearson, J. V., Nelson, S. F., and Craig, D. (2008). Multimarker analysis and imputation of multiple platform pooling-based genome-wide association studies. *Bioinformatics*, 24(17):1896–1902.

[Hornik, 1991] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

[Hu et al., 2011] Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., and Raychaudhuri, S. (2011). Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *American Journal of Human Genetics*, 89(4):496–506.

[Huttlin et al., 2010] Huttlin, E. L., Jedrychowski, M. P., Elias, J. E., Goswami, T., Rad, R., Beausoleil, S. A., Villén, J., Haas, W., Sowa, M. E., and Gygi, S. P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, 143(7):1174–1189.

[Jawaid et al., 2002] Jawaid, A., Bader, J. S., Purcell, S., Cherny, S. S., and Sham, P. (2002). Optimal selection strategies for QTL mapping using pooled DNA samples. *European Journal of Human Genetics*, 10(2):125–132.

[Johnson, 2007] Johnson, T. (2007). Bayesian method for gene detection and mapping, using a case and control design and DNA pooling. *Biostatistics*, 8(3):546–565.

[Karczewski et al., 2013] Karczewski, K. J., Dudley, J. T., Kukurba, K. R., Chen, R., Butte, A. J., Montgomery, S. B., and Snyder, M. (2013). Systematic functional regulatory assessment of disease-associated variants. *Proceedings of the National Academy of Sciences*, 110(23):9607–9612.

[Kichaev et al., 2014] Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, 10(10):e1004722.

[Kim and Fay, 2009] Kim, H. S. and Fay, J. C. (2009). A combined-cross analysis reveals genes with drug-specific and background-dependent effects on drug sensitivity in saccharomyces cerevisiae. *Genetics*, 183(3):1141–1151.

[Lander and Botstein, 1989] Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199.

[Lander and Schork, 1994] Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, 265(5181):2037–2048.

[Lee et al., 2013] Lee, K., Kim, S. C., Jung, I., Kim, K., Seo, J., Lee, H.-S., Bogu, G. K., Kim, D., Lee, S., Lee, B., et al. (2013). Genetic landscape of open chromatin in yeast. *PLoS Genetics*, 9(2):e1003229.

[Lee et al., 2009] Lee, S., Dudley, A. M., Drubin, D., Silver, P. A., Krogan, N. J., Pe'er, D., and Koller, D. (2009). Learning a prior on regulatory potential from eQTL data. *PLoS Genetics*, 5(1):e1000358.

[Lehner, 2013] Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics*, 14(3):168–178.

[Li and Durbin, 2010a] Li, H. and Durbin, R. (2010a). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595.

[Li and Durbin, 2010b] Li, H. and Durbin, R. (2010b). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595.

[Li et al., 2009a] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

[Li et al., 2009b] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009b). The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

[Li et al., 2011] Li, Z., Vizeacoumar, F. J., Bahr, S., Li, J., Warringer, J., Vizeacoumar, F. S., Min, R., VanderSluis, B., Bellay, J., DeVit, M., et al. (2011). Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nature Biotechnology*, 29(4):361–367.

[Liti et al., 2009] Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–341.

[Macgregor et al., 2008] Macgregor, S., Zhao, Z. Z., Henders, A., Nicholas, M. G., Montgomery, G. W., and Visscher, P. M. (2008). Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. *Nucleic Acids Research*, 36(6):e35.

[Magliani et al., 1997] Magliani, W., Conti, S., Gerloni, M., Bertolotti, D., and Polonelli, L. (1997). Yeast killer systems. *Clinical Microbiology Reviews*, 10(3):369–400.

[Magwene et al., 2011] Magwene, P. M., Willis, J. H., and Kelly, J. K. (2011). The statistics of bulk segregant analysis using next generation sequencing. *PLoS Computational Biology*, 7(11):e1002255.

[Maller et al., 2012] Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M., Auton, A., Myers, S., Morris, A., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, 44(12):1294–1301.

[Mancera et al., 2008] Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203):479–485.

[Maurano et al., 2012] Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195.

[McPeek and Strahs, 1999] McPeek, M. S. and Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *American Journal of Human Genetics*, 65(3):858–875.

[Michelmore et al., 1991] Michelmore, R. W., Paran, I., and Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences*, 88(21):9828–9832.

[Morris et al., 2000] Morris, A. P., Whittaker, J. C., and Balding, D. J. (2000). Bayesian fine-scale mapping of disease loci, by hidden markov models. *American Journal of Human Genetics*, 67(1):155–169.

[Murphy, 1999] Murphy, K. (1999). Filtering, smoothing and the junction tree algorithm. *University of California, Berkeley Technical Report*.

[Murphy, 2012] Murphy, K. P. (2012). *Machine Learning: a Probabilistic Perspective*. MIT press.

[Pagé et al., 2003] Pagé, N., Gérard-Vincent, M., Ménard, P., Beaulieu, M., Azuma, M., Dijkgraaf, G. J., Li, H., Marcoux, J., Nguyen, T., Dowse, T., et al. (2003). A saccharomyces cerevisiae genome-wide mutant screen for altered sensitivity to k1 killer toxin. *Genetics*, 163(3):875–894.

[Parker et al., 2013] Parker, S. C., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., Black, B. L., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences*, 110(44):17921–17926.

[Parts et al., 2011] Parts, L., Cubillos, F. A., Warringer, J., Jain, K., Salinas, F., Bumpstead, S. J., Molin, M., Zia, A., Simpson, J. T., Quail, M. A., Moses, A., Louis, E. J., Durbin, R., and Liti, G. (2011). Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Research*, 21(7):1131–1138.

[Paul et al., 2013] Paul, D. S., Albers, C. A., Rendon, A., Voss, K., Stephens, J., van der Harst, P., Chambers, J. C., Soranzo, N., Ouwehand, W. H., Deloukas, P., et al. (2013). Maps of open chromatin highlight cell type–restricted patterns of regulatory sequence variation at hematological trait loci. *Genome Research*, 23(7):1130–1141.

[Paul et al., 2011] Paul, D. S., Nisbet, J. P., Yang, T.-P., Meacham, S., Rendon, A., Hautaviita, K., Tallila, J., White, J., Tijssen, M. R., Sivapalaratnam, S., et al. (2011). Maps of open chromatin guide the functional follow-up of genome-wide association signals: application to hematological traits. *PLoS Genetics*, 7(6):e1002139.

[Pickrell, 2014] Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, 94(4):559–573.

[Quinlan and Hall, 2010] Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

[Ray and Craven, 2005] Ray, S. and Craven, M. (2005). Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 697–704. ACM.

[Raykar et al., 2008] Raykar, V. C., Krishnapuram, B., Bi, J., Dundar, M., and Rao, R. B. (2008). Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *Proceedings of the 25th International Conference on Machine Learning*, pages 808–815. ACM.

[Robinson et al., 2010] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386.

[Sadhu et al., 2016] Sadhu, M. J., Bloom, J. S., Day, L., and Kruglyak, L. (2016). Crispr-directed mitotic recombination enables genetic mapping without crosses. *Science*.

[Schacherer et al., 2009] Schacherer, J., Shapiro, J. A., Ruderfer, D. M., and Kruglyak, L. (2009). Comprehensive polymorphism survey elucidates population structure of saccharomyces cerevisiae. *Nature*, 458(7236):342–345.

[Schaub et al., 2012] Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Research*, 22(9):1748–1759.

[Schlötterer et al., 2014] Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals [mdash] mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, 15(11):749–763.

[Schmitt and Breinig, 2006] Schmitt, M. J. and Breinig, F. (2006). Yeast viral killer toxins: lethality and self-protection. *Nature Reviews Microbiology*, 4(3):212–221.

[Schneeberger, 2014] Schneeberger, K. (2014). Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nature Reviews Genetics*, 15(10):662–676.

[Schneeberger et al., 2009] Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., JÃÿrgensen, J., Weigel, D., and Andersen, S. U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods*, 6(8):550–551.

[Sham et al., 2002] Sham, P., Bader, J. S., Craig, I., O'Donovan, M., and Owen, M. (2002). DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics*, 3(11):862–871.

[Sinha et al., 2008] Sinha, H., David, L., Pascon, R. C., Clauder-Münster, S., Krishnakumar, S., Nguyen, M., Shi, G., Dean, J., Davis, R. W., Oefner, P. J., et al. (2008). Sequential elimination of major-effect contributors identifies additional quantitative trait loci conditioning high-temperature growth in yeast. *Genetics*, 180(3):1661–1670.

[Steinmetz et al., 2002] Steinmetz, L. M., Sinha, H., Richards, D. R., Spiegelman, J. I., Oefner, P. J., McCusker, J. H., and Davis, R. W. (2002). Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, 416(6878):326–330.

[Strope et al., 2015] Strope, P. K., Skelly, D. A., Kozmin, S. G., Mahadevan, G., Stone, E. A., Magwene, P. M., Dietrich, F. S., and McCusker, J. H. (2015). The 100-genomes strains, an s. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Research*, 25(5):762–774.

[Sutskever et al., 2013] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1139–1147.

[Treusch et al., 2015] Treusch, S., Albert, F. W., Bloom, J. S., Kotenko, I. E., and Kruglyak, L. (2015). Genetic mapping of mapk-mediated complex traits across s. cerevisiae. *PLoS Genetics*, 11(1):e1004913.

[Trynka et al., 2013] Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45(2):124–130.

[Trynka et al., 2015] Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B. E., Klein, R. J., Han, B., and Raychaudhuri, S. (2015). Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *American Journal of Human Genetics*, 97(1):139–152.

[Wenger et al., 2010] Wenger, J. W., Schwartz, K., and Sherlock, G. (2010). Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from saccharomyces cerevisiae. *PLoS Genetics*, 6(5):e1000942.

[Wessel and Flügge, 1984] Wessel, D. and Flügge, U.-I. (1984). A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Analytical Biochemistry*, 138(1):141–143.

[Wickner, 1992] Wickner, R. B. (1992). Double-stranded and single-stranded rna viruses of saccharomyces cerevisiae. *Annual Reviews in Microbiology*, 46(1):347–375.

[Williams and Rasmussen, 2006] Williams, C. K. and Rasmussen, C. E. (2006). Gaussian processes for machine learning. *MIT Press*, 2(3):4.

[Yan Tong and Boone, 2006] Yan Tong, A. H. and Boone, C. (2006). Synthetic genetic array analysis in saccharomyces cerevisiae. *Yeast Protocol*, pages 171–191.

[Zelin et al., 2012] Zelin, E., Zhang, Y., Toogun, O. A., Zhong, S., and Freeman, B. C. (2012). The p23 molecular chaperone and gcn5 acetylase jointly modulate protein-dna dynamics and open chromatin status. *Molecular Cell*, 48(3):459–470.

[Zeng et al., 2015] Zeng, H., Hashimoto, T., Kang, D. D., and Gifford, D. K. (2015). Gerv: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics*, page btv565.

[Zhang et al., 2005] Zhang, C., Platt, J. C., and Viola, P. A. (2005). Multiple instance boosting for object detection. In *Advances in Neural Information Processing Systems*, pages 1417–1424.