Quantitative genetics of complex traits: solutions for studying the genetic basis of variation in yeast

Thesis submitted for the degree of
Doctor of Philosophy
at The University of Leicester

By
Yue Hu BSc
Department of Genetics and Genome Biology
University of Leicester

4th March 2020

QUANTITATIVE GENETICS OF COMPLEX TRAITS : SOLUTIONS FOR STUDYING THE GENETIC BASIS OF VARIATION IN YEAST

Yue Hu

Department of Genetics, University of Leicester
Thesis submitted for the degree of Doctor of Philosophy, September 2019

ABSTRACT

Recent advances in high-throughput techniques for DNA sequencing and phenotyping have greatly facilitated the identification of genetic variants underlying traits at a genome-wide level. In this study, a large amount of yeast genetic resources and phenotypic data were collected for the study of natural genetic variation in yeast under different environment conditions. Quantitative trait locus (QTL) analysis and epistasis analysis have been applied to *Saccharomyces cerevisiae* on 6 groups of 1st generation bi-parental inter-cross segregants and 12th generation multi-parental high resolution segregants. Using yeast as model organism, growth under stress conditions of a variety of conventional genotoxic agents was measured. Different QTLs were mapped to causative genes that are related to DNA repair and protein transport. In addition, by comparing the genes identified under 19 different agents, 14 frequently occurring genes producing effect on the growth of yeast, were further analysed. QTL output was clustered through a changepoint model for improving the selection of candidate genes in large gene sets.

Furthermore, Temporal QTL analysis was applied to study the dynamic development of yeast growth under X-ray irradiation that expands the phenotype in the time dimension. By comparing the QTL in different time spans, genes that only exhibit effects for a certain period of time rather than continuously through, or at the end of, the experiment were found.

One of the major industrial applications of yeast is brewing. In this project, whole genome sequencing analysis were performed on a highly diverse 12th generation *de novo* hybrid population. Variant calling was applied for these pool sequencing and identification of genetic variants. Pool QTL analysis was applied to compare the allele frequency difference of extreme pools under the same condition. Multiple QTL intervals responding to the brewing environment were identified. This provides useful genetic insights for brewing yeast breeding and improvement.

Acknowledgements

This project has inspired me in every aspect of my PhD journey in Leicester. I would like to thank the following people who influenced my development in science and life. First and foremost, I would like to my sincere gratitude to my supervisor, Professor Ed Louis, for his guidance with patience and continuous encouragement. During the past four years, my mindset was developed from data analysis to bioinformatics, without my supervisor, this could not be possible. Professor Louis provides his expertise in genetics and spends invaluable time for answering me genetic queries. Under his guidance, my knowledge bridge between biology and statistical analysis has been built. I would also like to thank my second supervisor Professor Alexander Gorban, for providing helpful advice and direction of mathematics study. The profound knowledge from them throughout my study gives me the biggest support through the research.

High tribute shall be paid to my colleagues and friends. During my PhD at Leicester, Thomas Walsh, my colleague and good friend, stands out the continuous impact and providing constant help on bioinformatics. He encouraged me to involved in developing the R package with him that influenced my coding style and motivate me to choose bioinformatics in academics as my future career. I would also like to thank Matt Blades for giving me advice on bioinformatics. I would like to thank my probation review panel Dr Steven Foster and Dr Evgeny Mirkes for giving me direct criticism and helpful advice. Thanks to all the members of the Department of Genetics, their kindness and support have given me great encouragement.

Danae, Yishen, Salwa, Aga and Alex did hard work and performed high quality experiments that allowed me to touch real data. I can't start my analysis and research if without their great job. Chatting with them always inspired me and showing me the ways to understand data. Danae's hard work and passion for research gave me the greatest motivation to complete many challenges. I am also very fortunate to meet my fellow and great friends Liliane, Nikola, Monica and Charlotte, the time shared with them is the unforgettable memory.

Finally, I am indebted to my family. I feel so lucky to have the love and care of my partner Dawen, without his company and encouragement, I can't imagine how I can overcome all kinds of difficulties and survive from the stress. Last my thanks would go to my parents, their selfless support and love are always the biggest motivation in my life.

Table of Contents

List of Tables

List of Figures

Abbreviations

DSF: Disulfiram
EPA: Eicosapentaenoic Acid
MET: Metformin
PQ: paraquat
RAP: rapamycin
MMS: methyl-methane sulfonate
Phleo: phleomycin
HU: Hydroxyurea
AIL: advanced intercrossed lines
WA: West Africa
NA: North American
WE: Wine European
SA: Sake
DDR: DNA damage response
SGD: Saccharomyces Genome Database
QTL : Quantitative trait locus
TR: Treatment Ratio
MS: max slope
DOX: Doxorubicin
MAT: Mating Type Locus
OD: Optical Density
CNV: copy number variation
SNP: single nucleotide polymorphism
YPD: yeast extract, peptone and dextrose
PCC: Pearson correlation coefficient
CNHCC: Centralised normalised Hamming correlation coefficient
RIG: relative information gain

Chapter 1 Introduction

1.1 Preface

Everyone on this planet holds their own codes for their life. In nature and in human life, we can observe that there are differences between individuals, sometimes more, sometimes less. For example, there are more than 200 types of breeding cats from the smallest to the largest sizes, from no hair to long hair and with diverse coat patterns as well as fur colours. For humans, there are the well-studied differences of height, Body Mass Index (BMI) and many other traits. (Mayhew & Meyre, 2017). Heredity, the similarity between parents and offspring, is due to the transmission of information, i.e. genes. To understand and have a better view of how the genome landscape shapes the diversity of differences has become a major research topic of genetics. The in-depth understanding of this issue can greatly help the understanding of the generation of species and identify variations within populations that can lead to improved animal and plant breeding through selection. In public health, there is a wealth of possibilities for improving the diagnosis of diseases and the development of personalised medicine (Ware, et al., 2012). Recent advances in high-throughput techniques for DNA sequencing and phenotyping have greatly facilitated the identification of genetic variants underlying the inheritance of complex traits at a genome-wide level (Ansorge, 2009). Despite these developments, our understanding of the genetic basis of complex traits is still limited. Several recent studies recognise that many polymorphisms are involved in complex traits, and detecting these variants and accurately predicting the contribution of genes to complex traits remains a challenging task (Sirugo, et al., 2019). Genomic data resources for model organisms, such as *Arabidopsis thaliana*, *Drosophila melanogaster, Caenorhabditis elegans,* and mice etc., have been expanding in recent decades. Quantitative Trait Locus (QTL) analysis utilises biomarker information to identify genomic regions associated with quantitative traits in a population. With the advancement of biotechnology, the data complexity of QTL mapping has been continuously developed. Thousands of markers associated with human disease

phenotypes have been observed through human genome wide association studies (Suh & Vijg, 2005). Yeast has some key advantages compared to humans, including fast generation time, controlled genetic background, reproducible genotypes, and a variety of experimental validation techniques. As a wealth of genomes in yeast have been fully sequenced, here I use *Saccharomyces cerevisiae* as a model organism for the discovery with QTL analysis for better understanding the genetic variation and the determinants of complex traits. QTLs have also been explored for the hybrids between *Saccharomyces cerevisiae* with *Saccharomyces eubayanus, Saccharomyces kudriavzevii and Saccharomyces jurei* to dissect the association in genetic variation that influences changes in complex traits through whole genome sequence data analysis. Besides looking at markers for observing genetic variation, structural variation, such as copy number variation (CNV), has also been discovered between sub-genomes of *Saccharomyces* yeast hybrids (Van den Broek, et al., 2015). Overall, this thesis shows the complexity of the genetics of quantitative traits and provides some insight into the analysis to understand the complex genetic basis of traits.

As this thesis is an interdisciplinary project, at the beginning of this chapter I will give a brief introduction of necessary genetic elements in quantitative traits. I will then review the main genetic characteristics of yeast with the genetic diversity and the achievements for performing the analysis of complex traits through genome wide data.

1.2 General Introduction of Genetics

1.2.1 A brief history of genetics

As early as centuries ago, humans have already utilised and manipulated genetic problems, such as domestication of livestock and pets, to develop suitable vegetables, fruits, meat and fermented products. The actual establishment for studying mechanism of genetics in scientific theory are based on the experimental and theoretical results from Gregor Mendel in the mid-19th century, although the term 'Genetics' was named by William Bateson and was widely used after 1906 to describe genetics research (Falconer, 1996). In the Mendelian period, there was an informal accepted hypothesis that the genetic characteristics of an individual are derived from the mixed average of the characteristics of their parents (i.e. blending inheritance) (Falconer, 1996). Another popular hypothesis, ‘use and disuse’ theory proposed by Lamarck, also known as inheritance of acquired characteristics. Lamarck believes that organs that are often used by organisms are gradually developed, organs that are not used are gradually degraded, and the acquired traits can be inherited by the offspring (Burkhardt, 2013). However, Mendel’s experimental results negate these hypotheses. His famous pea experiments found that the parents passed specific factors to the offspring and proposed the Mendelian inheritance, law of segregation and law of independent assortment. His results show that genetic characteristics are the result of a comprehensive manifestation of discrete inheritance rather than blending or acquired inheritance, and the genetic laws of many traits can be demonstrated and explained by simple numeric rules and ratios. In the 1910s, based on the discovery of sex linkage caused white eye mutation found in the fruit fly *Drosophila*, Thomas Hunt Morgan and his students revealed that genes located on the same chromosome are linked together (Morgan, 1910). Alleles at a pair of genes can be exchanged between homologous chromosomes, that is, law of linkage and crossing-over. These observations identified chromosomes as the genetic material for explaining Mendelian inheritance. After this finding, many experiments then proved that the chromosome was composed by deoxyribonucleic acid (DNA). In 1953, James Watson and Francis Crick

successfully determined the double helix structure of DNA, which contains two DNA strands, and the chains are paired by nucleotides (Watson & Crick, 1953). The DNA structure expounded how genetics are carried out. Since this, many studies explored DNA functions and structures at the molecular level. The genetic molecular mechanism, central and peripheral dogmas was developed and completed, that is, DNA as a template to generate paired messenger RNA. The nucleotide information on the messenger RNA is used to produce the amino acid sequence on the protein (Lodish, 2016). This translation procedure from the nucleotide sequence to the amino acid sequence is based on the genetic code. In the 1980s, chain termination DNA sequencing (Sanger sequencing) and polymerase chain reaction (PCR) technology enabled scientists to begin reading and duplicate DNA sequences. With the advancement of sequencing technique development in 1990s, the next generation sequencing (NGS) allowed the entire genome to be sequenced at once (Metzker, 2010). The first human whole genome sequencing mapping through Human Genome Project (HGP) was declared complete in 2004 which included around 3.3 billion base pairs and have identified more than 20,000 genes. Up to now, a massive amount of genetic data has been collected and an increasing number of genomic databases have been established and are growing exponentially. In the post genome era, mining these genomic data and extracting useful information is revolutionising the study of genomics and molecular genetics.

1.2.2 Genomes and Genetic Variation

The life process is in an extremely complex system comprised of many components requiring the support of matter and energy. The biological system is also an information system which forms specific life activities by storing, modifying, interpreting genetic information and performing genetic instructions to control the inheritance, metabolism, growth and development of the organism. As the basic element of an organism, cells coordinate and cooperate with the functional expression of organisms. The genome is the complete set of genetic information for a cell comprising the nucleic acid sequences on the chromosomes. Each genome consists of one or more chromosomes that store the DNA sequences for thousands of genes. The genomes of various organisms have basic structural features, however, the genomes of different species differ and the size of the genome varies. The smallest known genome is from the Porcine circovirus, which is a circular genome only 1759 base pairs long (Finsterbusch & Mankertz, 2009). Interestingly, the largest genome is not from the most advanced organism humans, but a lungfish Protopterus aethiopicus (Pellicer, et al., 2010). This lungfish has the largest known vertebrate genome at 130 Gb in size (Pellicer, et al., 2010). The human genetic code consists of more than 3.2 Gb of nucleic acid, containing approximately 20,000 genes located on 23 pairs of chromosomes (Speicher, et al., 2010). The human genome data is continuously refined by whole genome sequencing and most of the genes are assembled and given annotations. The ploidy of chromosomes and the amount of genes in different individuals of the same species could also be different. The location and structure of some parts may change on the chromosomes and even lead to functional changes.

The nucleotide sequence of the DNA stores information on the amino acid sequence encoding of the protein, stores information on the regulation of gene expression, and stores genetic information. The genetic information of the organism is stored in DNA sequence that formed by four nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). Genes have been passed down from generation to generation, and because of this inheritance, species can be

stabilized and optimized (Falconer, 1996). On the one hand, replication of DNA allows genetic information to be passed from parents to offspring. On the other hand, the genetic information of genes are expressed by transcription and translation process, so that the progeny exhibits similar traits to those of the parent. During gene expression, genetic information is first transmitted along the nucleotide sequence of the DNA strand to RNA sequence and then from RNA translated into various proteins to perform specific biological functions (Lodish, 2016). Genes encode the synthesis of proteins. The relationship between the nucleotide sequence and the amino acid sequence of the protein is determined by simple cell translation rules, collectively referred to as the genetic code which is formed by a sequence of three nucleotides. Some DNA sequences also play a structural organization role in the chromosome. For example, telomeres usually contain few protein-coding genes, but are important for chromosome function and stability (Blackburn, 1991). With the development of sequencing technology, a large amount of DNA sequencing data and number of genetic databases have accumulated. DNA sequences are the most basic and intuitive genetic data, however, the translation of genetic information into phenotype is still not fully understood. For example, the function of most DNA non-coding regions is still not clear. To understand the association between genetic information and the key aspects is expressed in the organism are critical. Among those questions, the most primary one is how the characteristics or traits of organisms are affected by genes.

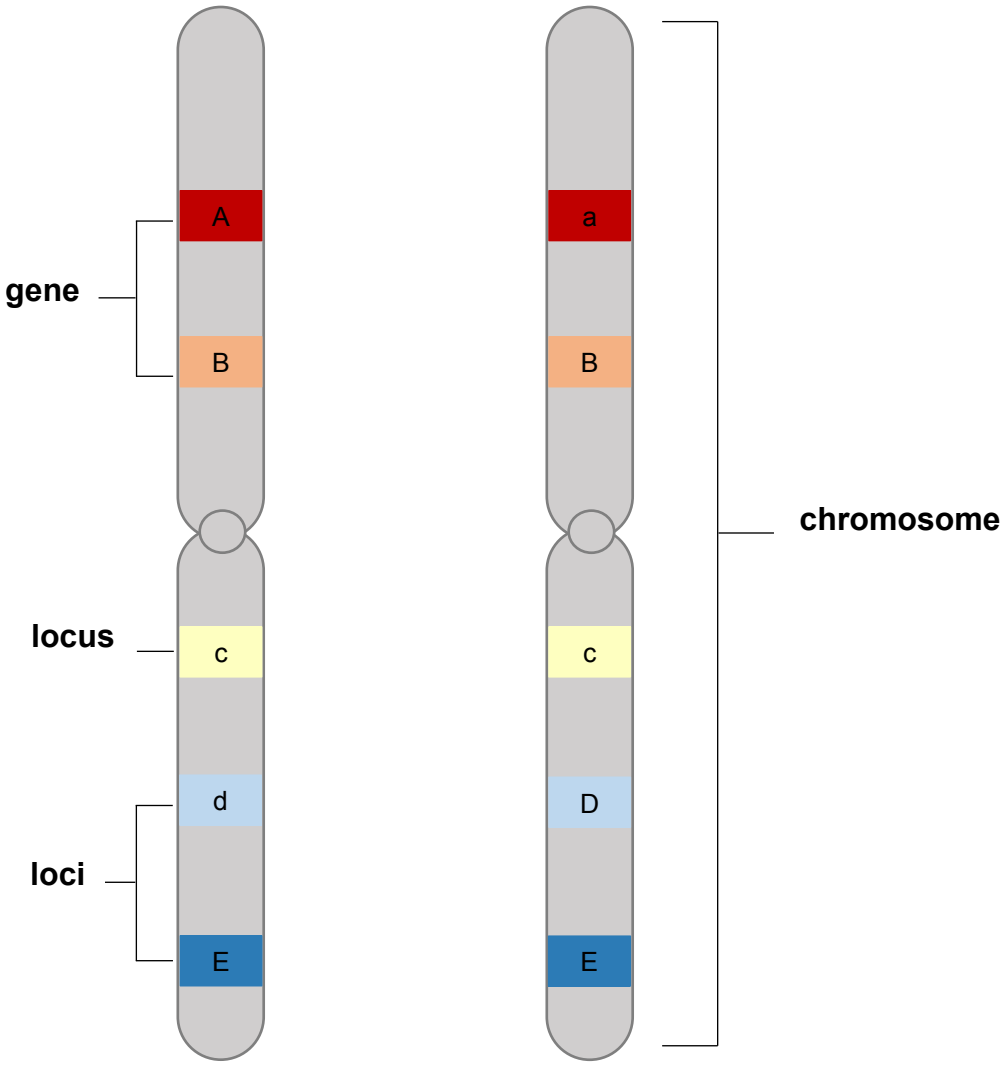


Figure 1-1 Illustration of genetic identification of DNA sequences.

The whole DNA sequence for an individual is called genome. Sequences hold the information of position and base content. DNA sequences are stored in different contiguous groups identified as chromosomes. Each chromosome consists of a continuous sequence is in the same group. A gene is a region of sequence on a chromosome that makes up together a functional unit such as a translated protein. A locus is the position on the homologous chromosomes among different individuals and an allele is the base contents at that locus position. Alleles are different forms of a gene occupying the same position (locus) on a chromosome.

When looking at individual genomes within a species, it can be seen that their sequences exhibit high relatedness to each other. But there are very small differences that make each one unique. Among individuals, mutation is the major source of new genetic variation that results in differences in the nucleotide sequence of the genomes of different individuals. In many organisms, recombination can generate new combinations of variants at different locations, increasing the differences between individuals. Genetic variation is the difference between genomes among a large set of sequences. A locus or genetic marker provides the position of the gene located on the chromosome and at the given locus different alleles could be segregating within the population. Figure 1-1 shows the identification of chromosomes with genes, alleles that located in the same locus. Different alleles are formed as a result of polymorphisms in the genes.

Genetic variants can be classified in terms of their frequency within a population, with common variants defined as when the minor allele frequency of greater than 5% in the population and rare being less than 5%. Single-nucleotide polymorphism (SNP) is the most frequent form of genetic variants among individuals. A SNP is a substitution of a single nucleotide at the locus in the genome that normally occurs greater than 1% frequency in a population. Each SNP represents the differences of two or more nucleotides in A,G,C or T at a given site and could occur not only within the protein-coding regions but more frequently in non-coding regions (Zuo, et al., 2015). SNPs among human genomes can be observed in most individuals. For example, there are about 3.3 million SNPs in a human genome when compared with others (Shen, et al., 2013). However, this doesn't mean all the SNPs are harmful and a large fraction of them do not correlate with disease or affect phenotype. For any particular condition, only a few of SNPs might lead to the causal factors that contribute to the phenotype differences. For example, SNP A1708E located on BRCA1 gene in human that highly linked to breast cancer risk. From SNPedia, a website for storing the human SNPs records that related to common diseases, 110824 SNPs have been recorded that could be the cause of human disease (Cariaso &

Lennon, 2012). Some SNPs among individuals may have different bases without changing the amino acid sequence of the protein as the genetic code could be different for the same protein. SNPs that don't change the protein coding may still contribute to gene expression, gene regulation by transcription factor binding and etc. Figure 1-2 shows the SNP classification by different locations on the chromosome. Besides single substitution of bases, changes in genome structure are also a great resource that contributes to the genetic variation. Several recent studies have pointed to the association between the structural variance and diseases. Insertion/deletion (Indel) polymorphism, copy number variation and translocations are the major structural variation in the population as illustrated in Figure 1-3. Copy number variation (CNV) are large DNA sections that vary in repeats numbers among individuals. CNVs which cover genes can potentially alter gene transcription factors or affect gene expression levels. For example, Schaschl's study identified the association between CNV in genomic regions harbouring dosage-sensitive genes and Autoimmune diseases in humans (Schaschl, et al., 2009).

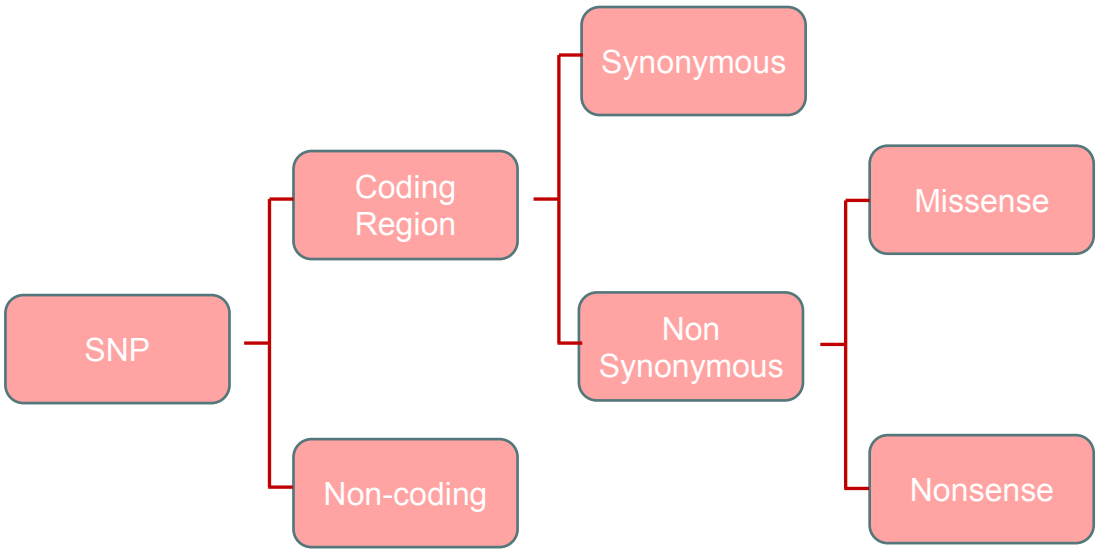


Figure 1-2 SNP Classification

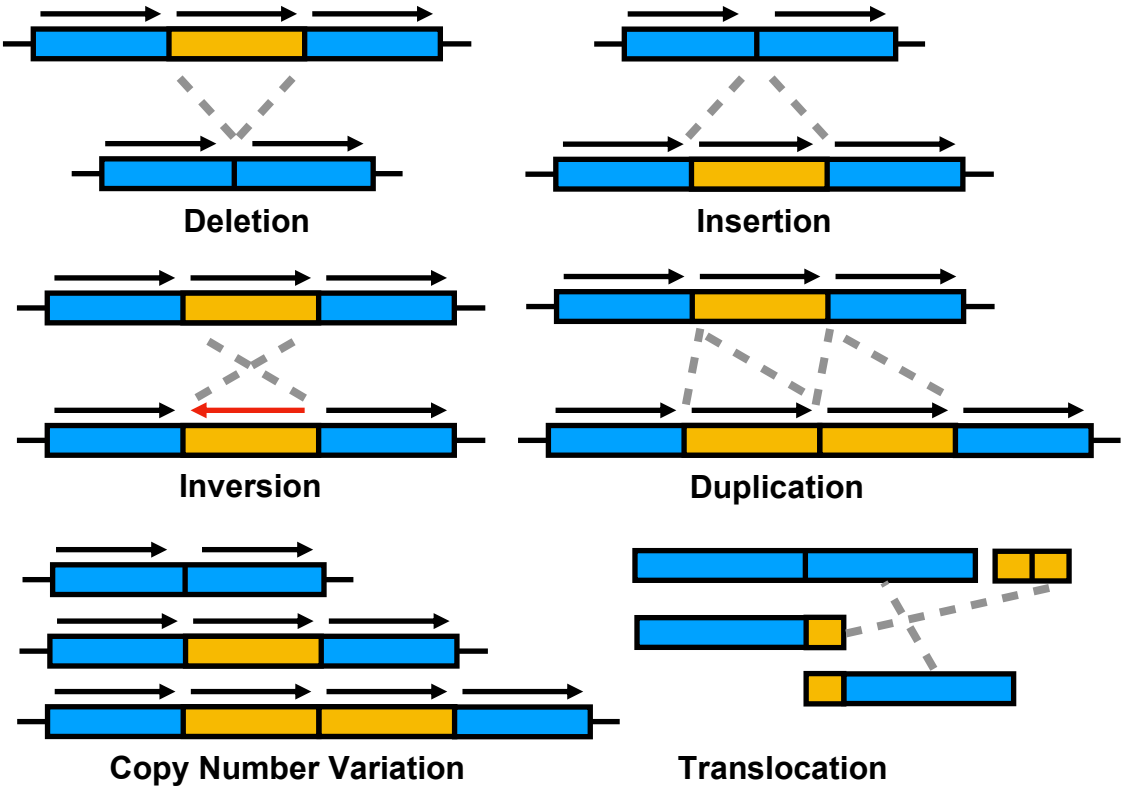


Figure 1-3 Illustration of structural variation.

1.2.3 Genetic traits and phenotypes in a population

The most amazing aspect of nature is the diversity among individuals. Behaviour, physiology or even gene expression can be called a phenotype or trait. Observing the composition of the organism's features that cause phenotypic differences between species or individual representatives of the same species that are encoded in their genomes has always been a major challenge in modern biological science. The sum of individuals of the same species that can interact within a certain time and space can be called a population. Phenotypic variation within the population, resulting from genetic variation and interaction with the environment, is a basic prerequisite for evolution through natural selection as the genetic contribution to the variation will be transmitted into the next generation. Since Darwin published his theory of evolution by natural selection in 1859 (Darwin, 1859), and the contemporary Mendel elucidated the laws of genetics just a few years later (Mendel, 1865), it was not until the early twentieth century

that these ideas were integrated, culminating in the modern evolutionary synthesis (Fisher, 1930).

Classical Mendelian traits, which have qualitative phenotypes that measured into classes, are often caused by the action of one major genetic variant or gene. Much progress has been made in identifying the molecular basis of Mendelian traits, such as the classical experiment of pea colour by Mendel (Mendel, 1865) and blood types of humans (Chong, et al., 2015). By contrast, many of the traits vary within a population exhibit continuous variation, usually shaped in a normal distribution, which means they cannot be classified into qualitative classes (Lynch & Walsh, 1998).

Quantitative traits, also known as complex traits, are typically measured numerically or in ranges, such as body height, Body Mass Index and blood pressure in humans. The phenotype of a quantitative trait depends on the cumulative influence of many genes working interactively along with the effect of the environment (Falconer, 1996). It is a great challenge to find out the exact number of genes involved in quantitative trait and to explain the total contribution of heritability. Furthermore, in addition to the combined effect of genes, it is also possible to find interactions among genes, referred to as epistasis, and the interactions between genetic variations and the environment.

1.3 Mapping the relationship between genotype and phenotype

1.3.1 Genetic Map and Physical Map

Although genes are discretely located on the chromosomes, SNPs can be inherited together rather than independently when they are located closely in a nearby region on a chromosome which means they are linked. The process of using observed genotypes of these loci known for the genome to infer recombination frequencies during crossover is called genetic map. Unlike the physical map that indicates the base position (bp) on the chromosome, the unit of genetic linkage is the centimorgan (cM). A distance of 1 cM between two loci means that the markers are exchanged to with each other in 1% of meiotic products. Complete independence or lack of linkage results in 50% recombinant meiotic products. With linkage among DNA sequences, haplotype mapping clusters these markers so that a single SNP can identify many linked SNPs through linkage association.

1.3.2 Genomics based approaches

Quantitative trait loci (QTL) are specific segments on the chromosome that affects phenotypic variation of traits and can locate genes that control complex quantitative traits. A quantitative trait locus can include multiple SNP sites in a genomic region associated with a quantitative trait. With the advancement of whole-genome sequencing technology, many studies on the identification of QTLs for various quantitative traits in animals and plants have been reported. Although QTL positioning technology has made significant progress in genetic screening, due to the complexity of complex traits, there is still the challenge of identifying causal relationships. Complex traits are affected by genetic and environmental factors, and the uncontrollability and complexity of environmental factors in human studies make it difficult to study a large amount of complex diseases. In addition to humans, the growth environment of many organisms may also be unstable. Therefore, many QTLs can only be detected under very special environmental conditions. Identification of genes or quantitative trait loci that

exhibit significantly different phenotypic effects in different environments by linkage or association mapping.

With the two types of information for individuals, i.e. phenotypic data and genotypic data, the question comes is to discover the action, interaction, number, and precise location of the genetic regions and the variants at these regions that are responsible for the phenotype. Recent advances in genomics and marker-assisted selection have greatly facilitated and strengthened biological research processes such as genetic screening, breeding, etc. (Nadeem, et al., 2018).

Linkage analysis is a method that tests the variable intervals on chromosomes among the population that contribute to the expression of traits. When a new mutation arises on a particular chromosome, initially there is a large shared segment of DNA with a particular combination of variants, a haplotype, and hence it is in linkage disequilibrium with these variants. With each subsequent generation, this region of linkage disequilibrium becomes smaller as a result of meiotic recombination. The basic approach in parametric linkage analysis is to determine if alleles at a genotyped marker segregate with the alleles at a putative trait locus together more often than one would expect by random assortment or chance. This can be assessed by comparing the frequency of recombinant chromosomes in which a crossing over event has rearranged the parental chromosomes to the frequency of non-recombinant chromosomes.

Quantitative trait loci (QTL) analysis involves finding the association of a genetic variant with the variation in a quantitative trait through an experimental cross or by association within populations (Miles & Wayne, 2008). The linkage of complex trait and polymorphisms can give a genetic explanation in human disease mechanism, agriculture, and evolutionary theory (Brem & Kruglyak, 2005). Historically, the availability of adequately dense markers (genotypes) has been the limiting step for QTL analysis. Recent advances in high-throughput methods for DNA sequencing and molecular linkage mapping construction have a greatly facilitated the determination of quantitative trait genes (QTG) and quantitative trait nucleotides (QTN).

Genome-wide association studies (GWAS) are becoming increasingly popular in genetic research, and they are an excellent complement to QTL mapping. Whereas a QTL can contain many linked genes, which are then challenging to dissect, GWAS produces many unlinked individual genes or even nucleotides, but these studies are riddled with large expected numbers of false positives. Though GWAS remains limited to organisms with genomic resources, combining the two techniques can make the most of both approaches and help provide the ultimate deliverable: individual genes or even nucleotides that contribute to the phenotype of interest.

1.3.3 Heritability

Genetic recombination generates random reshuffling of genetic information on the homologous chromosomes from parents during meiosis, forming new DNA sequences that are passed on to offspring. With the recombinant chromosome sequences among various individuals, each locus may be made up with different genotypes, i.e. alleles. Primarily meiotic recombination results in different linear combinations, haplotypes, of alleles of various genes along the chromosomes. Sometimes new alleles are generated when recombination occurs between variants within a gene.

In addition to the genetic factors that make up the phenotypic variation, the environment can have an effect on the trait, which could be simply adding to measurement error but could also interact with genetic variation. Heritability is a measurement to estimate the contribution of genetic variation among individuals towards phenotypic variation. There are two types of heritability that can be estimated, broad sense heritability and narrow sense heritability. The broad sense heritability estimates the ratio of total genetic variance to total phenotypic variance. When accounting the effect of gene-gene interactions, the narrow sense heritability estimates the ratio of additive genetic variance to the total phenotypic variance (Evans, et al., 2018). When genome-wide genotype data and phenotypes from large population samples are available, one can estimate

the relationships between individuals based on their genotypes and use a linear mixed model to estimate the variance explained by the genetic markers (Bloom, et al., 2013). This gives a genomic heritability estimate based on the variance captured by common genetic variants.

1.4 Background of yeast

1.4.1 Cell cycle of yeast

Saccharomyces yeast is a unicellular eukaryote that has a nucleus and an endomembrane system. Yeast cells can divide as fast as 90 minutes vegetatively (mitotically), and diploid cells can undergo meiosis in as little as a day with meiotic haploid products following the Mendelian Laws of Segregation. Haploid yeast cells could be mating type a or α, i.e. MATa and *MAT*α. The yeast life cycle can be asexual and sexual as illustrated in Figure 1-4. In asexual growth, these cells can undergo mitotic cell division through budding, producing daughter cells with the same mating type. In laboratory strains, the mating type of haploid cells is stable due to the absence of a functional HO endonuclease. The two cell types release pheromones, initiating the formation of shmoos and subsequent mating, resulting ultimately in a stable diploid MATa / *MAT*α (a / α cell). Diploid cells also divide mitotically by budding to produce genetically identical daughter cells. Under nitrogen starvation this diploid then produces four haploid spores called a tetrad through meiosis. Each tetrad consists of two copies genetic information of each of the two parents and the four haploid spore cells consist of two MATa cells and two *MAT*α cells with recombination of chromosomes between the parents. The Meiotic and Mitotic processes are illustrated in Figure 1-4.

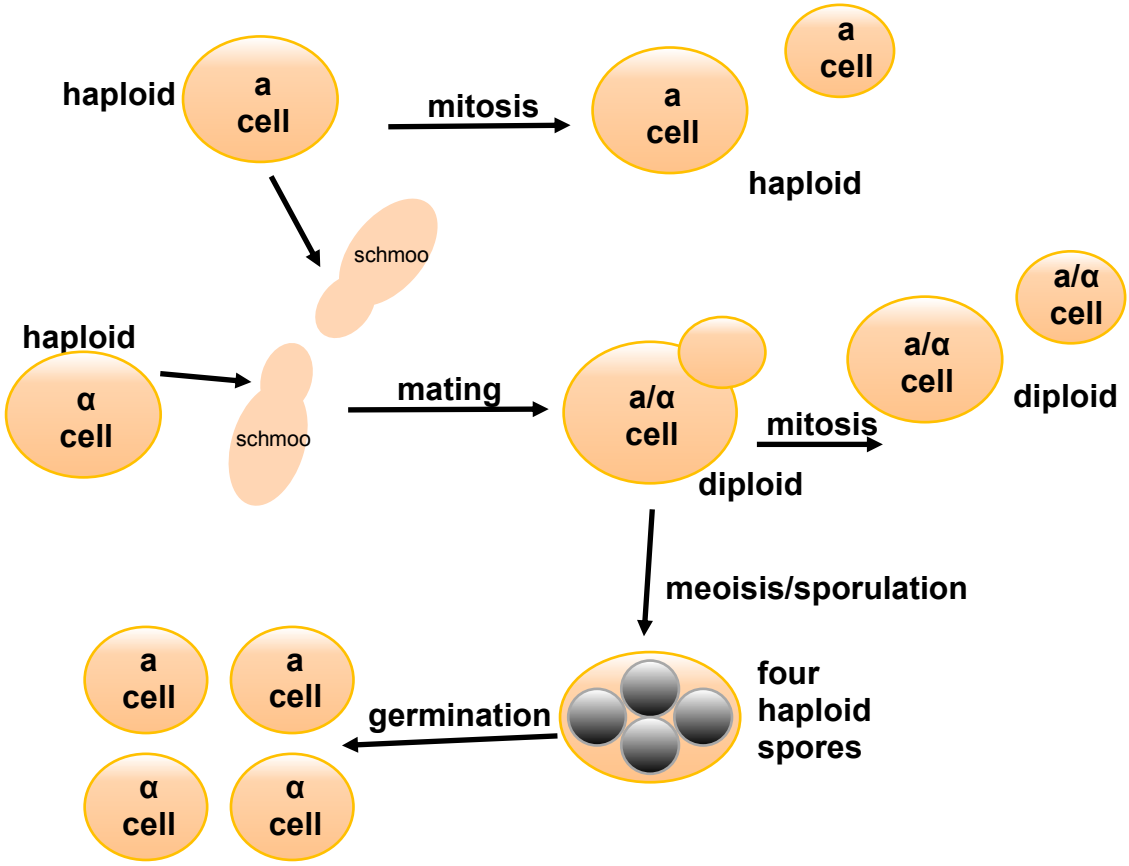


Figure 1-4 A simplified life cycle diagram of budding yeast.

Mitosis process can occur in both haploids and diploids. During the asexual process, identical daughter cells are produced through budding, either haploid (1n) or diploid (2n). Haploid a and haploid α cells can shmoo in response to each other to mate (a/α dioloid). During the sexual process under starvation, diploid cells (2n) go through the meiotic process to generate a (4n) cell and then this cell will be separated into 4 haploid spores (1n) during sporulation resulting in two a cells and two α cells. The figure was adapted from (Duina, et al., 2014).

1.4.2 Genetic Diversity of *Saccharomyces* yeast

The most widely used *Saccharomyces* yeast species is *Saccharomyces cerevisiae,* which is also the first eukaryotic genome that has been fully sequenced and well annotated (Goffeau, et al., 1996). Since then, an increasing number of full genome sequences have been completed, which provide great resources to understand genome diversity and evolution. The reference genome of the laboratory strain S288C consisted of about 12,000,000 base pairs and has over 6000 open reading frames (ORF) arranged among 16 chromosomes (based on R64-2-1, SGD). Almost all the genome features have been recorded in the SGD database and over 5000 genes have been verified. After S288C, more wild isolates and lab strains have been sequenced and aligned to the reference genome, for example, RM11-1a. Also, *Saccharomyces cerevisiae* has been the most widely domesticated yeast for centuries and S288C is the most important reference for assessing the genetic variation of other yeast strains. Besides S288C, 99 characterized Chinese isolates have been sequenced and well-studied for finding the human-associated domestication (Wang, et al., 2012). Moreover, there are over 1000 diverse strains isolated all over the world have been fully sequenced (Peter, et al., 2018) which created the largest data resources to study the budding yeast evolutionary process. Figure 1-5 gives the phylogenetic tree of *S. cerevisiae* strains.

With more and more *S. cerevisiae* strains being discovered, a large number of yeast strains isolated in the *Saccharomyces* clade have been defined and sequenced as well. Figure 1-6 illustrates the clade of closely related *Saccharomyces* species as currently is known. It is composed of *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. jurei*, *S. kudriavzevii*, *S. arboricola*, *S. eubayanus* and *S. uvarum,* along with their interspecies hybrids. *S. paradoxus* is a wild yeast that is currently the most closely relative to *S. cerevisiae*. It is a good resource for tracing the human activities by comparing with *S. cerevisiae* (Yue, et al., 2017). From this, both *S. cerevisiae* and *S. paradoxus* show diversity in the subtelomeric regions and exhibit phenotypic diversity among different strains.

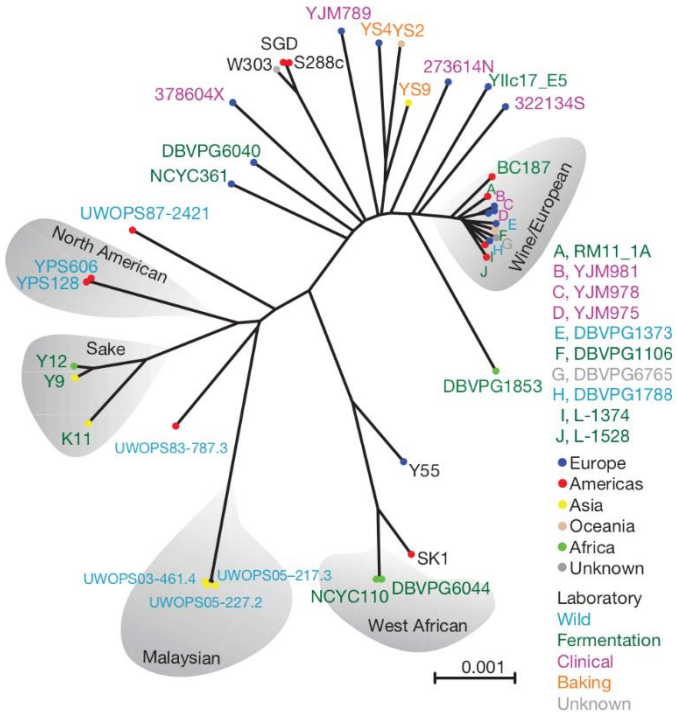


Figure 1-5 Phylogenetic tree

Phylogenetic tree of S. cerevisiae. S. cerevisiae strains in North American, Wine/European, Sake, West African and Malaysian showed clean lineages highlighted in grey. Different colours gave the information about the name and origins. *(adapted from G. Liti et al. 2009)*

Unlike *S. cerevisiae* and *S. paradoxus*, other species have only a few isolates. *S. kudriavzevii* was first isolated in Japan (Naumov, et al., 2000) and then the European strain was found in Portugal (Scannell, et al., 2011). There are two species identified recently. One of them is *S. eubayanus* - the missing parent of the hybrid *S. pastorianus* – which has been searched for over a long time (Nguyen, et al., 2011). It was first isolated in Argentina (Libkind, et al., 2011) and then discovered in North America, west of China and New Zealand (Peris, et al., 2014). In addition to being a parent of *S. pastorianus*, the identification of *S. eubayanus* as a pure species also affected the species definition of *S. bayanus*, as it appears that many members of this species complex are actually hybrids of *S. eubayanus* and *S. uvarum* (Nguyen, et al., 2011). Another newly identified species is *S. jurei* (Naseeb, et al., 2018), which currently is represented by only two strains isolated from France.

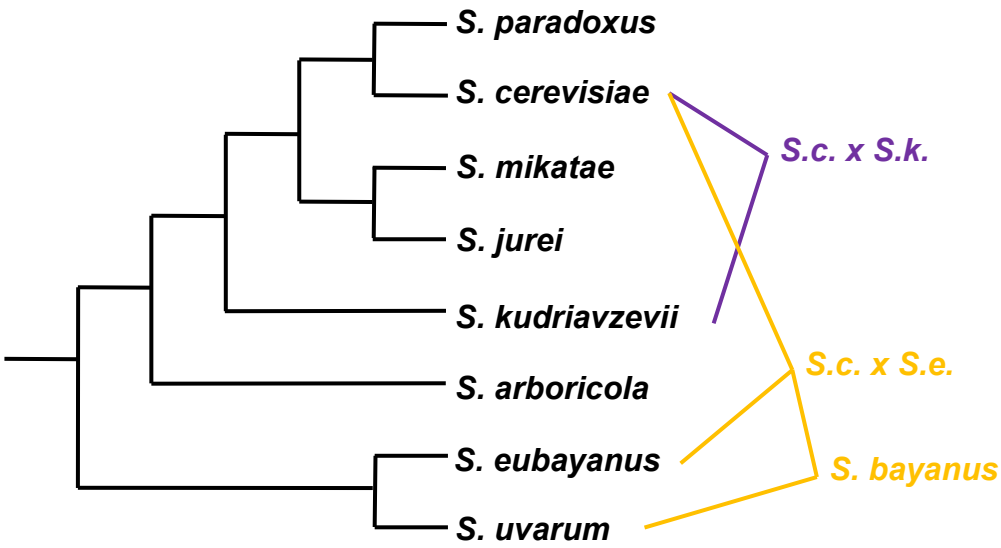


Figure 1-6 Phylogenetic tree of Saccharomyces yeast.

The Saccharomyces sensu stricto group is composed of eight biologically distinct yeast species, namely *S. cerevisiae*, *S. paradoxus, S. cariocanus, S. uvarum, S. mikatae, S. jurei, S. kudriavzevii, S. arboricola* and *S. eubayanus*, and two natural hybrids, namely *S. pastorianus* and *S. bayanus*. Adapted from (Dujon & Louis, 2017).

1.5 Yeast Application in biotechnology and human health

1.5.1 Using yeast for alcohol production

Yeast has many attractive features and it is widely used in industrial, commodity production, environmental protection and production of biofuel. Brewing alcohol and bread baking are the most common and oldest ways to use yeast. As *Saccharomyces cerevisiae*'s name is original from the meaning of beer, *Saccharomyces cerevisiae* plays a major role in beer fermentation. In addition, many yeasts can be used to produce a variety of feeds as well as industrial nutrients such as yeast extracts. Some yeasts are resistant to acids, high permeation, decomposition and absorption of toxic substances, and are widely used in the field of sewage treatment.

Yeast is the main fermenter and is well applied in the production of alcoholic beverages, such as beer, fruit wine, and distilled spirits. Yeast consumes carbohydrates such as sugars from grains and fruits under anaerobic conditions or low oxygen concentration to provide energy while producing alcohol and carbon dioxide. The most common species used in beer and wine brewing is *Saccharomyces cerevisiae*. In the past years, yeast hybrids between *S. cerevisiae* and non-*S. cerevisiae* have been found to have advantages in brewing alcohol. For example, *S. pastorianus*, which is the hybrid between *S. cerevisiae* and *S. eubayanus*, exhibits hybrid vigour in larger brewing conditions. Moreover, hybrid strains combining the genomes of *S. cerevisiae* and *S. kudriavzevii* have been found in certain fermentation environments, mostly from wine and cider fermented at low temperatures.

1.5.2 Using yeast as a model organism to study human complex traits

Many physiological and disease related traits in humans display complex genetic landscapes, such as DNA damage responses to genotoxic agents. Mapping the genetic contribution of such human traits has been well explored, however, the studies have been limited due to the complexities of human environmental and ethnic differences. Based on the theory of evolution, many of the basic properties of life are conserved among the various biological species on earth, and biologists can reveal some generalization by conducting scientific research on selected simple representative biological species, i.e. model organisms. Model organisms have been widely studied for most human complex trait research and thousands of gene variants and pathways were identified using model organisms, such as mice, fruit flies etc. Yeast, as the simplicity of its small genome size, rapid division time (90 minutes), and high recombination rate, has been greatly used to understand the genetic and cellular defects behind human complex diseases (Liti & Louis, 2012). *Saccharomyces cerevisiae* is a simple single cell eukaryote with only 16 chromosomes but 6000 genes included. The reference genome of *Saccharomyces cerevisiae* has a complete full genome sequence and well-annotated genes for the lab strain S288C discovered in the 1950s, and it has become an ideal tool for studying pathways of life processes in higher eukaryotes. As a well-studied model organism, yeast and human have high similarity in amino acid sequence in many genes so its genetic studies provide a good comparison for research in human genetics. Many genes and cellular signaling pathways are 30% conserved from yeast to humans. Gene deletions, tagging and mutagenesis are easier and faster than in human cells, and thus deciphering gene function is more easily accomplished in yeast. When a novel human gene with unknown function is discovered, it can be quickly retrieved from any yeast genome database to obtain a yeast gene with a hopefully known function, and obtain information about its function, thereby accelerating the human functional study of genes. Furthermore, the similarity between yeast genes and genes associated with human polygenic hereditary diseases will provide important assistance in improving our diagnostic and therapeutic levels.

1.6 QTL analysis for studying yeast population

A great number of large-scale sequencing studies have investigated the genetic diversity within a species, revealed a sight into the genetic characteristics. With more and more genomic data and phenotype data at the yeast population level being generated, QTL analysis has become a major tool for identifying the causative genes or gene variants. The phenotypic diversity observed by different strains of yeast lead to a large number of QTL analyses applied for various studies. Many yeast crosses between different strains were designed, and their progeny were cultured in the environment of interest to characterise their offspring segregation of phenotype differences.

Advanced intercross QTL mapping uses individual genome sequences to call variances and identifying the variants that contributes to phenotype variance. The generation of advanced intercross lines (AIL) is a powerful approach for high-resolution fine mapping of QTLs, as they accumulate many more recombination events compared with F1 intercross populations. Further generations of intercrossing breaks the linkage disequilibrium among closely linked variants, increasing the genotype/haplotype variation to a sufficient scale for fine-scale mapping of quantitative trait loci (QTLs). Advances in genotyping technology and techniques for the statistical analysis of AILs have permitted rapid advances in the application of AILs. Lab strains BY and RM11-1a were used as founders to generate a huge population of second generation offspring for characterizing expression difference in yeast. In the same cross design, QTL analysis was applied to study carbon source change, transcript levels, protein abundance, ethanol production, etc. (Ehrenreich, et al., 2009). However, the experimental strains have limited variation among these two founders and small effect QTLs cannot be detected due to lack of power in genotype information. In addition, industrial wine strain EC1118 crossed with S288C was used for several QTL mapping of fermentation features, such as wine aroma (Steyer, et al., 2012), multi stress condition (Brion, et al., 2013), etc. In order to expand the genetic and phenotypic diversity available to explore, four natural *S. cerevisiae* isolates with

clean lineages, that were sequenced in the *Saccharomyces* Genome Resequencing Project (SGRP), were selected to generate first generation progenies for each of the six pairwise crosses (Liti, et al., 2009), (Cubillos, et al., 2011). These four founders are YPS128 [North American ("NA")], DBVPG6044 [West African ("WA")], Y12 [Sake ("SA")] and DBVPG6765 [Wine/European ("WE")]. The crosses were used to map the QTLs related to heat sensitivity, paraquat, and arsenite (Cubillos, et al., 2011). Similar crosses were used to map fermentation related features including acetic acid, glycerol, residual sugar, and succinic acid production (Salinas, et al., 2012). The SGRP4-X population is a 12^{th} generation advanced intercross line by a multi-parental cross of these four lines, designed to reshuffle their genomes through meiotic recombination so that a large number of progenies were highly recombinant, reducing linkage disequilibrium (Cubillos, et al., 2013). This large population has been used in numerous QTL mapping studies such as nitrogen consumption and it has proved the power for detecting phenotypic diversity and catching the small effect QTLs (Cubillos, et al., 2017).

The extreme QTL (X-QTL) method, uses the traditional bi-parental cross, exhibiting high variation in the progeny to expand the phenotype diversity. It starts with generating a large pool of offspring segregants with genotype and phenotype diversity. Segregants were then selected by extreme phenotype conditions (Ehrenreich, et al., 2010). There are many ways to design the crosses and experimental selections to produce pools for comparison. The two main methods for pool conditions are a control pool with normal growth population compared with an extreme phenotype pool or individuals from the two opposite extreme thresholds, such as 5% lowest fitness versus 5% highest fitness are compared. The two pools then sequence to identify polymorphic loci by analyzing the allele frequency differences among the two groups. Under selection, the allele frequency of no effect loci for traits in the progeny should be equal as they do not contribute to survival. While the causal loci that affect the trait will show very different allele frequency distribution as it gives the main influence for survival during pool selection. Using this method, a cross between lab strain BY and

clinical isolate 3S was used to define the smooth and rough morphological phenotype in yeast (Taylor, et al., 2016). Although X-QTL does not locate the precise causal gene at a base level, pooling segregants with extreme phenotypes from a large population directly targets the intervals showing allele frequency differences and it increases the sensitivity of multiple loci while also being cost saving in sequencing.

1.7 Aims and Objectives

The objective of this project is to study the effects of yeast genotypic variation on the quantitative trait through QTL analysis. This thesis mainly focuses on the following issues:

Firstly, QTL analysis was performed for *S. cerevisiae*'s bi-parental cross F1 generated by the 6 populations derived from the four founders in SGRP4 and F12 4-way cross offspring to determine significant sites that could explain changes in growth traits under different conditions.

Moreover, Fine mapping and solutions were applied to find the causative genes and variation with the modelling of cluster analysis of genetic changes, the non-inherited effects, and the changes in significant sites at different time points to measure the QTL.

Furthermore, to understand interspecies hybrid causes of phenotypic variation under different external environments and extreme conditions, whole genome sequencing analysis was performed with 12th generation *de novo* hybrids including *S. cerevisiae* × *S. kudriavzevii*, *S. cerevisiae* × *S. eubayanus* and *S. cerevisiae* ×*S. jurei*. De novo assembly, Variant Calling, and Pool QTL analysis were applied.

1.8 Data Statement

All the data analysis and bioinformatics analysis in this thesis were done by the author. This section introduces the sources of experimental data analysed in this thesis.

- In Chapter 2, the experiment mentioned in Section 2.2.2 was performed by Yishen Li, the University of Leicester.
- In Chapter 3, the experiments mentioned in Section 3.2.2 were performed by Yishen Li, Danae Georghiou, and Salwa Almayouf, the University of Leicester.
- In Chapter 5: the experiment mentioned in Section 5.2 was performed by Danae Georghiou, the University of Leicester.
- In Chapter 6: the experiments mentioned in Section 6.2 were performed by Alex Hinks Roberts, Agnieszka Maslowska from the University of Leicester and Dr. Samina Neesab from the University of Manchester.
- In addition, all phenotype data check, preprocessing and data analysis for these experiments in this thesis were performed by the author.

1.9 Structure of this thesis

This thesis is structured into two major parts. The content of each part is roughly divided into three groups: Description of Data, Analysis, and Assessing the Results.

The first part studies the genetic basis of *S. cerevisiae*'s F1 two-way cross and F12 4-way cross to identify QTLs for growth traits under different conditions. The second chapter is the collecting and pre-processing of the genotype and phenotype data corresponding to 6 groups of F1 crosses of *S. cerevisiae*, the third chapter mapping QTLs under different treatment for studying the drug response of F12 4-parental segregants with epistasis identification. The fourth chapter is the evaluation of the QTL analysis with clustering method. And the fifth chapter is to expand phenotype data to time dimension to consider the detection of temporal QTL for growth traits under the dynamic development with time.

The second part is mainly for mapping pool QTLs for different diploid *de novo* hybrids, considering the characteristics of interspecies to cope with species living under different conditions. The content of this part is written in Chapter 6. The final chapter is the discussion and future work. The appendix contains the code, chart and supplementary content of the analysis.

Chapter 2 Analysing genetic variation affecting growth in a recombinant intercross yeast population

2.1 Introduction

The risk of diseases affecting human health are mainly caused by multiple interacting factors with complex mechanisms underlying obesity, cancer, hypertension, etc. The phenotype of such complex disease-related traits doesn't show obvious Mendelian inheritance but is characterised by continuous quantitative trait variation. However, many variants have small marginal effects and are coordinated by multiple genes and the environment. The genetic mechanism for these complex diseases is very complicated and making it difficult to make a clear diagnosis in medicine. The inherent complexities of human make the identification of disease-associated genes and the mapping of related genetic loci underlying the quantitative trait variation a challenging task. The use of model organisms provides ideal experimental conditions for the understanding of phenotypic variation because of the ability to control the environment, gene composition, etc. The budding yeast, *S. cerevisiae*, a simple genetic system with small genome size, rapid division time, and high recombination rate, is a great resource in high-throughput studies for modelling the genetic characteristics of complex traits.

This chapter aims to study and explain the relationship between genetic variation and phenotypic responses to the chemical treatment Doxorubicin in different yeast populations, quantifying the effect of genetic factors among six large sets of recombinant populations of haploid yeast. Doxorubicin is a widely used chemotherapy agent to treat different types of human cancers, for example, breast cancer, stomach cancer, and others. F1 segregants genotyped using high resolution melting PCR are provided by the previous study (Cubillos, et al., 2011).

The founders of these six group F1 offspring are based on four strains in the yeast genome sequencing project SGRP-4 in (Liti, et al., 2009). Each of the offspring has been genotyped and has an accurate genetic map compared to the reference genome S288C. Using the data available from these individuals, QTL Analysis were performed to identify treatment-related loci associated with yeast growth under the Doxorubicin environment condition. By comparing the results of the offspring of different parental combinations, 44 common QTL intervals that overlap among crosses and 45 unique QTL intervals were found. 8 candidate genes were selected through functional annotation for the genes within the QTL peak. These genes or regions are screened to account for differences among phenotypes.

2.2 Materials and methods

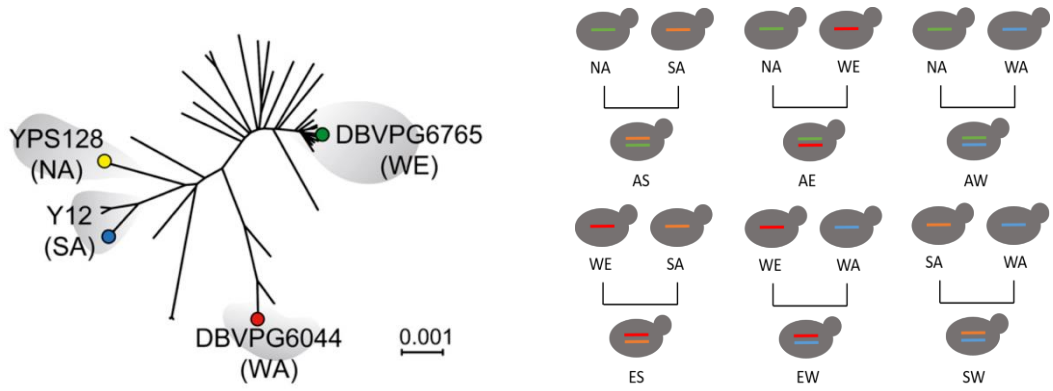
2.2.1 Strains

The yeast haploid strains used for the QTL mapping experiment were progeny of F1 hybrids. Four haploid strains in SGRP4: North American (NA): YPS128, West African (WA): DBVPG6044, Sake (SA): Y12, Wine ／ European (WE): DBVPG6765 were selected as founder strains (Liti, et al., 2009). They represent much of the diversity of natural populations of *S. cerevisiae* and represent clean lineages, that have not interbred with each other. Previously pairwise combinations of the strains with different mating types (MATa and *MATα*) were crossed to generate six different F1 diploid hybrids (Cubillos, et al., 2011). These diploid hybrids were sporulated and tetrads containing the four haploid meiotic products were dissected. From each cross 24 tetrads were dissected generating 96 haploid segregants. The segregants were constructed as shown in Figure 2-1 (b). 576 segregants used in total were obtained in (Cubillos, et al., 2011) . Figure 2-1 (a) illustrated the clean lineage of these four founders and the cross procedures for F1 segregants. Table 2-1 summarised the marker characteristics in the 4 founder strains which existed before crossing.

Table 2-1 Founder strains marker characteristics

ho, ura3/URA3, ly2, HygMX, KanMX were gene names where selected marker located. 'x :: Y' means x gene was replaced by the Y gene in the related strain. (derived from (Cubillos, et al., 2011))

Strain Name	Collection ID	Marker Information
NA (A)	YPS128	ho::HygMX, ura3::KanMX
WE (E)	DBVPG6765	ho::HygMX, ura3::KanMX, lys2::URA3
SA (S)	Y12	ho::HygMX, ura3::KanMX, lys2::URA3
WA (W)	DBVPG6044	ho::HygMX, ura3::KanMX



(a) (b)

Figure 2-1 Genetic diversity and cross design of the founder strains.

(a) Phylogenetic tree includes all the *S. cerevisiae* strains sequenced in the SGRP project (Liti, et al., 2009), the four founder strain were highlighted in separate clusters (Adapted from (Cubillos, et al., 2013)). (b) F1 cross design for SGRP-4X pairwise combination.

2.2.2 Experiment Preparation and Phenotyping

The experiment of QTL mapping under Doxorubicin environment condition using F1 segregants - a total of 12 yeast arrays for phenotyping with control and treatment were performed by Yishen Li, the University of Leicester. Yeast strains were kept on solid YPD media under normal conditions for incubation. The temperature was set at 30 °C as the most suitable growing condition of yeast cells in the incubator overnight for solid media cultivation as required. After sufficient growth they can be stored temporarily at 4 °C in a refrigerator or 24hrs in a cold room. For precise phenotyping the growth of each segregant, rectangular media plates - Singer Plus Plate from Singer Instruments company - were used. Under normal growing condition, yeast strains were stocked separately by MATa and *MATα* in 96 spot density arrays on 20g/L yeast extract, peptone and dextrose (YPD) solid agar media. Transferring stock strains onto strains used for experiment were achieved by robotic micro-cultivation using ROTOR HDA rotor from Singer Instruments company. The robot made 96 or 384 arrays though 96 or 384 Long-pin repads (Barton, et al., 2018). Agar media was prepared by pouring on the leveller to ensure a flat surface for the repad pinning with even pressure. Each strain was then replicate 3 times from stock plates and 4 replicates of each were recorded on another YPD agar with 384 densities plates. Two identical strains arrangement arrays with the same mating type were generated for the comparison between control and treatment. The 384 array plate was replicated onto 6.5g/L YPD soft agar plates first which was then replicating onto 20g/L YPD hard agar control plates and YPD hard agar treatment plates with Doxorubicin added for later phenotyping purposes.

For experiment treatment, identical arrays as the control were used with a total dose of 100uM Doxorubicin concentration applied. All 12 plates of 384 colony arrays on YPD (control) or YPD with Doxorubicin (treatment) were assayed for cell growth using FLUOstar Omega multi-detection micro-plate reader. The plate reader recorded the growth of yeast segregants individually on the array by measuring the absorbance through optical density (OD) values at a wavelength of 600nm every 20 minutes for 65 hours. The temperature of the incubator in this

plate reader was set at 30 °C for keeping the identical standard growth environment. Before printing the array stored with yeast strains, the empty plate was recorded for measuring the absorbance of agar on the empty media and initial printed mass absorbance. The growth values are all filtered out the absorbance of agar and initial plate mass caused by the robot.

PHENOS (PHENotyping On Solid media) is an optimised pipeline for recording the growth of *Saccharomyces* yeast developed under Python 2.7 by Dr. Dave Barton (Barton, et al., 2018). After growth measurements, a directory was generated for each plate by PHENOS which included a comma delimited file of strain identities and each time point reading for strain growth value. The animation for yeast growth and the growth curves for 65-hour yeast growth were also included in the output directories. As the same two strain layout arrays were control and treatment, they were then combined and compared at the end time points for optimising the growth treatment ratio (TR ratio = Average OD value with treatment / Average OD value without treatment) which is computed by PHENOS. In addition, there are some optimising growth features for assessing yeast growth curves could be analysed through PHENOS which are lag (Lag) and speed (Max Slope) under logarithmic phase.

In this chapter, the default features TR ratio, treatment Lag and treatment Max Slope were collected as attributes for each strain. In addition, the control Lag, the control growth Max Slope were also collected and analysed as the growth features for assessing and comparing the growth status under YPD and YPD with Doxorubicin environment condition.

2.2.3 Genotyping

The genotype data of F1 segregants from biparental intercrosses between pairs of the four representative founder strains (NA, WE, SA, WA) of *S. cerevisiae* were obtained in previously (Cubillos, et al., 2011). Different pairwise combinations of these four founders generated 6 crosses in total for next stage experiment. A total of 179 bi-allelic markers were SNP-genotyped for all 576 segregants among whole population without missing data. Marker ID was labelled with format that consists of the chromosome information ('c' + 01:16) and base position with 7 digits which mapped the markers to reference genome of the budding yeast *S. cerevisiae* S288C (version R64-2-1) (downloaded from *Saccharomyces* Genome Database (SGD) (https://www.yeastgenome.org/)). It also contains the information of the standard gene name if the marker is located within a gene based on the genomic feature annotation gff3 file supplied with the reference genome. For example, marker c01:0038000 (CNE1) was assigned to chromosome 1, base position 38000 located in the region of gene CNE1. Genetic mapping (in centiMorgan units) was performed with the R/qtl package using its standard method in the R computing environment. Table 2-2 gives the marker label information table for chromosome I markers. The full table is attached in Appendix C. Table 2-3 describes the marker numbers and average density in each chromosome based on the statistics of the reference genome S288c. After requiring each segregant to have both genotype data and phenotype data in at least one environment, all 576 segregants were retained for subsequent analysis.

Table 2-2 Marker information of F1 population in chromosome I.

Marker ID	Chromosome	Positon(bp)	Genetic map(cM)
c01:0038000 (CNE1)	1	38000	0.00
c01:0064000 (CDC24)	1	64000	10.36
c01:0078000 (FUN12)	1	78000	44.99
c01:0095000 (SAW1)	1	95000	59.36
c01:0114000 (ATS1)	1	114000	67.23
c01:0158000 (RFA1)	1	158000	82.96
c01:0170000 (ADE1)	1	170000	88.43
c01:0191000 (YAT1)	1	191000	119.08

Marker ID column includes the information of chromosome, base position and gene as the attribute names (column features) to combine with genotype data. Chromosome and Physical map column show the location assigned to reference genome S288C. Genetic map column stored the output for genetic map computing by R/qtl.

Table 2-3 Summary statistics of marker numbers and average marker density in each chromosome of F1 segregants genotype data.

Chromosome	Total length (bp)	Marker numbers	Marker density (bp / marker number)
I	230218	8	28777
II	813184	10	81318
III	316620	10	31662
IV	1531933	22	69633
V	576874	8	72109
VI	270161	8	33770
VII	1090940	14	77924
VIII	562643	9	62516
IX	439888	7	62841
X	745751	11	67796
XI	666816	11	60620
XII	1078177	15	71878
XIII	924431	11	84039
XIV	784333	9	87148
XV	1091291	14	77949
XVI	948066	12	79006

Total length column was calculated through R environment with the reference genome fasta file of S288C in version R64-2-1 as input. Marker numbers column was counted through R environment with the genotype data of F1 segregants. Marker density column was calculated as the ratio between total length and marker numbers.

2.2.4 Dataset Preparation

After collecting phenotyping data and genotyping data, the combined database eventually kept the genotype dataset includes 180 attributes (1 identifier of segregant and 179 markers) and 4 replicates each of the 96 haploid yeast strains (number of individuals) for each cross dataset. The first attribute is the identifier of strain names which stored the information of founder cross, generation and the strain number. For example, AE01fc101 is the identifier of the individual who is the progeny of A and E in first generation labelling number is 101. Other attributes are the 179 markers located in 16 chromosomes at different position and contains founder allele genotype which were encoded as 'A', 'E', 'S' and 'W' which based on the raw genotype bases in founder sequence which were extracted using makegeno function in r/shmootl (see below). All the records are fully genotyped without any missing values. The phenotype data were matched by the identifier of strain names. The marker information data were matched by the maker identifier. Table 2-4 describes the example of the datasets for AE cross. Full input datasets for all crosses were available to access through Rdrive.

Table 2-4 Sample input datasets of AE cross population.

Treatment Ratio	ID	c01:0038000 (CNE1)	c01:0064000 (CDC24)	c01:0078000 (FUN12)	c01:0095000 (SAW1)	c01:0114000 (ATS1)	c01:0158000 (RFA1)
		1	1	1	1	1	1
		0.00	10.36	44.99	59.36	67.23	82.96
0.955339	AE01fc101	E	E	A	A	A	A
0.965771	AE01fc102	A	A	A	A	A	A
0.955144	AE01fc103	A	A	E	E	E	E
1.01001	AE01fc104	E	E	E	E	E	E
0.997644	AE01fc105	E	E	E	E	E	E
0.953489	AE01fc106	A	A	E	E	E	E
0.984708	AE01fc107	E	E	A	A	A	A
0.998196	AE01fc108	A	A	A	A	A	A
1.04751	AE01fc109	A	A	A	A	A	A
0.993652	AE01fc110	E	A	A	E	E	E
…	…	…	…	…	…	…	…

ID columns stored the labels for each strain with the cross information, generation information, and strain number. ID were kept consistence in phenotype data for matching each individual phenotype data to genotype information in row. Each marker column records the genotype value of founder allele. 'A' in this example datasets means kept same genotype base with North America (NA) sequence call and 'E' means kept the same genotype base with Wine/European (WE) sequence call. Row 2 and 3 are the chromosome number and genetic map by matching marker id column to marker information table.

2.2.5 Statistical Analysis

Descriptive statistics of phenotype attributes were performed in R environment through IDE Rstudio. Violin plots were generated for comparing growth features in R with ggplot2 package (Wickham, 2016).

QTL analysis were performed with shmootl package in R environment developed by Yue Hu and Dr. Thomas Walsh for yeast QTL analysis through run_scanone pipeline (package available through https://github.com/gact/shmootl). The run_scanone pipeline was performed for each cross to find the additive QTLs by using interval mapping. The linkage LOD scores were used to calculate the significance of each marker on all chromosomes, calculated as the log10 likelihood ratio comparing the null hypothesis and alternative hypothesis that a QTL exists at that locus (Broman, et al., 2003). A LOD score of 3 or higher is generally indicated as significance. Significance level alpha was set as 0.05 to estimate the LOD threshold. 1000 permutation tests were applied for shuffling the phenotypes and use the 95^{th} percentile of the maximum LOD score in each permutation test as the LOD threshold for QTL scan. Step size for genotype probabilities was set as 1. The significant markers are identified as QTL by the LOD threshold for each experiment. LOD interval estimation are located by 1.5 LOD drop for each QTL as support region to find candidate genes (Figure 2-2). The run_scantwo was then performed to find interactions among markers, i.e. epistasis. The QTL analysis process and output were stored in hdf5 file format for each run. Figure report was stored as a pdf file and a table report was stored as an xlsx file. All the QTL analyses were performed on ALICE, a High Performance Computing (HPC) cluster under CentOS Linux operating system. R Scripts and pipeline qsub were supplied in Appendix 1. All the QTL analysis records are accessible through the GACT research Rdrive on the University of Leicester.

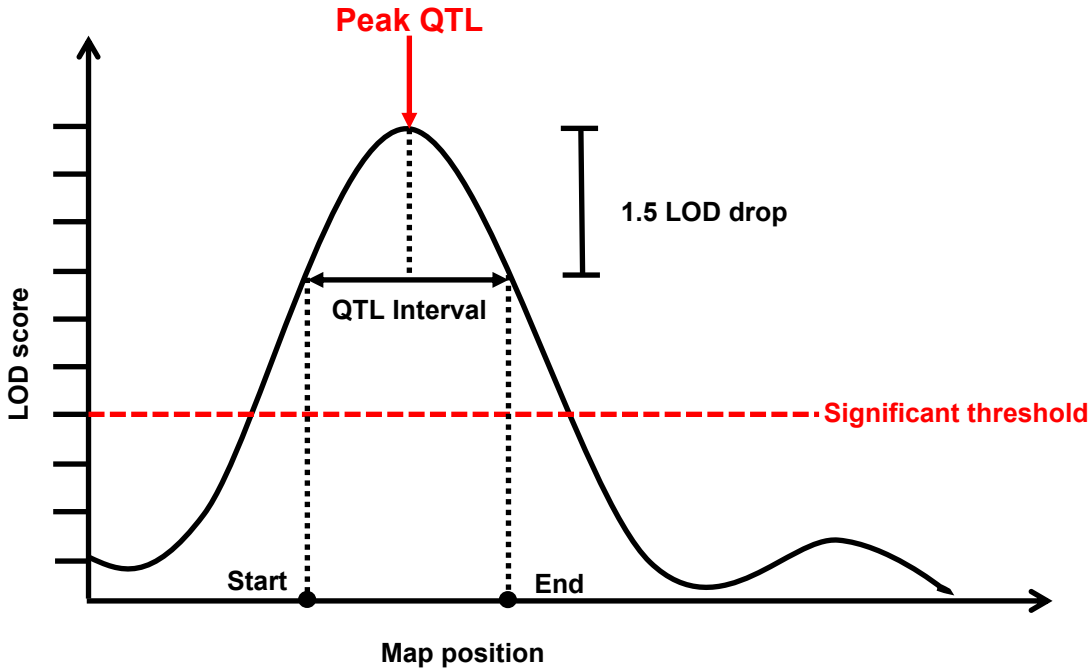


Figure 2-2 Illustration of QTL Interval

The red dash line marks the 5% significance threshold of LOD score based on the permutation. Peak QTL is the highest significant LOD score. QTL Interval was estimated the QTL range given by the chromosomal position corresponding to the highest significant LOD score with 1.5 drop to locate the start and end of the interval.

2.3 Results

2.3.1 PHENOS output growth curves of F1 yeast arrays

F1 haploid yeast segregants were arranged on 384 arrays on YPD media as the control and Doxorubicin YPD media as the treatment. 6 groups of progeny with bi-parental lines for the four founders combination which are AS, AE, AW, SW, ES, and EW. 12 plates of experimental array in total were incubated for 65 hours in a plate reader with measurements of growth data recorded by OD value for each segregant. The growth data were stored and normalised by PHENOS to illustrate growth curves for each plate. Figure 2-3 and Figure 2-4 illustrate the growth curves of the control with standard YPD conditions and treatment with YPD and Doxorubicin for the 6 groups of F1 segregants. The F1 yeast segregants under different parental crosses differ both in standard growth and in responses to Doxorubicin treatment. The growth curves under YPD condition as the control plates showed similar growth pattern within parental crosses. All 6 control arrays reached to the stationary phased around the same time period (20 hours). Variation is seen among progeny generated by different parental crosses. In contrast, a significant difference of growth pattern could be identified from treatment growth curves for each group. In order to get a better understanding of these segregants' growth differences, descriptive statistics analysis was performed in the next section for evaluating growth and phenotype features.

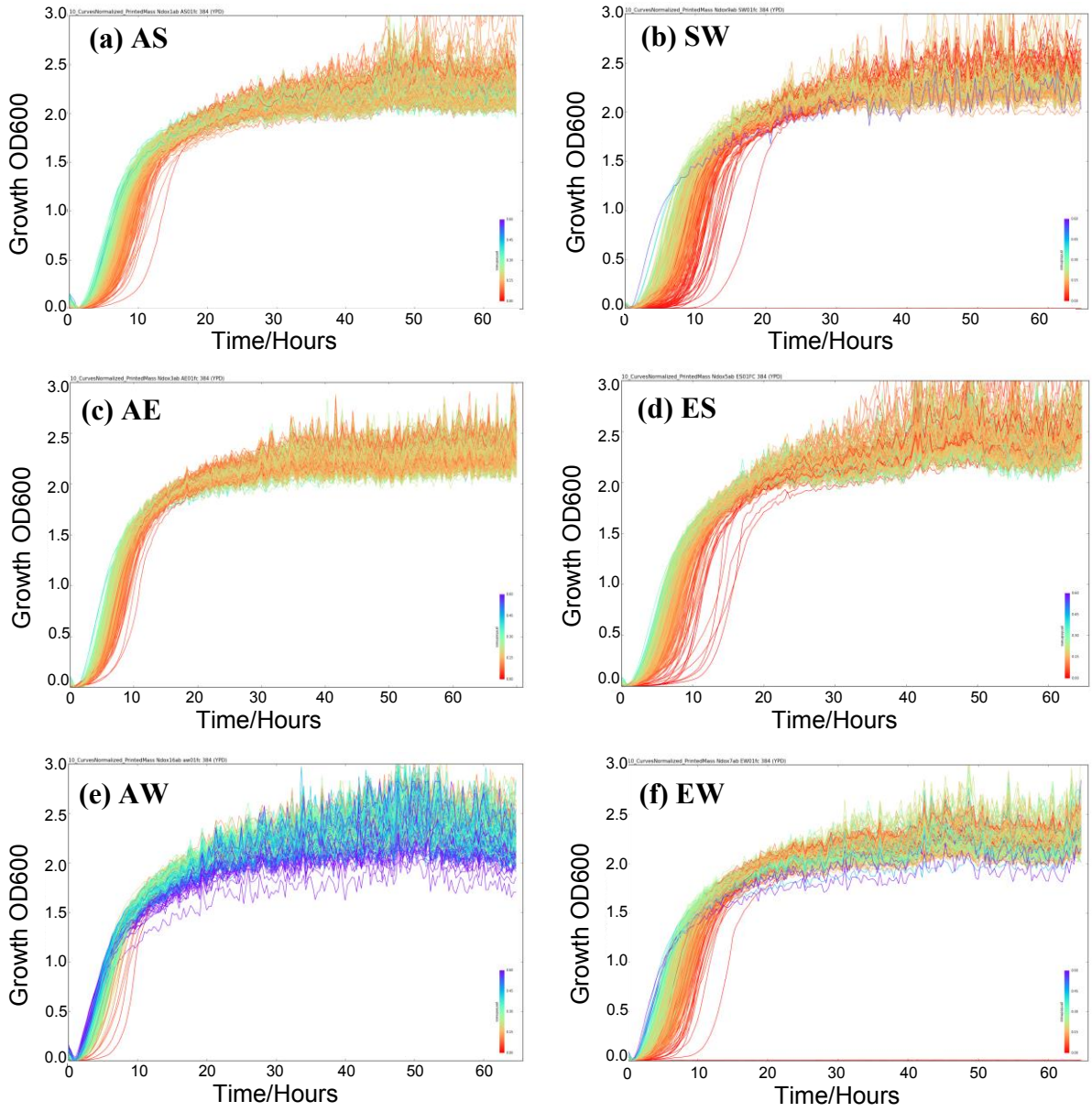


Figure 2-3 F1 yeast control arrays: PHENOS growth curves.

Each figure represented growth curves for individuals under the identical condition on YPD for each cross. Each growth curve illustrates the growth condition with measurement value changes from minimum (Y axis) over 65 hours of growth in the plate-reader. The curves are printed without agar absorbance. The curve colour is based on the initial printed mass on the array with rainbow theme from red to purple indicating small to large printed mass. Three strains (sw01fc524, ew01fc411, ew01fc412) which were not recorded with value in SW and EW that might due to bad printing quality or edge effects were filtered out from the analysis.

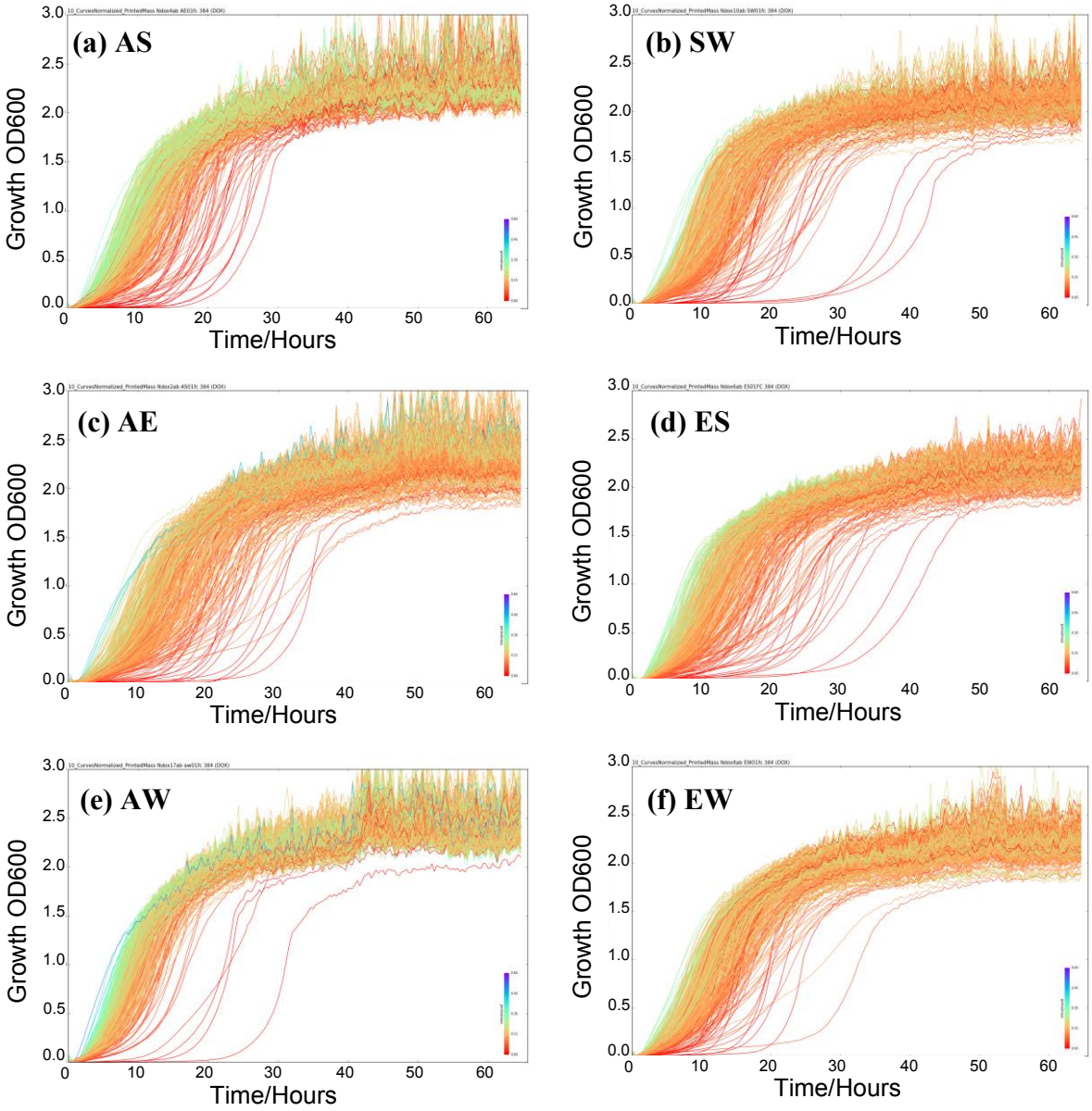


Figure 2-4 F1 yeast treatment arrays: PHENOS growth curves.

Each figure represented growth curves for individuals under the condition of YPD with Doxorubicin for each cross. Each growth curve illustrates the growth condition with measurement value changes from minimum (Y axis) over 65 hours growing in plate-reader. The curves were printed without agar absorbance. The curve colour were based on the initial printed mass on the array with rainbow theme from red to purple indicating small to large printed mass.

2.3.2 Comparative Analysis of growth features for F1 segregants

In order to view the segregants' growth phases of each F1 cross with different parental lines in detail, several growth features under standard YPD as control and treatment condition with YPD and Doxorubicin were compared in this section. Under the standard YPD condition for yeast growth, the lag phase of all the segregants lasts approximately 4 hours on average. The lengths of each group's lag phase duration are all relatively short which is expected under the conditions for normal growth. However, there are different responses among groups. The average lag period on segregants growth in the AW group is 2.36 hours, which is about 3 hours shorter than segregants in the SW (average lag is 5.44 hours) (shown in Figure 2-5, labelling with AW and SW). There are three groups of F1 yeast segregants having the North America strain as one of the parental lines, AE, AS, and AW. The average lag phase of these three were 3.48 hours which is shorter than the average duration of the other three groups. The range of lag is smaller as well in these three meaning the distribution was dense around average, that is the time to enter the exponential growth phase is overall short. Under the treatment with YPD and DOX, the average lag phase of all segregants is approximately equal 6.50 hours, 2.50 hours longer than the control condition. The lag phase of all the F1 segregants was on a longer duration than the lag phase under the control condition on average by group, and exhibited various responses within the group (Figure 2-5). Segregants in AW group still hold the shortest average lag (4.98 hours). Segregants in AS group have the longest average lag phase (8.29 hours). This is interesting given the NA containing crosses all having the shortest lag phase in their controls. The average lag phase of F1 segregants in AS, SW, and ES group is 7.50 hours, which is longer than the other three groups without Sake strains as the parental lines. These results showed the difference in the initial adaption responses to DOX among F1 segregants of these 6 crosses.

In addition to the assessment of the Lag phase, the speed during exponential phase (Max Slope) is also an important factor that affects the growth of segregants. Variation in each cross group was seen for max slope features. Under standard YPD control conditions, the average max slope for all F1 segregants is 0.27 OD change per hour, and the average difference of Max Slope for these six cross groups was very small (0.26 - 0.28). This satisfies the assumptions of the experimental design, because the growth rates should be stable and similar under the conditions suitable for growth. Under DOX treatment conditions, the average speed during exponential phase of all segregants is 0.16 OD change per hour, which is 0.1 slower than the speed under control condition. A study confirmed that the toxic effect of DOX can inhibit the growth of yeast (Buschini, et al., 2003). The average Max Slope of the segregants in AW group is 0.20 OD change per hour, which is the fastest growth rate among this six groups. The average Max Slope of the segregants in EW group is approximately equal 0.15 which inhibited the most comparing to the speed of the segregants under control. These results indicate that the growth rate of segregants is affected by DOX. Treatment Ratio, as a consideration for the final growth of the phenotype, also shows differences between groups. The median of the treatment ratio for F1 segregants in AE, AS, and AW groups was slightly greater than 1.00, while the other three groups were approximately 0.97. This result suggests that 50% of segregants with North America strain as one of the parental lines grow more under the treatment than control at the end point, that is, they were more resistant to DOX. The segregants in ES group had the ratio at 0.92 on overage and more than 75% of segregants grew less than under control, that is, they were more sensitive to DOX. Overall, these results demonstrate that the growth of these segregants was affected by DOX by different degrees.

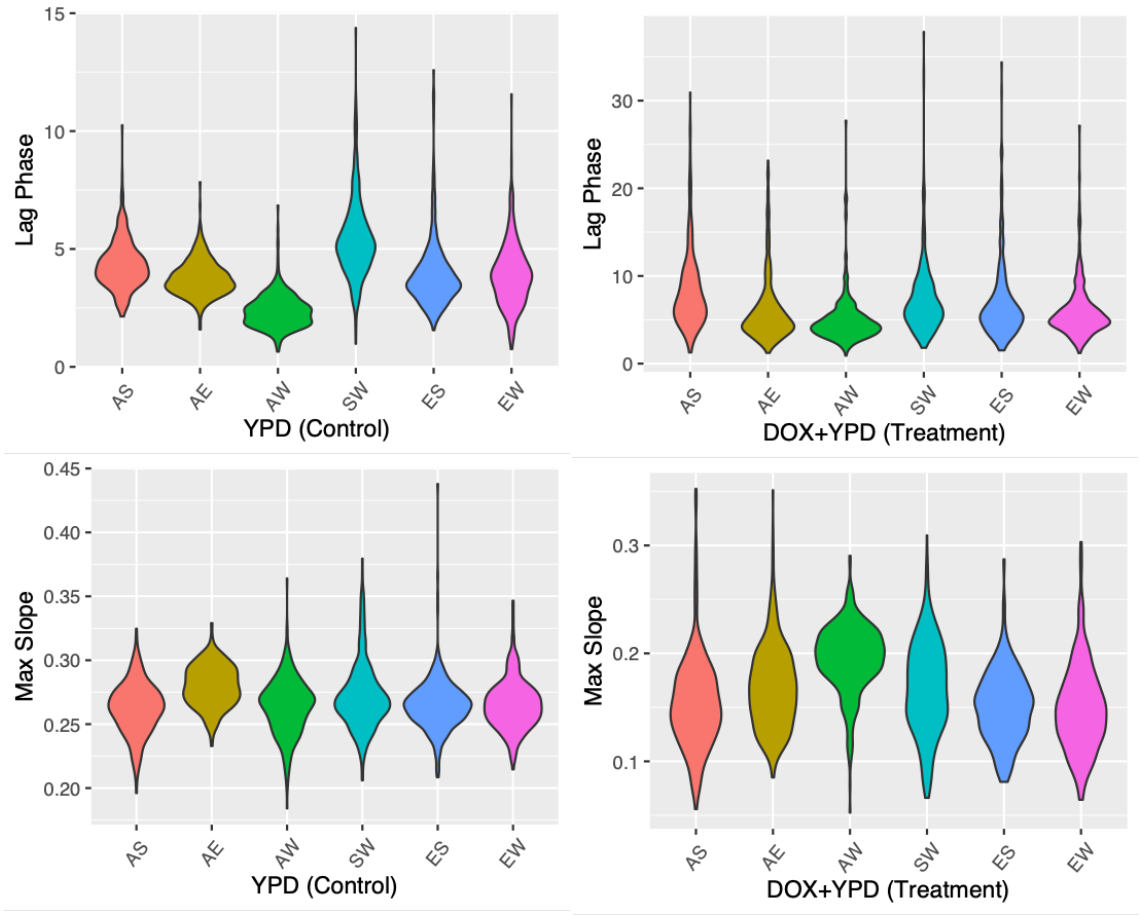


Figure 2-5 Growth feature distributions of F1 segregants' growth.

Violin plots illustrate different distribution of Lag phase and Max slope (growth speed) among 6 crosses for both Control and Treatment. X axis is AS, AE, AW, SW, ES, EW. Y axis in Lag plots is Lag phase time by hours. Y axis for speed plots is Max slope for the growth curve. Shape in red is AS, brown is AE, green is AW, mint is SW, blue is ES and purple is EW. Violin plots generated through ggplot2 package in R environment.

`

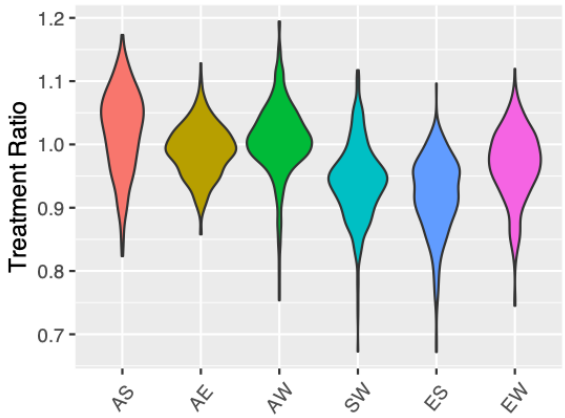
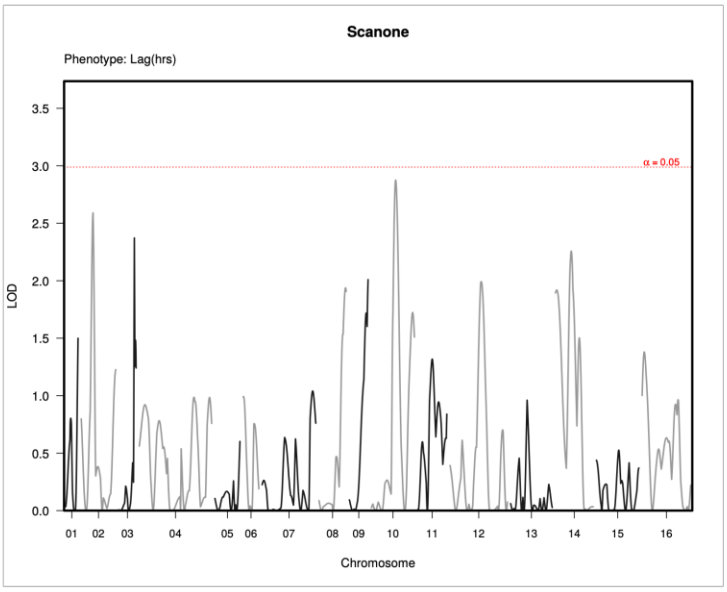


Figure 2-6 Phenotype feature Treatment Ratio distribution of F1 segregants.

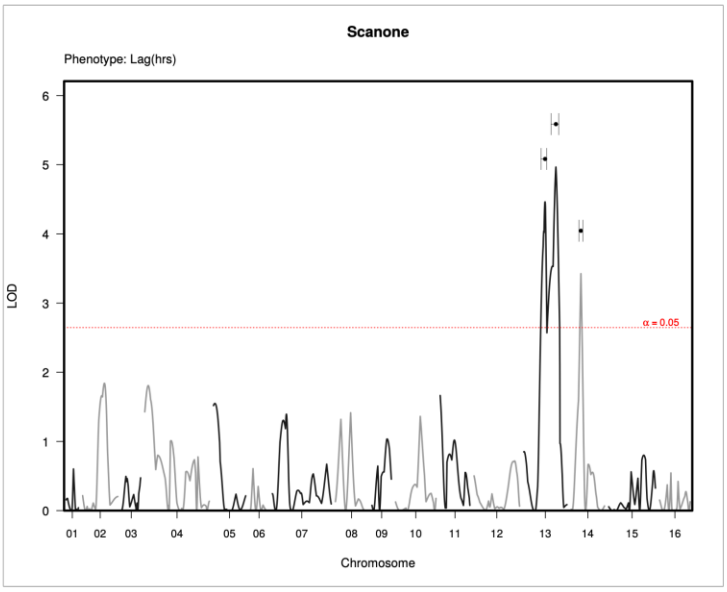
Violin plot illustrate different distribution of Treatment Ratio among 6 crosses which compare the final growth of control and treatment. Same labelling and colouring as Figure 2.3.3. Y axis is Treatment Ratio.

2.3.3 Single QTL Analysis

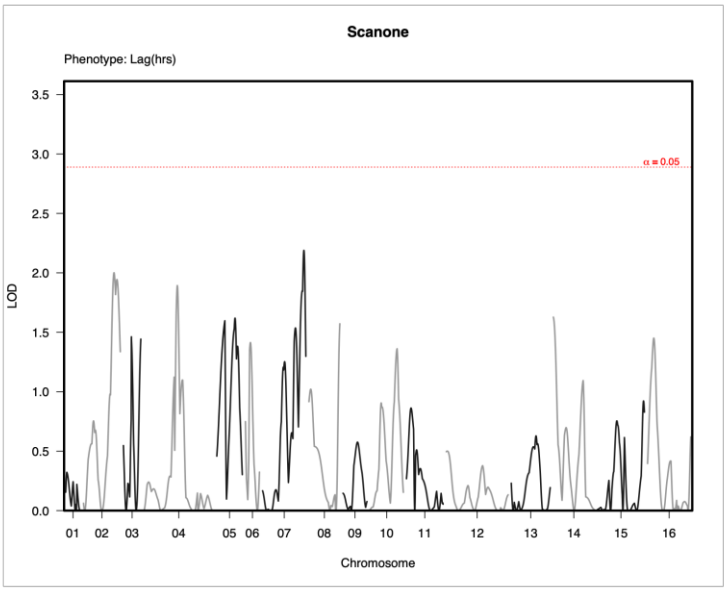
Additive QTL analysis was performed on each of the three growth phenotypes for each group of F1 segregants by interval mapping methods as the markers span large ranges. The LOD scores were computed for each marker and all the markers were mapped with LOD scores through whole genome. The results of the QTL mapping were plotted for all markers with LOD scores by genome-wide and the interval mapping of the chromosome with peak LOD. The significance level is set at 0.05. The LOD threshold obtained by 1000 permutation tests which is indicated by a red dashed line in each plot. Marker regions that exceed the threshold are marked with interval if the size is visible through Rplot. However, due to a bug in R, there are a few intervals that cannot be plotted. See Figure 2-7 to Figure 2-9 for each F1 cross QTL mapping and for each phenotype by whole genome.


Scanone
Phenotype: Lag(hrs)
LOD
Chromosome
α = 0.05
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16

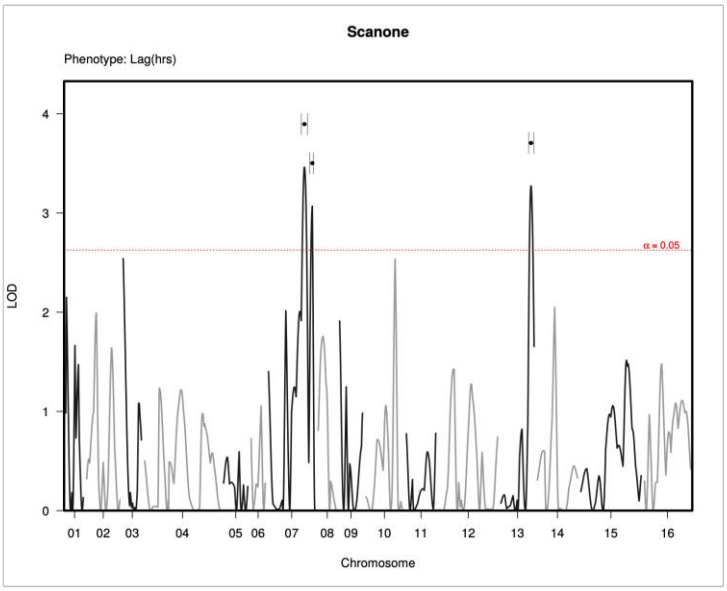
(a) AS


Scanone
Phenotype: Lag(hrs)
LOD
Chromosome
α = 0.05
0
1
2
3
4
5
6
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16

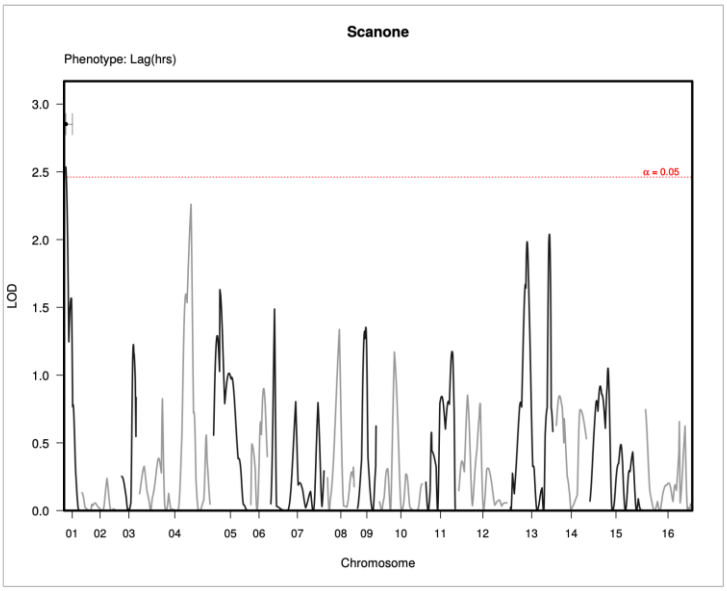
(b) SW


Scanone
Phenotype: Lag(hrs)
LOD
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
α = 0.05
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16
Chromosome

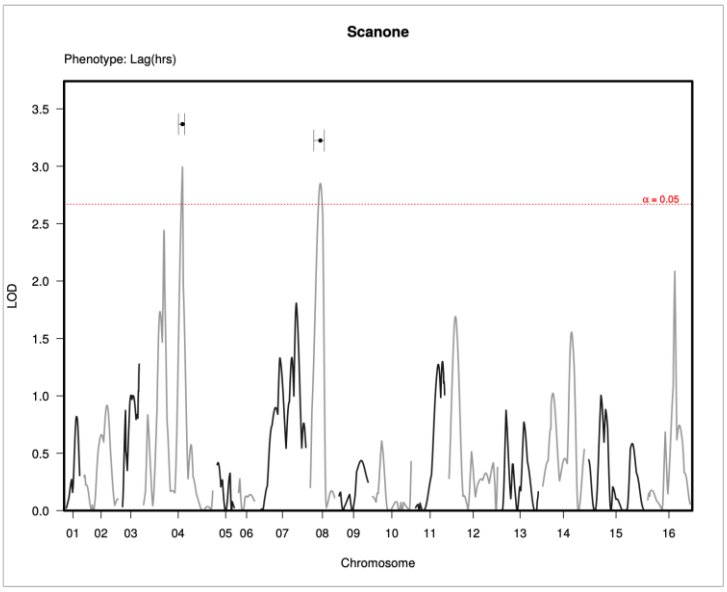
(c) AE


Scanone
Phenotype: Lag(hrs)
LOD
4
3
2
1
0
α = 0.05
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16
Chromosome

(d) ES



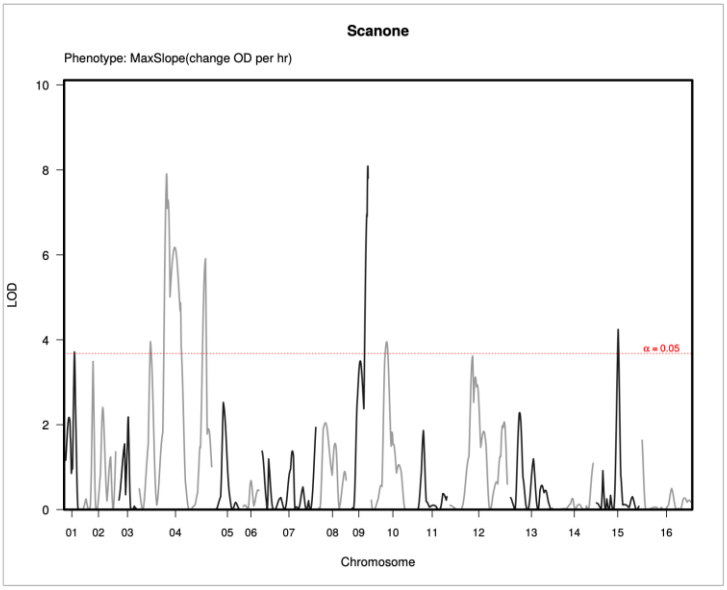
(e) AW



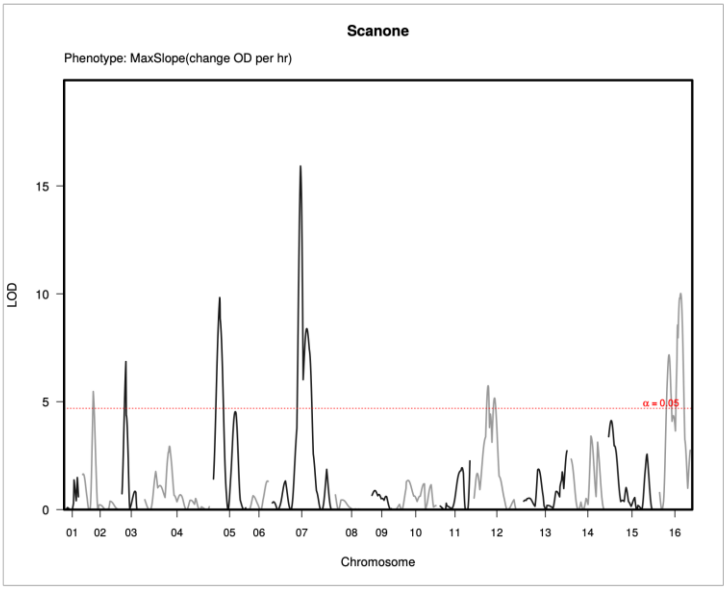
(f) EW

Figure 2-7 Illustration of QTL analysis of Lag phase for all 6 F1 crosses in DOX

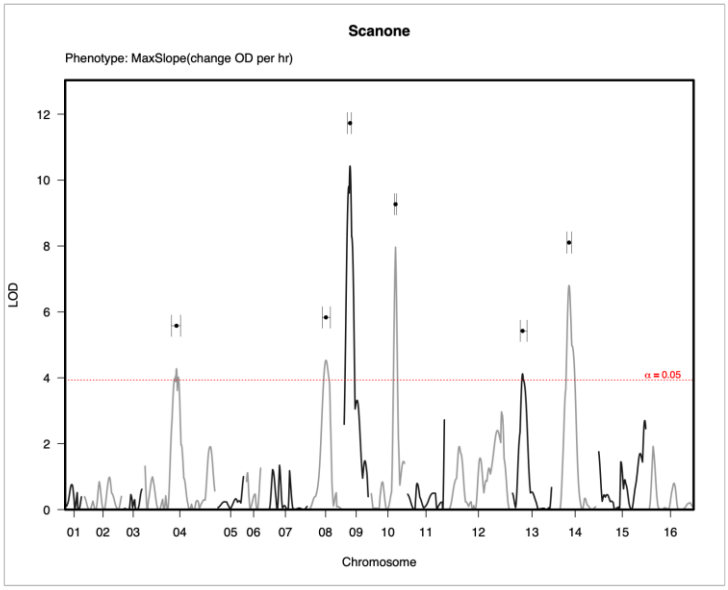
Genome-wide QTLs for Lag phase of each F1 cross among 16 chromosomes. The x-axis represents the chromosome number and the y-axis represents the LOD score. Single QTL LOD scores were illustrated for each chromosome with black coloured on odd number chromosomes and grey coloured on even number chromosomes. The red dashed line indicated the LOD threshold determined at the significance level at 0.05. The interval mapping of the peak QTL regions which over the threshold were plotted in range where the peak positions were highlighted with black dot if visible. (a) cross NA × SA (b) cross SA × WA (c) cross NA × WE (d) cross WE × SA (e) cross NA × WA (f) cross WE × WA


Scanone
Phenotype: MaxSlope(change OD per hr)
LOD
Chromosome
α = 0.05
0
2
4
6
8
10
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16

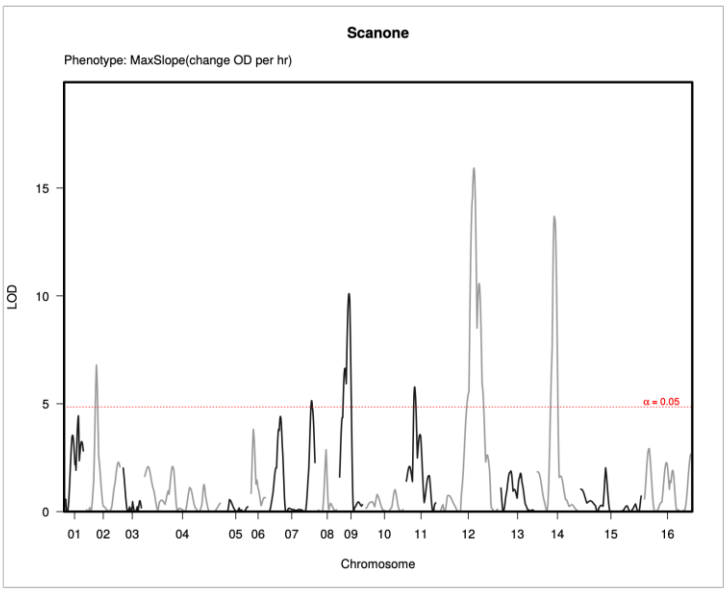
(a) AS


Scanone
Phenotype: MaxSlope(change OD per hr)
LOD
Chromosome
α = 0.05
0
5
10
15
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16

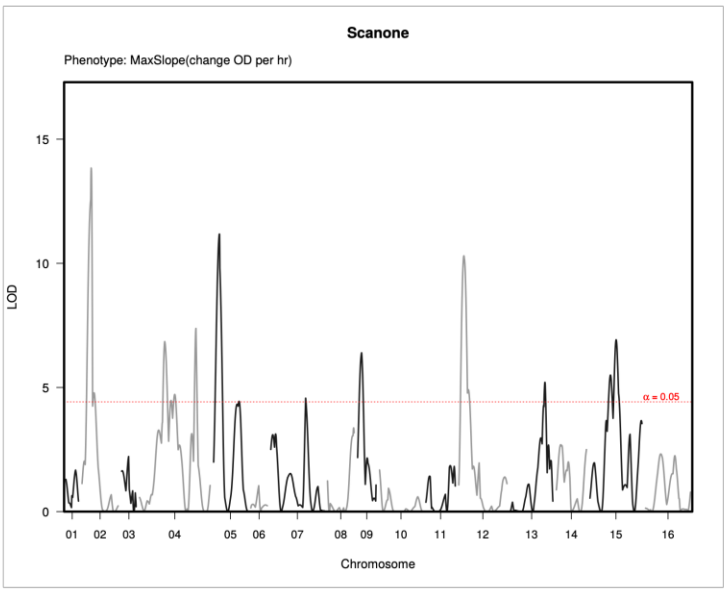
(b) SW


Scanone
Phenotype: MaxSlope(change OD per hr)
LOD
12
10
8
6
4
2
0
α = 0.05
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16
Chromosome

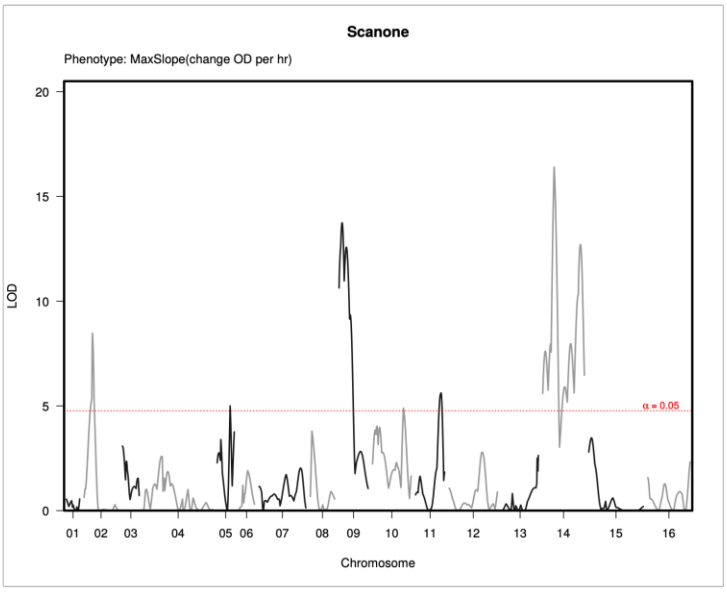
(c) AE


Scanone
Phenotype: MaxSlope(change OD per hr)
LOD
15
10
5
0
α = 0.05
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16
Chromosome

(d) ES



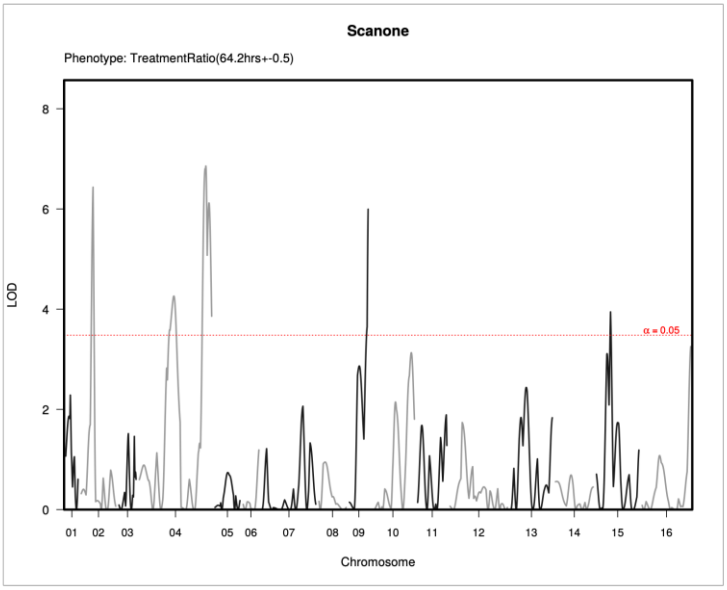
(e) AW



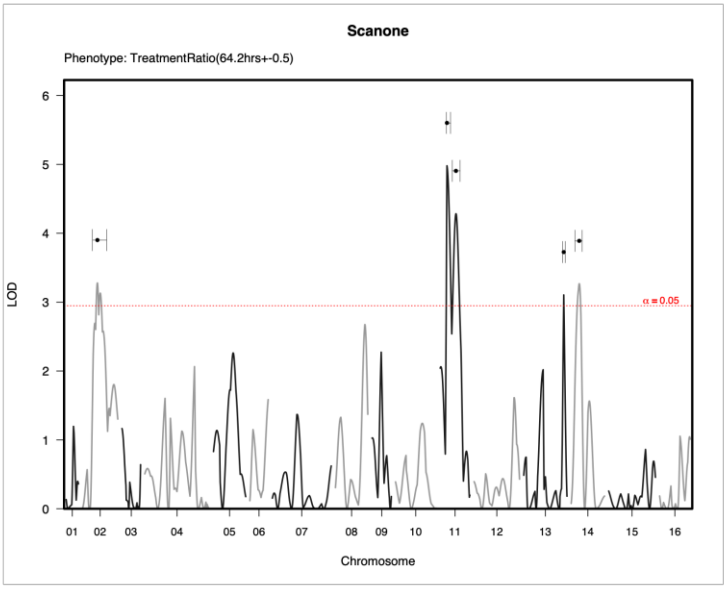
(f) EW

Figure 2-8 Illustration of QTL analysis of phenotype in max slope for all 6 F1 crosses in DOX

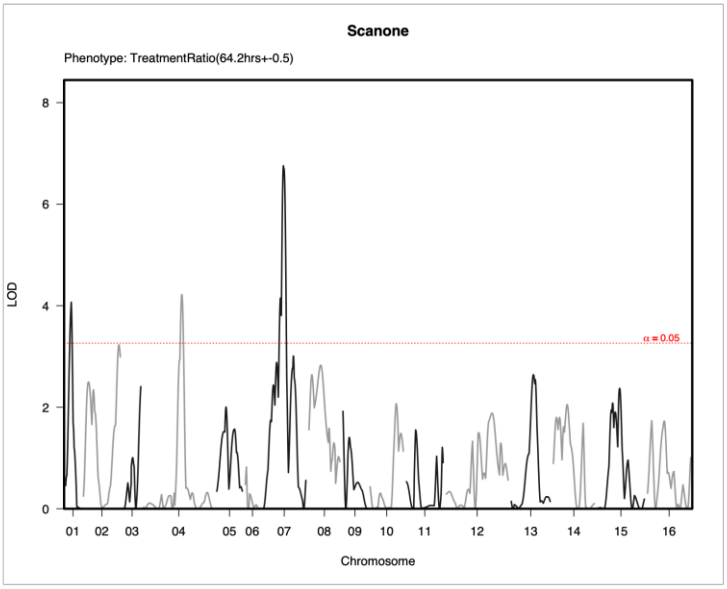
Genome-wide QTLs for max slope of each F1 cross among 16 chromosomes. The x-axis represents the chromosome number and the y-axis represents the LOD score. Single QTL LOD scores were illustrated for each chromosome with black coloured on odd number chromosomes and grey coloured on even number chromosomes. The red dashed line indicated the LOD threshold based on the significance level at 0.05. The interval mapping of the peak QTL regions which over the threshold were plotted in range where the peak positions were highlighted with black dot if visible. (a) cross NA × SA (b) cross SA × WA (c) cross NA × WE (d) cross WE × SA (e) cross NA × WA (f) cross WE × WA


Scanone
Phenotype: TreatmentRatio(64.2hrs+-0.5)
LOD
Chromosome
α = 0.05
0
2
4
6
8
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16

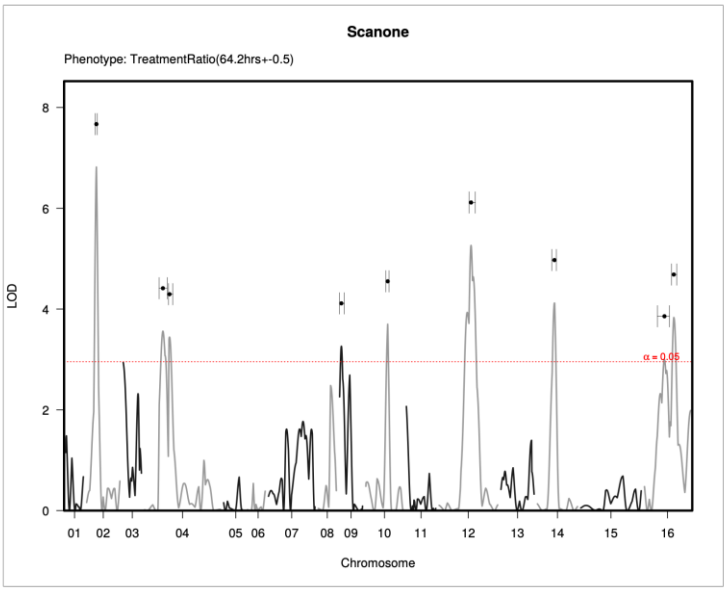
(a) AS


Scanone
Phenotype: TreatmentRatio(64.2hrs+-0.5)
LOD
Chromosome
α = 0.05
0
1
2
3
4
5
6
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16

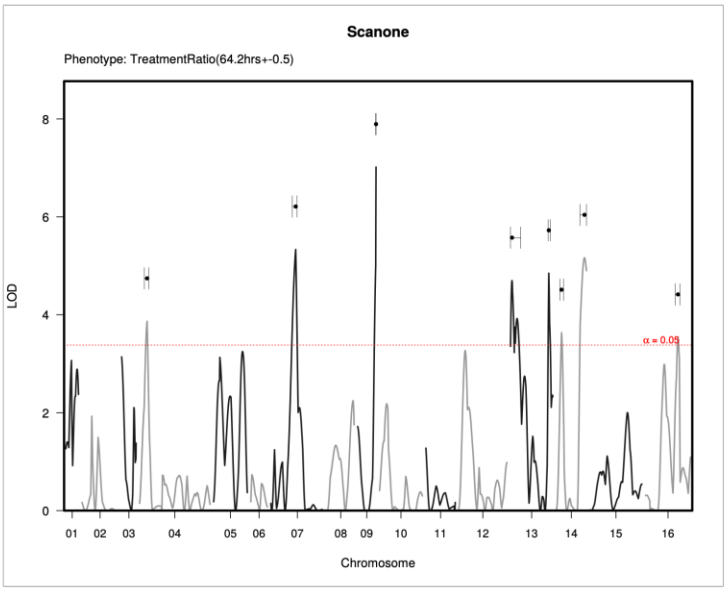
(b) SW


Scanone
Phenotype: TreatmentRatio(64.2hrs+-0.5)
LOD
Chromosome
α = 0.05

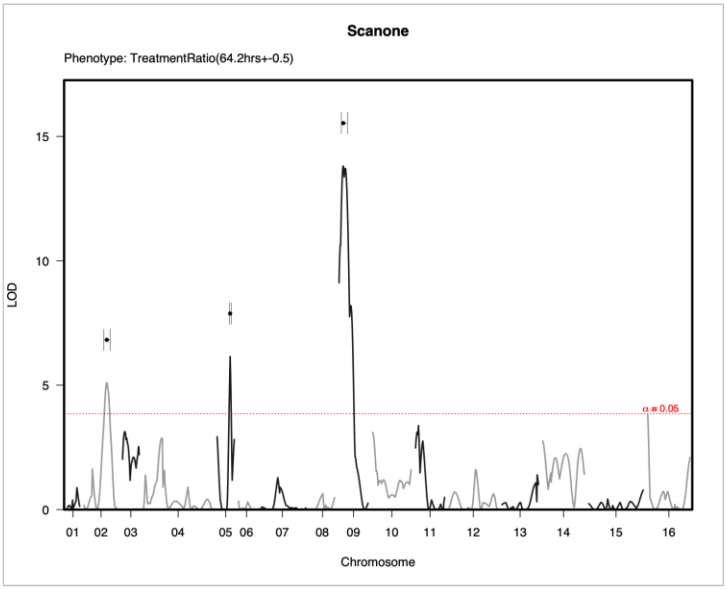
(c)AE


Scanone
Phenotype: TreatmentRatio(64.2hrs+-0.5)
LOD
Chromosome
α = 0.05

(d) ES



(e)AW



(f) EW

Figure 2-9 Illustration of QTL analysis of growth rate for all 6 F1 crosses in DOX

Genome-wide QTLs for Lag phase of each F1 cross among 16 chromosomes. The x-axis represents the chromosome number and the y-axis represents the LOD score. Single QTL LOD scores were illustrated for each chromosome with black coloured on odd number chromosomes and grey coloured on even number chromosomes. The red dashed line indicated the LOD threshold based on the significance level at 0.05. The interval mapping of the peak QTL regions which over the threshold were plotted in range where the peak positions were highlighted with black dot if visible. (a) cross NA × SA (b) cross SA × WA (c) cross NA × WE (d) cross WE × SA (e) cross NA × WA (f) cross WE × WA

2.3.4 Identification of QTL intervals shown strain dependent overlaps among crosses

There are different subsets of markers showing evidence of a candidate QTL. A total of 89 significant QTL intervals for the growth phenotype features were identified for all 6 crosses. Table 2-5 lists the summary of QTLs present for each cross obtained from QTL analysis. Among them, 9 loci were identified through QTL mapping that controlling the time of leaving lag phase and started growing under DOX conditions. Hardly any QTL of significance were found for the groups with same founder strain NA. There are no QTLs mapped for NA cross with SA or WE. Only 1 QTL which locate at chromosome I was identified in cross between NA and WA. 7 QTLs were present in the cross combination involving founder strain West African (WA). 2 QTLs were present in the cross combination between SA and WE. Two crosses SW and ES which both containing the SA strain showed a major QTL overlap in a region of 14kb on chromosome XIII 802000-815789 with highest LOD scores in each cross (SW:4.79; ES: 3.27). Figure 2-10 shows the gene features included in this overlap region. This result suggested that a Sake variant (allele) has a strong effect on the leaving lag phase.

Table 2-5 Summary of QTLs present for each cross obtained from QTL analysis.

None of the markers were over the threshold for Lag phenotype of AS and AE (displayed as '-'). Total QTL numbers for each phenotype feature and cross are summarised from QTL analysis output. The peak LOD score is displayed as the maximum score for each QTL run of the phenotype with the cross. The chromosome and position are the location of the Peak LOD for each run.

Cross	Phenotype	Number of QTLs	LOD threshold	Peak LOD	Chr	Position	Region Size
AS	Lag	0	2.99	-	-	-	-
AS	Max Slope	7	3.68	8.09	IX	371394	53249
AS	TR Ratio	5	3.48	6.86	IV	1345823	124809
AE	Lag	0	2.89	-	-	-	-
AE	Max Slope	6	3.94	5.38	V	88000	66528
AE	TR Ratio	3	3.26	6.76	VII	501606	120497
AW	Lag	1	2.46	2.53	I	38000	72111
AW	Max Slope	12	4.42	13.84	II	235927	164633
AW	TR Ratio	8	3.38	7.02	IX	380000	27995
SW	Lag	3	2.76	4.79	XIII	802000	104839
SW	Max Slope	9	4.46	15.98	VII	515000	57021
SW	TR Ratio	5	2.95	4.97	XI	173000	93544
ES	Lag	3	2.63	3.46	VII	834291	133811
ES	Max Slope	6	4.85	15.92	XII	637472	69113
ES	TR Ratio	9	2.95	6.82	II	308000	80863
EW	Lag	2	2.67	2.99	IV	902000	169372
EW	Max Slope	7	4.77	16.4	XIV	281000	43083
EW	TR Ratio	3	3.85	13.81	IX	115406	82773

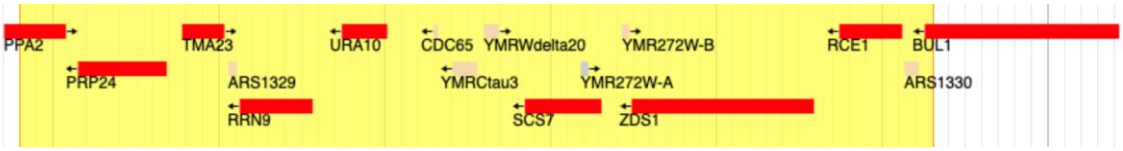


Figure 2-10 Annotated sequence features in QTL overlap region.

Screenshot of genes located on chromosome XIII 802000 to 815789 from JBbrowser based on the annotation of reference sequence S288C.

As the growth speed varies among populations under DOX conditions, there are 47 QTL intervals identified for Max slope phenotype across genome. Two thirds (60%) of the QTLs for max slope phenotype were identified in cross combinations containing WA. There are 12 QTL intervals identified over the LOD threshold in cross AW, with the largest amount of marker hits. 9 markers were identified as significance in the SW cross and 7 QTLs were mapped in the EW cross. In addition, there are 7 QTLs mapped for AS cross, 6 QTLs for AE cross and 6 QTLs for ES cross.

Nearly half (47%) of the QTL intervals with close QTL peaks overlapped across different crosses containing the same founder strain and all the six crosses have overlaps with other crosses' QTL interval. The same markers were mapped with the peak LOD score between cross SW and ES located at chromosome II: 308000. There are 4 crosses (AW, SW, ES, EW) with a QTL interval overlap of 15kb region between 283757 and 299053 on chromosome II. Figure 2-11 shows the gene features included in this overlap region.

Figure 2-11 Annotated sequence features in QTL overlap region.

Screenshot of genes located on chromosome II from 283757 to 299053 highlighted with yellow colour from JBbrowser based on the annotation of reference sequence S288C.

In addition, another QTL interval overlapped among three crosses (AW, SW, EW) that involve strain WA are located at chromosome V between 379703 and 424637. The QTL intervals of AW and SW located a second large overlap on chromosome V between 71511 and 164667. The three crosses (AS, AE, AW), which have founder strain NA, shared an overlap on chromosome IV between 493958 and 592378. AS and AW located another overlap on chromosome XV between 574461 and 594775. Interestingly, an overlap QTL interval among AE, AW, EW, the three crosses without founder Sake strain, are narrowed at a 35kb

region on chromosome IX between 60422 to 96127 (displayed on Figure 2-12). However, there is one overlap between SW and ES which have common founder Sake strain located at chromosome XII between 598839 and 656174.

Although there is not a large variation among treatment ratio as much as max slope under DOX condition, 33 QTL intervals were identified across the genome. 9 QTLs for treatment ratio phenotype were detected in cross combination ES and 8 QTLs in AW which are the largest two significant marker sets. There are 5 QTL intervals each identified in cross combination of AS and SW. In addition, three markers were identified as significance in AE cross and three in EW. 22 QTL intervals overlapped in more than one combination. Three of the overlaps were shared among AS, AE and AW. AE and AW shared an overlap on chromosome VII at a 20kb region from 454669 to 475473. AS and AW have two overlaps on chromosome IV at a large region from 1296282 to 1421091 and on chromosome IX span from 354621 to 380000 where the peak LOD scores were shown on the same markers which located at end of both intervals 380000. There is another overlap located on chromosome IX which was shared by ES and EW at the region between 83365 to 101606. Interestingly, there is an overlap among AS, AE and ES which were the crosses without founder strain WE narrowed at chromosome IV from 596537 to 610102. Conversely, 2 overlaps were found between AW and SW who contain the founder strain WE locate on chromosome XIII from 853803 to 870948 and chromosome XIV from 99605 to 192445. One overlap shared between SW and EW on chromosome II from 474622 to 621830. Another overlap located on chromosome II from 260402 to 325763 which were between AS and ES contain same founder strain Sake.

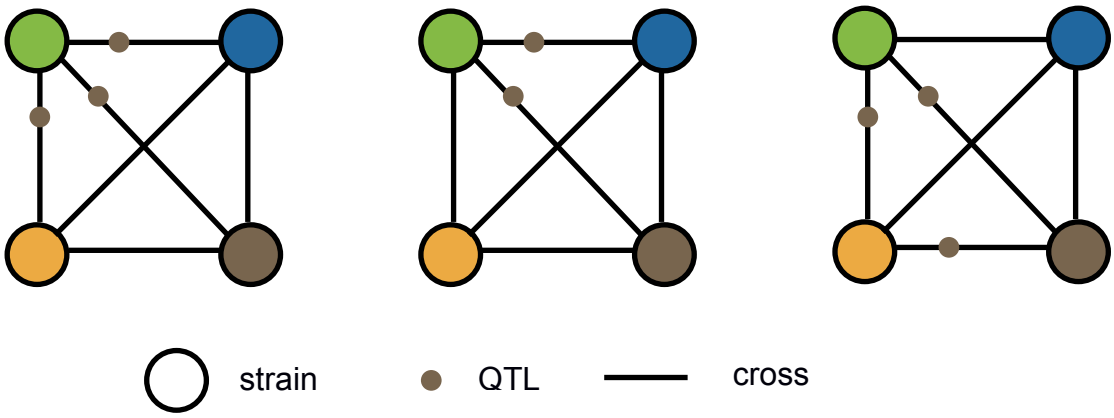


Figure 2-12 Overlap types among 6 crosses of 4 parental lines.

Circles indicated the founder strain that different colour means different founder. lines indicated the cross combinations and the brown dots indicated the QTL intervals for the cross. Apart from overlaps located for same founder (2 or 3 dots with same founder), overlaps shared for the crosses who without specific founder (3 dots in triangle region) were also present in this QTL sets. Adapted from (Cubillos, et al., 2011).

2.3.5 Unique QTL intervals detected for each cross

Although a large proportion of QTL intervals were involved in overlaps, half (45 of 89) of the QTL intervals were only identified on one cross. When looking at all the highest QTL scores for each analysis, 13 of 16 QTL intervals are detected with the large effect for the specific F1 cross combination. Table 2-6 (a)-(f) lists of genes present within the unique QTLs obtained from QTL mapping result for each phenotype and cross.

Table 2-6 QTL mapping results

Phenotype	CHR	LOD score	Interval Length (bp)	Start (bp)	Peak (bp)	End (bp)	Peak Gene Feature
MS	I	3.71	43086	129995	158000	173081	RFA1
MS	IV	3.95	100798	125816	189000	226615	NOP14
MS	IV	5.91	72095	1294256	1333000	1366351	PPM1
MS	**IX**	**8.09**	**53249**	**326751**	**371394**	**380000**	**EGH1**
MS	X	3.95	76856	221508	267526	298364	SIP4
TR	XV	3.95	94743	377783	442000	472526	C000A2

(a) AS

Phenotype	CHR	LOD score	Interval Length (bp)	Start (bp)	Peak (bp)	End (bp)	Peak Gene Feature
Lag	XIII	4.79	104840	717007	802000	821846	PPA2
Lag	XIV	3.60	93416	140464	208000	233881	SIN4
MS	III	6.80	42595	82785	99000	125380	BUD3
MS	**VII**	**15.98**	**57021**	**496582**	**515000**	**553603**	**SNU71**
MS	XII	5.77	83801	304532	360998	388334	MDN1
MS	XVI	7.10	106100	195142	250383	301241	-
MS	XVI	9.93	60010	632694	666300	692704	SMK1
TR	XI	4.98	93544	141692	173000	235236	AVT3
TR	XI	4.28	56960	259488	286730	316448	SMY1

(b) SW

Phenotype	CHR	LOD score	Interval Length (bp)	Start (bp)	Peak (bp)	End (bp)	Peak Gene Feature
MS	VIII	4.53	60217	156599	177762	216816	PIH1
MS	**IX**	**10.42**	**66529**	**59950**	**92674**	**126479**	**TMA108**
MS	X	7.96	46182	528042	552000	574224	MNN14
MS	XIII	4.12	124459	222844	287554	347303	-
TR	I	3.71	20698	70324	78000	91022	FUN12

(c) AE

Phenotype	CHR	LOD score	Interval Length (bp)	Start (bp)	Peak (bp)	End (bp)	Peak Gene Feature
Lag	VII	3.46	133812	758332	834291	892143	YGR168C
Lag	VII	3.07	86164	943315	1009000	1029479	RAD2
MS	VII	5.14	95548	946768	991655	1042315	YGR250C
MS	**IX**	**10.10**	**50427**	**150449**	**179902**	**200876**	**FMC1**
TR	IV	3.56	165430	270677	340521	436107	-
TR	X	3.70	58349	357116	390000	415465	VPS53
TR	XII	5.26	134155	533797	577159	667953	CCC1
TR	XIV	4.12	123911	276120	366648	400031	-
TR	XVI	3.00	237605	303395	448484	541000	PDR12

(d) ES

Phenotype	CHR	LOD score	Interval Length (bp)	Start (bp)	Peak (bp)	End (bp)	Peak Gene Feature
Lag	I	2.54	72111	38000	38000	110111	CNE1
MS	IV	4.72	167887	629607	748615	797494	SWI5
MS	IV	7.38	35180	1158247	1170000	1193427	MRP1
MS	VII	4.57	53482	655122	674000	708605	VAS1
MS	IX	6.40	42482	53645	72000	96127	SLN1
MS	**XII**	**10.30**	**74764**	**87684**	**117527**	**162449**	**BPT1**
MS	XIII	5.21	67605	770360	821785	837965	FCP1
MS	XV	5.50	51706	450483	477074	502189	TGL5
TR	XIV	5.16	102535	617465	689710	720000	ABZ1

(c) AW

Phenotype	CHR	LOD score	Interval Length (bp)	Start (bp)	Peak (bp)	End (bp)	Peak Gene Feature
Lag	IV	2.99	169373	855950	902000	1025323	RAD9
Lag	VIII	2.85	101238	132987	180453	234224	BRL1
MS	**IX**	**13.74**	**63722**	**60422**	**95016**	**124144**	**-**
MS	X	4.89	55930	580835	612000	636765	TPC1
MS	XIV	12.70	59408	626149	657268	685557	ACC1
TR	V	6.16	46744	379703	395000	426448	SPR6

(e) EW

2.3.6 Functional annotation among peak QTL genes

Functional clustering that annotate the genes associated with related biological process or molecular function was analysed through DAVID 6.8 with the peak genes. Interestingly, there are two major annotation clusters shown for these genes with P-value < 0.005 and False discovery rate (FDR) < 0.25. Gene Ontology (GO) terms for these two clusters are summarised in Table 2-7.

Table 2-7 Lists of GO terms with involved peak genes.

Nucleotide binding and ATP binding are from the first cluster and Metal ion biding are from the second cluster. Genes which shown on all columns are highlighted in bold. FDR value based on Benjamini adjustment.

GO Term	GO ID	Gene Count	P-value	FDR	Gene Features
Nucleotide binding	GO:0000166	20	0.002	0.14	FUN12, GAL1, MAK5, **SPF1, SLN1,** SMY1, SNQ2, VAS1, RIE1, BPT1, REA1, YEF3, POL2, **ACC1**, PDR12, MCM4, MSF1, SMK1
ATP binding	GO:0005524	17	0.003	0.15	GAL1, MAK5, **SPF1, SLN1,** SMY1, SNQ2, VAS1, BPT1, REA1, YEF3, **ACC1,** PDR12, MCM4, MSF1, SMK1
Metal ion binding	GO:0046872	19	0.002	0.25	FUN12, RFA1, **SPF1,** TMA108 **SLN1,** GAT4, GLT1, VMS1, SWI5, MRP1, MIG1, RAD2, SIP4, PPA2, POL2, **ACC1,** CAT5

Gene network analyses were further generated for the peak Genes through GeneMANIA database. Gene networks among these genes showed a complex relationship with functionally associated and protein-protein interactions. Three of the genes (SPF1, SLN1, AAC1) were shortlisted through functional clusters. Gene PMR1, CCC1, HFA1, UPS1, CCP1 were chosen as candidate genes by the linkages with SPF1, SLN1 and AAC1 of shared protein domain obtained from network Figure 2-14. Description of these genes were summarised in Table 2-8 through SGD database. Most of the candidate genes involved in metal from the SGD database. Most of the candidate genes are involved in metal transporter for Fe^{2+}/Ca^{2+}, Ca^{2+} homeostasis and mitochondria. A recent study suggested that DOX and DOX metabolites affect a major mechanism of cardiotoxicity with

mitocondria dysfunction and loss of iron homeostasis leading to congestive heart failure as illustrated in Figure 2-13 (Thorn, et al., 2011).

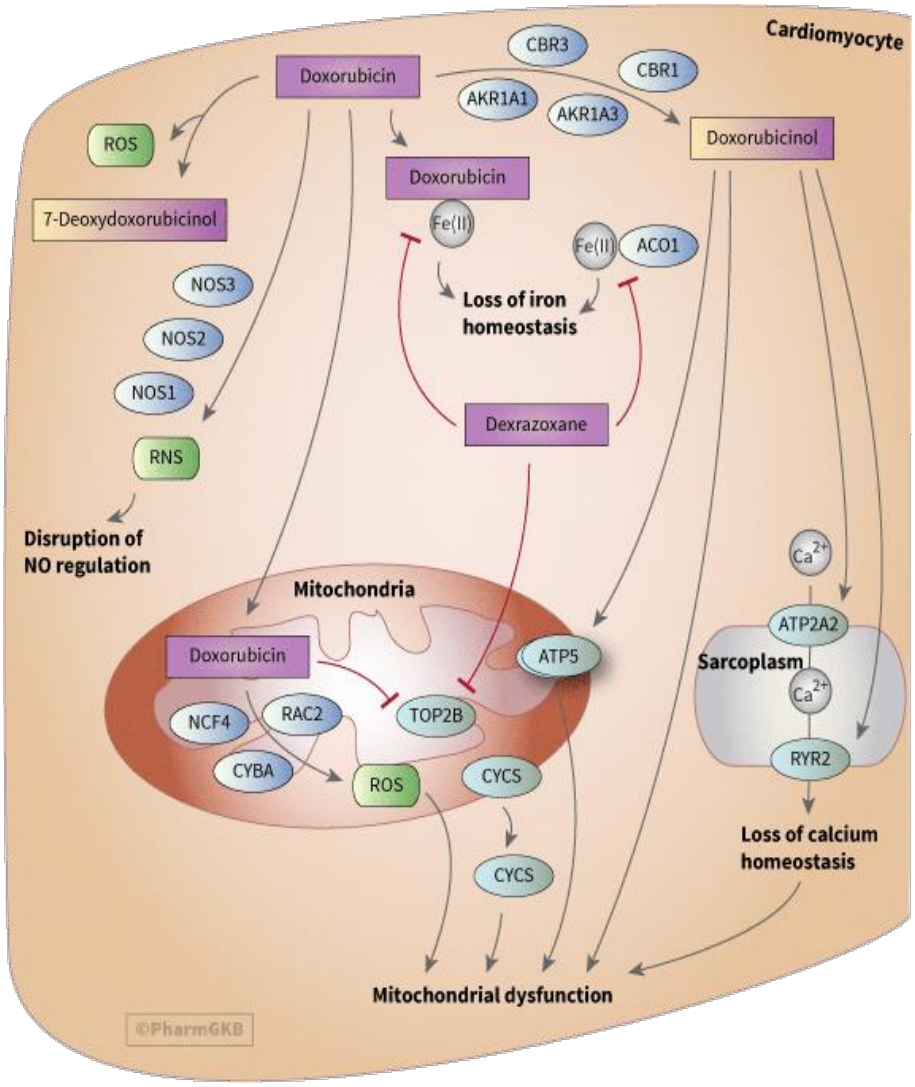


Figure 2-13 Pathways involved in adverse effects of doxorubicin.

The mechanisms of the effect of DOX implicated in loss of iron homeostasis, loss of calcium homeostasis and mitochondrial dysfunction which also targeted in this study. Derived from (Thorn, et al., 2011) .

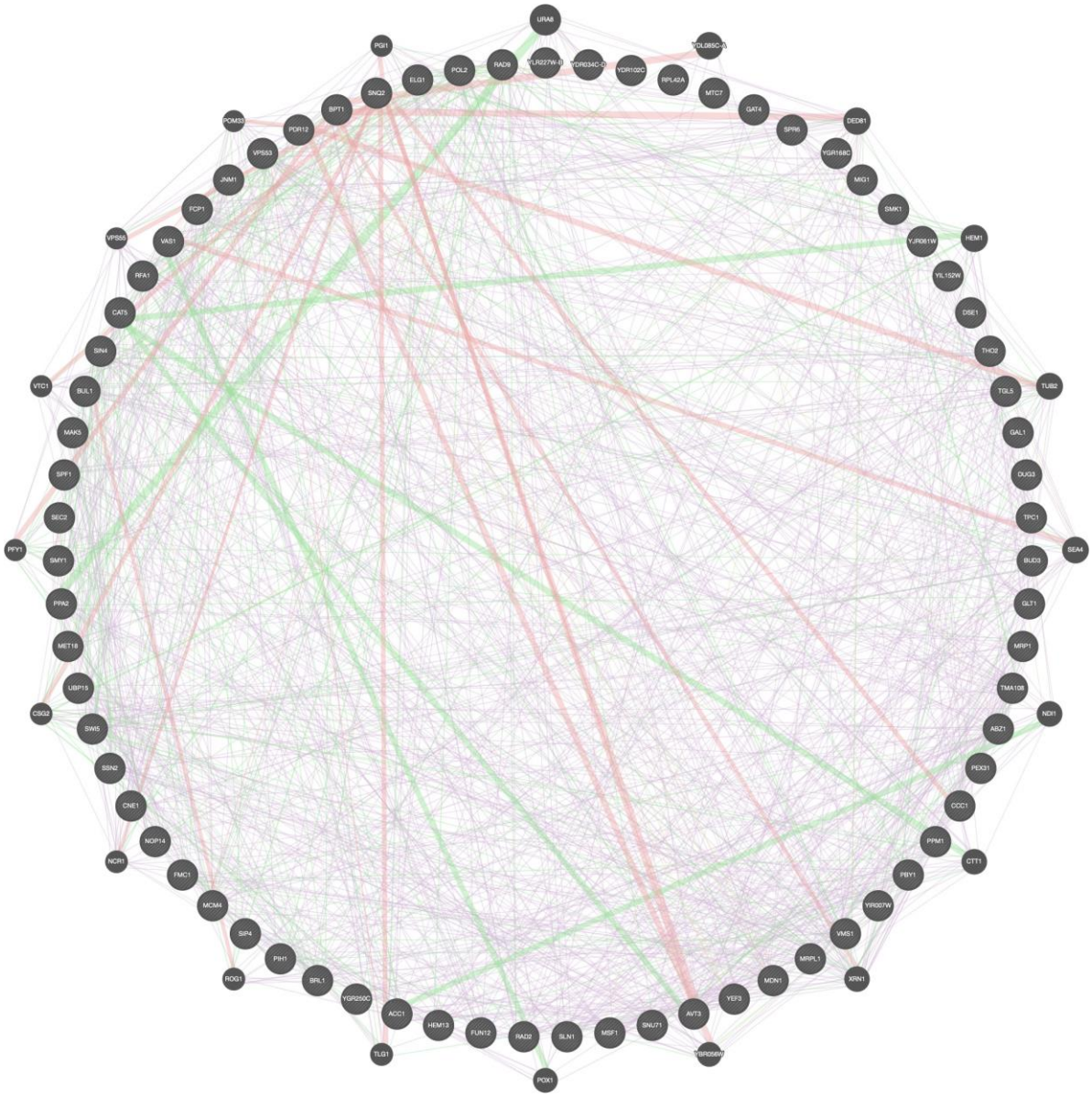


Figure 2-14 Gene network figures for selected genes.

Pink line connections if there are studies showing protein-protein interaction between genes. Purple line connections if the gene expression levels were similar and have been published. Green line connections if two genes were functionally associated. Yellow line connections if the gene products have the same protein domain. Figures were generated through GeneMANIA database (Warde-Farley, et al., 2010). Dynamic figure could be viewed through link:
https://genemania.org/search/saccharomyces-cerevisiae/

Table 2-8 Lists of candidate genes identified by gene function analysis collected from SGD database.

Candidate Gene	Standard Name	Gene Name	Description
SPF1	YEL031W	Sensitivity to Pichia Farinosa killer toxin	P-type ATPase, ion transporter of the ER membrane; required to maintain normal lipid composition of intracellular compartments and proper targeting of mitochondrial outer membrane tail-anchored proteins; **involved in ER function and Ca2+ homeostasis**; required for regulating Hmg2p degradation; confers sensitivity to a killer toxin (SMKT) produced by Pichia farinosa KK1
PMR1	YGL167C	Plasma Membrane ATPase Related	**High affinity Ca2+/Mn2+ P-type ATPase; required for Ca2+ and Mn2+ transport into Golgi; involved in Ca2+ dependent protein sorting**, processing; D53A mutant (Mn2+ transporting) is rapamycin sensitive, Q783A mutant (Ca2+ transporting) is rapamycin resistant; Mn2+ transport into Golgi lumen required for rapamycin sensitivity; mutations in human homolog ATP2C1 cause acantholytic skin condition Hailey-Hailey disease; human ATP2C1 can complement yeast null mutant
SLN1	YIL147C	Synthetic Lethal of N-end rule	Transmembrane histidine phosphotransfer kinase and osmosensor; regulates MAP kinase cascade; transmembrane protein with an intracellular kinase domain that signals to Ypd1p and Ssk1p, thereby forming a phosphorelay system similar to bacterial two-component regulators
CCP1	YKR066C	Cytochrome c Peroxidase	**Mitochondrial cytochrome-c peroxidase**; degrades reactive oxygen species in mitochondria, involved in the response to oxidative stress
UPS1	YLR193C	UnProceSsed	Phosphatidic acid transfer protein; plays a role in phospholipid metabolism by transporting phosphatidic acid from the outer to the **inner mitochondrial membrane**; localizes to the mitochondrial intermembrane space; null mutant has altered cardiolipin and phosphatidic acid levels; ortholog of human PRELI
CCC1	YLR220W	Cross-Complements Ca(2+) phenotype of csg1	**Vacuolar Fe2+/Mn2+ transporter**; suppresses respiratory deficit of yfh1 mutants, which lack the ortholog of mammalian frataxin, by preventing mitochondrial iron accumulation; relative distribution to the vacuole decreases upon DNA replication stress
AAC1	YMR056C	ADP/ATP Carrier	**Mitochondrial inner membrane ADP/ATP translocator**; exchanges cytosolic ADP for mitochondrially synthesized ATP; phosphorylated; Aac1p is a minor isoform while Pet9p is the major ADP/ATP translocator; relocalizes from mitochondrion to cytoplasm upon DNA replication stress
HFA1	YMR207C		**Mitochondrial acetyl-coenzyme A carboxylase**; catalyzes production of malonyl-CoA in mitochondrial fatty acid biosynthesis; relocalizes from mitochondrion to cytoplasm upon DNA replication stress; genetic and comparative analysis suggests that translation begins at a non-canonical (Ile) start codon at -372 relative to the annotated start codon

2.3.7 QTL Analysis indicated QTL regions with high LOD scores are located near centromeres

In addition to the candidate genes mapped by the peak markers, 19 QTL intervals located around centromere region were found. Table 2-9 summarises the QTL intervals that located near centromeres on different chromosomes. Eight of the overlaps are located with the centromere regions included and 6 of QTL intervals are overlapped close to centromere region. However, the QTLs under F1 segregants can only mapped in limited resolution. The span of each QTL interval is more than 25000 bp. Among these region, DDR related genes (MPH1, RTT109, RAD57) and metal ion binding related genes (LEU1, PRI1, VPS27) were also located around this region. Hence, further validation needed to verify the causative genes. A recent study in fission yeast revealed that several DOX resistance proteins can affect the centromeric localisation which result in the centromeric defects. However, the QTLs under F1 segregants can only be identified with typed markers which have limited sizes as only one round recombination through meiosis for F1 population (Nguyen, et al., 2015).

Table 2-9 Lists of QTL intervals that locating around centromere.

Overlap records were highlighted in bold.

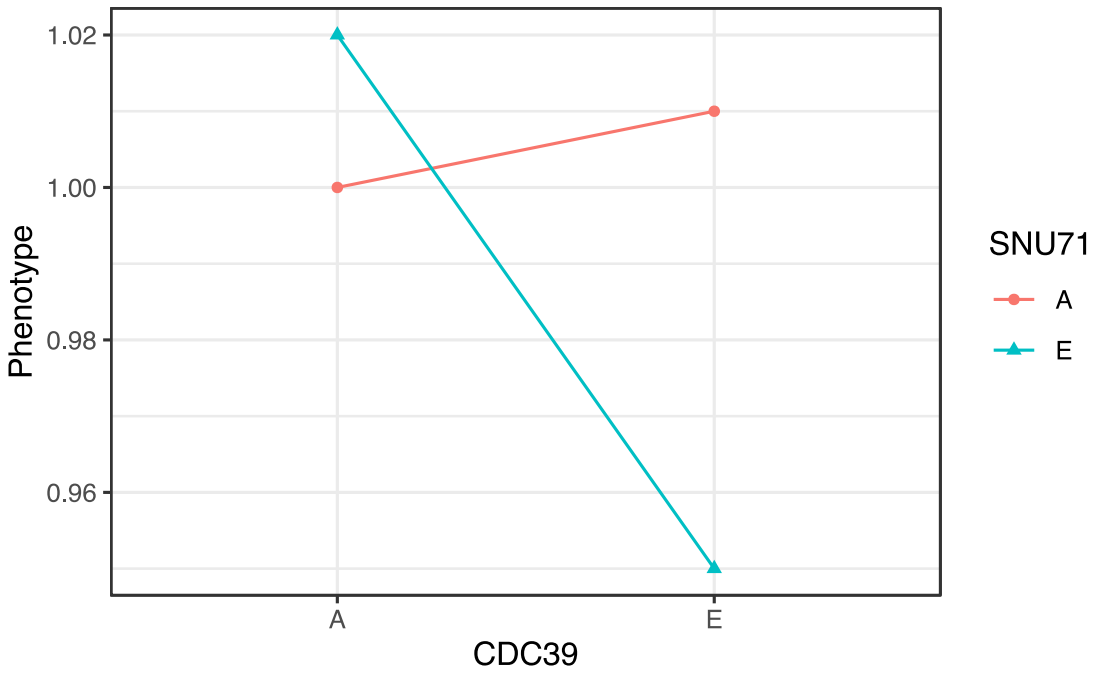
Cross	Phenotype	chromosome	LOD score	CEN
AE	**MS**	**IV**	**4.27604725**	**CEN4**
AE	MS	XIII	4.11976257	CEN13
AE	TR	VII	6.91803379	CEN7
AS	MS	I	3.71192492	CEN1
AS	**MS**	**IV**	**7.90659665**	**CEN4**
AS	**MS**	**IX**	**8.09019987**	**CEN9**
AS	TR	IX	5.99516246	CEN9
AW	**MS**	**II**	**13.8357888**	**CEN2**
AW	**MS**	**V**	**11.1789263**	**CEN5**
AW	MS	XII	10.3002976	CEN12
AW	**TR**	**IX**	**7.01710027**	**CEN9**
AW	TR	XIII	4.69846583	CEN13
AW	TR	XIV	5.16426512	CEN14
AW	TR	XVI	3.53893531	CEN16
EW	**MS**	**II**	**8.46532305**	**CEN2**
EW	MS	XIV	12.7002804	CEN14
SW	MS	III	6.79920416	CEN3
SW	**MS**	**V**	**9.87343058**	**CEN5**
SW	MS	VII	15.9760303	CEN7

2.3.8 Two-dimensional genome scan

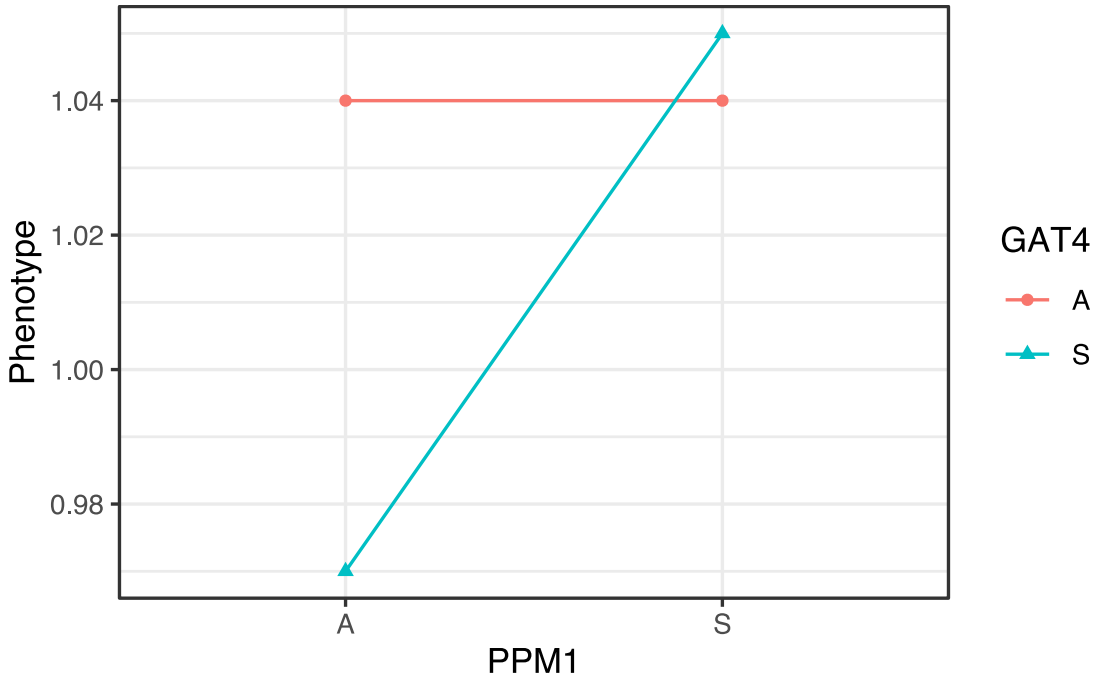
Two-dimensional genome scan under R/qtl and R/shmootl were performed for these 6 crosses and hundreds of QTL interactions event were detected with 1000 times permutation test with significant level at 0.001 (Broman & Sen, 2009). The large amount of QTL pairs came out as significance is due to the reason that the genetic linkage is very high for the F1 population. In this situation, both of the markers were selected as significance only when they were not detected as QTLs in the single scan and were located as the largest LOD scores over threshold. Phenotype effect under the combination of alleles were illustrated in the Figure 2-15. Both positive epistasis and negative epistasis effect presented. Gene SNU71 related to mRNA processing was present in AE and AW interacted. CDC39 involved in pathway of RNA degradation. In addition, gene SEC23 were negative interact with SNU71 that was related to metal ion binding. Gene GAT4 for the regulation of transcription from RNA polymerase II was present in AS and AW interacted with PPM1 and SEC2. PPM1 also regulate transcription of RNA polymerase II. The interaction QTL for cross EW linked to gene MGR1 which forms subcomplex for mitochondrion (Dunn, et al., 2006) and POG1 forms Promoter-binding protein (Demae, et al., 2007).

Table 2-10 List of Two-dimensional QTL scan output

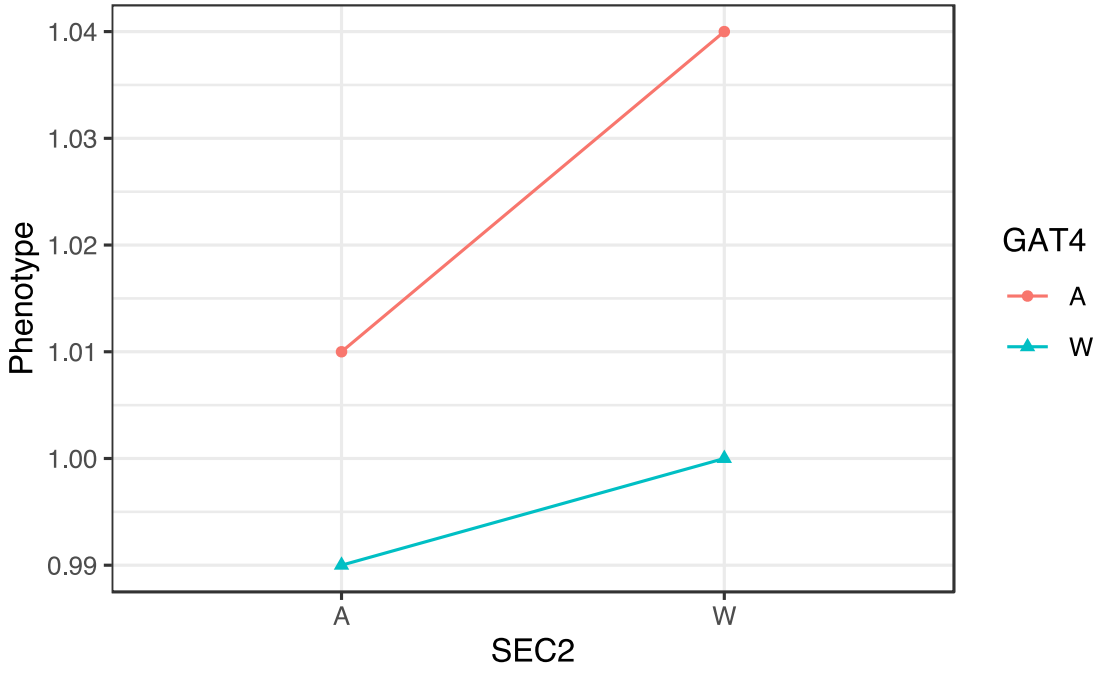
Cross	Phenotype	CHR 1	CHR 2	POS 1	Gene 1	POS 2	Gene 2
AE	TR	III	VII	283000	CDC39	515000	SNU71
AS	TR	IV	IX	1333000	PPM1	380000	GAT4
AW	TR	IX	XIV	380000	GAT4	128000	SEC2
ES	TR	VII	XVI	515000	SNU71	899000	SEC23
EW	TR	III	IX	48000	MGR1	131000	POG1


1.02
1.00
0.98
0.96
Phenotype
A
E
CDC39
SNU71
A
E

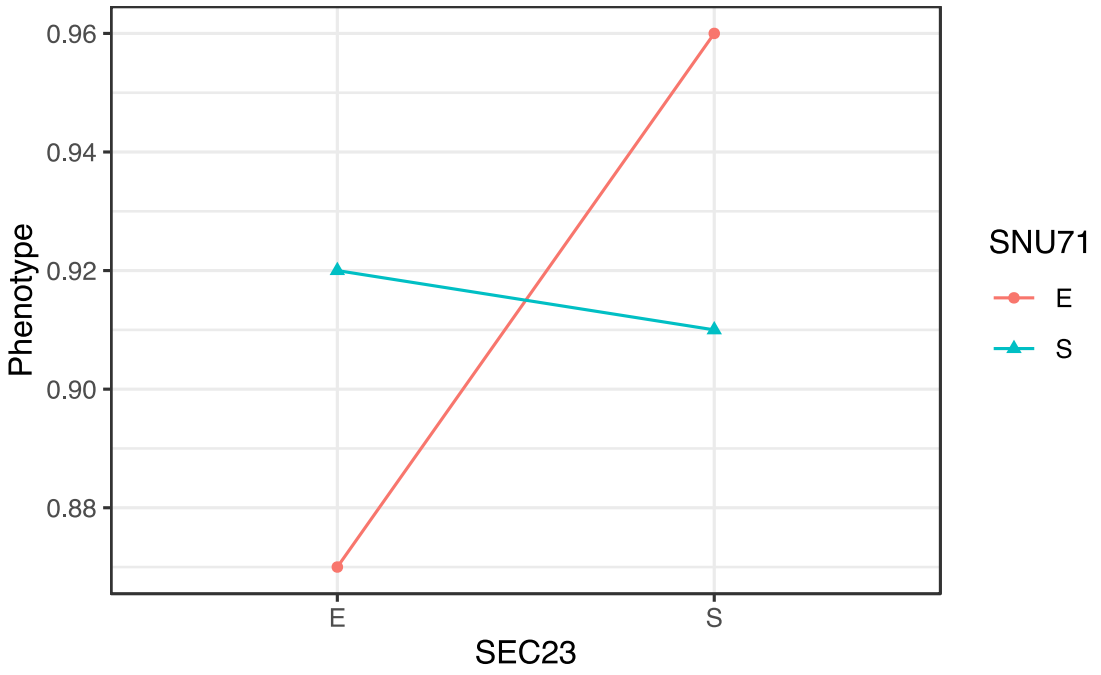
(a) AE


1.04
1.02
1.00
0.98
Phenotype
A
S
PPM1
GAT4
A
S

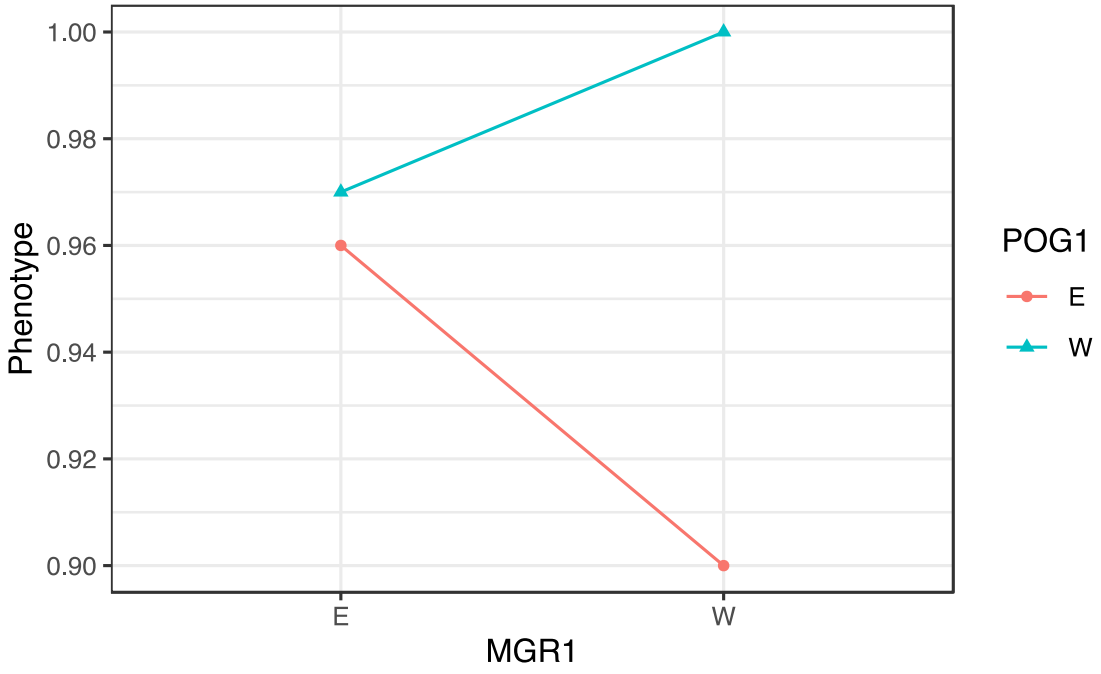
(b) AS


1.04
1.03
1.02
1.01
1.00
0.99
Phenotype
A
W
SEC2
GAT4
A
W

(c) AW


0.96
0.94
0.92
0.90
0.88
Phenotype
E
S
SEC23
SNU71
E
S

(d) ES



(e) EW

Figure 2-15 Illustration of significant epistasis effect between two markers among F1 population for treatment ratio under DOX.

Average phenotype value was calculated for each combination of allele between two markers. Each marker with two alleles indicated as 1 and 2. Red lines are for allele 1 in the other marker and blue lines are for allele 2. (a) Significant nonlinear effect between CDC39:283000 and SNU71:515000 in AE group. There is no difference in the effect of the two alleles at SNU71 in the genotype A of CDC39 but shown large difference with genotype E. (b) Significant nonlinear effect between PPM1:1333000 AND GAT4:380000. No difference showed of the two alleles at PPM1 under genotype A of GAT4 but shown difference when GAT4 with allele S. (c) Significant nonlinear effect between GAT4:380000 and SEC2:128000 in AW group. No difference shown in the effect of the two alleles at SEC2 under GAT4 marker with allele W. Large difference shown under GAT4 with allele A. (d) Significant nonlinear effect between SNU71: 515000 and SEC:899000 in ES group. Larger difference of the effect in SNU71 marker with allele E than with W. (e) Nonlinear effect between mgr1:48000 and POG1:131000 in EW group. Larger effect difference in MGR1 marker with allele W than with allele E.

2.3.9 Discussion

In this section, QTL analysis were performed on 6 F1 crosses with SGRP4 as founders to identify gene variants with significant effect under exposure to DOX. As the point shown in (Cubillos, et al., 2011), different combinations of parental lines demonstrate the power to detect more QTLs than only generated experiment on two parental lines. From the one-dimensional QTL scan, the QTL intervals were overlapped among same founder strain which shown large effect and most of the candidate genes are located within overlaps. An additional overlap type detected in all the crosses without a specific founder strain were shown from the analysis. Eight candidate genes (SPF1, SLN1, AAC1, PMR1, CCC1, HFA1, UPS1, CCP1) were selected through functional annotation and gene network analysis, most of them have been involved with metal binding and mitochondrion processes. These results matched with the Doxorubicin pathways which were shown this drug have the side effect cardiomyopathy may cause the congestive heart failure on human (Chatterjee, et al., 2010). Besides of the overlap QTL intervals, there are also half of the QTL intervals were hold by the specified cross combination and interacted with the overlap QTLs. Moreover, enrichments of QTL intervals were identified around centromere region. These intervals were mapped as DDR related genes (MPH1, RTT109, RAD57) and metal ion binding related genes (LEU1, PRI1, VPS27), which potentially caused the centromere region to be included.

In addition, two-dimensional genome scan was also applied for studying the epistasis effect. Several genes were missed in single QTL scan. Both positive epistasis effect and negative epistasis effect were observed under different crosses. SNU71 presented in AE, ES founders with CDC39 and SEC23. GAT4 presented in AS and AW founders with PPM1 and SEC2. RNA degradation pathways were target among these genes. However, as it is only one round of mating of yeast, the segregants have limited number of genotypes due to the low recombination rate. The peak markers were located with the interval mapping estimation which might cause false positives. It is challenging that locate genes among intervals as the span could be large. A high resolution population with

dense markers could be the ideal population for summarising the candidate genes located not in the peak and able to detect small effect for the complex trait. In the next 3 chapters, F12 population with high resolution that generated with SGRP 4 parental crosses will be used for QTL analysis.

Chapter 3 High resolution mapping of genetic variation underlying growth differences under different treatment in a 4-parent intercross population

3.1 Introduction

The traditional QTL analysis is based on large population generated by 2-parental lines with bi-allelic marker data. In the Chapter 2, 6 groups of first-generation lines for pairwise crosses of SGRP4 strains were used to identify QTLs under DOX exposure. With the comparative analysis among the QTL intervals of the 6 crosses, a few gene features were detected. However, there are still many genes that cannot be identified due to linkage associations in only one generation and the limited variation of only 2 parents. A multiple founder cross design with multiple generations expands genetic and phenotype diversity and breaks the linkage blocks to shuffle alleles. Recent advances in high-throughput techniques for DNA sequencing and phenotyping have greatly facilitated the identification of genetic variants underlying traits at a genome-wide level (Wilkening, et al., 2014). A high rate of recombination results in a high density physical and genetic maps, allowing greater resolution in mapping quantitative trait loci (Glazier, et al., 2002).

A 12 generation 4-way cross population of *S. cerevisiae* with the same four natural yeast isolate founders (NA, SA, WA, WE) used in the F1 crosses, but designed with 4-parental lines, SGRP4-X, was generated and segregants were sequenced in (Cubillos, et al., 2013). Phenotyping and sequencing analysis of those strains revealed high degree of variation in phenotype under different environmental and chemical conditions and genomes with more than 245 thousand polymorphisms (Liti, et al., 2009). A total of 15217 bi-allelic markers

were selected from the high-quality SNPs for 166 F12 haploid segregants with no missing genotype data that could place makers at the individual gene level.

Responses to chemical drugs vary among individuals which might due to the complexities underlying genetic architecture. Using these powerful F12 yeast segregants as a test bed for genetic differences in response to drugs, 18 treatment conditions under different chemicals or combinations of drugs were studied in this chapter to identify novel genetic variants and loci responsible for growth differences (both favourable alleles and disadvantageous alleles) with high-resolution mapping. Analysis of the candidate genes can then further localise potential molecular mechanisms for drug responses. For detecting the parental origin of alleles, a 4-parental genotype model was developed for tracing back the inheritance from the original parental founders.

3.2 Materials and Methods

3.2.1 Strains

In order to examine the effect of natural genetic variation on heritable traits of yeast, the yeast haploid strains used for QTL mapping experiment were 12th generation advanced intercross lines (SGRP4-x). Four haploid strains in SGRP4: North American (NA): YPS128, West African (WA): DBVPG6044, Sake (SA): Y12, Wine/European (WE): DBVPG6765 were selected as founder strains which were the same four founder strains as Chapter 2. The procedure starts with four homozygous parental strains to create two heterozygous F1 diploid hybrids. Then these diploid hybrids were sporulated and went through meiosis to produce recombinant haploid offspring. The F1 haploid progenies are used to create F2s by mating and meiosis. For getting high resolution QTL mapping, 11 more rounds of random mating and meiosis are required to reduce the average linkage block size through homologous recombination. Figure 3-1 illustrates the process of this intercross. The segregants were constructed as described above. Each segregant is represented by four replicates in experiments.

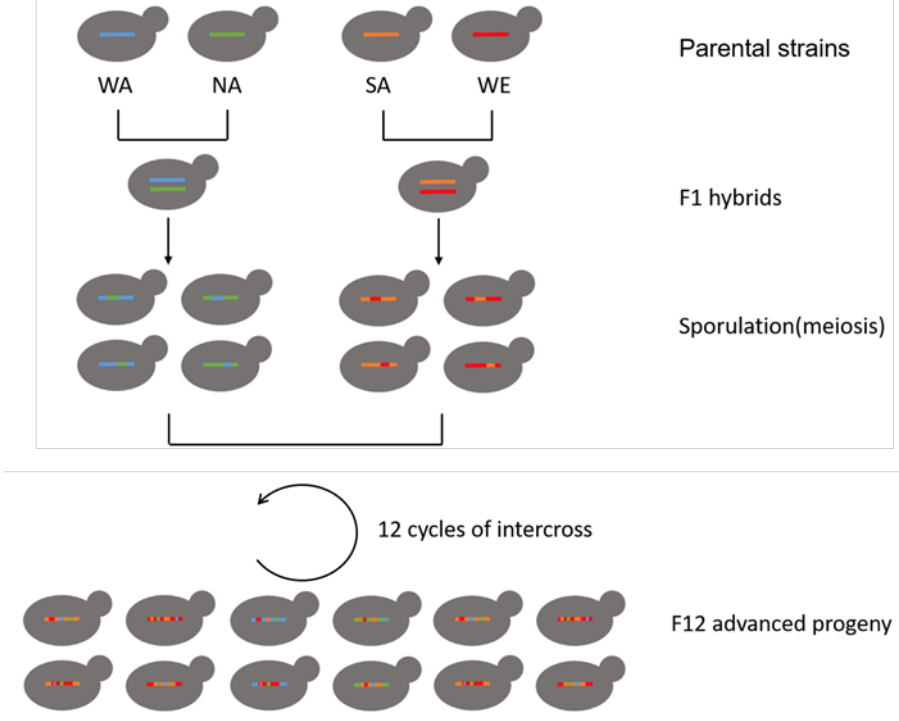


Figure 3-1 The process of generating F12 population

3.2.2 Agents and experiments

The experiment of QTL mapping under 18 distinct environment conditions with different chemicals using F12 segregants for phenotyping with control and treatment were performed by Danae Georghiou, Salwa Almayouf and Yishen Li, University of Leicester. The chemicals used and their concentrations are given in Table 3-1. Yeast plates, preparation and Phenotyping methods are the same as described in Chapter 2. QTL mapping experiments of the F12 yeast generation were prepared for each agent with four yeast arrays for phenotyping. MATa and *MATα* strains were grown on separate plates under standard YPD as control and treatment with chemicals added.

3.2.3 Genotyping

The raw sequencing data of multi-parental F12 segregants among the cross of four representative founder strains (NA, WE, SA, WA) of *S. cerevisiae* were obtained in a previous study (Cubillos, et al., 2013). Genetic markers were further filtered through SNP calling output generated by Dr. Thomas Walsh without duplicate markers that kept the identical bases for each segregant. After filtering, I eventually have the genotype dataset which includes 15217 bi-allelic markers genotyped for all 166 F12 segregants with no missing data. This genotype dataset is 85 times denser than the F1 segregant markers. Marker IDs were labelled with same format as F1 genotype dataset described in Chapter 2. Table 3-1 describes the marker numbers and average density in each chromosome based on the statistics of the reference genome S288C. Except for the smallest chromosome in S288C, CHR I, all other chromosomes have marker density under 900bp on average between markers.

Table 3-1 Summary statistics of marker numbers and average marker density in each chromosome of F12 segregants genotype data.

Total length column was calculated through R environment with the reference genome fasta file of S288C in version R64-2-1 as input. Marker numbers column was counted through R environment with the genotype data of F12 segregants. Marker density column was calculated as the ratio between total length and marker numbers.

Chromosome	Total length (bp)	Marker numbers	Marker density
I	230218	201	1145
II	813184	1157	702
III	316620	446	710
IV	1531933	1730	886
V	576874	731	789
VI	270161	428	631
VII	1090940	1356	805
VIII	562643	824	683
IX	439888	612	719
X	745751	940	793
XI	666816	899	742
XII	1078177	1276	845
XIII	924431	1121	825
XIV	784333	1077	728
XV	1091291	1347	810
XVI	948066	1072	884

3.2.4 Dataset Preparation

For each chemical agent, phenotype data and genotype data were combined into a QTL input dataset that includes 15218 attributes (1 identifier of each segregant and 15217 markers) and 166 records with 4 replicates of haploid yeast strains for the F12 dataset. Some experiments had reduced sets for the analysis due to contamination observed. The first attribute is the identifier of the 4-way cross (AESW), generation and the strain number. For example, AESW12fc201 is the identifier of the individual of 4-way cross in 12 generation with labelling number 201. Other attributes are the 15217 markers located in 16 chromosomes at different position and contains enumerated allele genotype which were encoded as '1' and '2', an arbitrary raw genotype (any two of base in 'A', 'C', 'G' and 'T') were computed using makegeno function in r/shmootl. All the records are fully genotyped without any missing values. The phenotype data were matched by the identifier of strain names. Marker information table were generated for storing Mating type, LYS2 +/-, URA3 +/- as covariate data for each strain and matched by the identifier. Table 3-2 an example of the datasets for F12 segregants' QTL analysis.

Table 3-2 Sample input datasets of F12 segregants

Treatment Ratio	ID	c01:0038000 (CNE1)	c01:0064000 (CDC24)	c01:0078000 (FUN12)	c01:0095000 (SAW1)	c01:0114000 (ATS1)	c01:0158000 (RFA1)
		1	1	1	1	1	1
		12.70	12.74	12.74	13.16	13.46	14.36
0.943787	AESW12fc201	1	1	1	1	1	1
0.934674	AESW12fc202	2	2	2	1	1	2
0.879095	AESW12fc203	2	2	2	1	1	2
0.875925	AESW12fc204	2	2	2	2	2	2
0.855349	AESW12fc205	1	1	1	1	1	1
0.892152	AESW12fc206	1	1	1	1	1	1
1.19754	AESW12fc207	1	1	1	1	1	1
1.06395	AESW12fc208	2	2	2	1	1	2
1.11133	AESW12fc209	1	1	1	1	1	1
1.14976	AESW12fc210	2	2	2	1	1	2
…	…	…	…	…	…	…	…

3.2.5 Statistical and Bioinformatics Analysis

Descriptive statistics of phenotype attributes were performed in the R environment through Rstudio. Box plots were generated for comparing growth features with ggplot2 package. The interquartile range was calculated for comparing the phenotype distribution of F1 segregants and F12 segregants through IQR function in R (IQR = Q3 − Q1) between upper (75%) and lower quartiles (25%). The comparison of 12th generation progeny genotypes to founder genotypes was analysed in R. The founder genotypes were obtained from SGRP known SNP dataset in from (Bergström, et al., 2014) based on the S288C reference genome and aligned by using internal pipeline 'founder align' in R for finding the genotypes derived from each parental background. Coloring for a match is based on wesanderson package in R with Darjeeling. QTL analyses were performed through run_scanone pipeline in r/shmootl with marker regression given the large number of markers covering the whole genome. The significance level alpha was set at 0.05 to determine the LOD threshold. 1000 permutation tests were applied for each QTL scan. Covariate parameter is set with the covariate table, LOD interval estimation are located by 1.5 LOD support intervals (the regions have LOD scores within 1.5 to peak) for each QTL as support region to find candidate genes. Clustering of functional annotations of candidate genes within intervals was performed on DAVID 6.8. Genetic variant annotation was performed using SNPEff V4.0 with reference genome S288C R64-1-1. All the QTL analyses were performed on ALICE. R scripts for F12 QTL analysis and pipeline qsub files are supplied in Appendix A. All the QTL analysis records can be accessed through the GACT research R-drive at the University of Leicester.

3.3 Results

3.3.1 Genetic diversity and phenotype variation around F12 population

Table 2-3 and Table 3-1, show the increasing genetic marker density across the whole genome with a multi-parental background and high recombination rate as seen in the mosaic genotypes in the F12 population. With such dense markers as attributes, a narrower region can be delimited between two markers. With the large genetic diversity in the 12^{th} generation segregants (Figure 3-2), the phenotype distribution of growth exhibited larger variation and span compared to the F1 population. Under the DOX condition, experiments were performed in both F1 and F12 generation which enabled me to compare the phenotype distribution of F1 and F12. AE, AS, AW, ES, EW and SW were the F1 haploid population that having the same founder strain with the F12 population. The bee swarm plots for growth features show clearly the phenotype distribution difference between F1s and F12. The F12 population has a larger interquartile range (0.14 for treatment ratio and 0.07 for max slope compared to F1 value) showing the increase in spread. The F12 sergeants also exhibit a larger span of the distribution with more extreme growth features than the F1 sergeants in these six groups (Figure 3-3). This comparison in genotype and phenotype data underpin the advantages for detecting QTLs in the F12 population rather than the F1s.

Figure 3-2 Genetic diversity of 12 generation segregants on chromosome IX.

Genome shuffling that breaks linkage disequilibrium of F12 compared to founder strain genotypes. Coloring is based on the match to founders, where red is WA, mint is WE, yellow is SA and orange is NA.

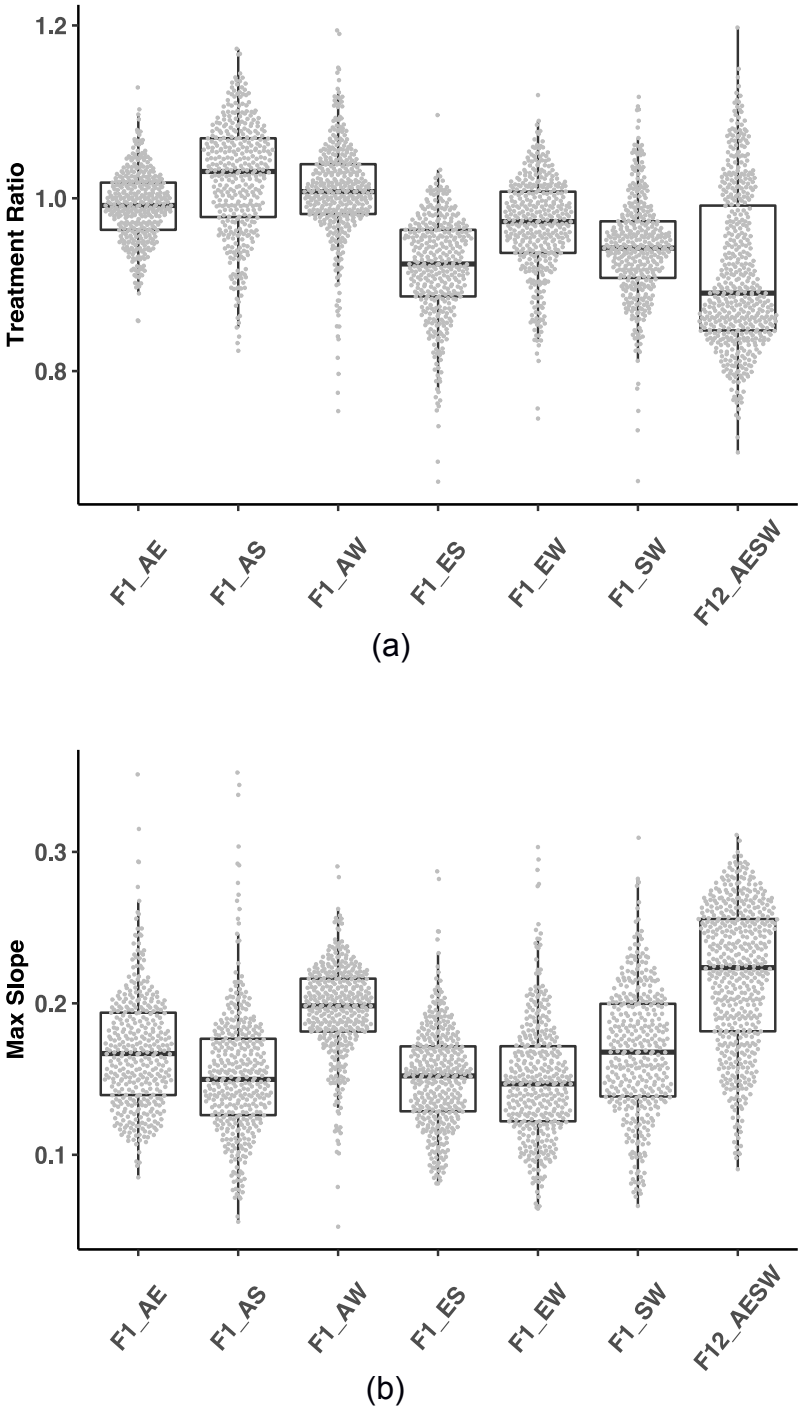


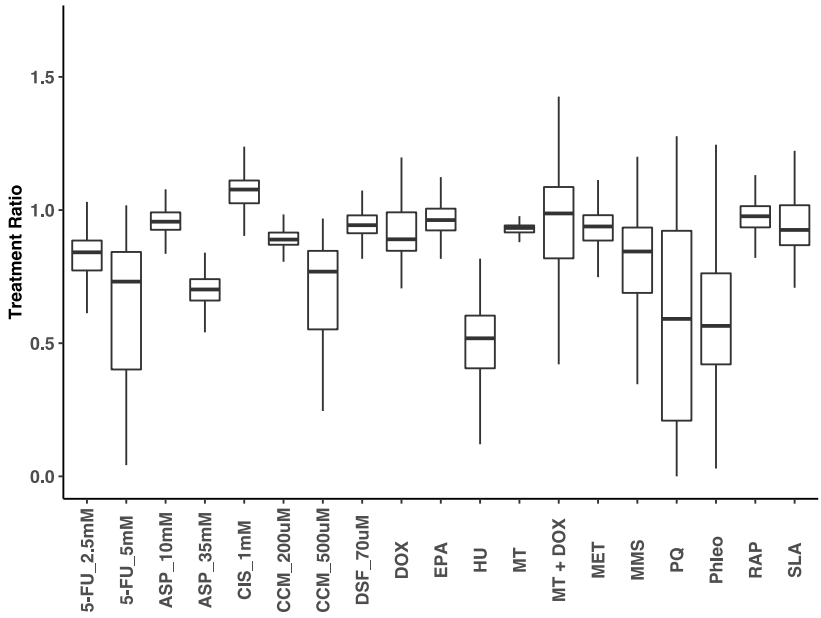
Figure 3-3 Bee swarm and boxplots of phenotype distribution under DOX for F1 segregants and F12 segregants showed difference in mean and spread.

AE, AS, AW, ES, EW, SW were the six group F1 bi-parental segregants and F12 was the four-parental segregants annotated as F12_AESW. Y-axis is the phenotype value. Grey dots in each bee swarm plot represented the value for each individuals with group. a) Treatment Ratio b) Max Slope

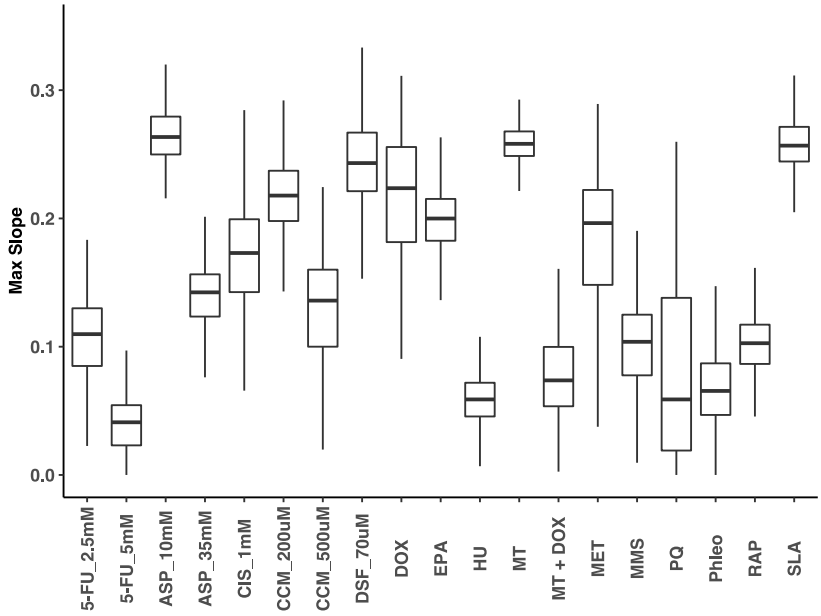
3.3.2 Phenotype distribution varies among F12 yeast segregants under different agents.

The quantitative changes in F12 yeast segregants grown under 19 different chemical compounds were collected for high-resolution phenotypic analysis. As these chemicals are mostly related to cancer therapy or genotoxic agents, most of segregants (82.4%) have growth rates under 1 (normalized to YPD control growth). With cultivation under different conditions, F12 yeast growth patterns varied with different growth levels (Treatment Ratio) and different growth speeds during exponential phase (Max slope). The growth phenotypes are shown with box plots in Figure 3-4.

Among the treatments, HU exposure results in the lowest average growth level, 0.5 of control, which means that most of the yeast segregants were inhibited under this condition. Other than HU, the same agents with different concentrations exhibited similar trends with the average growth attained for the higher concentration were less than the lower concentrations (5-FU 2.5mM 0.82 VS 5mM 0.63, Aspirin 10mM 1.00 VS 35mM 0.69, CCM 200mM 0.89 VS 500mM 0.69). Interestingly, the majority of segregants (80%) have better performance with growth level over 1 under Cisplatin exposure. High levels of phenotypic diversity were seen in the growth of F12 segregants under each condition. Extreme values could be observed in most of the agents. The growth levels attained under Paraquat were distributed with the interquartile range more than 0.7 and under 5mM 5-FU with more than 0.4.



(a) Treatment Ratio



(b) Max Slope

Figure 3-4 Boxplots showing phenotype distribution of F12 segregants under different chemical treatments.

X-axis is the agents name arranged in alphabet order. Y-axis is the phenotype value for (a) is treatment ratio and for (b) is Max slope.

Besides the growth ratio being different, growth speed under different agents also exhibited more diversity among segregants. Similar trends were seen where the average speeds for the higher concentrations were inhibited compared to the lower concentrations for each agent. The higher concentration would cause the segregants to take longer to catch up to the growth speed making growth slower. From Figure 3-4(b), it is clear that 10mM Aspirin, Matrine and Salicylic acid have the fastest growth rate of over 0.25 OD change per hour. Segregants exposed to Paraquat exhibit various performances in growth rate with a low speed of 0.08 on average. Interestingly, segregants which grew fast under DOX or Matrine singly, grew slowly with combination of both DOX and Matrine. Overall, different growth patterns were found with different treatments among each segregant, suggesting that different genetic landscapes might be discovered for explaining these complexities.

3.3.3 Different numbers of QTLs were detected under different chemical conditions

Table 3-3 Summary of QTL mapping for each agent.

Total QTL numbers in response to each agent treatment are obtained from QTL analysis output. The peak LOD score is displayed as the maximum score for each QTL run of the phenotype for the agents. The chromosome and peak name are the locations of the Peak LOD for each mapping. Peak gene is annotated based on S288C gff file where the marker located within. Gene name coloured in grey indicated that the marker is intergenic which not located in the coding region but is around 250bp upstream/downstream gene region.

(a) Max Slope

Agents	Total QTL numbers	LOD threshold	Peak LOD	CHR	Peak Name	Peak Gene
5-FU_2.5mM	246	4.33	11.00	X	c10:0173029	YJL130C
5-FU_5mM	102	4.44	11.07	X	c10:0626155	YJR106W
ASP_10mM	61	4.48	8.65	IX	c09:0143924	YIL116W
ASP_35mM	135	4.41	10.35	VII	c07:0896840	YGR198W
CIS_1mM	31	4.48	5.83	X	c10:0254068	YJL094C
CCM_200uM	134	4.44	12.16	XIV	c14:0507665	YNL063W
CCM_500uM	64	4.38	10.57	VIII	c08:0037489	YHL032C
DSF_70uM	51	4.41	10.24	VII	c07:0807938	YGR160W
DOX	229	4.46	21.19	II	c02:0470054	YBR114W
EPA	18	4.24	6.77	XII	c12:0405942	YLR131C
HU	182	4.34	11.71	IV	c04:1487442	YDR523C
MT	114	4.35	9.96	XVI	c16:0503459	YPL024W
MT + DOX	117	4.36	12.39	IV	c04:0125665	YDL186W
MET	203	4.44	13.43	X	c10:0554884	YJR062C
MMS	38	4.57	7.91	IV	c04:0639565	YDR096W
PQ	181	4.38	13.44	XIII	c13:0570462	YMR156C
Phleo	65	4.42	8.67	XII	c12:0706510	YLR284C
RAP	165	4.42	18.00	X	c10:0559883	YJR066W
SLA	34	4.44	9.80	XII	c12:0792277	YLR332W

(b) Treatment Ratio

Agents	Total QTL numbers	LOD threshold	Peak LOD	CHR	Peak Name	Peak Gene
5-FU_2.5mM	211	4.47	10.47	XIII	c13:0155781	YML059C
5-FU_5mM	183	4.42	12.03	IX	c09:0181158	YIL097W
ASP_10mM	25	4.48	7.44	III	c03:0234004	YCR067C
ASP_35mM	81	4.92	10.97	X	c10:0559922	YJR066W
CIS_1mM	7	4.72	6.86	XIV	c14:0710649	YNR047W
CCM_200uM	95	4.37	15.21	IX	c09:0371543	YIR007W
CCM_500uM	33	4.6	8.17	IX	c09:0088045	YIL139C
DSF_70uM	68	4.46	7.74	VII	c07:0799904	YGR155W
DOX	151	4.36	33.46	III	c03:0204507	YCR042C
EPA	20	4.44	10.83	XIV	c14:0437003	YNL101W
HU	194	4.39	12.68	XVI	c16:0617214	YPR026W
MT	18	4.77	11.42	XV	c15:1009827	YOR357C
MT + DOX	204	4.65	14.66	IV	c04:0125665	YDL186W
MET	164	4.99	14.1	VII	c07:0278844	YGL122C
MMS	160	4.4	14.52	XI	c11:0031821	YKL213C
PQ	170	4.37	10.6	XV	c15:0108439	YOL111C
Phleo	140	4.32	11.63	XI	c11:0324419	YKL062W
RAP	105	4.63	9.53	V	c05:0325466	YER082C
SLA	33	4.53	11.29	VI	c06:0084187	YFL025C

3.3.4 QTL analysis reveals the complex factors affecting yeast growth under changes of environment

Marker regression was performed to map QTLs in order to understand the genetic basis, for response to each of the agents. Two growth features were used as phenotypes: max slope and end point growth level (TR64). A large number of QTLs were identified as significantly associated with different agents with LOD above the threshold at 0.05 significance level. Table 3-3 summarises the numbers of QTL hits for each agent with the peak markers within genes listed. A total of 3153 out of 4198 unique QTLs across genomes were characterised for all 19 agents growth responses. Under 2.5mM 5-Fluorouracil, the largest subset of markers was mapped as QTLs in both growth phenotypes with 246 QTLs for max slope and 211 QTLs for treatment ratio at the end point. Fewer QTLs were obtained at the higher dose (Fluorouracil 5mM) compared to the lower dose. The same trend was also found for Curcumin in that the lower dose yielded more QTLs then the higher dose. In contrast, for Aspirin, a higher number of QTLs were identified under the higher dose in both growth phenotypes with 135 QTLs of max slope and 81 QTLs of treatment ratio compared to X and Y for the lower dose. Many QTLs were detected for DOX treatment in both phenotypes. 229 QTLs were detected for max slope and 151 QTLs for treatment ratio at the end point. Smaller numbers of QTLs were obtained for Matrine where 141 QTLs were detected for max slope and only 18 QTLs were identified for treatment ratio at the end point as there is limited differences in growth under Matrine exposure. Perhaps surprisingly more QTLs were detected for treatment ratio under the combined treatment of both DOX and Matrine (204 QTLs) than under each of the single agents. However, for max slope, the number of QTLs is lower than in both of the single agents. The smallest number of QTLs for the max slope phenotype was found for EPA which only has 18 associated QTLs. For treatment ratio at the end point of segregants under Cisplatin only 7 markers were detected as significance which is the smallest number of QTLs among all experiments. Markers were further annotated to determine the highest LOD score in each mapping. When comparing the QTL results of all agents, their largest QTLs were not located at the same marker, the LOD scores of the QTLs are also different.

The QTL intervals were further summarised into one dataset to examine overlaps among all the agents. Gene annotation and functional clustering analysis indicates the pathways involved and possible mechanisms of yeast growth under different agents. QTL analysis defines a large number of QTL intervals across the agents and overlaps were found between agents. Some of the overlapping genes were involved in responses to DNA damage, DNA repair, cell membranes, nucleotide and protein transport and others.

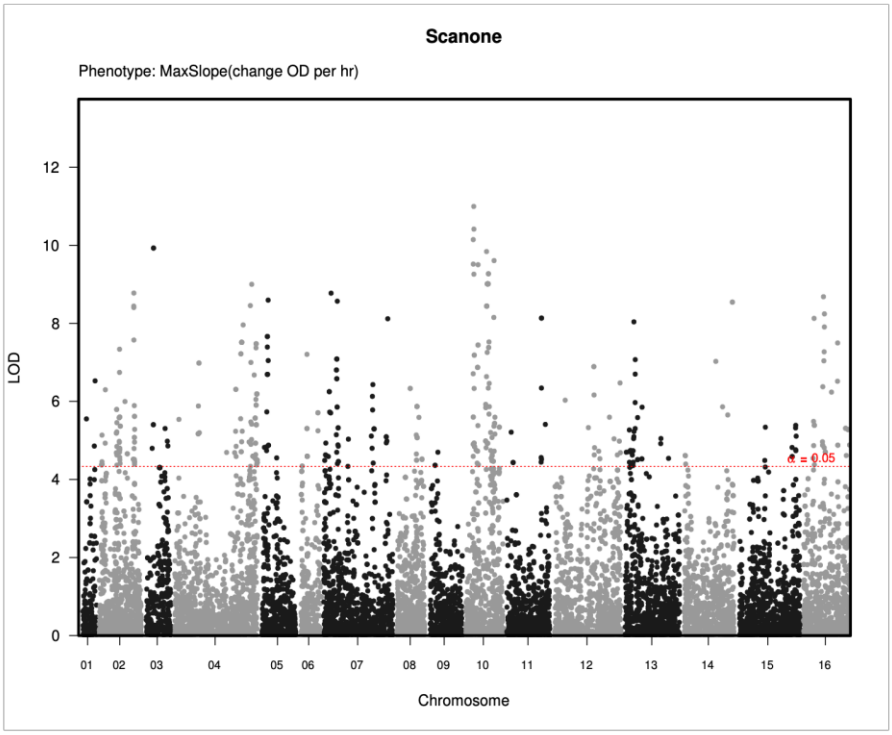
5-Fluorouracil (5-FU) is a genotoxic agent used as an antimetabolite for cancer treatment. Two concentrations (2.5 mM and 5mM) were applied to F12 segregants to explore the genetic association with growth (Figure 3-5). The expectation was that the identified markers should be seen in both sets. Interestingly, different genetic landscapes were found between these two concentrations. When tracing the overlap between the analyses, 25 genes were identified in both high and low concentrations. These genes are related to target functions for responses to environmental changes, such as regulation of transcription, DNA binding, membrane transport etc. (summarised in Table 3-4). Among these genes, HPC2 (YBR215W) overlapped in both concentrations and is located on chromosome II. It is involved in chromatin remodelling and the regulation of histone gene transcription (Zhang, et al., 2013), (Eriksson, et al., 2012). The overlapping gene TOM1 (YDR457W), has a human homolog HUWE1 which is involved in DNA repair and histone modification. Top ranked QTL intervals and peak markers were further annotated to explore the potential shared function. The peak marker at the lower dose is located on chromosome X linked to the gene URA2 (YJL130C) which has a human homolog, CAD, involved in drug metabolism. Three more peak markers which have LOD scores in the top rank are linked to genes (POL31, OST1, TDH2) clustered into relatedness with metabolic pathways. In addition, 10 significant markers linked to genes (PIF1, SRS2, TEL1, RDH54, RAD10, REV3, SPT10, SPT16, TOR1, RAD57) are involved in the DNA damage response pathway. At the higher dose, the peak marker is also located on chromosome X but at gene ECM27 (YJR106W) whose function involves Ca^{2+} exchange and carbohydrate storage (Klukovich &

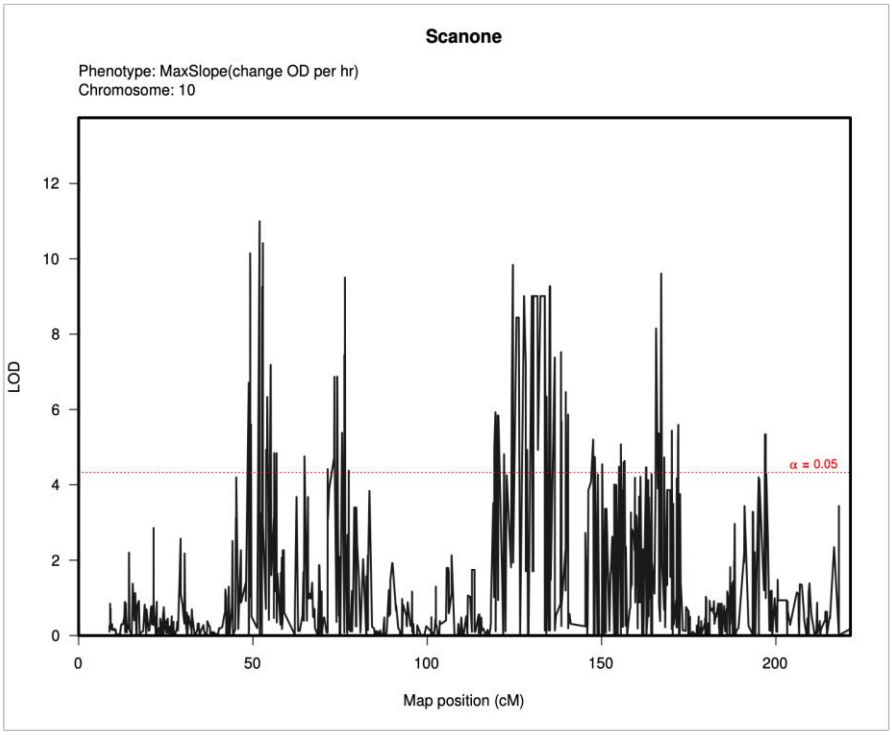
Courchesne, 2016). In addition to the overlap genes, 22 peak genes were related to nucleotide binding. 10 genes (MET6, TRP2, PRO1, HIS3, HIS7, GPM1, PYC1, PYC2, LYS1, LYS9) are involved in the biosynthesis of amino acids. A further 9 QTL intervals were further located (RRP8, BRE2, FUN30, FYV6, SDC1, SPT21, STN1, ADA2, DOT1) involved in chromatin silencing at telomere (Fahrenkrog, 2016).

Table 3-4 Overlap gene features annotation between two concentration for 5-Fluorouracil agents.

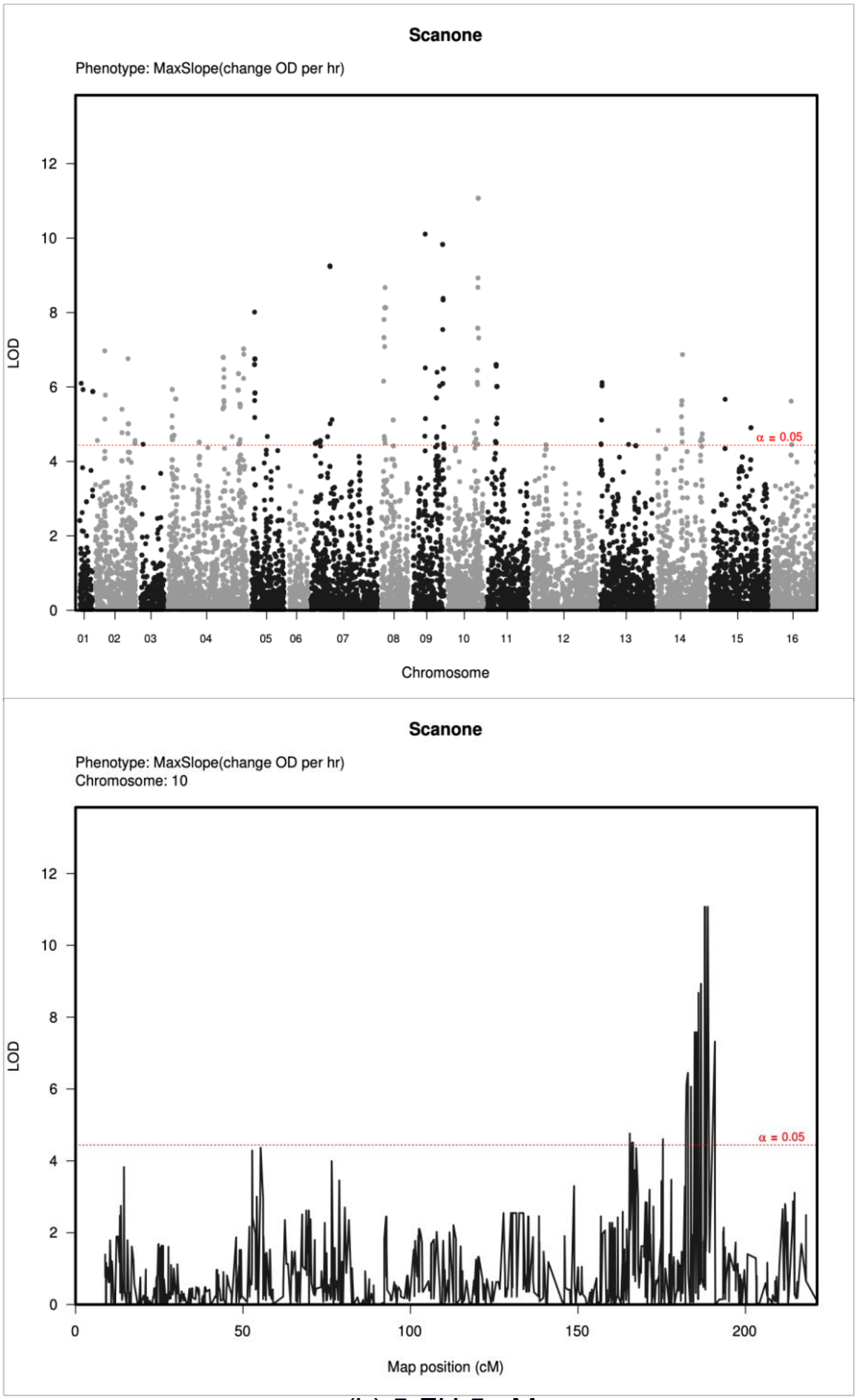
Genes were clustered through DAVID functional clustering analysis that column function were obtained from the output with involved genes. Genes which involved in different clusters were highlighted in bold. Two phenotype features were clustered separately.

Agent 1	Agent 2	Phenotype	Function	Gene Features
5-FU_2.5mM	5-FU_5mM	Max slope	Regulation of transcription	TOM1, **HPC2**, TRA1, SPT16, **GCR1**
5-FU_2.5mM	5-FU_5mM	Max slope	DNA binding	**HPC2**, PYC2, **GCR1**
5-FU_2.5mM	5-FU_5mM	Max slope	Phosphorprotein	CDC19, **GCR1**, COG1, YRA1, SDS24, TOM1, **HPC2**, SPT16
5-FU_2.5mM	5-FU_5mM	TR64	Transcription	**HPC2**, MED1, SPT7, NDT80
5-FU_2.5mM	5-FU_5mM	TR64	Membrane	ANES1, GEA1, IST2, RAV1, KRE2, EPT1, GNP1, YJL132W, YPR071W
5-FU_2.5mM	5-FU_5mM	TR64	Nucleotide binding	UBC4, YHR127W, RPT3, PKH3, PYC2


Scanone
Phenotype: MaxSlope(change OD per hr)
LOD
Chromosome
α = 0.05
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16


Scanone
Phenotype: MaxSlope(change OD per hr)
Chromosome: 10
LOD
Map position (cM)
α = 0.05

(a) 5-FU 2.5mM



(b) 5-FU 5mM

Figure 3-5 QTL mapping for 5- Fluorouracil agents under different concentration.

Manhattan plots for markers LOD score of 16 chromosomes on the left side. Interval mapping of the chromosome have the peak LOD score were illustrated on the right. The red dash line indicates the LOD threshold for each QTL run at significance level 0.05. Each point represents a marker with the LOD score.

In addition to 5-FU, cisplatin, HU, phleomycin, MMS, and rapamycin are also used in the treatment of cancer. Cisplatin yields the smallest set of QTLs among these five agents, with few genes overlapping those in the other treatments. Among the genes in the QTL intervals for Cisplatin, 8 genes (HOT13, GAP1, RER2, LRO1, YNR048W, FPK1, TCD2, COQ1) are relevant to the membrane and FPK1 was also significant under 5-FU. A number of QTLs were overlapping between HU, phleomycin, MMS and rapamycin treatments. Among the overlapping genes, five genes (CEP3, CHL4, SLK19, SPC105, STU2) were physically near the centromeres region and related to chromosome kinetochore. Three genes (SMC5, LIF1, RAD51) are involved in DNA repair which might indicate chromosomal instability related to these drugs (Zhang, et al., 2016). In addition, three genes are involved in metabolic pathways, ALD2, ALD3, HIS3 for Histidine metabolism and gene ALD2, ALD3 for Phenylalanine metabolism, already shown to be involved in mice for Histidine metabolism and cancer therapy (Frezza, 2018). HU yields the largest number of QTL intervals and these overlap with phleomycin QTLs that identified 46 genes. In addition to the overlapping genes around centromere regions, 9 genes (ATP14, GEP3, MPM1, MRM1, SPC105, DNF1, LSP1, YJL070C, PRE6) are involved in mitochondria. For HU, a further 9 genes (MPH1, ARP8, HNT3, EAF5, IES4, MLP1, TTI2, NTG2, MLH1) are involved in DNA repair. The rapamycin target gene TOR1 also overlaps with HU and rapamycin treatments, also occurred in 5-FU sets.

The results of experiments using aspirin at 10mM, 35mM and salicylic acid are shown in Appendix B. Multiple genes were found associated in all of these three agents. The overlapped QTLs include a number of genes involved in DNA repair (APN1, ABF1, MET18, SPT16, MSH6, RAD27, TOR1, RAD52). The marker located in MSH6 (YDR097C) overlapped in all three conditions is involved in the DNA mismatch repair pathway (Antony, et al., 2006). In addition, MSH6 has a human homolog MSH6 which is a tumour suppressor gene with mutations of this gene resulting in increased risks of cancer (Leenders, et al., 2018). TOR1 also identified in the sets of overlap between high dose aspirin and salicylic acid which were also hit in overlap genes for DDR relatedness in 5-FU. APN1 (YKL114C) is

needed in DNA repair for damage by oxidising agents. In addition, three overlapping genes (GSC2, MID2, SSK2) are involved in the MAPK (mitogen-activated protein kinase) signaling pathway were obtained. A previous study in mice have shown that these genes can the phosphorylation of MAPK (Zhang, et al., 2017). Moreover, several overlapped genes between aspirin in high dose and salicylic acid were involved in mitochondrial transport (MSP1, TOM5, ERV1, TIM21, TIM23). When looking QTLs for each condition, multiple cellular functions were found and all of them show relation to phosphoproteins. For the lower dose of Aspirin, the peak marker was located on chromosome IV at gene VHS1 (YDR247W) which encodes a cytoplasmic protein kinase (Simpson-Lavy, et al., 2017). 24 significant markers were linked to genes for phosphatase activity and 9 genes were involved in hydrolase activity. The peak marker for the higher dose of aspirin is in gene YPP1(YGR198W) on chromosome VII. Three genes (CIT2, IDH1, IDP1) are involved in 2-Oxicarboxylic acid metabolism pathway. For salicylic acid, in addition to the majority of genes involved in phosphatase activity, four genes (COG1, ALG12, NUP2, ATO3) clustered as related to transport and membranes.

Curcumin is a natural chemoprevention agent that can be extract from Curcuma species (Maulina, et al., 2019). Curcumin was applied at 2 concentrations to the F12 segregants at 200mM and 500mM. From the phenotype distribution shown in Section 3.3.1, the growth patterns are very different from each other. Although it is expected that the same agent would show a similar genetic pattern of association, the QTL result between these two conditions have few overlaps which reflects the different genetic architectures under different phenotype responses to the two doses. RRP7 (YCL031C) was shared between high and low doses of curcumin and is an essential gene involved in rRNA processing and exhibits responses to DNA replication stress (Tkach, et al., 2012). At 200mM, multiple functional annotation terms are found in the identified QTLs. The majority of markers were linked to genes related to phosphoprotein and hydrolase activity. 20 markers were in genes annotated to be involved in metal ion binding, such as SAL1(YNL083W) which binds Ca^{2+} ions (Laco, et al., 2010). In addition, 7

markers are in genes (PMS1, CSM2, RAD18, NEJ1, SLX8, SPT16, SLX1) that involved in responses to DNA damage and are involved in DNA repair. For the high dose at 500mM, 10 genes were involved in mitochondria (PFK1, BEM2, MMM1, EXO5, LMO1, SAM50, GUT1, OXA1, ATG32, ERG6) and 6 genes (PDE2, PFK1, EXO5, FAP1, HMX1, SDD4) also function in metal ion binding. Disulfiram, EPA and metformin are also used as chemoprevention agents. The only overlap between EPA and metformin are two consecutive QTL intervals linked to genes COS111(YBR203W) and TAF5(YBR198C). 6 genes overlapped between DSF and MET. Three of them are involved in nucleotide binding (NRP1, PTK2, IFM1).

DOX and matrine are both genotoxic agents. Combination treatment was also tested with DOX and matrine. The result shows that overlaps between DOX and matrine are related to metal ion binding and membranes. The combination treatment (DOX and matrine) shared more overlaps with DOX than matrine. A number of overlapped genes were involved in mitochondria between DOX and the combination treatment. In addition, the centromere region (CEN13) in chromosome XIII was overlapped, this is also being identified in the F1 analysis. Apart from CEN13, two more centromere regions were significant QTL intervals under DOX (CEN2, CEN3) and 7 genes (DAD1, HIR1, SGO1, STU2, NDC80, MCD1, PSH1) were located in the centromere regions. When mapping the QTLs under DOX, a further 19 genes were involved in DNA repair that including SMC5, RAD16, MSH6 which were found associated with other treatments. QTLs were also shared between DOX and Paraquat and these are involved in the regulation of transcription (BUR2, MED8, MSA1, SOK2, SWC3) and ATP binding (PRP22, SMC5, ACC1, MSE1, YPK3). For Paraquat, gene OTU1 was also involved in the regulation of transcription that has been located in previous QTL study for the same drug (Cubillos, et al., 2013). In addition, a number of QTLs are linked to integral components of membranes.

The most frequently occurring genes (those including pleiotropic QTLs that found more than 5 times among all agents conditions) were further annotated with nucleotide changes and amino acid changes (Table 3-5). The variant type for the SNPs around F12 segregants were computed through SNPeff that include missense and synonymous SNPs in the codings region as well as intergenic SNPs in non-coding regions. 16 genes were selected for multiple agents including 54 markers that located on 5 chromosomes. Among these selected genes, four of them (SMC5, SPT16, MSH6 and TOR1) responded to DNA damage. In addition, 34 out of 54 markers were typed into synonymous variants. Although these variants' nucleotide information was different, they hold the identical amino acid. A recent study reveals that synonymous mutation are not silent and can affect the stability of mRNA and the efficiency of the translation so that the individuals can have the different growth rate and fitness (Kristofich, et al., 2018). Moreover, 18 markers were typed into missense which means that different nucleotides codes for difference amino acid. Among these 18 missense variants, the alternative alleles which were different with reference for gene MSH6, PAM1, PXP1, SPT16, and BFA1 were only from the contribution of founder WA. For gene UTR1 and SMC5, the alternative alleles were only from founder Sake. In addition, two markers were typed as missense but contributed from different founders in gene NRP1, PDE1, CDC8 and MNN14.

Table 3-5 SNP markers in the overlapping genes

Marker	Gene	Gene Name	Nucleotide Position	Nucleotide	Founder Alleles	ALT Alleles	Substitution Type	Amino acid Position	Amino Acid
c04:0163031	YDL167C	NRP1	124	A>G	A-E; G-A,W,S	E	missense	42	Thr>Ala
c04:0162933	YDL167C	NRP1	222	C>T	T-S; C-A,E,W	S	synonymous	74	Asn>Asn
c04:0162745	YDL167C	NRP1	410	C>T	T-W; C-A,E,S	W	missense	137	Ser>Phe
c04:0162498	YDL167C	NRP1	657	G>T	T-E; G-A,W,S	E	synonymous	219	Ala>Ala
c04:0640367	YDR097C	MSH6	3471	T>A	T-W; A-A,E,S	W	missense	1157	Asp>Glu
c04:0642755	YDR097C	MSH6	1083	C>T	C-S; T-A,E,W	S	synonymous	361	Arg>Arg
c04:0643772	YDR097C	MSH6	66	A>G	G-S; A-A,E,W	S	synonymous	22	Gln>Gln
c04:0960374	YDR251W	PAM1		G>A	A-S; G-A,E,W	S	intergenic		
c04:0960570	YDR251W	PAM1		C>T	T-W; C-A,E,S	W	intergenic		
c04:0962232	YDR251W	PAM1	1650	A>G	G-A; A-E,S,W	A	synonymous	550	Arg>Arg
c04:0962280	YDR251W	PAM1	1667	A>G	G-W; A-A,E,S	W	missense	556	Gln>Arg
c05:0119201	YEL020C	PXP1	1099	G>A	G-E,S; A-W,A		missense	367	Ala>Thr
c05:0119388	YEL020C	PXP1	912	G>A	G-E; A-A,S,W	E	synonymous	304	Gly>Gly
c05:0120062	YEL020C	PXP1	238	G>T	T-W; G-A,E,S	W	missense	80	Ala>Ser
c07:0099326	YGL207W	SPT16	358	G>A	A-W; G-A,E,S	W	missense	120	Val>Ile
c07:0099361	YGL207W	SPT16	393	G>A	G-S; A-A,E,W	S	synonymous	131	Val>Val
c07:0100684	YGL207W	SPT16	1716	A>G	G-E; A-A,S,W	E	synonymous	572	Pro>Pro
c07:0101050	YGL207W	SPT16	2082	A>G	A-S; G-A,E,W	S	synonymous	694	Val>Val
c07:0101815	YGL207W	SPT16	2847	T>C	T-S; C-A,E,W	S	synonymous	949	Gly>Gly
c07:0079586	YGL223C	COG1	780	G>T	G-S; T-A,E,W	S	synonymous	260	Arg>Arg
c07:0079943	YGL223C	COG1	423	T>C	C-A; T-E,S,W	A	synonymous	141	Asn>Asn
c07:0080345	YGL223C	COG1	21	G>A	G-E,S; A-W,A		synonymous	7	Leu>Leu
c07:0035709	YGL248W	PDE1	57	A>G	G-W; A-A,E,S	W	synonymous	19	Gly>Gly
c07:0035993	YGL248W	PDE1	341	C>G	G-A; A-E,S,W	A	missense	114	Thr>Ser
c07:0036448	YGL248W	PDE1	796	G>A	A-E; G-A,W,S	E	missense	266	Glu>Lys
c09:0089987	YIL137C	TMA108	2802	C>T	T-S; C-A,E,W	S	synonymous	934	Ser>Ser
c09:0091352	YIL137C	TMA108	1437	T>C	T-A; C-E,S,W	A	synonymous	479	Ile>Ile
c09:0091400	YIL137C	TMA108	1389	A>G	G-S; A-A,E,W	S	synonymous	463	Pro>Pro
c09:0092065	YIL137C	TMA108	724	A>G	A-A; G-E,S,W	A	missense	242	Ile>Ile
c10:0527002	YJR049C	UTR1	1475	C>T	T-E; C-A,W,S	E	missense	492	Thr>Ile
c10:0527586	YJR049C	UTR1	891	A>G	G-A; A-E,S,W	A	synonymous	297	Thr>Thr
c10:0528102	YJR049C	UTR1	375	G>A	G-E,W; A-A,S		synonymous	125	Leu>Leu
c10:0528153	YJR049C	UTR1	324	G>A	A-W; G-A,E,S	W	synonymous	108	Ala>Ala
c10:0534212	YJR053W	BFA1	186	G>A	A-E; G-A,W,S	E	synonymous	62	Thr>Thr
c10:0534224	YJR053W	BFA1	198	T>C	A-E; G-A,W,S	E	synonymous	66	Asn>Asn
c10:0534536	YJR053W	BFA1	510	G>A	A-A; G-E,S,W	A	synonymous	170	Arg>Arg
c10:0534618	YJR053W	BFA1	592	G>A	A-W; G-A,E,S	W	missense	198	Glu>Lys

c10:0535622	YJR053W	BFA1	1596	T>C	C-S; T-A,E,W	S	synonymous	532	Phe>Phe
c10:0535642	YJR053W	BFA1	1616	T>C	C-W; A-A,E,S	W	missense	539	Ile>Thr
c10:0535682	YJR053W	BFA1	1656	G>A	A-E; G-A,W,S	E	synonymous	552	Thr>Thr
c10:0544361	YJR057W	CDC8	300	G>A	A-W; G-A,E,S	W	synonymous	100	Val>Val
c10:0544624	YJR057W	CDC8	563	G>A	A-E; G-A,W,S	E	missense	188	Gly>Asp
c10:0544677	YJR057W	CDC8	616	A>G	G-S; A-A,E,W	S	missense	206	Thr>Ala
c10:0545026	YJR057W	CDC8	150	G>A	T-W; C-A,E,S	W	synonymous	50	Gln>Gln
c10:0550553	YJR061W	MNN14	43	T>A	A-A; G-E,S,W	A	missense	15	Ser>Thr
c10:0551614	YJR061W	MNN14	1104	A>G	A-E; G-A,W,S	E	synonymous	368	Glu>Glu
c10:0552137	YJR061W	MNN14	1627	G>A	A-S; G-A,E,W	S	missense	543	Val>Ile
c10:0554464	YJR062C	NTA1	1068	G>A	G-E; A-A,S,W	E	synonymous	356	Glu>Glu
c10:0559883	YJR066W	TOR1	468	T>C	T-E; C-A,W,S	E	synonymous	156	Pro>Pro
c10:0559922	YJR066W	TOR1	507	A>G	A-E; G-A,W,S	E	synonymous	169	Leu>Leu
c10:0560828	YJR066W	TOR1	1413	A>G	G-W; A-A,E,S	W	synonymous	471	Leu>Leu
c10:0562764	YJR066W	TOR1	3349	T>C	C-S; T-A,E,W	S	missense	1117	Ser>Pro
c10:0562775	YJR066W	TOR1	3360	G>A	G-E,W; A-A,S		synonymous	1120	Arg>Arg
c10:0563075	YJR066W	TOR1	3660	T>C	C-S; T-A,E,W	S	synonymous	1220	Ser>Ser
c10:0565712	YJR066W	TOR1	6297	G>A	A-S; G-A,E,W	S	synonymous	2099	Val>Val
c10:0565724	YJR066W	TOR1	6309	G>A	A-S,A; G-W,E		synonymous	2103	Lys>Lys
c15:0260399	YOL034W	SMC5	477	G>A	G-E; A-A,S,W	E	synonymous	159	Glu>Glu
c15:0260841	YOL034W	SMC5	919	G>T	T-E; G-A,W,S	E	missense	307	Ala>Ser
c15:0261959	YOL034W	SMC5	2037	A>C	C-W; A-A,E,S	W	synonymous	679	Ala>Ala

3.4 Two-dimensional genome scan

Epistasis QTL scan were performed for F12 segregants under DOX treatment. The scantwo pipeline was used to identify the significant QTL pairs with interaction effect under shmootl and R/qtl (Broman, et al., 2003). 1000 times permutation test and significant level 0.001 were used for the analysis. Due to the large sets of QTL pairs, only the pairs with highest LOD in each chromosome and not shown in the single scan were summarised. Apart from the QTLs involved the present genes in additive effect, almost all the largest interaction effect were targeted the locus at 204507 base pair on chromosome III associated with multiple markers. These markers were missed in the single QTL scan. This locus is linked to gene TAF2 which is housekeeping factor and involved in RNA polymerase II transcription (Tora, 2002), (Weiss, et al., 2018). Phenotype effect were further compared under two markers' allele combination (Figure 3-6). Although it is not known for multiple parental lines whether the epistasis effect is positive or negative, Figure 3-6 shows the clear non-additive effect between TAF2 and other genes allele. Table 3-7 summarised the QTLs interaction that contain TAF2. Among the markers interacting with TAF2, 5 genes (LDS2, CAK1, CIN8, DCR2, ALK1) were shown relatedness with ATP binding and cytoplasm that might indicates coordination between them involved in response to stress (Shalem, et al., 2011). The founders' alleles were further traced back to check if there is any founder predominance. SA an WA present in most alternative alleles compared to reference and NA, WE also contributed some markers alternative allele. For marker c03:0204507, the allele T among segregants are inherited from NA, WE and allele A inherited from WA and SA. For marker c16:0026869, the allele A among segregants are inherited from WA, NA and allele G inherited from WE, SA (Table 3-7).

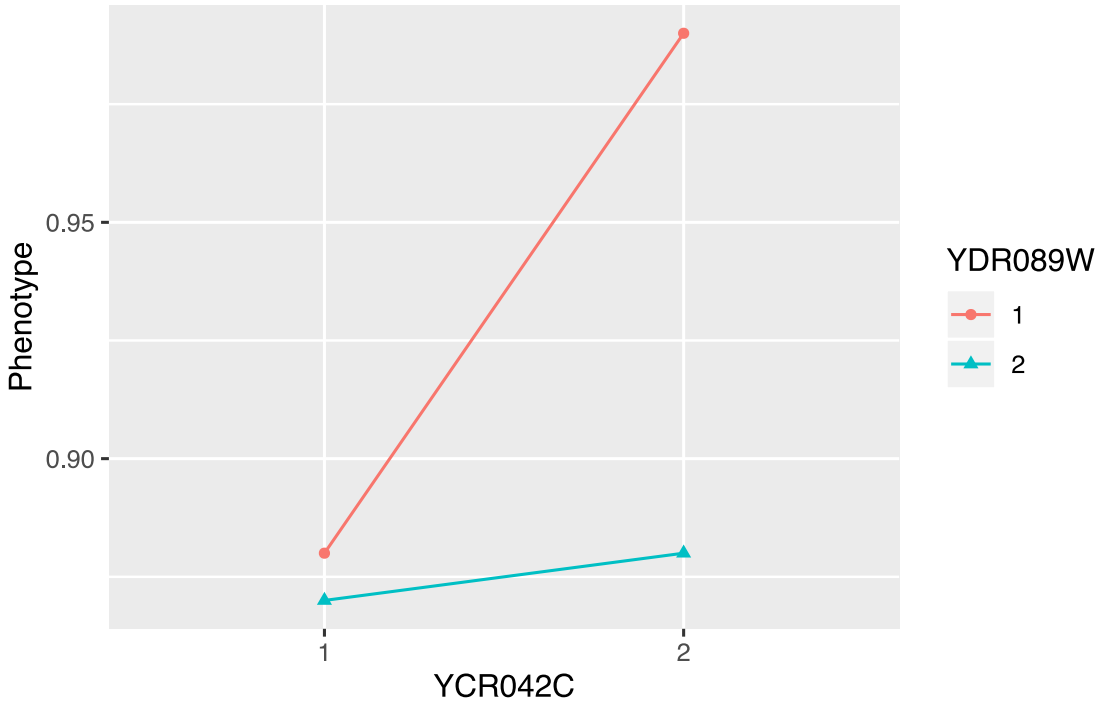
Table 3-6 Lists of two genome scan results with YCR042C

Epistasis CHR1 and CHR2 are the location of chromosome for each marker. Peak 1 and Peak 2 are the peak marker ID. Peak 1 gene and Peak 2 gene are the genes that peak marker located in. Full LOD is the LOD scores for two genome scan.

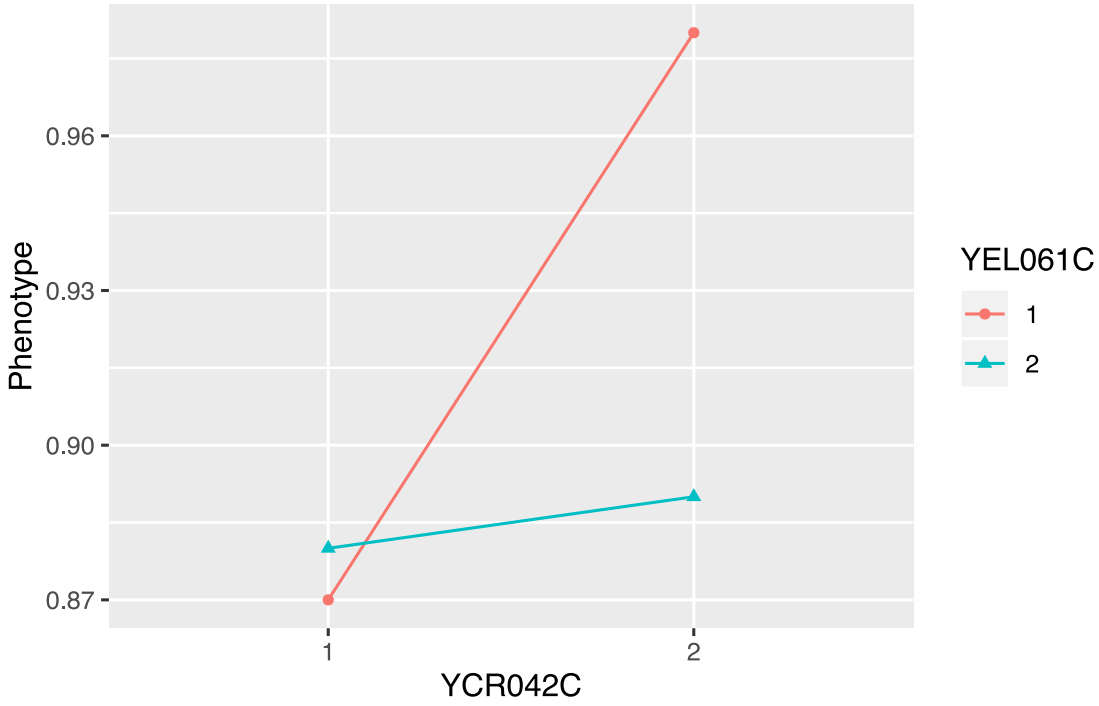
CHR1	CHR2	Full LOD	Peak 1	Peak 1 gene	Peak 2	Peak 2 gene
3	4	50.80712	c03:0204507	YCR042C	c04:0624565	YDR089W
3	5	41.53163	c03:0204507	YCR042C	c05:0037041	YEL061C
3	6	40.67485	c03:0204507	YCR042C	c06:0079038	YFL029C
3	7	48.10782	c03:0204507	YCR042C	c07:0455007	YGL021W
3	9	45.92585	c03:0204507	YCR042C	c09:0197949	YIL088C
3	10	49.89025	c03:0204507	YCR042C	c10:0162714	YJL132W
3	12	48.68101	c03:0204507	YCR042C	c12:0848842	YLR361C
3	14	43.61631	c03:0204507	YCR042C	c14:0043498	YNL315C
3	15	48.90262	c03:0204507	YCR042C	c15:0242488	YOL047C
3	16	51.60865	c03:0204507	YCR042C	c16:0026869	YPL272C
1	3	42.07359	c01:0169629	YAR015W	c03:0204507	YCR042C

Table 3-7 list of alleles of markers with founder information

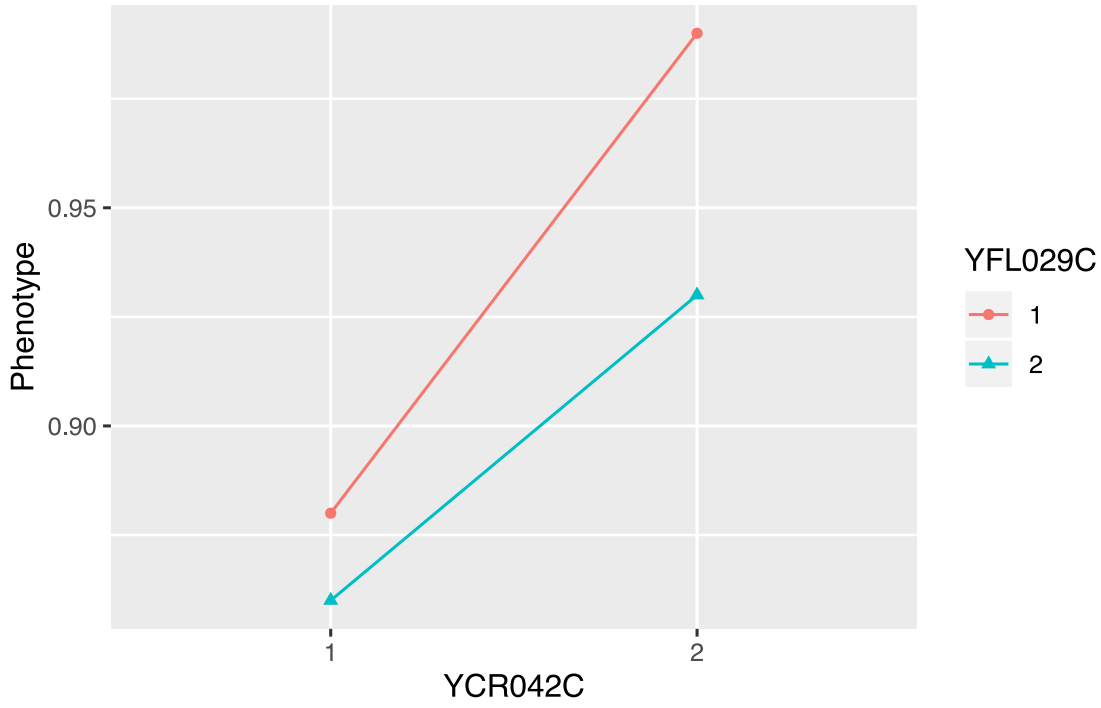
Marker	Gene feature	Allele Frequency	REF	ALT
c04:0624565	YDR089W	0.22	T	SA-A
c05:0037041	YEL061C	0.89	C	WA-T
c06:0079038	YFL029C	0.36	G	WA-A
c07:0455007	YGL021W	0.31	G	SA-A
c09:0197949	YIL088C	0.65	A	SA-G
c10:0162714	YJL132W	0.26	G	NA-A
c12:0848842	YLR361C	0.75	A	WE-G
c14:0043498	YNL315C	0.71	C	NA-T
c15:0242488	YOL047C	0.25	T	WA-C
c16:0026869	YPL272C	0.45	G	WA, NA-A
c01:0169629	YAR015W	0.28	T	SA-C
c03:0204507	YCR042C	0.57	T	WA, SA -A



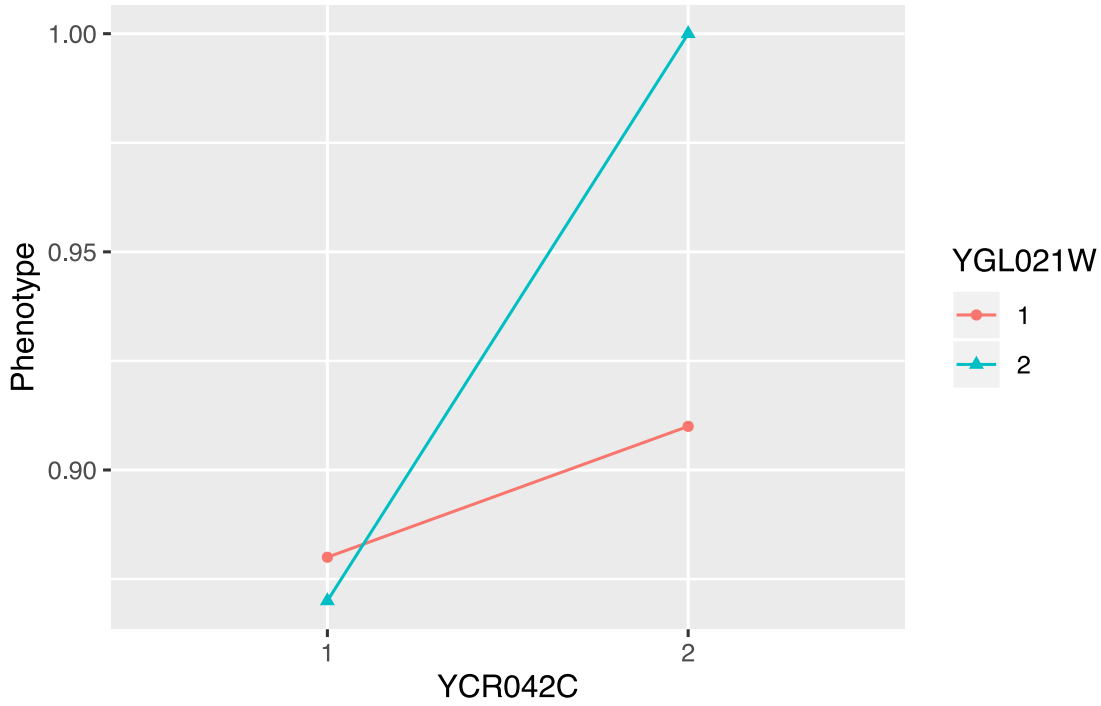
c03:0204507 × c04:0624565



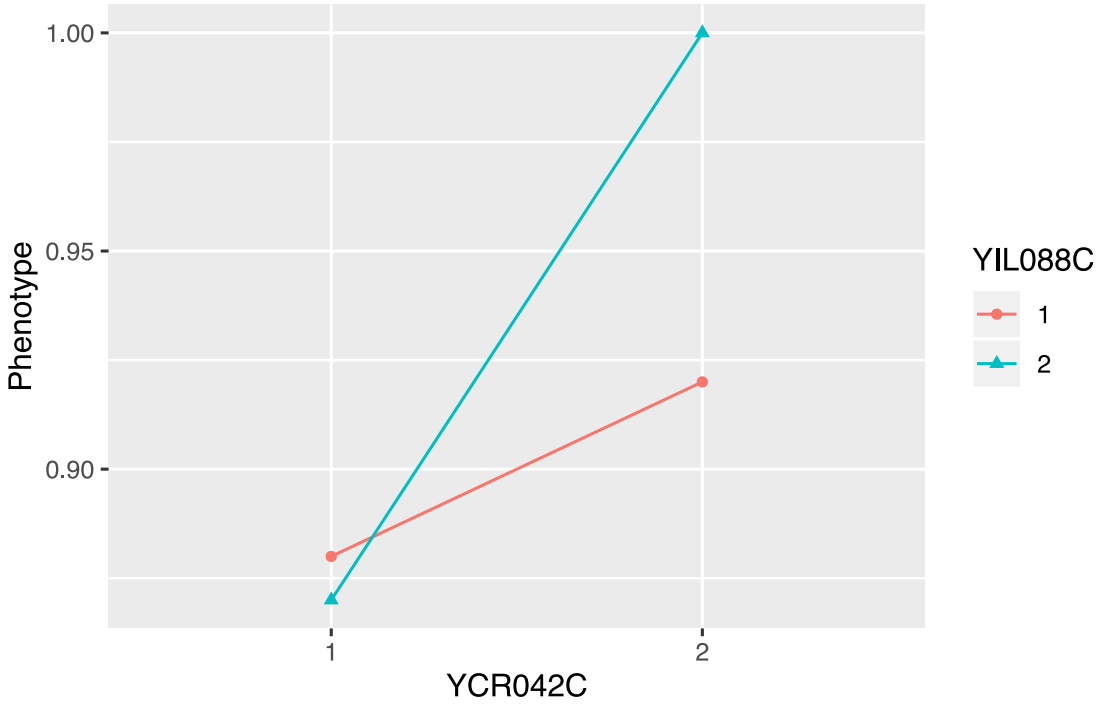
c03:0204507 × c05:0037041



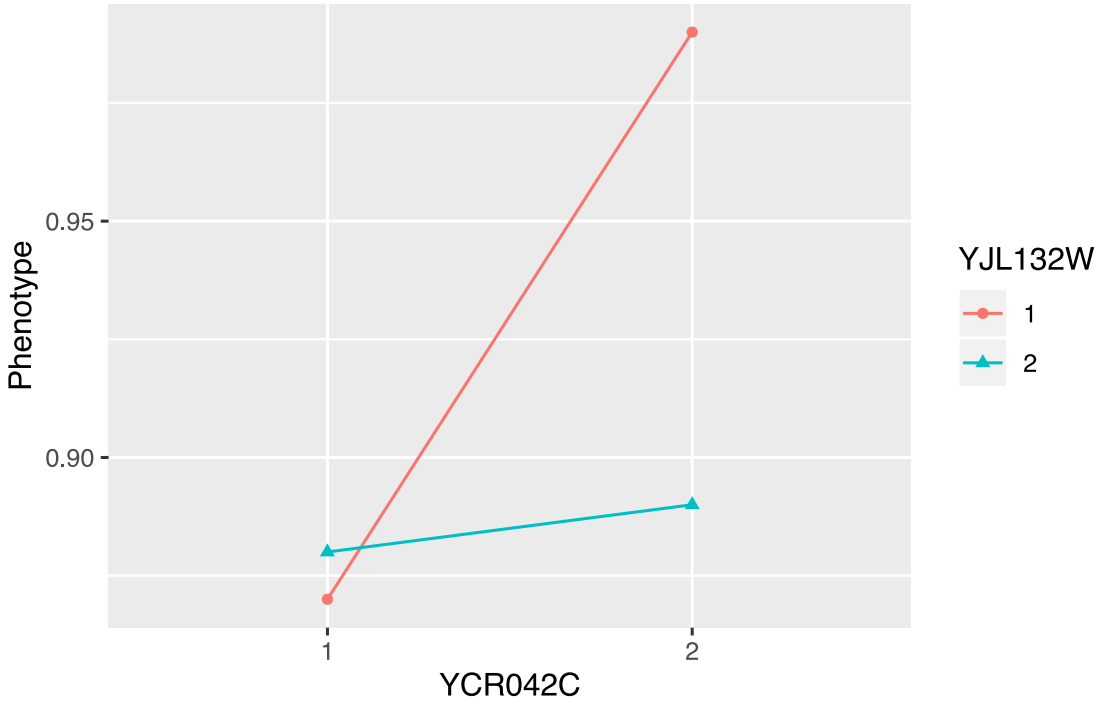
c03:0204507 × c06:0079038



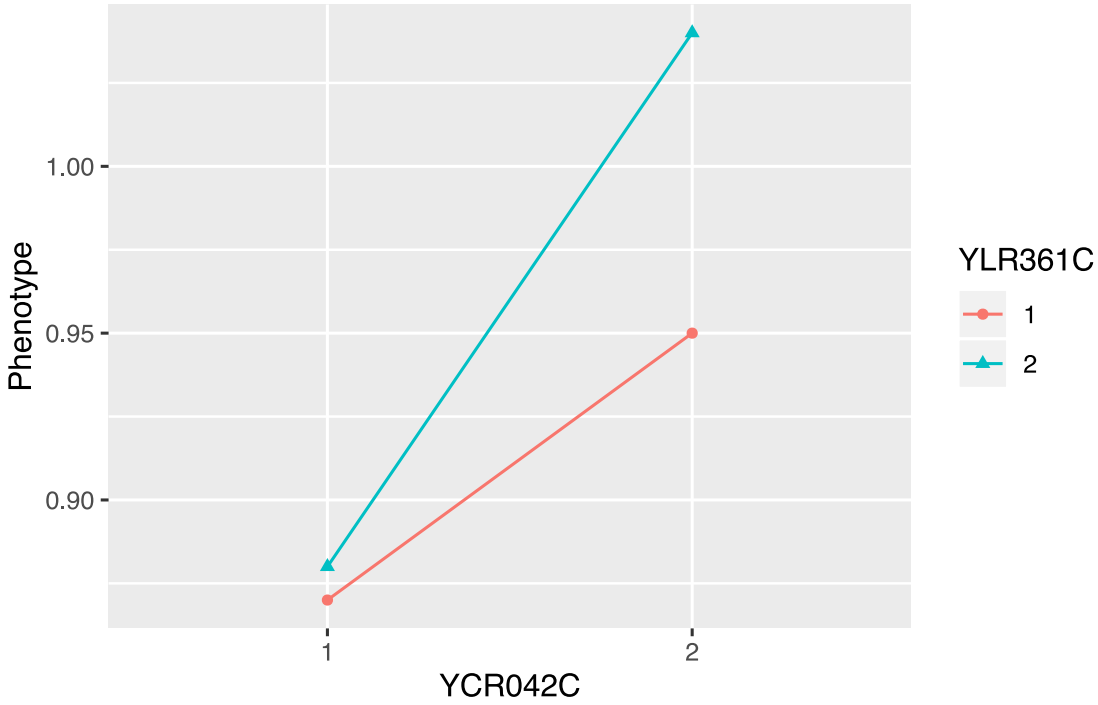
c03:0204507 × c07:0455007



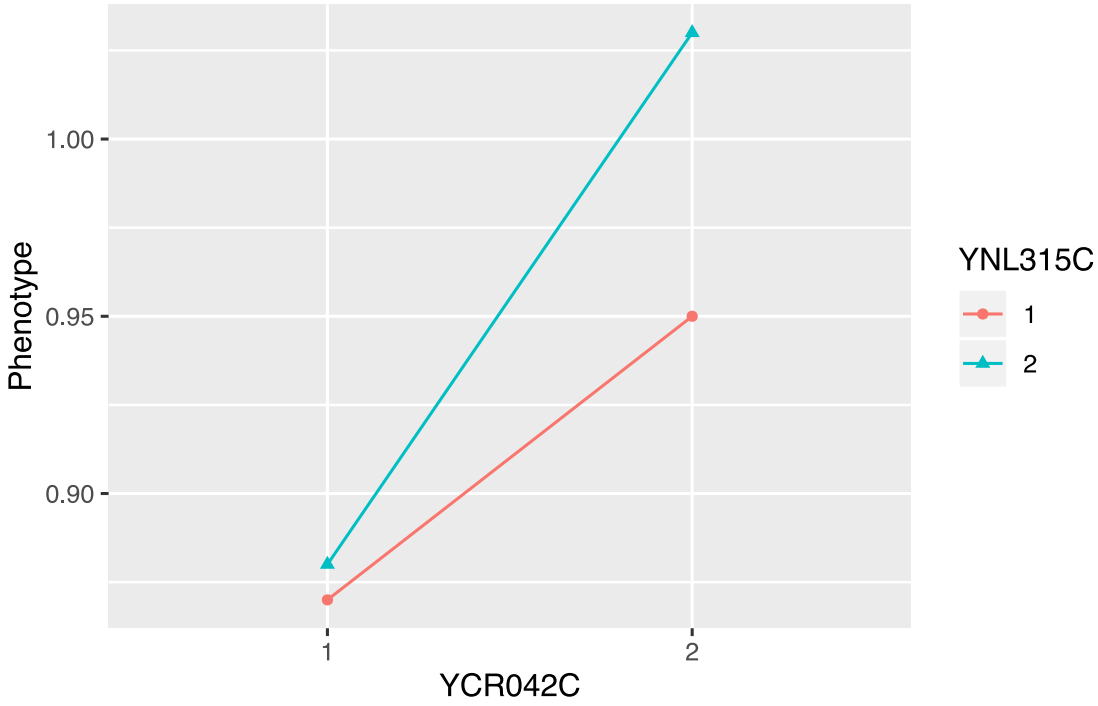
c03:0204507 × c09:0197949



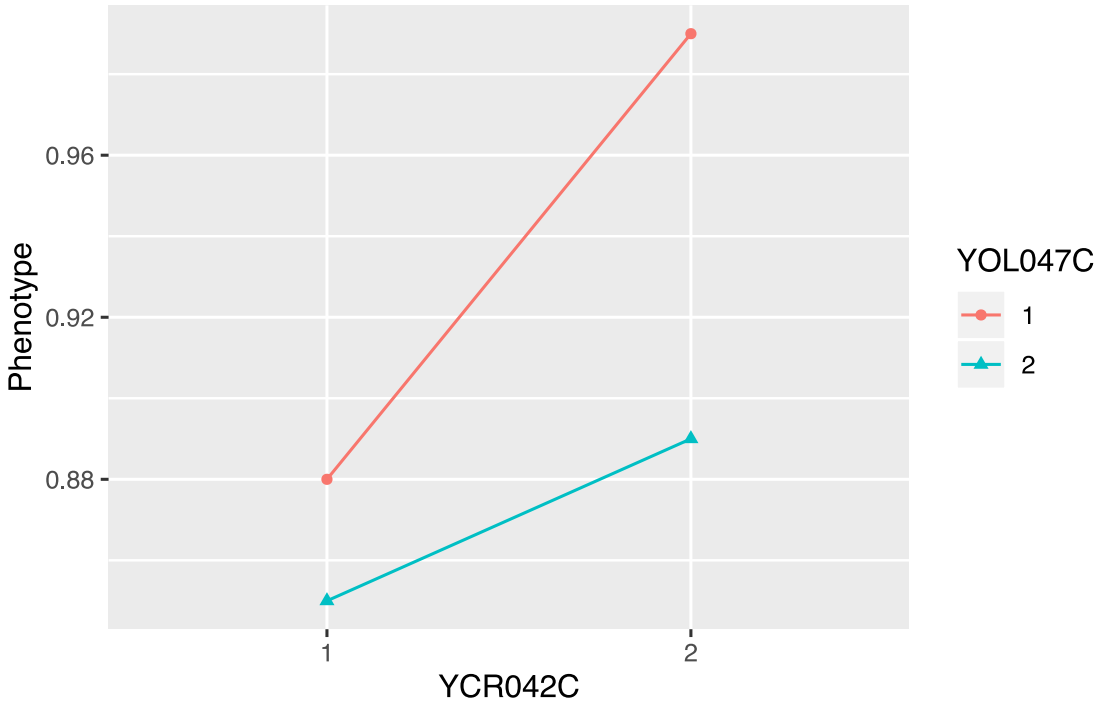
c03:0204507 × c10:0162714



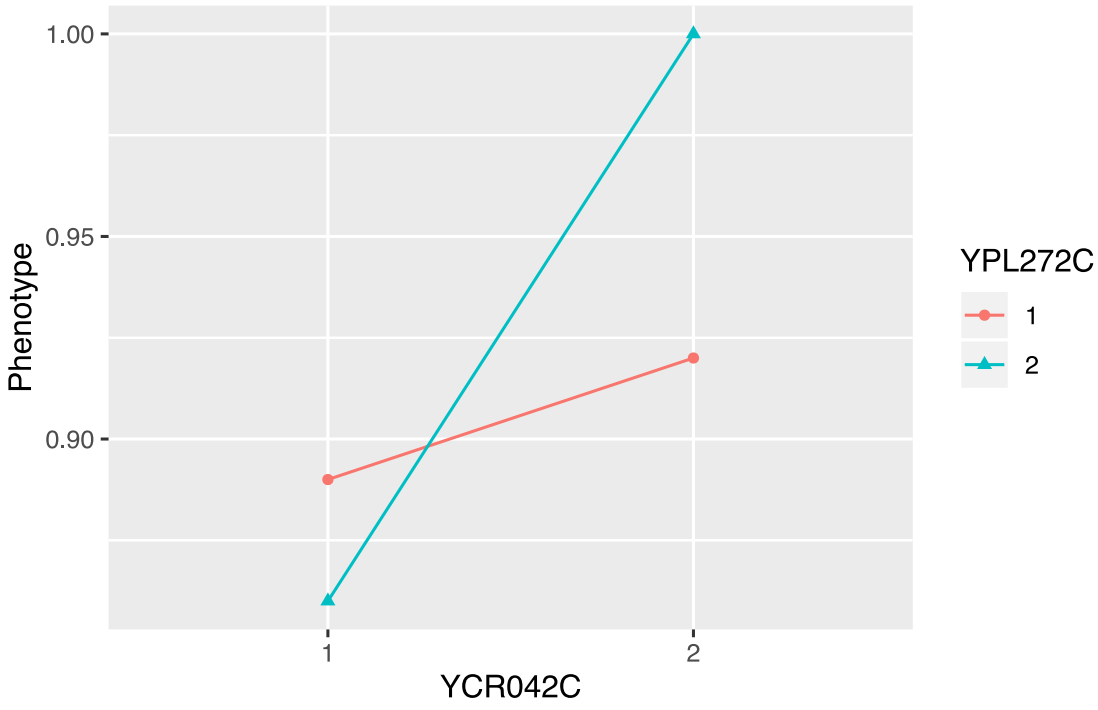
c03:0204507 × c12:0848842



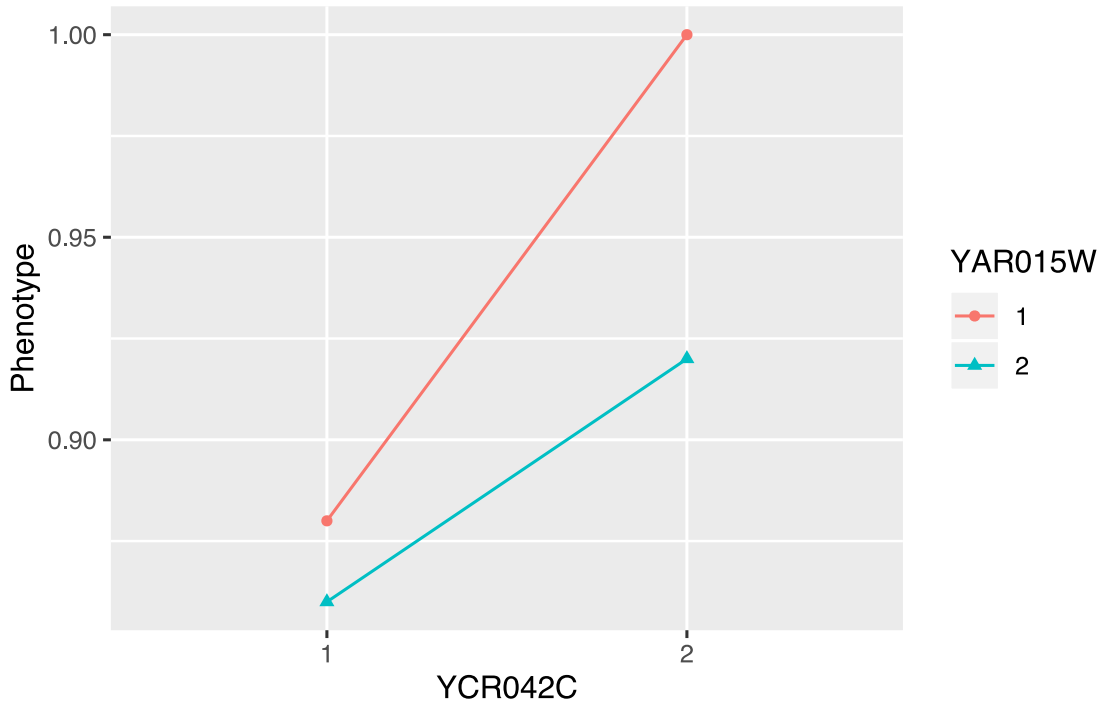
c03:0204507 × c14:0043498



c03:0204507 × c15:0242488



c03:0204507 × c16:0026869



c03:0204507 × c01:0169629

Figure 3-6 Effect plots of marker alleles

Average phenotype value was calculated for each combination of allele between two markers. X axis is the allele for c03:0204507, y axis is treatment ratio. Red lines are for allele 1 in the other marker and blue lines are for allele 2

3.5 Discussion

In this chapter, the data for 19 different treatments on 15217 bi-allelic markers of the 12th generation population were analysed. Growth phenotypes were taken at max slope and the growth ratio in the end point of each experiment compared between control and treatment. In F1 population of the 6 groups used in Chapter 2, the genomes for each group have two parental backgrounds. In F12 population, four parental backgrounds were all included. Moreover, multiple rounds of intercross were applied to produce F12 population which gave higher recombination chances than F1s (Cubillos, et al., 2013). This provides the genome of F12 population more diverse because of the genomes are in high resolution with more genotype features than F1. The genotype data for each group in F1 has 96 strains for QTL analysis with 179 markers where F12 has 166 strains with 15217 markers. The large number of markers makes F12 population with the high sensitivities for QTL mapping and can map the QTLs finer to locate genes.

This emphasis of this chapter is on potential QTL overlaps among different agents in F12 segregants. QTL mapping in F12 multi-parental lines demonstrates the ability to explore the larger range of phenotype distribution and genetic diversity. Under DOX treatment, the total QTL number detected in the F12 population is greater than the F1 population. A limited number of overlaps presented between the F1 and F12 analyses. Epistasis analysis were performed for DOX with F12 segregants. Housekeeping gene TAF2 presented in the two-dimensional QTL scan but missed in the single QTL output. TAF2 also targeted pathways of degration as F1 targeted. The functions of QTLs are overlapping with F1, such as DNA damage pathway related genes. This suggests that the phenotype is likely affected by the coordination for the regulation of multiple genes. Also, the change of environment might affect the strain growth with complex mechanisms, even the similar agents with different concentration the response could achieved differently. The high recombination rate of F12

population not only increases the genetic diversity, but also improves the detection of QTLs.

Under all the different agents, there is no single gene that appears as overlapping QTL. The complexity underlying QTLs affecting phenotypes makes the possibility of having one gene accounting for a certain phenotype very low. Nevertheless, overlapping gene functions exist across agents that have the similar effects. For all genotoxin agents, DNA damage repairing related genes were identified. This suggests that genetic background for F12 segregants have distinct response to the change of environment, which are important for implicating QTL results. The result also means that the overlapping genes might have a seeming unrelated function to the response, but the function could affect the organism response indirectly. The most nested overlapping gene NRP1 among all agents is an RNA binding protein which might be involved in ribosome biogenesis. Ribosome biogenesis plays an essential role in growth control and the coordination in the cell-cycle in yeast and all life (Lempiäinen & Shore, 2009). A recent study confirmed that ribosome biogenesis has an essential association to cancer with multiple interactions with other factors (Penzo, et al., 2019). This suggests that ribosome biogenesis could be a causal factor underlying molecular mechanisms of cancer. It is interesting to notice that for the same agent, different concentrations can lead to very different QTL detection sets. In a previous study, the drug haloperidol is highly dose-dependent which caused the genetic loci identified in only low doses or high doses (Wang & Kruglyak, 2014). This suggests that different doses of treatment could play an important role in cell cycle and growth control. This would affect processes like personalised medicine, in a way that a certain phenotype could be triggered by accurately controlling the dosage of a medicine.

The analysis presented here only considered the max speed and the end point growth at 64 hours. The growth of a few strains under high concentrations of agents were extremely suppressed without growth and some of the strains reached the same level in the end time but exhibited different patterns during

growth. The change in dynamic time cannot be considered only by the max speed and the end point growth rate. It is worthwhile to take consideration of multiple time points so that if the QTL detection during each stage is different from the end point, the underlying genetic information could be further gained. The relationship between genes that are involved in different stages might be identified through the comparison over many time points. In Chapter 5, temporal QTL analysis is performed on cancer therapy X-ray radiation.

Chapter 4 Identifying potential causal variants within QTL regions

4.1 Introduction

Through the analysis performed in Chapter 3, a large number of QTL intervals in the F12 generation under different agents was identified at the gene level with narrow QTL intervals on average of 2400 bp wide. The large number of QTL mappings allows the identification of shared genomic features or regions associated with yeast growth diversity under different agents. However, when there is a large number of QTL intervals identified for an agent, it is challenging to distinguish the causative markers from genes. By setting the level of significance strictly, only top rank markers with large LOD values will be identified. In addition, it is possible that the SNP markers were strongly correlated with high effect markers located in the adjacent region, resulting in indirect but significant associations between these adjacent locus and phenotype changes. The causal variation has a risk of not being the top-ranking significant markers with large LOD values having small effects in reality. Validation of all candidate genes in the intervals is unrealistic. Although the 12^{th} generation multi-parental segregants have greatly reduced the linkage disequilibrium between adjacent genes among the population, there are still haplotype blocks spanning one or more genes. When these markers were screened as genotype attributes to find associations with phenotypes, QTL analysis is not able to distinguish similar distributed data with different actual effects.

For better understanding the correlation between the genotype marker in the F12 population, three types of correlation tests (absolute Pearson, Centralised Normalised Hamming Correlation Coefficient, and Relative Information Gain) were applied on the bi-allelic genotype attributes. From these correlation tests, moderate linkage clusters among adjacent regions were identified along chromosomes. A few of strong correlation were also shown between markers

located on different chromosomes. To solve this complexity and to locate the causative genes for validation, a segmentation based detection method was proposed to determine the causal association of a genomic region and pick the peak that most associated with the trait in the cluster for yeast growth based on the QTL output. Causal QTL mapping was performed on the four genotoxin agents HU, MMS, Phleomycin and Paraquat for the end growth rate where large numbers of QTLs were identified in the previous chapter.

4.2 Materials and Methods

4.2.1 Phenotyping and Genotyping

Phenotype data and genotype data were the identical sets that explained in Chapter 3 for 12^{th} generation segregants.

4.2.2 Marker correlation analysis among F12 segregants

For evaluating the correlation between markers in F12 segregants, three distance methods: Pearsonss correlation coefficient, centralised normalised Hamming correlation coefficient and Relative Information Gain were computed. These are described for the analysis of marker interaction in (Mirkes, et al., 2015).

For the marker genotype data in F12 segregants, two random values X and Y with values 1 and 2 for representing bi-allelic markers. Number of observations in this data in n. The i^{th} observation is (x_i, y_i) and following notation:

- n_x is the number of ones in all observations of random variable X.
- n_y is the number of ones in all observations of random variable Y.
- n_1 is the number of observations with $x_i = y_i = 1$.
- n_2 is the number of observations with $x_i = y_i = 2$.
- p_{xy} is the fraction of observations with $x_i = y_i = 1$.
- p_x is the fraction of observations with $x_i = 1$.

- p_y is the fraction of observations with $y_i = 1$.

(a) The absolute Pearson's correlation coefficient

For two random variables, the PCC is

$$\rho = \frac{cov(X,Y)}{SD(X)SD(Y)}$$

Then the absolute PCC is

$$r = \left| \frac{\frac{n_1}{n} - p_x p_y}{\sqrt{p_x(1 - p_x)p_y(1 - p_y)}} \right|$$

(b) Centralised Normalised Hamming Correlation Coefficient (CNHCC)

CNHCC can be wrote as

$$CNHCC(x,y) = 4(p_{xy} - p_x p_y)$$

The absolute CNHCC is

$$h = |CNHCC(x,y)|$$

(c) Relative Information Gain (RIG)

RIG is based on the entropy function which can be defined as

$$Entropy(X) = -p_x \log p_x - (1 - p_x) \log(1 - p_x)$$

$$RIG = \frac{IG(X|Y)}{Entropy(X)} = \frac{Entropy(X) + Entropy(Y) - Entropy(XY)}{Entropy(X)}$$

4.2.3 QTL analysis and causal QTLs detection pipeline

QTL analysis results were the single scan outputs for end point growth rate obtained from the chapter 3 for HU, Phleomycin, MMS and Paraquat. LOD score for each marker were used as input for identifying clustering. Changepoint analysis is used for detecting single or multiple changes within a given sequence which is matching the aim of this study. The analysis was performed by changepoint package in R with function cpt.var with method PELT algorithm (Killick & Eckley, 2014). The peak markers in the each detect clusters were selected for effect estimation.

4.3 Results

4.3.1 Correlation shown within chromosome

For the correlation test between different markers, three types of distance information Pearson distance, Hamming distance and Relative Information Gain (RIG) were applied. The clusters with obvious relationship are displayed for each distance method and they exhibited similar cluster patterns (Figure 4-2). As the same information could be shown for the correlation of markers, only the Person correlation was used in the summary for this section. The average correlation for each chromosome level (e.g. chromosome I vs chromosome II) was calculated and shows that markers along the same chromosome are more correlated than markers between chromosomes. Among all the comparisons, chromosome I and chromosome III showed strong correlation among markers within with an average of 0.14 (Figure 4-1). On chromosome I, markers clustered into 4 major correlation groups where the adjacent regions exhibited higher association of the markers than those located in non-adjacent regions. This shows that the major clusters could be potentially consisted of nested subgroups. The nesting could be obviously detected in chromosome III and chromosome VI (Figure 4-3). Notably, the largest chromosomes in S. cerevisiae are chromosome IV, XV and VII. The average correlation for these chromosomes is not obviously high, but the adjacent region also shown relationships into clusters (Figure 4-4). In addition, the markers from different chromosomes were also shown high correlation where two regions were detected between the subtelomere region in the right arm of chromosome X (c10:0737966-0737980) and the subtelomere region in the left arm of chromosome IV (c04:0018180-0027090). The genes involved in these two regions were MPH3 and AAD4.

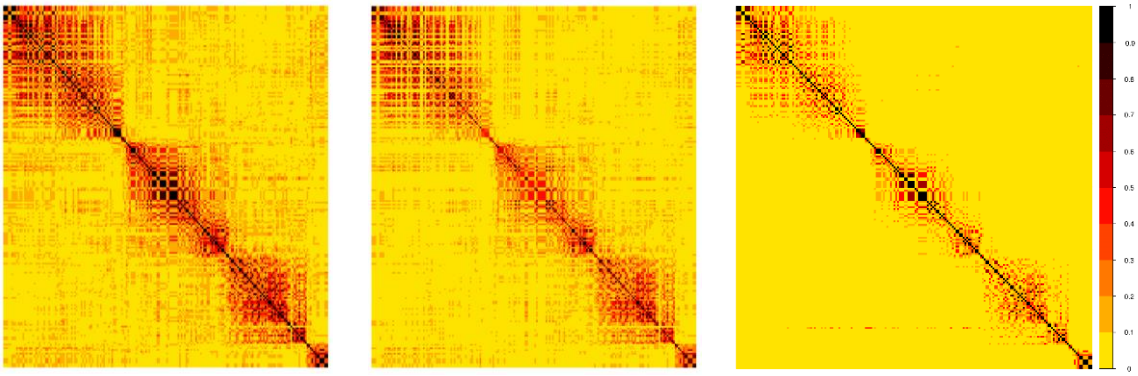


Figure 4-1 Heatmap of correlation between each marker in Chromosome I

Three distance methods were calculated and interpreted through R. Color from red to black indicated high correlation between the pair of markers.

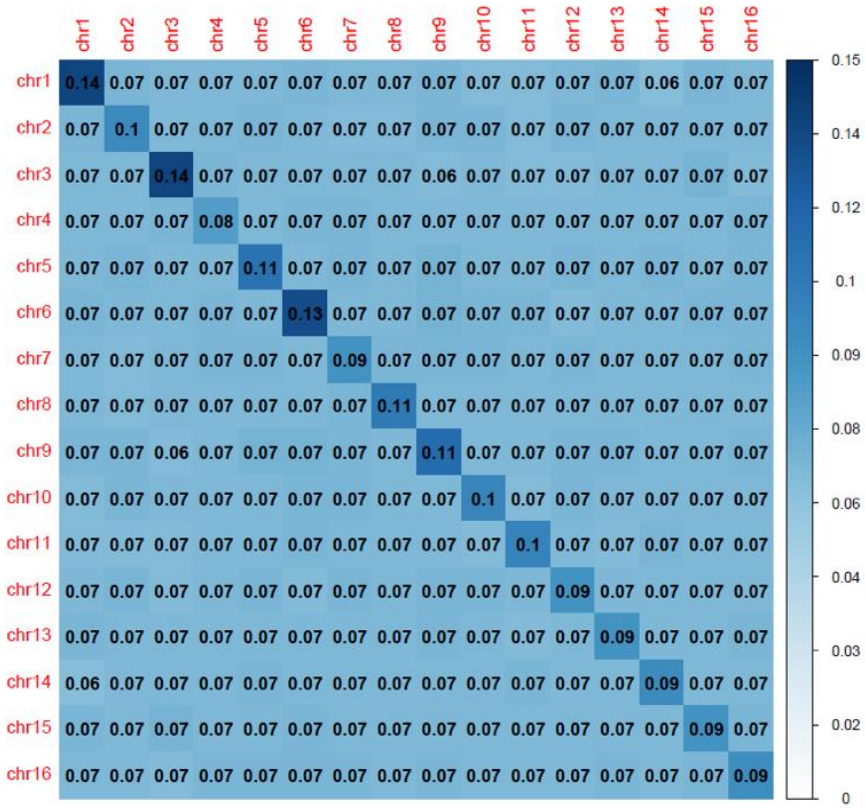
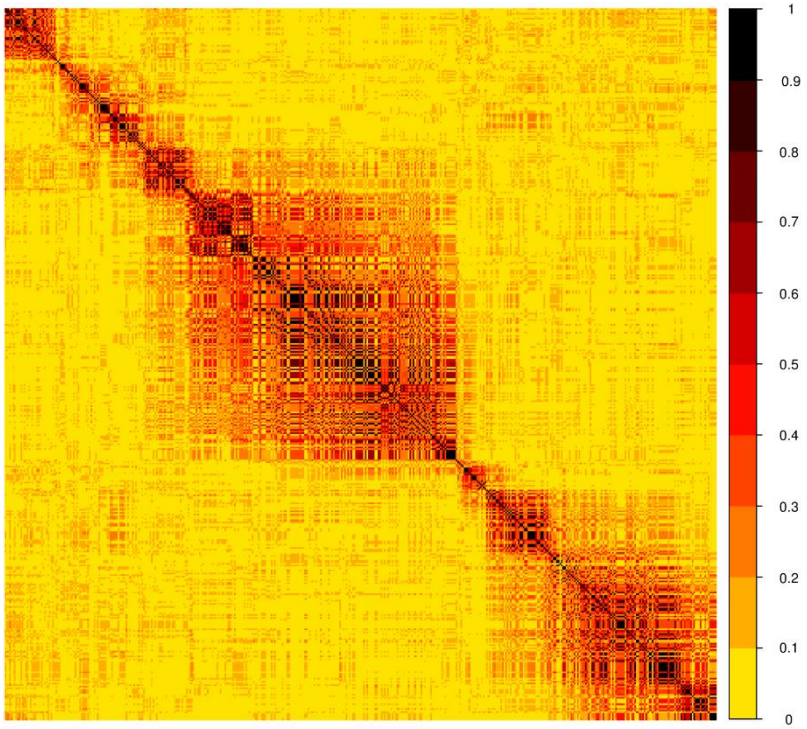
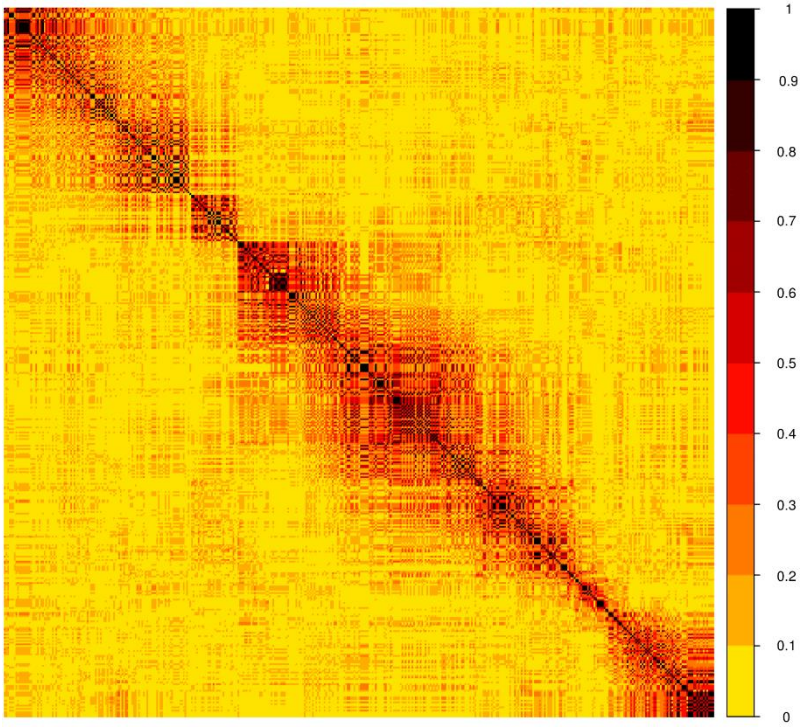


Figure 4-2 Average correlation score across chromosomes

For the same chromosome pair, same markers correlation was not counted into. The score from 0 - 0.15 coloured in white to dark blue. The grid with darker colour indicated a higher degree of correlation.

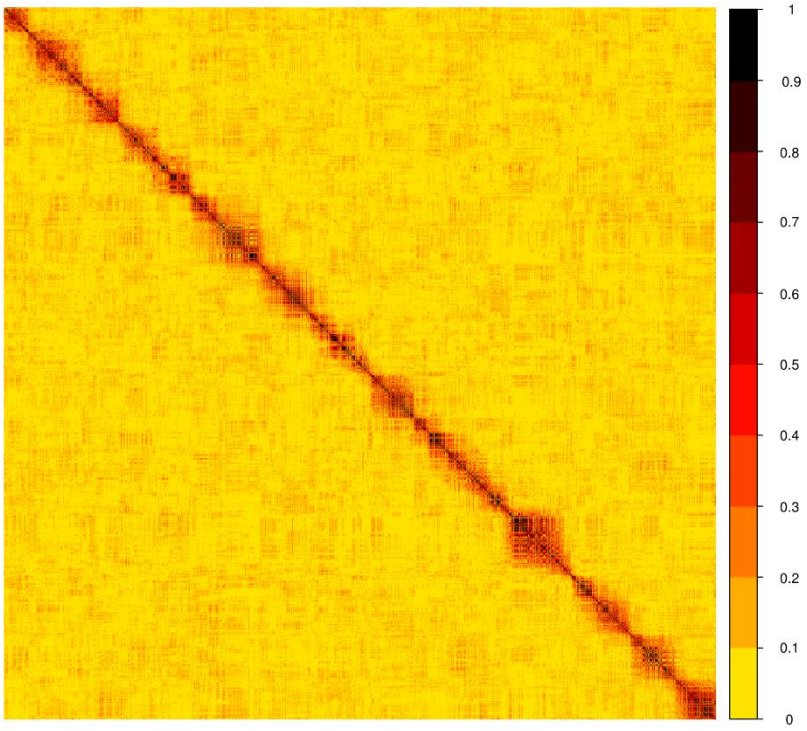


(a) chromosome III

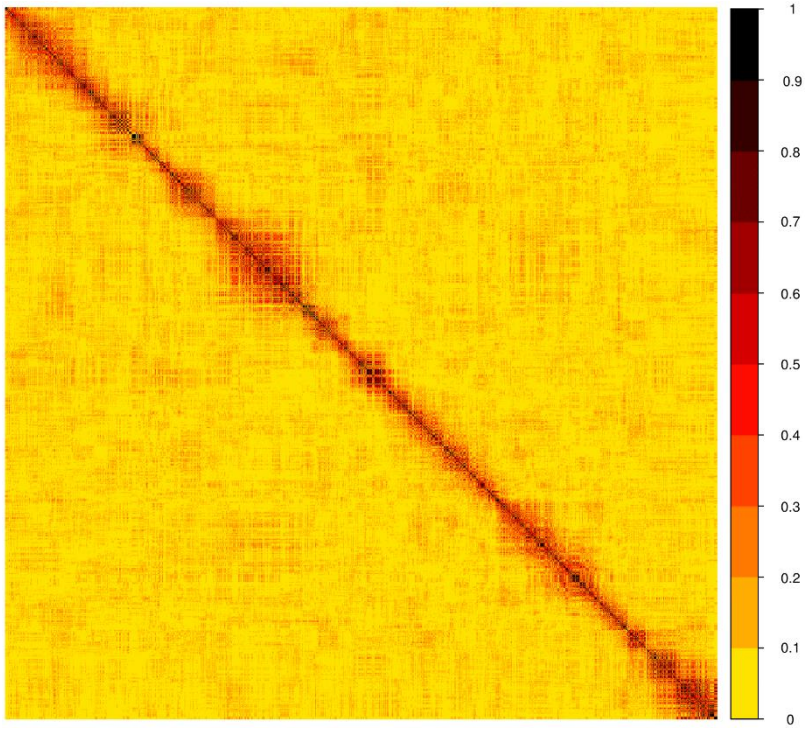


(b) chromosome VI

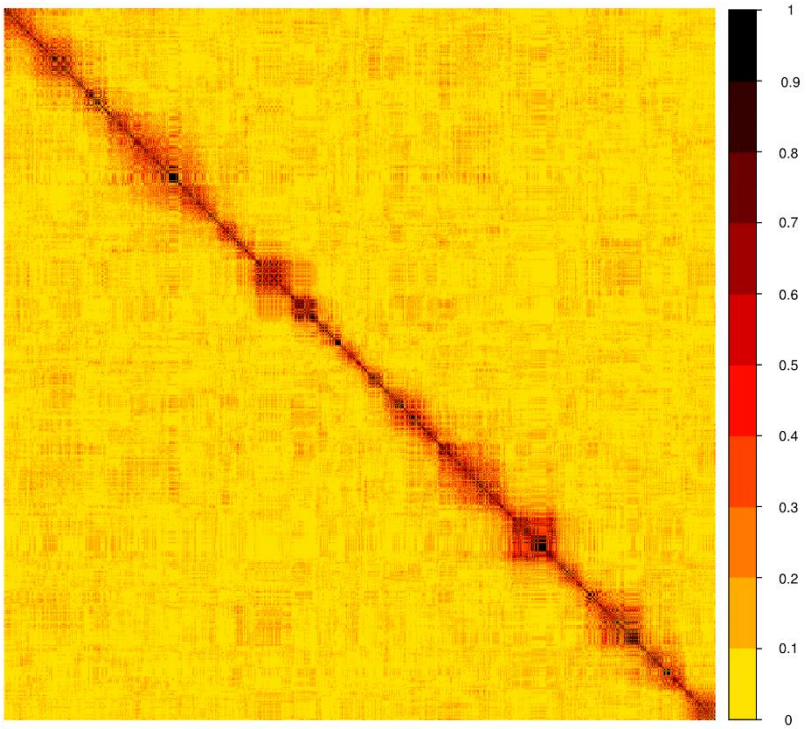
Figure 4-3 Heatmap of PCC correlation between each marker in Chromosome III and VI.



(a) chromosome IV



(b) chromosome XV



(c) chromosome VII

Figure 4-4 Heatmap of PCC correlation between each marker in large length chromosomes

4.3.2 Assessing QTL features clusters

Through the checkpoint model, the QTL outputs were split into different groups. For HU, 194 QTLs were identified originally. With checkpoint segmentation, the QTL detection was reduced to 67 while maintaining the major genes of interest. For MMS, 52 QTLs were chosen from the original 160. 46 QTLs of 140 QTLs were selected for Phleomycin. Among the clusters, some peak loci were solitary in significance. In these segments, the gene annotated by the solitary peak remained in the result sets. For example, In the analysis of the QTLs for MMS treatment, MFG1(YDL233W) was shown clustered in the narrow region with only one marker having significance.

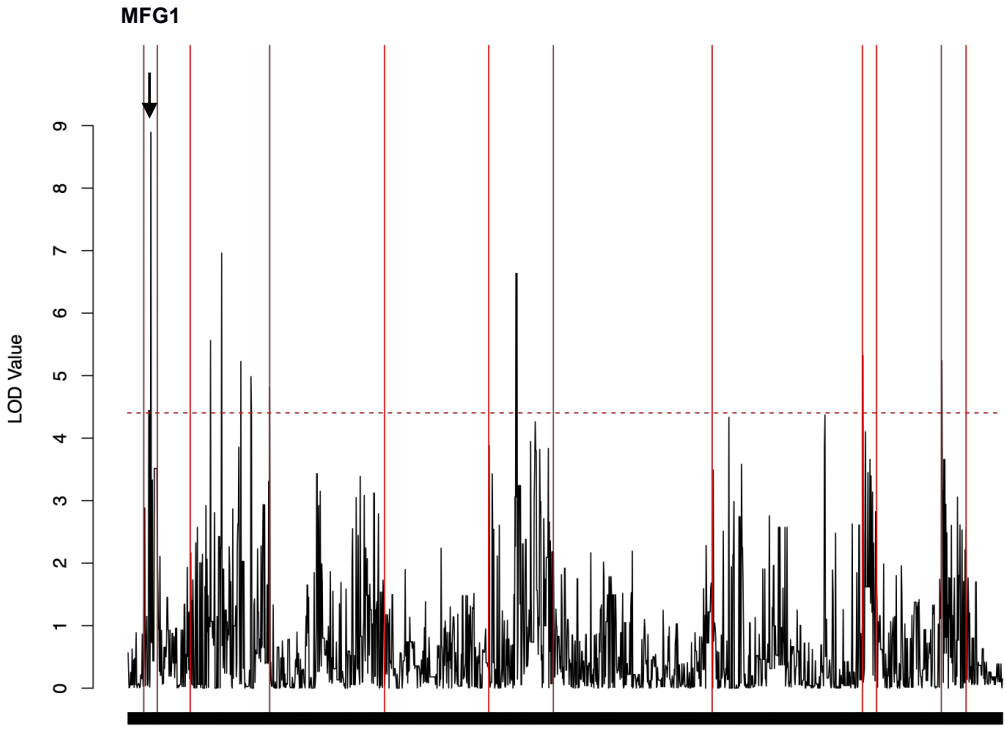


Figure 4-5 LOD plots of chromosome IV with feature clusters for analysis of MMS

The black sharp lines plotted the LOD score for each marker under chromosome IV. The red dashed horizontal line is the LOD threshold under significant level at 0.05. The red vertical lines are the segmentation borders for the cluster. The loci of MFG1 is pointed with arrow.

In addition, there are clusters including multiple LOD peaks that emerged in the adjacent region. In this case, the redundant markers which carry similar gene information were filtered. Only the peak among the clustering region was selected. For example, in the analysis of HU, chromosome VII has 8 QTLs with significance in the region from 308830 to 344395 detected as a result of these QTLs being located closely to the centromere region (Figure 4-6). The selected markers in this cluster region have the minor allele frequency from 0.31 to 0.33. High linkages were shown between the peak marker at 337618 and other markers in this region (0.81 – 0.94). After applying the methods, the detection was reduced to the peak marker which contains the gene of interest SPC105(YGL093W).

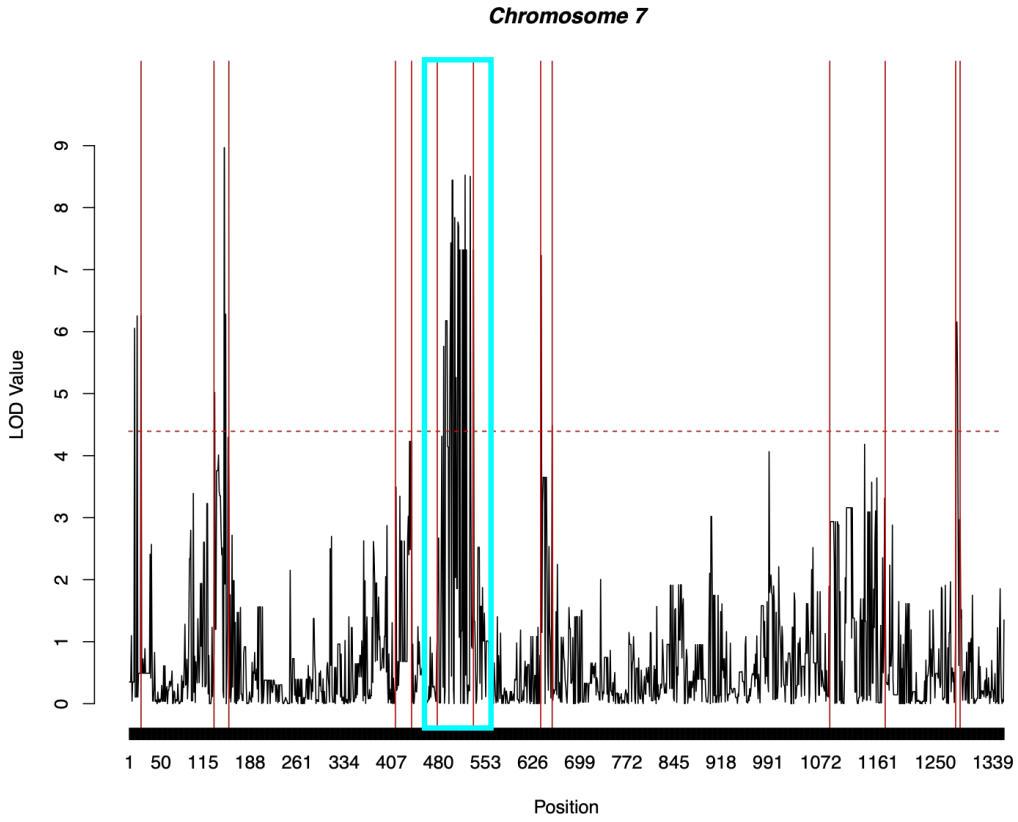


Figure 4-6 LOD plots of chromosome VII with feature clusters for analysis of HU

Genomic positions are shown on the x-axis and LOD value for each marker on the y-axis. The cyan box highlighted the clustering region. Clusters were segmented by the red vertical lines.

4.4 Discussion

From the above analysis, multiple clusters are present in adjacent regions for correlation across the genome. In many cases for QTL analysis, multiple LOD peaks emerged in the same vicinity. Hence, it is necessary to consider QTL identification within clusters. Using the assessment of QTLs with the changepoint model, the QTL numbers have been greatly reduced so that the experimenter can further validate genes of interest with a narrowed but still potentially causative sets of genes. This method offered a modelling-based selection for large sets of QTL output, which keeps more information than just increasing the strictness of the significance level. This method eliminated redundant peak information while keeping the significance level, which allows markers having lower LOD scores to be retained. From current validate results, RAD57 and BMH2 detected from the reduced set of genes have been validated through reciprocal hemizygosity analysis (Almayouf, 2018). However, the efficiency of this method needs to be verified by further validation such as through gene deletion experiments.

Chapter 5 Temporal quantitative trait locus analysis for yeast response to cancer therapies

5.1 Introduction

Recent advances in high-throughput techniques for DNA sequencing and phenotyping have greatly facilitated the identification of genetic variants underlying traits at a genome-wide level (Wilkening, et al., 2014). A great deal of research has been focused on DNA damage responses to work out possible therapies for human complex diseases (Rainey, et al., 2008), (Massey & Jones, 2018), (Pilzecker, et al., 2019). Because of the complexity of quantitative traits, analysis in humans and animal models is extremely challenging. Using yeast to study the response to therapies gives the opportunity to control both time and space from an experimental perspective more efficiently. Growth in yeast is an important complex trait for measuring performance in different environments. Dissecting the genetic basis of growth in yeast is a major challenge as multiple factors are involved in growth. Usually, growth in yeast is studied through measurements of the final growth density or growth rate at maximum doubling time. However, growth is a time dependent feature that is dynamic in the life of yeast and genetic variants could drive phenotypes at different time stages rather than the just the end of the growth profile. For example, a recent study shows that the variants could regulate gene expression differences over time (Strober, et al., 2019).

In the last two chapters, the analysis revealed that a few of the QTLs were identified for the differences of growth speed (rate) but were not associated the final growth level as all samples reached a similar level of growth at the end. Only focusing on final growth could therefore potentially lose genetic information about different responses to the environment over time. This chapter aims to explore

time dependent QTL mapping for the growth development of the F12 yeast strains over 65 hours under X-ray irradiation. With the large number of markers in F12 multi-parental lines (15,217) and the fine time scaling for growth phenotype (every 20 minutes), the QTL temporal structure for yeast growth with X-ray radiation was mapped. Surprisingly, some genetic variants only exhibited significant associations at certain periods of time rather than continuously through the experiment. Among the QTLs detected for X-rays, most of the genes that respond to radiation show their effects in the earliest time points after radiation with the significant association disappearing later during growth, becoming undetectable at the end point. Overall, the results suggest that detecting QTL mapping under different time points might be an effective way for assessing the genetic variants that contribute to yeast growth in response to various treatments/environments.

5.2 Methods and Materials

5.2.1 Experiment Preparation and Phenotyping

The experimental data for temporal QTL analysis were generated by Danae Georghiou. Control experiments are arrayed at 384 densities on YPD agar media with normal growth. Treatment experiments are grown on the same media with a treatment added. Growth of the 4 founders was also recorded under control and treatment conditions for comparative analysis. For the X-ray experiment, an X-Strahl instrument was used for delivering the X-ray treatment. One day after printing onto soft agar arrays the cells were incubated for 24hrs in 4°C to suppress growth. Straight after printing onto the experimental plate, the array was placed within an ice bucket to slow enzymatic DNA repair during the course of irradiation. The ice-treated plate was placed on the irradiation stage central in the 30cm diameter and at 30cm perpendicular to the X-ray source. A 1mm Cu filter was applied to divert the low energy produced photons while keeping the high energy photons to irradiate the samples. The treatment regime was: 300kV 10mm for 60mins to reach a total dose of 200Gy. Two concomitant regimes were applied to reach the total dose of 400Gy.

The growth of yeast colonies was measured by absorbance (OD value) and recorded every 20 minutes. Control and treatment experiments were run for approximately 65 hours in a micro-plate reader. The treatment measurement used for time-dependent QTL analysis is the treatment ratio which is the calibrated treatment phenotype value in each hour compared to the calibrated control OD value in each hour. Standard QTL analysis uses the treatment ratio of the final OD values.

5.2.2 Genotyping

The genotype data were the identical marker sets explained in Chapter 3 for the F12 yeast population. Strain AESW12fc232 was excluded from the strains list for the analysis due to a technical problem where the apparent measurements were outside the limits of the reader. The irregular growth was likely a contamination of this strain with a fast growing bacteria during the experiment (Figure 5-1).

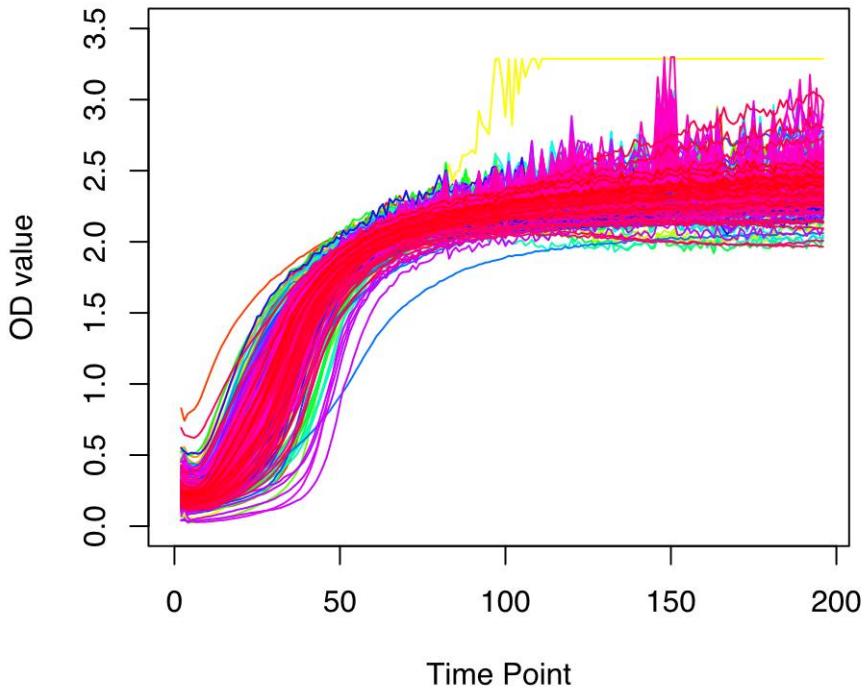


Figure 5-1 Growth curves of F12 segregants with raw readings

The yellow growth curve is the strain AESW12fc232 which illustrates the irregular growth compared to others and over the plate reader recording limits.

5.2.3 QTL analysis and bioinformatics analysis

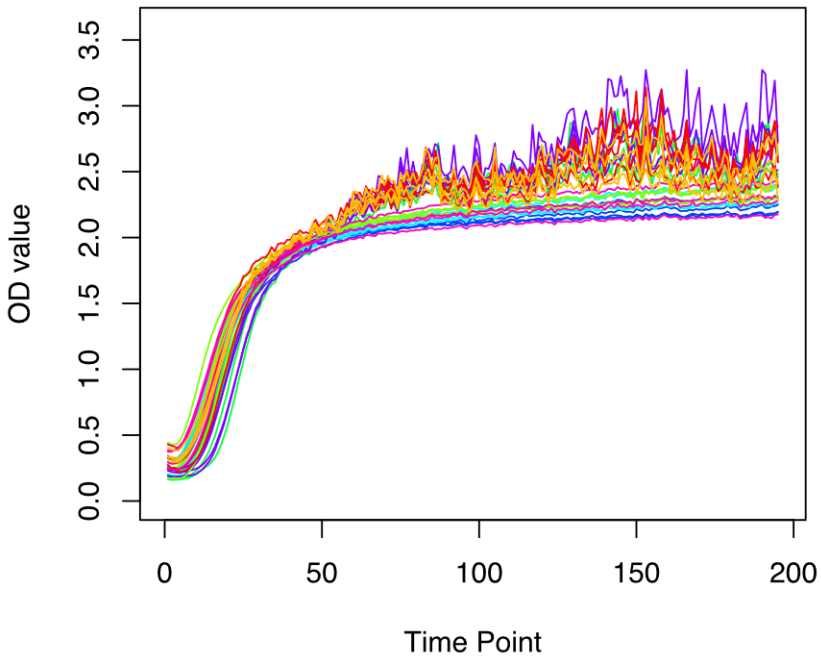
The raw growth curves were plotted in R by illustrating the growth under each time point (20 minutes). Smoothing was applied with mean filter and cubic spline for the average growth value in each hour to reduce the fluctuation. Single QTL analysis was performed for phenotype data in each hour. The parameters for QTL analysis were kept the same as in Chapter 3. Comparative analysis was performed in each hour. QTLs appearing in overlapping times and throughout the whole growth process were selected as candidates for downstream analysis. Growth curve calibration analysis was performed with R scripts on a local computer which is supplied in Appendix A. QTL analyses were performed on ALICE with 16 threads with R scripts. The QTL analysis output was recorded in the GACT research Rdrive. For finding the function and possible pathways, YeastMine and DAVID 6.8 were used for analysing the candidate lists with homologue identification and functional annotation clustering (Dennis, et al., 2003). GeneMANIA was used to find connections among genes (Warde-Farley, et al., 2010).

5.3 Results

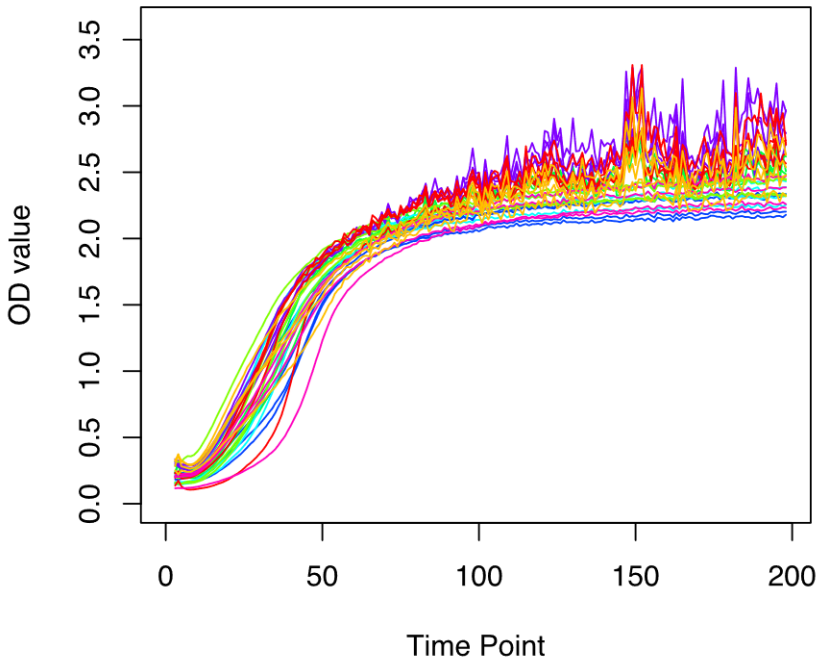
5.3.1 Growth behaviour varies in the F12 population and founders under X-ray and cold temperature conditions

Growth measurement is a key component for yeast analysis as yeast is unicellular organism. PHENOS records the differential growth curve of the F12 segregants every 20 minutes over 65 hours generating 195 growth OD values for each segregant. The readings of growth recorded by PHENOS shows that both F12 offspring and founders exposed to X-ray radiation exhibit more variability than the standard growth during growing phase but reached nearly the same level in stationary phase. Figure 5-2 illustrates the growth curves for founders under cold temperature and 400Gy X-rays with raw readings obtained from PHENOS. From the comparison among all four founder's growth, Wine European strains have a better performance on average both under control and treatment conditions. Different responses among the four founders were also seen from the comparison between the growth under control and X-ray radiation as the speed and duration of growth are very distinct among them. When looking at the growth curves of the F12 segregants, a similar trend with clearer variance were seen in the comparison between control and treatment (shown in Figure 5-4). After removed the strain AESW12fc232 (coloured in yellow in Figure 5-1), the growth phase looks cleaner with growth varying after radiation and reaching stationary phase slower than the control. However, the readings from the plate reader exhibits oscillation at stationary phase for all the populations. This could be caused by microscopic bubbles. To clean the growth data for analysis and comparison over time, calibration and correction was applied to raw PHENOS records. For further comparison of temporal QTLs at hourly intervals, the average value in each hour was calculated and smoothing by cubic spline was applied to reduce the fluctuation in stationary phase. Each strain has 65 hourly interval records for growth with the initial growth time point subtracted for calibration. Calibration figures for segregants under control and treatment conditions are illustrated in Figure 5-3. Furthermore, in order to perform temporal QTL analysis

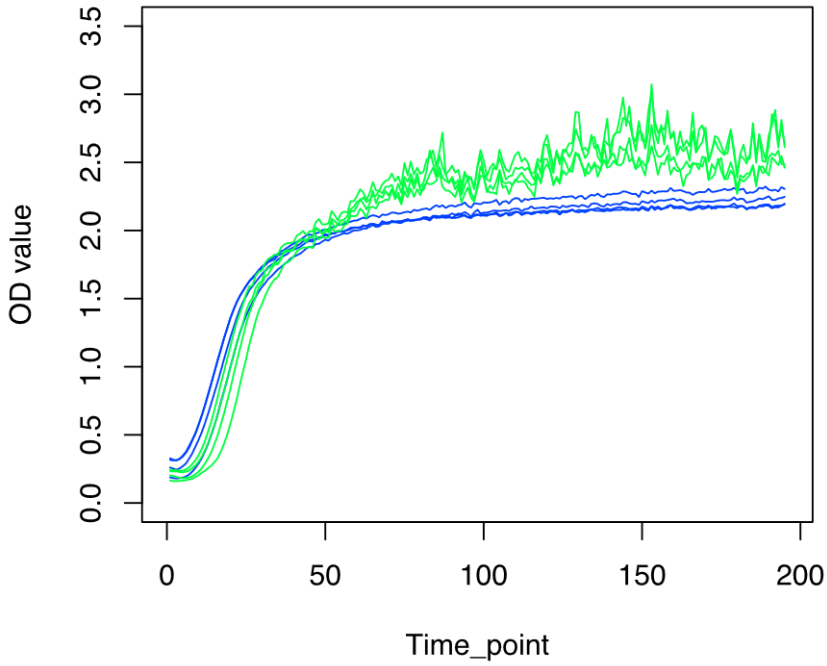
for detecting the dynamic responses to X-rays, the treatment ratio measures were computed as a phenotype for each hourly interval. The treatment ratio obtained as the average treatment OD value in each hour compared to the average control OD value allows the growth dynamics to be assessed as well as reduce the effect of cold temperature by the records of control growth. Figure 5-6 shows the illustration of treatment ratio over the growth developing time. From the figure for illustrating the growth ratio patterns, major variance is seen in the early stages up to 30 hours and with the remaining time the ratio is around 1.0 without much variance.

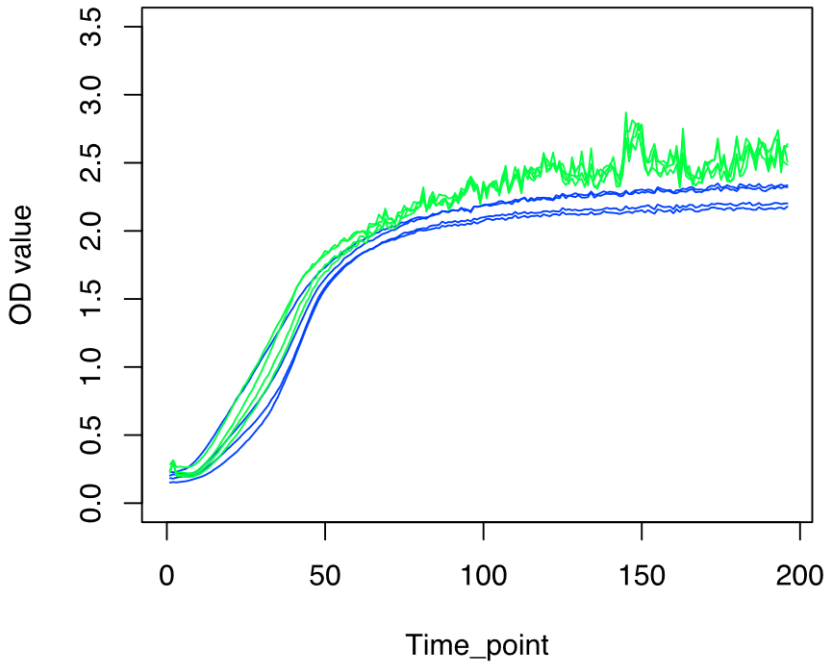


(a) Four founders' growth curves under control

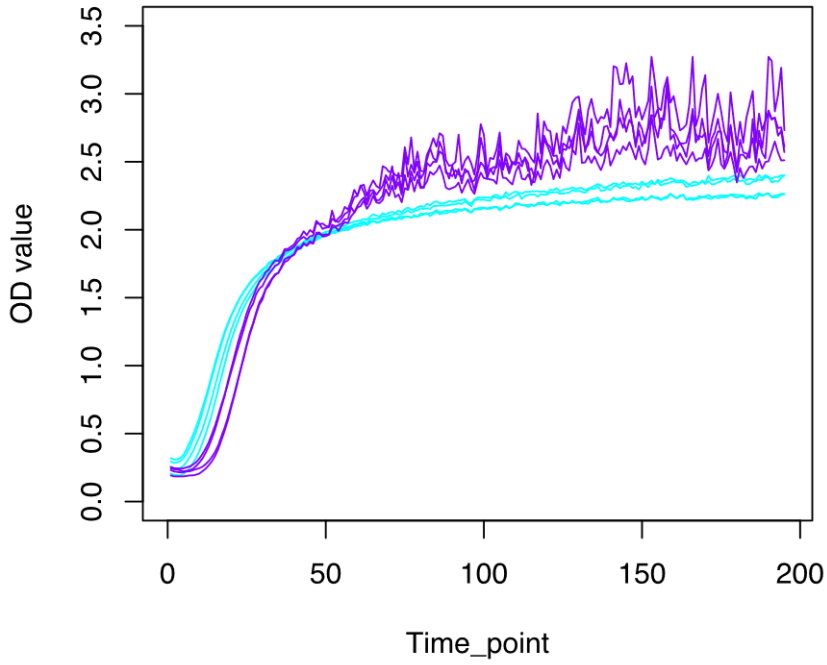


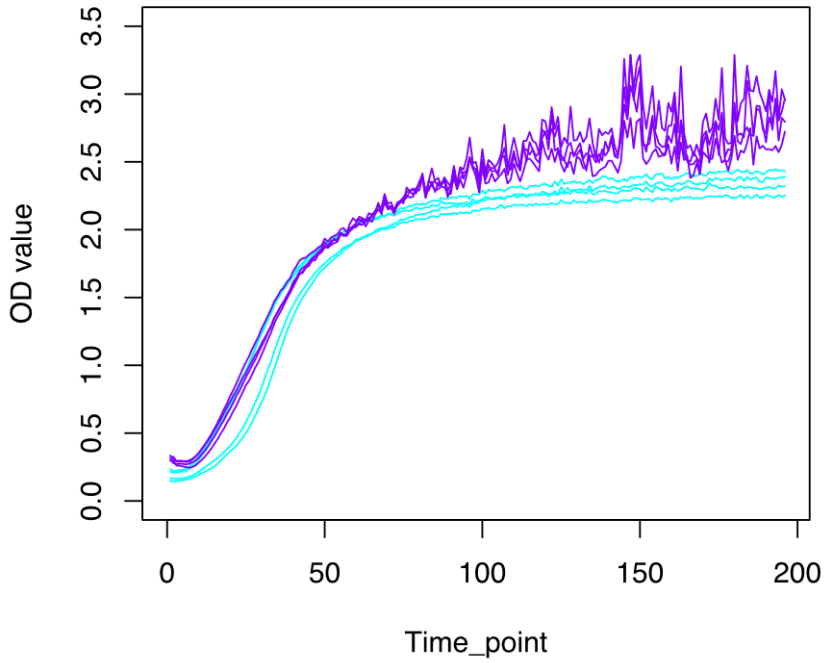
(b) Four founders' growth curves under X-ray Treatment


OD value
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
50
100
150
200
Time_point

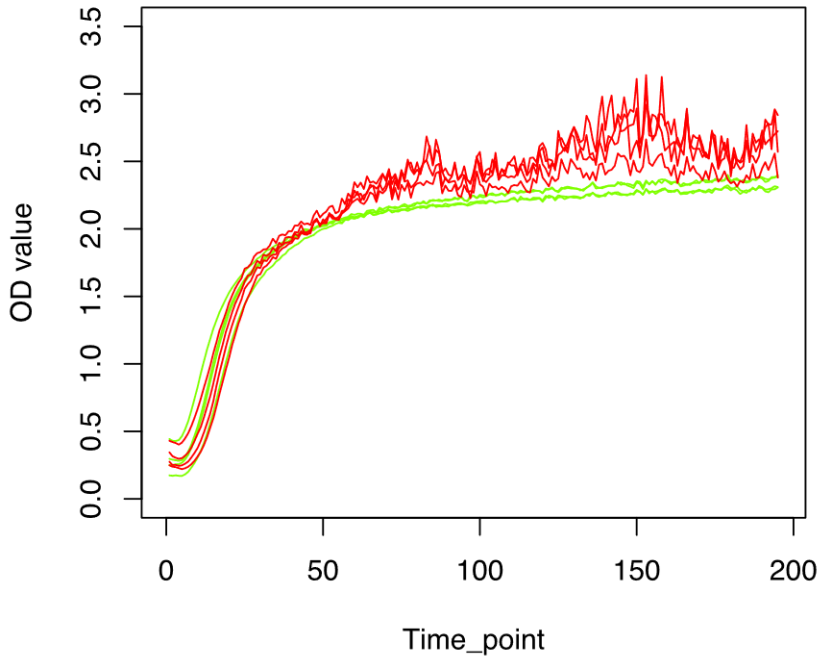

OD value
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
50
100
150
200
Time_point

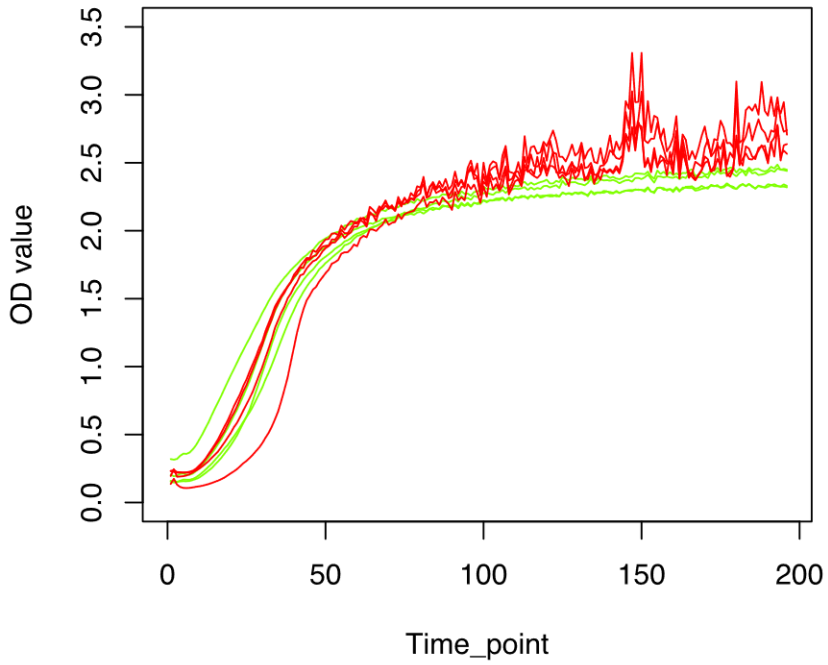
(c.1) NA


OD value
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
50
100
150
200
Time_point

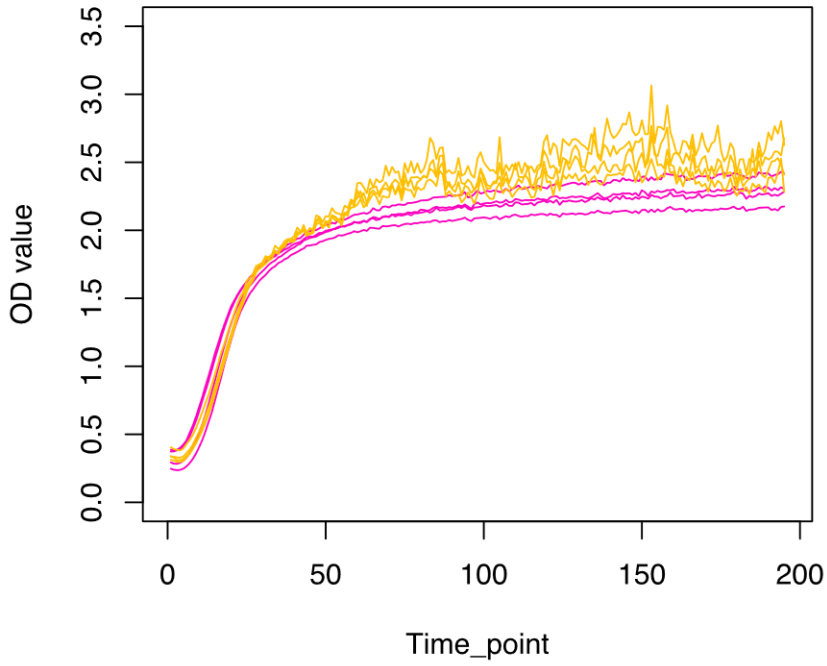

OD value
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
50
100
150
200
Time_point

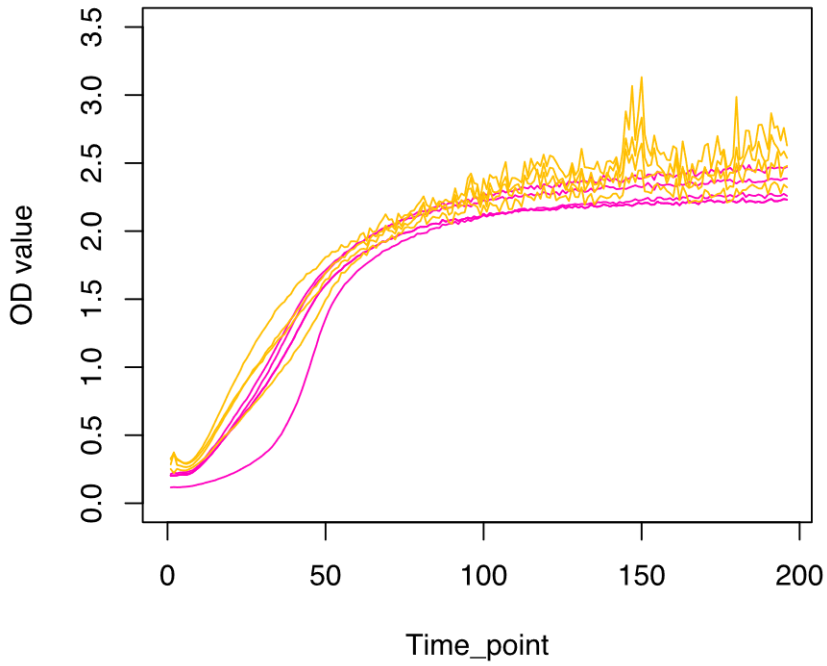
(c.2) WE


OD value
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
50
100
150
200
Time_point


OD value
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
50
100
150
200
Time_point

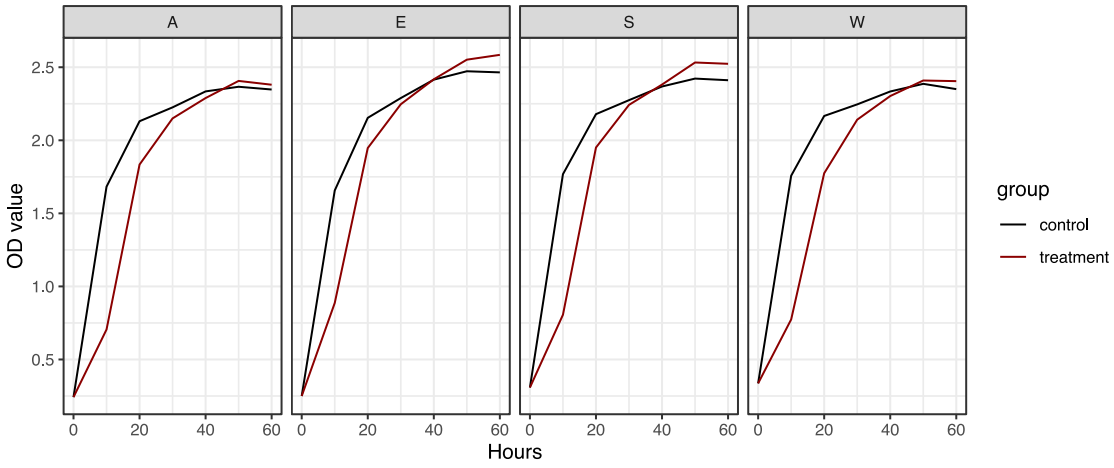
(c.3) SA





(c.4) WA

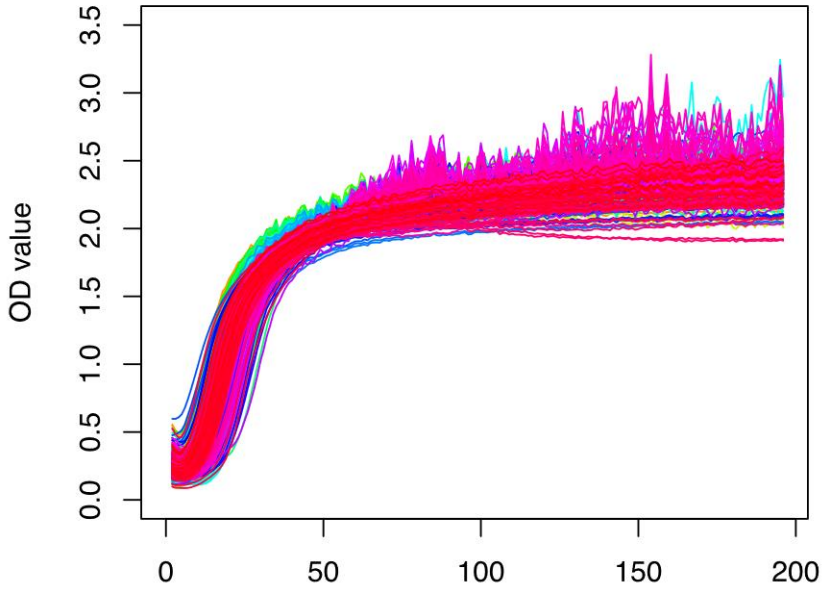
(c) Growth curves of each founder under control (top) and treatment (bottom)



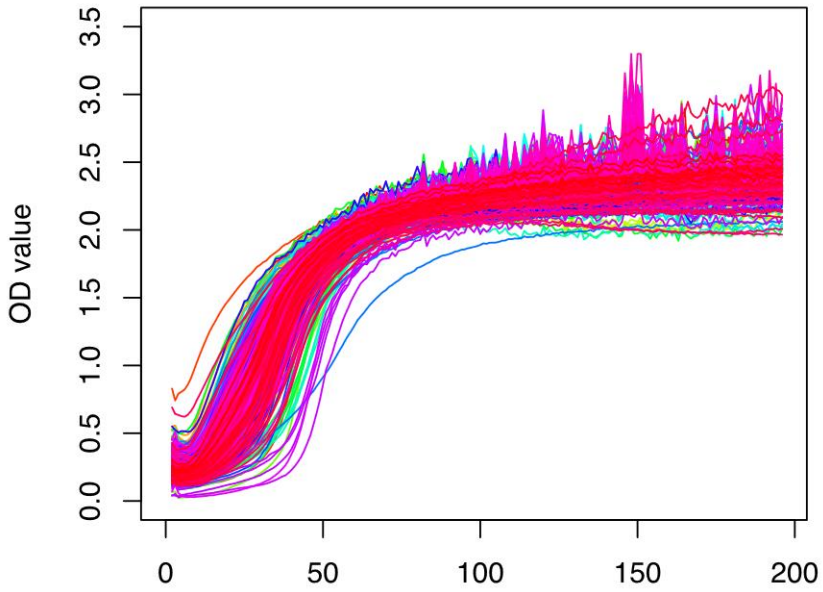
(d) Average growth of each founder

Figure 5-2 Growth curves of founder strains.

(a) - (b) stand for the plots of all founder strains in replicates growing on control and treatment with X-ray radiation. (c) stand for the plots of each founder strains with replicates ((c.1) NA, (c.2) WE, (c.3) SA, (c.4) WA). For each sub section in (c), growth under control are shown on the top and growth under treatment are shown in the bottom for each founder strain. In subgraph (d), the average growth of each founder under control and treatment. The growth curves of control are coloured in black. The treatment curves are in red.



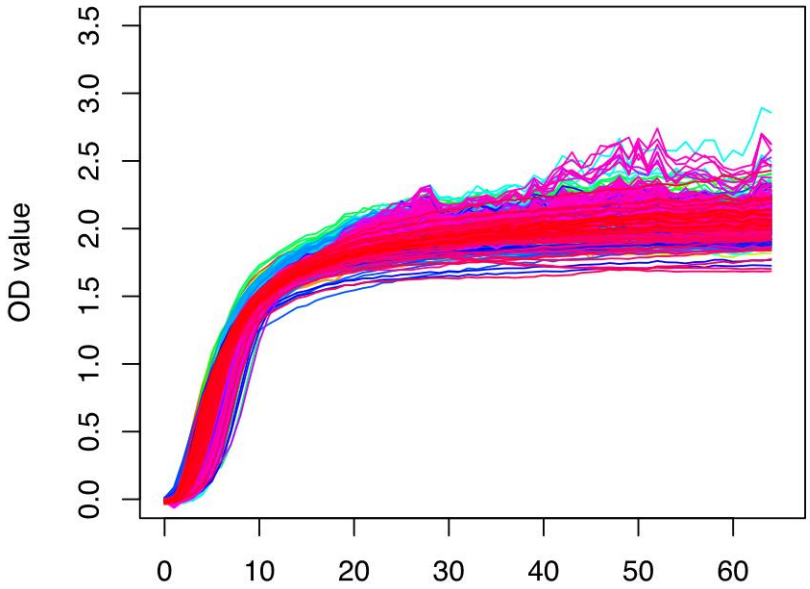
(a) Control



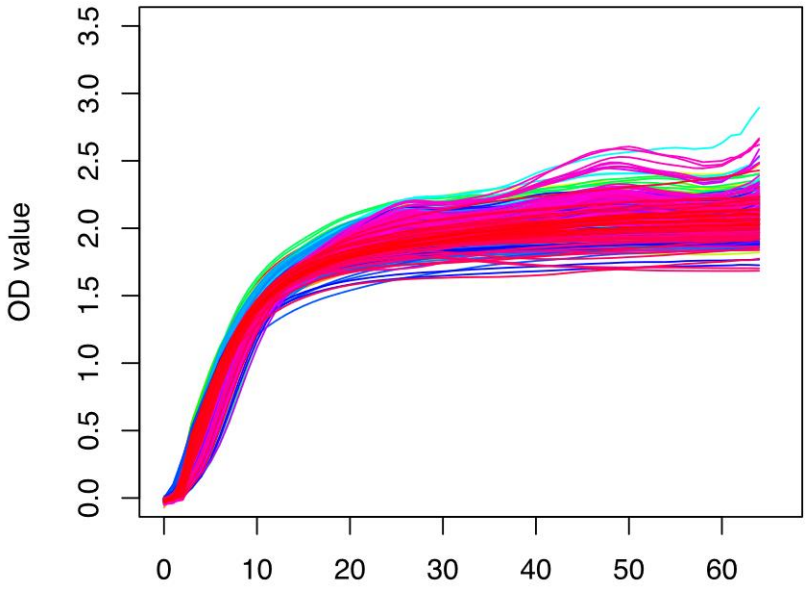
(b) Xray treatment

Figure 5-3 Growth curves of F12 segregants

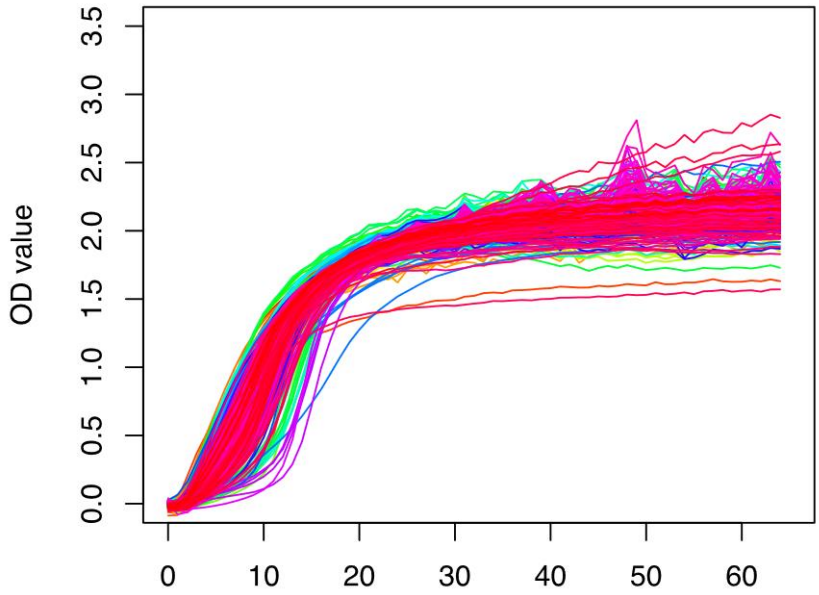
(a) stand for the plot of F12 strains growing with control condition and (b) stand for the plot of F12 strains growing with treatment condition that with X-ray radiation.


OD value
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
10
20
30
40
50
60

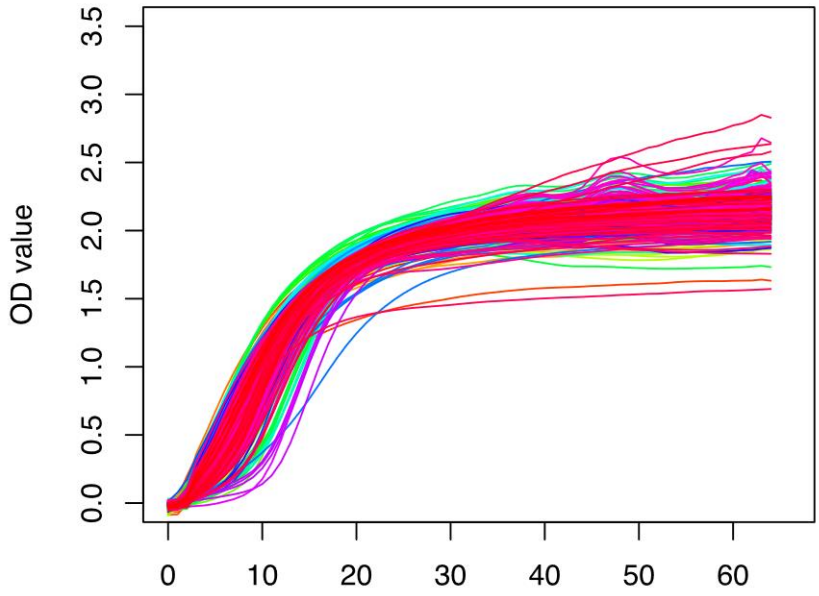
(a)


OD value
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
10
20
30
40
50
60

(b)



(c)



(d)

Figure 5-4 Calibrated growth curves of F12 segregants

(a) stands for the plots of F12 strains OD values with control condition in each hour. (b) stand for the smoothed curves of F12 strains OD values under control. (c) stands for the plots of F12 strains OD values with treatment condition in each hour. (d) stand for the smoothed curves of F12 strains OD values under treatment.

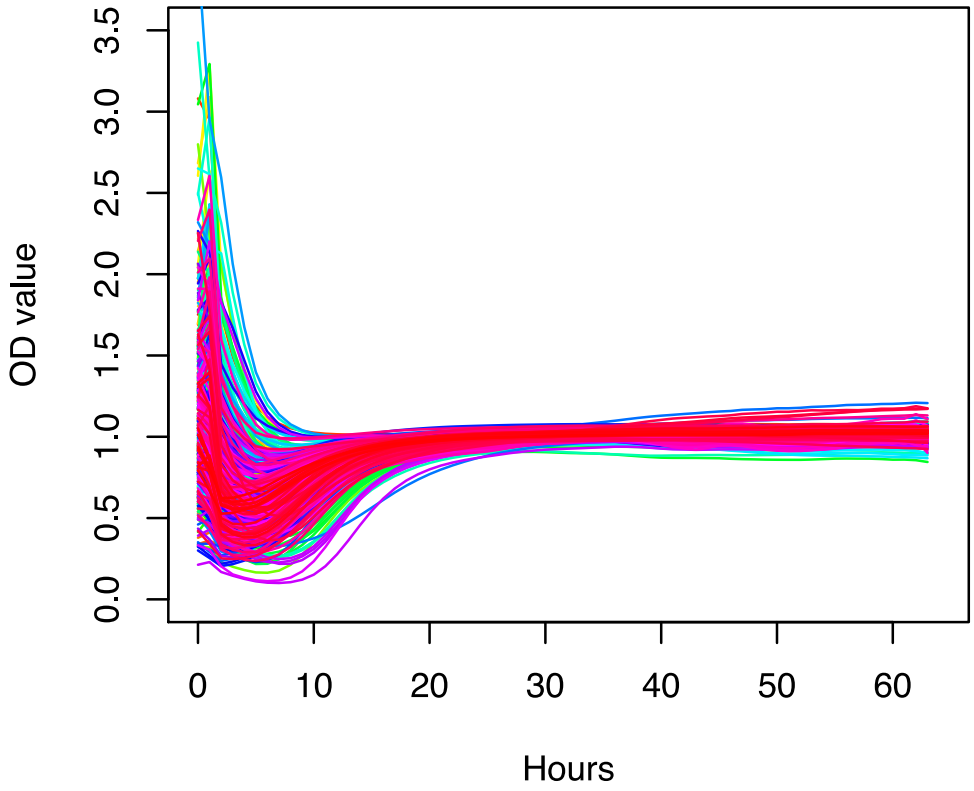


Figure 5-5 Dynamic developments of growth rate for each F12 segregants

Growth rate was calculated under each hour.

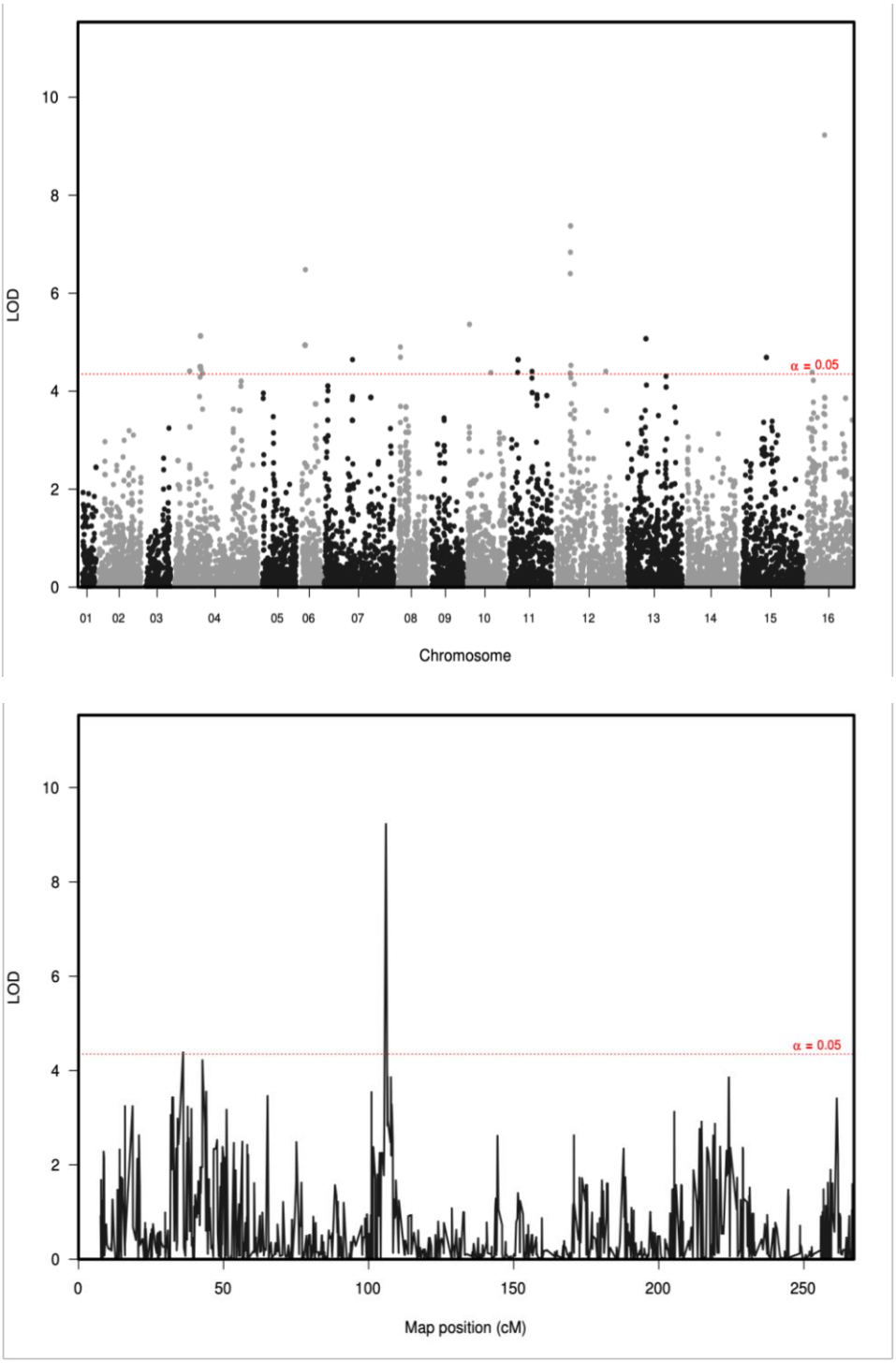


Figure 5-6 LOD plots for end point growth rate phenotype

QTL analysis mapped only a few markers as significant for the treatment ratio at the end of growth. The red line was the LOD threshold at 0.05 significance level. (a) is the Manhattan plot of the LOD scores on the y-axis with the distribution at genome level in 16 chromosomes with black and grey points. (b) plots the LOD score and Map position for each marker of the chromosome with peak LOD.

5.3.2 Standard QTL analysis detecting a few markers for phenotypic variation in response to X-ray treatment

As PHENOS provides the treatment ratio at the end point, QTL analysis was generated using end point treatment ratio as a phenotype to study the response of the F12 population under 400 Gy X-rays through shmootl/R. Figure 5.6 illustrates the QTL results over the whole genome and interval mapping of the chromosome with peak LOD. 31 markers included in 27 QTL intervals were identified across the genome with LOD threshold 4.35 based on 1000 times permutation at 0.05 significant level. Table 5-1 summarises the significant markers with the region of 1.5 LOD drop on each side of the peak that were used to define the QTL intervals. The range of each interval was determined and 20 genes with significant markers in them are listed. The peak marker across the genome that has the largest QTL score was located on chromosome XVI within gene MSY1 (YPL097W) which encodes a mitochondrial tyrosyl-tRNA synthetase (Arnez & Moras, 1997). Besides the peak gene, GCV1 (YDR019C) also has a role in the mitochondrial. In addition, three genes that located on chromosome IV have been confirmed that are related to responses to DNA damage and responses to stress. RAD61 (YDR014W) has been studied confirming the sensitivity to X-ray radiation (Jordan, et al., 2007) and is involved in the regulation of chromosome condensation. VPS54 (YDR027C) is needed for mitosis after checkpoint arrest caused by DNA damage (Dotiwala, et al., 2013). Hence further gene deletion for verification will be performed to test the responses among these selected genes among founders.

Table 5-1 List of genes present within the QTL intervals for treatment ratio

Peak markers were highlighted in the marker name column. Gene features were annotated based on the reference genome S288C and the genes that have been studied related to DNA damage were highlighted in bold.

Marker Name	Chromosome	Peak LOD	Start (bp)	End (bp)	Peak QTL Features
c04:0288588	IV	4.41	285051	290077	YDL095W
c04:0475724	IV	4.50	475233	480846	**YDR014W**
c04:0480895	IV	4.50	480887	484026	YDR017C
c04:0484955	IV	5.13	484050	488180	**YDR019C**
c04:0495303	IV	4.45	492782	496589	**YDR027C**
c04:0522237	IV	4.36	521639	523401	YDR035W
c06:0067330	VI	4.94	66607	67487	YFL034W
c06:0070033	VI	4.94	69321	70372	YFL033C
c08:0050240	VIII	4.90	49600	50710	YHL028W
c08:0051470	VIII	4.69	50752	53336	YHL027W
c10:0064516	X	5.36	63189	64596	YJL197W
c10:0465678	X	4.38	465531	465795	YJR016C
c12:0226485	XII	4.36	226353	226566	YLR039C
c12:0229211	XII	6.84	229136	229597	YLR040C
c12:0236872	XII	4.53	236638	237673	YLR045C
c13:0314523	XIII	5.07	312892	315485	YMR019W
c13:0316612	XIII	5.07	316589	317880	YMR020W
c15:0440013	XV	4.69	438841	442178	YOR059C
c16:0124331	XVI	4.39	118878	125111	YPL226W
c16:0365403	XVI	9.23	363218	367091	YPL097W

5.3.3 Temporal QTL analysis reveals time-dependent markers occur in different stages of growth after X-ray treatment

As for the growth curves shown in section 5.3.1, the growth of treatment under X-ray has more variance than the control over time. To investigate the dynamic association between growth difference and markers, QTL analysis was generated for growth dynamics in each 64 hour interval to define the significant QTL intervals over time. Different numbers of QTLs were detected in different hourly intervals (shown in Figure 5-7). An interesting trend was observed on the QTL numbers with 2 peak stages of time intervals accumulating large numbers of QTLs. The first peak starts from 10 hours and ends at 20 hours. The second peak is between 42 hours and 48 hours. 283 markers were identified as significantly detected QTLs across all 64 hours (Figure 5-8). Different QTLs were detected during the maximum growth rate stage and stationary phase of growth for F12 segregants. For QTL mapping over the whole period of growth, the maximum number of QTLs was detected at 13 hours. In total, 87 markers were significant for contribution to this hour's growth pattern.

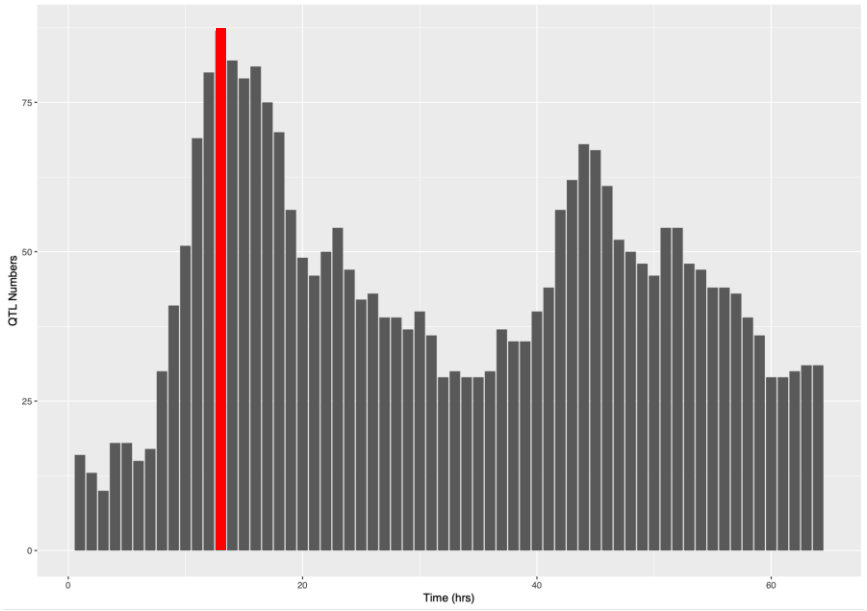


Figure 5-7 Number of QTLs for yeast growth at different times

y-axis is the QTL numbers on and the time point in each hour on x-axis. Each point displays the QTL number located among whole genome in 16 chromosomes which indicate different QTLs were detected in different hours in the F12 population. The bar with largest QTL number was highlighted with red colour.

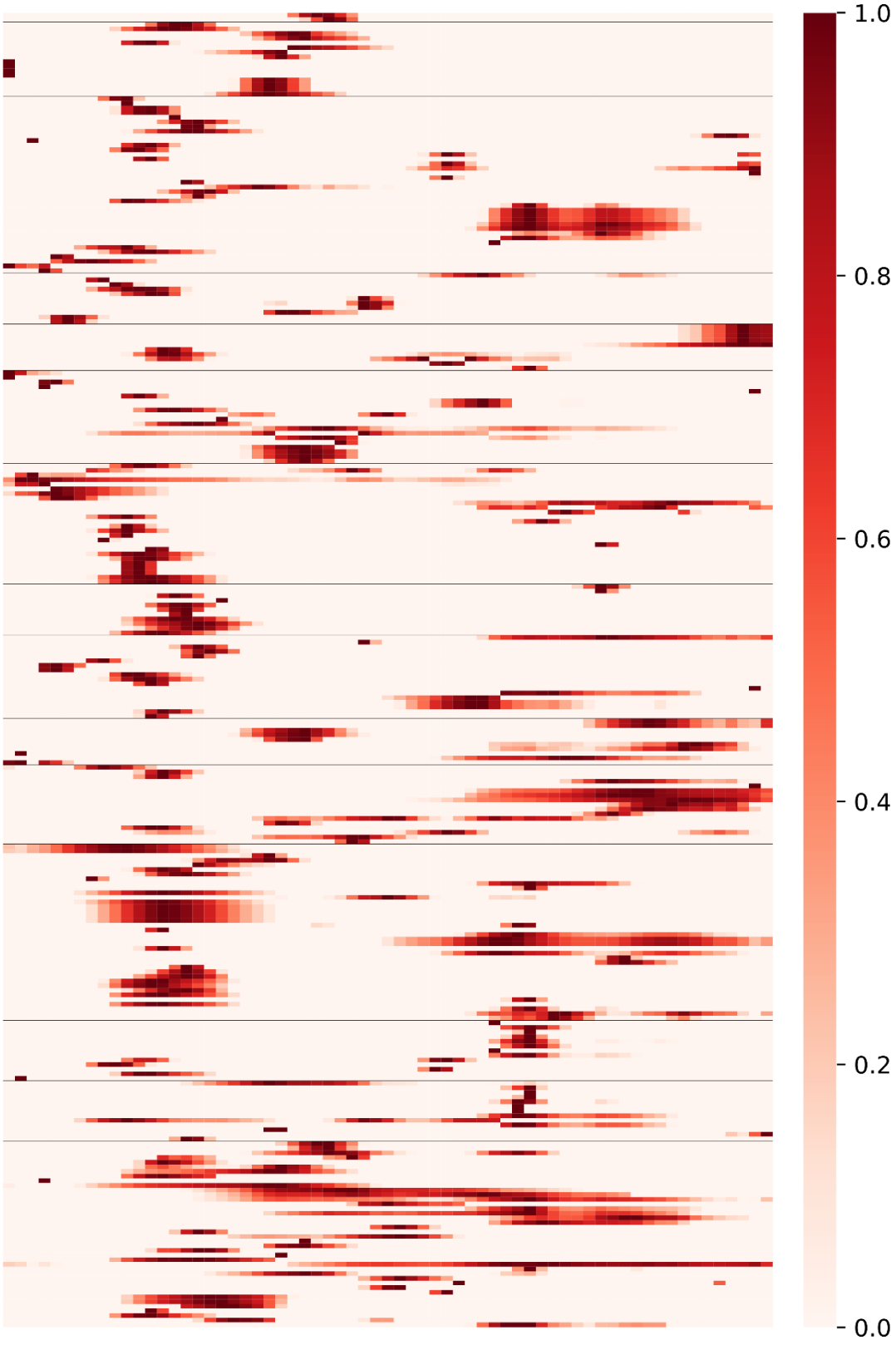


Figure 5-8 Heatmap of LOD value for all the QTLs identified over 64 hours.

Colour intensities represent LOD scores scaled by the maximum LOD value. The loci under threshold are colored in white and significantly detected QTLs loci in red. Chromosomal borders are indicated by horizontal lines.

For these 87 QTLs, the genes that included the peak markers which shown significance were annotated. 1.5 LOD interval was further delimited for defining the QTL interval boundary. A few of genes that are already known to be involved in responses to DNA damage and DNA repair processes are highlighted in the Table 5-2.

Other than RAD61 which was detected in the standard QTL analysis for the end point growth, 10 genes where the peak markers sit and 5 genes included in the interval range (summarised in Table 5-2) having a role in DNA replication were detected at 13 hours growth but did not exhibit a significant association at the end point of growth. Among the QTL intervals, the peak QTL interval was mapped to the chromosome XIII from 257090 to 274871 which linked to the gene MIX17 close to the centromere region. This might indicate the influence of X-ray radiation affecting chromatin structure. In addition, another QTL with a large LOD score was also located on chromosome XIII from 228514 to 233147 where three genes NSE5 (YML023C), APT1 (YML022W), UNG1 (YML021C) were included in this range. The peak marker was involved and located in the coding region of UNG1. UNG1 is the essential gene for repairing of uracil base damage in DNA (Chan, et al., 2012). Moreover, NSE5 is a component of the SMC5/6 complex involved in the DNA repair pathway and DNA damage responses (Xu, et al., 2013). There is significant evidence shown that SMC3 (YJL074C) found in the interval is needed for repair of double-strand breaks by sister-chromatid exchange (SCE) (Cortés-Ledesma & Aguilera, 2006). Four genes PSY4 (YBL046W), EDE1 (YBL047C), RRT1 (YBL048W) and MOH1 (YBL049W) were located in the QTL interval on chromosome II where the peak is at 129943 bp. The peak marker is in the coding region of EDE1 which related to regular non-apoptotic cell death (Kuilman, et al., 2015). Gene PSY4 in yeast involves in the regulation of DNA damage checkpoint (Fedorov, et al., 2013). This gene has a human homolog, PPP4R2, which is an essential gene involved in DNA double-strand break repair in human cell lines (Lee, et al., 2010). SIZ1 (YDR409W) shows evidence in assisting the survival of DNA damage (Horigome, et al., 2016)

Table 5-2 List of QTL intervals with contributed genes under 13 hours.

Table consists QTL intervals information with the chromosome, LOD score, start position, end position and localised genes. Gene features are annotated based on gff file for reference genome S288C. The candidate genes that related to DNA damage are highlighted in the QTL features column.

Marker Name	Chr	Peak LOD	Start (bp)	End (bp)	QTL Features
c02:0010136	II	4.79	10038	10463	**YBL107C**
c02:0129943	II	4.84	126921	133779	YBL049W YBL048W **YBL047C** YBL046W
c04:1290886	IV	5.80	1288071	1292138	YDR408C **YDR409W**
c04:0476181	IV	4.64	475724	477394	**YDR014W**
c05:0184684	V	5.21	181621	186649	YER013W YER014W **YER014C-A** YER015W
c08:0042255	VIII	5.16	37876	42291	YHL032C YHL031C YHL030W-A **YHL030W**
c10:0303490	X	5.23	300176	303556	**YJL074C** YJL073W
c12:0147547	XII	5.14	146262	149617	**YLL002W** YLL001W
c13:0230682	XIII	5.91	228514	230823	YML023C YML022W **YML021C**
c13:0593285	XIII	4.63	591204	598520	YMR165C YMR166C **YMR167W** YMR168C
c13:0612492	XIII	5.81	608255	612702	**YMR173W** YMR173W-A YMR174C YMR175W YMR175W-A YMR176W
c14:0628490	XIV	5.13	627710	628503	**YNL001W**
c16:0090910	XVI	5.00	88567	92875	YPL243W **YPL242C**
c16:0106300	XVI	7.01	104203	106953	**YPL235W** YPL234C YPL233W
c16:0797530	XVI	6.02	797112	802249	YPR133C YPR133W-A YPR134W **YPR135W**
c16:0889543	XVI	5.44	889063	890560	**YPR175W**

RTT109 (YLL002W) is involved in the cellular response to DNA damage and is essential for maintenance after repairing double strand DNA break (Tsabar, et al., 2016). From the study in Han J, et al. 2007, this gene also has genetic association with genes that are involved in DNA replication. Interestingly, among the genes selected by the peak marker, 7 genes NOC3 (YLR002C), ECM29 (YHL030W), RIX1 (YHR197W), ECM5 (YML176W), DOM34 (YNL001W), IQG1

(YPL242C) and DPB2 (YPR175W) are all involved in responses to DNA replication stress. Some markers are located in intergenic regions and adjacent genes are shown as having possible relevance. CTF4 (YPR135W) encodes a chromatin binding protein which is involved in DNA replication (Lengronne, et al., 2006). Gene RVB2 (YPL235W) which participates in chromatin remodelling has association with genes that respond to DNA damage and DNA repair (Radovic, et al., 2007). MLH1 (YMR167C) plays an important role in DNA mismatch repair which is located on chromosome XIII. Furthermore, DDR48 (YMR173W) encodes a DNA damage-responsive protein and is found in the same QTL interval as gene ECM5 where the peak marker is located. The aforementioned effect might due to a bimodal QTL that merging the peak effect in between but it could also indicate that the gene DDR48 shown contribution. Hence further validation with gene deletion are needed to assess the role of this gene.

In addition to the detection of candidate genes at 13 hours, overlap QTLs for the whole growth tracking for the treatment OD value were also examined. Figure 5-9 illustrates the dynamics for the temporal changes of the LOD score in each marker measured at three hourly intervals (10 hours, 20 hours and 30 hours) for representing three growth development stages. 51 QTL intervals were identified at 10 hours growth rate, 49 QTL intervals for 20 hours growth rate and 40 QTL intervals for 40 hours growth rate. The peak markers for 10 hours and 20 hours were both located on chromosome XIII with the largest LOD score but sit in different genes. Markers on chromosome XIII at 25114 bp are in PHO84 (YML123C) showed temporal changes with the improvement of growth that have the high QTL scores in the 10 hours and decrease the significance on 20 hours and faded out at 30 hours. The reverse trend in seen on chromosome XVI at 365403 with no significance at 10hour and 20hour but reached peak significance at 30 hours. Moreover, overlaps were compared among these three time points as illustrated in Figure 5-10. 29 genes were shared solely between 10 hour and 20 hour which involved 6 QTL intervals. 18 genes were shared solely between 20 hour and 30 hour which involved in 9 QTL intervals. None of the genes were shared solely between 10 and 30 hours, however, there are 9 genes identified in

all three intervals which involved in 5 QTL intervals. The difference of overlapping genes in different times shows that QTLs might not only be significant continuously in the development of the phenotype over time but could also contribute in temporal stages. This indicates that the relevant genetic information may be missed or under represented by only looking at one time point of growth. Table 5-3 summarises the genes that overlapped in these three time periods. The functional annotation in DAVID of these 9 genes were clustered into protein transport and ATP binding. RVB2, GUT1 and BUD32 were related to protein kinase and involved in ATP binding. APL6 (YGR261C), ATG19 (YOL082W) and ATG21 (YPL1002) are needed for protein transport and with cytoplasmic functions. Besides this major QTL that was significant in most time points, there are some QTLs that show associations only for specific stages. As a high dose of X-ray radiation was applied before the growth recording, there are a few markers identified as QTLs involved in the recovery mode after radiation which have immediate response to the growth in the early time points (1-10 hours). Surprisingly, almost none of these markers were as QTLs in the later time points. The QTLs involved in the early time growth for recovering from X-ray radiation are summarised in Table 5-3. From the first 10 hours, two markers overlapped over time and the genes involved are in responses to DNA replication stress. The first marker is located on chromosome IV at 1328769 bp which is in the intergenic region between YDR431W and NPL3 (YDR432W) which encodes an RNA-binding protein. Another marker is located on chromosome VIII at 43299 bp in the coding region of ECM29 (YHL030W) which is the gene shown DNA damage responses realising during DNA replication stress (Tkach, et al., 2012). Besides the QTL shown in all 10 hours, a further marker also shown overlap from 4 hours to 9 hours. The time overlapped marker c05:051338 as the peak localised the gene RAD4 (YER162C) is the radiation sensitive gene that response to DNA damage repair.

Table 5-3 List of QTLs with overlap during the early time growth.

Gene Features were annotated. Candidate features which were shown responses in DNA damage were highlighted in bold.

Time Period	Chromosome	Gene Features
4 - 9 hour	V	YER162C
1 - 8 hour	IV	YRD432W YDR433W YDR434W YDR435C
1 - 10 hour	VIII	YHL030W
4 - 8 hour	X	YJL079C YJL078C

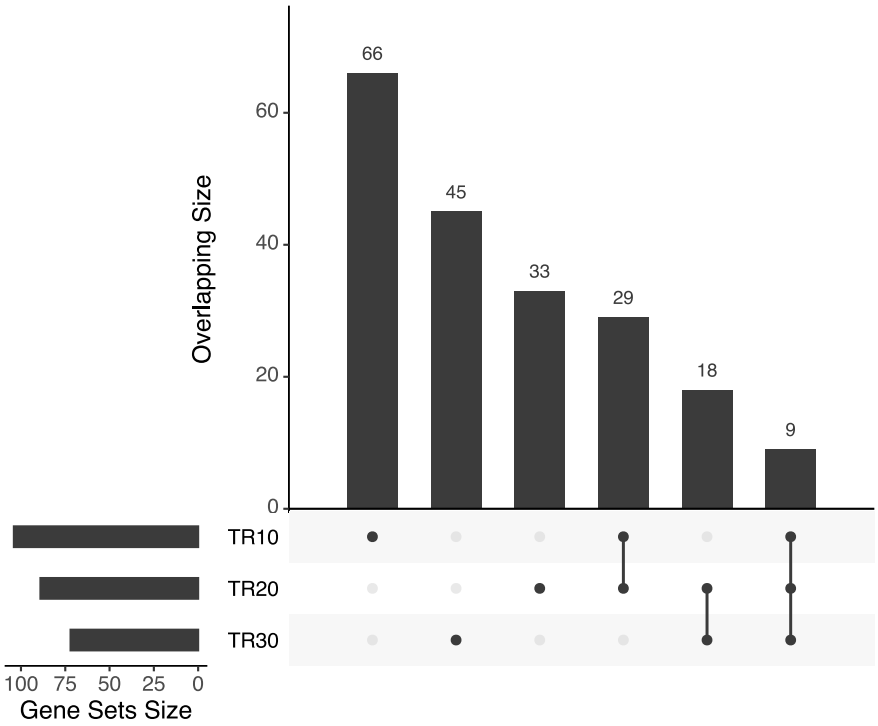


Figure 5-9 Summary of QTL intervals under 3 different time

The UpSet graph illustrates the number of overlapping genes in 10 hours, 20 hours and 30 hours for X-ray treatment phenotype. Left bottom is the histogram for the number of genes that located by QTL intervals for 10 hours, 20 hours and 30 hours. The main histogram shows the number of overlaps (intersections) between each set. The first three in the main histogram with single dots indicated the total number of overlaps for each time. 66 genes overlapped between 10 hours and others. 45 genes overlapped between 20 hours and others. 33 genes overlapped between 30 hours and others. Multiple dots with lines connection illustrate the overlap comparison in different sets. The histogram illustrates the number of the overlapping genes. 29 overlaps between the sets of 10 hours and 20 hours. 18 overlaps between 20 hours and 30 hours. 9 overlapping genes among 10 hours, 20 hours and 30 hours.

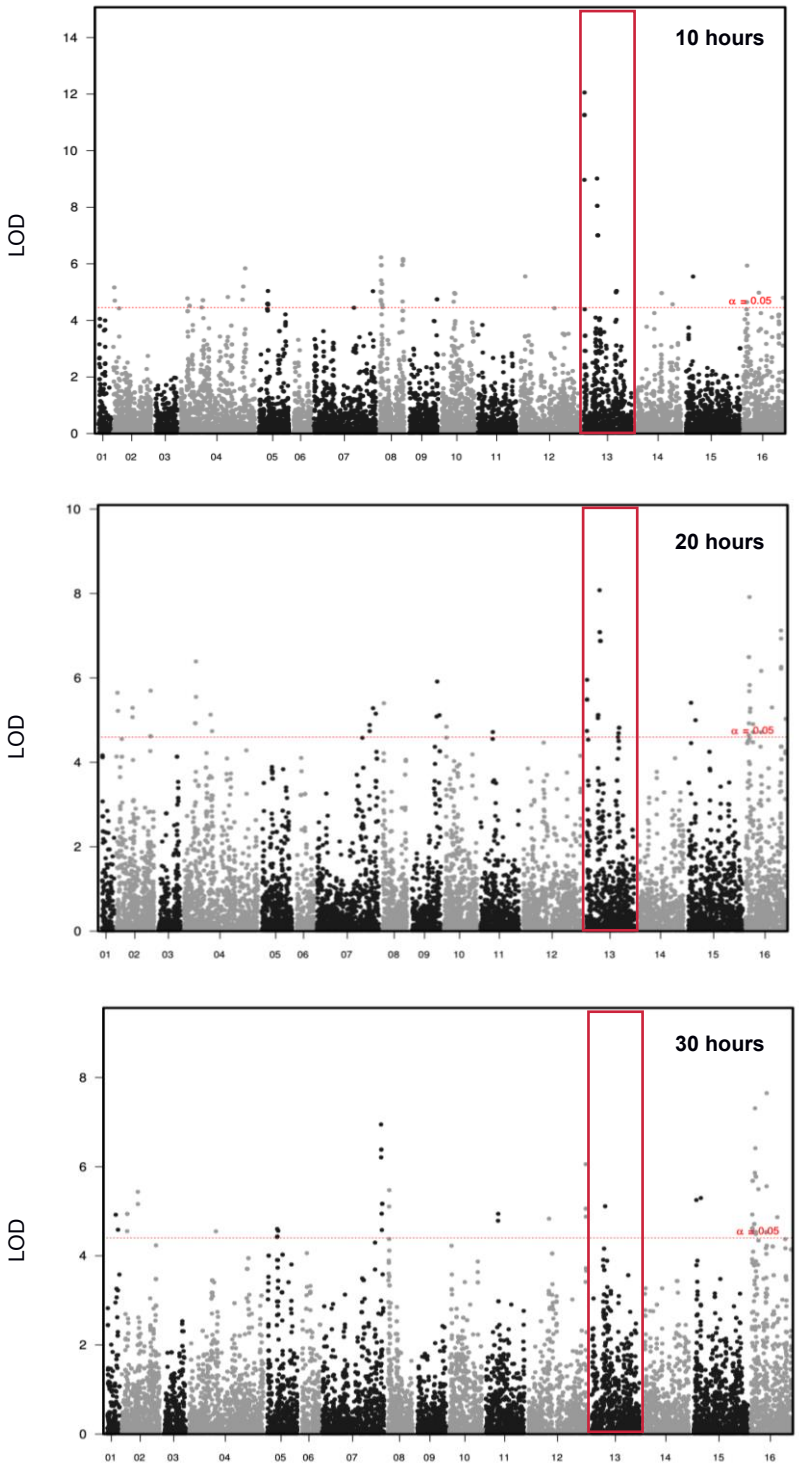


Figure 5-10 QTL mapping for growth difference between treatment and control

Manhattan plots for 10 hour, 20 hour and 30 hour QTL output of 16 chromosomes are displayed in this figure. The red dash line indicates the LOD threshold for each QTL run at significance level 0.05. Each point represents a marker with the LOD score. To trace the QTL changes among these three periods, the markers on chromosome XIII were highlighted with orange box. y-axis of the three figures are in different scale for arranging markers shown in the window.

5.4 Discussion

Many analysis aspects of the temporal QTL detection of this study have been highlighted in the section above. Quantitative research is a fundamental approach for better understanding cellular processes, to determine the cellular role of genes through screening the quantitative changes of growth phenotypes in a wide variety of growth conditions for yeast. Only focusing on the end point growth might result in the loss of a lot of information. In this chapter, yeast growth was tracked in every hour to investigate the temporal association of genetic variants. Temporal QTL analysis provides an extra dimension to complex traits to gain information and potentially unravel the dynamics of growth response. As the study involves growth changes over time, the growth curves were calibrated with respect to the initial print OD value and a mean filter and cubic spline were applied to smooth the curves without fluctuation. The more precise phenotype values enabled detection of QTLs in each hour and comparison over time. With two dynamic features, treatment value and treatment difference, compared to controls, many markers were identified as QTLs where the genetic variants exhibited a significant contribution. Several candidate genes were identified in a specific time period instead of being functional over the whole growth period. In comparison with the standard QTL analysis for the end point growth rate, we find many QTLs that would be missed by only considering the end point. Another important objective of this study was to find related cellular process and pathways in response to radiation. Besides the tracking of markers, QTL intervals were further located and compared at different timepoints. Several candidate genes in the overlap regions in different times appear to be involved in DNA damage pathway and contribute to DNA repair. Further study can be improved to use Bayesian methods to learn adapted interval drop. These could also be studied in patients and human cell lines. Moreover, several genes involved in transcriptional response which might indicate the dynamic control of complex traits. There are also some function unknown genes identified throughout the analysis which will be potential genes for validation.

To summarise, this chapter analysed huge amount of data in genotype scale that involved in thousands of markers among F12 multi-parental population and phenotype scale that tracking growth trait in each hour. The large-scale analysis generated a significant amount of data regarding the responses to X-ray radiation. By looking at DNA damage related genes over the time series, different QTLs were detected during different phases of the experiment. This indicates that the mapping of involved trait is possibly to be changing over time. Apart from DNA damage response related process, a lot of other biological process could be affected by time and environment. Hence, the analysis shows the importance of tracking growth traits with temporal development. However, validation by experiment are needed for the candidate genes.

Chapter 6 Mapping QTLs for *Saccharomyces* hybrid yeasts to improve brewing solutions

6.1 Introduction

In the previous chapters (chapter 2, 3, 4, 5), *Saccharomyces cerevisiae* were used as model organism to study human complex traits. By using natural variation in 6 F1 bi-parental cross and 12th generation multi-parental segregants generated with four founders from SGRP, the mosaic genetic background gives the opportunity to understand the genetic regulation for phenotypes in yeast growth under environment stress. Experiments were carried out on F1 segregants and F12 segregants using different drugs or cancer related therapies. A large number of significant locus were obtained through QTL analysis to show the complex association with yeast growth. Apart from applying yeast as model organism for scientific research, the earliest and most widely used application of yeast is the traditional production of fermenting foods and alcoholic beverages. Humans have been domesticating yeast for alcoholic drinks from thousands of years ago. Different sources with sugar can be fermented by yeast to produce alcohol and CO_2. In recent decades, the use of yeast in industrial production for new biotechnology applications has grown exponentially, such as biofuels in energy production, and using yeast for bioremediation of the contaminated environments (Buijs, et al., 2013), (Ojuederie & Babalola, 2017). However, using yeast in the production of alcoholic drinks, such as beer and wine, remains a major interest. This has prompted a lot of research in breeding and strain improvement for getting robust yeast for application. Stable and excellent performance traits are essential for the brewing industry for large scale fermentation, such as ethanol tolerance, cryotolerance, sugar tolerance, high flocculation and etc (Bokulich & Bamforth, 2013). Improving the growth and fermentation of yeast can improve yeast utilisation and reduce production costs.

In addition, interspecific hybrid between *S. cerevisiae* and other species in *Saccharomyces* yeast strain are commonly used in brewery industries and exhibit better performance in responses of stress (Borneman & Pretorius, 2015). For example, *Saccharomyces pastorianus* is a natural alloploidy yeast for brewing lager beer that has the genetic information of *S. cerevisiae* × *S. eubayanus*. This hybrid inherited advantageous features from each parent with cryotolerance and even outperformed its parents (Hebly, et al., 2015), (Gibson & Liti, 2015). Apart from hybrid strains for brewing lager beer, the natural hybrid between *S. cerevisiae* and *S. kudriavzevii* showed hybrid vigour that has better commercial performance to improve flavour and aroma in wine making (Borneman, et al., 2012).

Besides the existing natural interspecific *Saccharomyces* hybrids, recent studies have been focused on de novo interspecific *Saccharomyces* hybrids to improve brewing solutions compared to their founders (Snoek, et al., 2015), (Krogerus, et al., 2018). However, Because of the sterility of interspecific hybrids, it is challenging to perform mating to generate segregants with recombination. Dr. Agnieszka Maslowska, Dr. Alex Hinks Roberts from the GACT group and Dr. Samina Neesab from the Manchester group overcame these difficulties in breeding and successfully cultivated *de novo* interspecific yeast diploid hybrids between *S. cerevisiae* strains and other *Saccharomyces* species including *S. cerevisiae* × *S. eubayanus*, *S. cerevisiae* × *S. kudriavzevii*, *S. cerevisiae* × *S. jurei* under different experiment methods. The high-resolution hybrid strains in 12 rounds of random mating with mosaic genomes were further generated and constructed for carrying genetic information of four parental strains where each species contributed 2 founder strains. For *S. cerevisiae*, 2 of the 4 strains in SGRP (NA, WE, SA, WA) were chosen as parental strains based on the largest growth difference under the tested condition. Each type of hybrid that carried the same species founders was also further constructed with mitochondria inheritance. For example, there are two types of yeast hybrid between *S. cerevisiae* × *S. eubayanus* in this study: one with *S. cerevisiae* mitochondria, the other carried *S. eubayanus* mitochondria. These hybrids were further screened

under different environment condition which are essential factors in brewing and fermentation. Under the different conditions, contrast population pool for 12th generation hybrid progeny was selected with extreme phenotypes which have the high fitness and poor fitness for each type of cross. With the rapid development of high-throughput genome sequencing, whole genome sequencing was applied in mapping genetic variation with complex traits under selection. To understand the genetic mechanisms that influence the hybrid growth performance and the difference response under environment stress, 36 pools in total were sequenced.

In this chapter, the aim is to identify the genetic characterisation for the different performance pool under environment stress selection. The whole genome sequencing data for each pool pools of the extremes in arrayed progeny were collected. For each pool, sequence mapping with the concatenated reference and SNP variant calling were performed. The reference genome of *S. kudriavzevii* in the public database was under scaffold level (6671 contigs) which was too fragment to guide the mapping and alignment of these de novo hybrids. The genome of one founder strain *S. kudriavzevii* IFO 1802 was re-sequenced and assembled to high quality contigs to use as reference in this study. For comparing the allele frequency difference for each marker in contrasting fitness population pools (high performance pool and low performance pool) under the same environmental condition, pool QTL mapping analysis was applied for each type of hybrid through Multipool (Edwards & Gifford, 2012). 36 groups were analysed and various numbers of markers had significant differences between high and low fitness for the same selection. In this study, a large number of genetic regions and causal genes were revealed in close association with different responses to temperature, maltose, ethanol and acetic acid. The analysis reveals the complex factors involved in growth of interspecific hybrids under environment stress. These findings would provide information and validation ideas for developing better performance strains for brewing and industrial application.

6.2 Methods and Materials

6.2.1 Hybrid Generation

The fertile hybrids were generated in two strategies by Dr. Alex Hinks Roberts and Dr. Samina Neesab. Each cross involved two strains for *S. cerevisiae* and two strains for another *Saccharomyces* species. One way for generating this hybrid cross is firstly crossed two strains of the same species to create two intraspecific diploid strains, one for species A and one for species B. Then deleting *MATα* locus in the species A to create species A with mating type MATa. MATa was then deleted in species B to create species B with mating type *MATα*. Two interspecific diploids carried different mating time were crossed to create interspecific tetraploid with 4 founder strain genomes. Then diploid gametes were sporulated with each founder species genome. The other way for generating fertile hybrid is to cross two strains under different species to create two intraspecific diploid strain that each diploid carrying species A with MAT locus deletion and species B mating type either MATa / *MATα*. These two diploid strains were then crossed to generate interspecific tetraploid and sporulated into diploid gamers with each founder species genome (Hinks Roberts, 2019). Figure 6-1 illustrates the strategy for fertile hybrid generation. With the fertile hybrids, 12th generation multi-parental lines were further created through mating and crossing for generating wide variety of diploid hybrid segregants. In addition, each diploid hybrid inherited mitochondrial DNA bi-parentally inherited, but colonies rapidly lose one type of mitochondria, becoming homoplasmic. Hence, each type of cross was then designed with two group of hybrids, one group carries mitochondria of species A and another group carries mitochondria of species B. The experimental detail of the de novo yeast hybrids generation included founder selection and phenotyping arrays can be found in (Hinks Roberts, 2019).

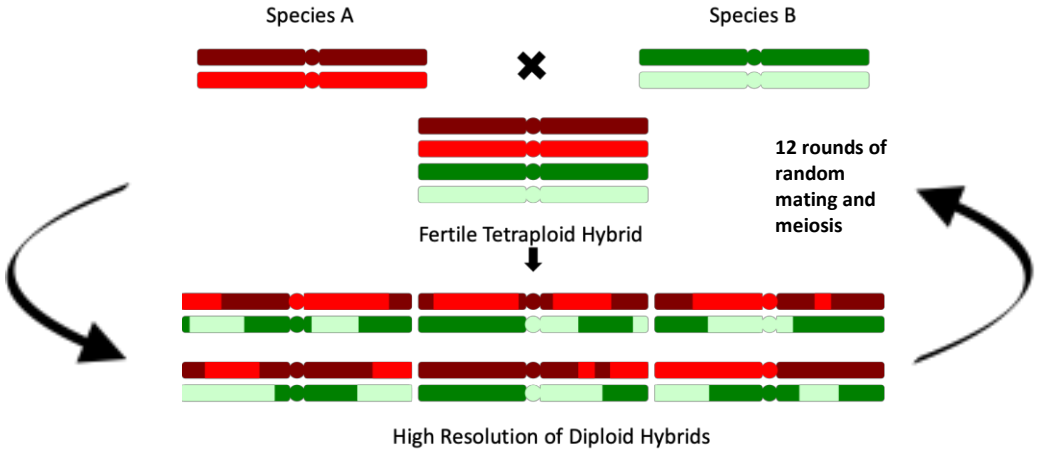


Figure 6-1 High resolution diploid hybrids

Intraspecific diploids are crossed generating a tetraploid hybrid. Sporulation produces diploid gametes, with crossovers occurring between sister chromosomes of the same parental species. F1 gametes can mate within the population creating a new tetraploid population. This is repeated eleven times to create a genetically diverse F12 population, with each individual genome unique, after extensive rearrangements between sister parental genomes.

6.2.2 Yeast strains and Pool selection

Table 6-1 Lists of Founder strains information

Species	Strain
S. cerevisiae	YPS128
S. cerevisiae	Y12
S. cerevisiae	DBVPG6765
S. kudriavzevii	IFO 1802
S. kudriavzevii	China
S. jurei	D5088
S. jurei	D5095
S. eubayanus	CBS12357
S. eubayanus	OS626 West China

Nine founder strains were involved for generating hybrids which included 3 strains of *S. cerevisiae*, 2 strains of *S. kudriavzevii*, 2 strains of *S. eubayanus* and 2 novel strains *S. jurei*. The founder strains for each hybrid group used in this study were listed in the Table 6-1. Three types of diploid hybrids between *S. cerevisiae* strains and other *Saccharomyces* species were generated including: *S. cerevisiae* × *S. eubayanus*, *S. cerevisiae* × *S. kudriavzevii*, *S. cerevisiae* × *S. jurei* under different cross methods. Each type of the diploid hybrid contains two hybrid groups carrying different mitochondria, which were generated from the same founder cross. 6 groups of hybrids were totally involved in this study which were named as HY1- HY6. HY1 and HY2 are the hybrids between *S. cerevisiae* × *S. jurei.* HY1 carried *S. cerevisiae* mitotype and HY2 carried *S. jurei* mitotype. HY3 and HY4 are the hybrids in *S. cerevisiae* × *S. kudriavzevii.* HY3 carried *S. cerevisiae* mitotype and HY4 carried *S. kudriavzevii* mitotype. These four groups of F12 diploid hybrids were pooled and selected under three conditions which were low temperature (12°C), maltose, and acetic acid. HY5 and HY6 are the hybrids in *S. cerevisiae* × *S. eubayanus.* HY5 carried *S. cerevisiae* mitotype and HY6 carried *S. eubayanus* mitotype. Table 6-2 summarised the founder strain and selection condition under each pool.

Table 6-2 Summary of pool samples information

Hybrid Group	Parental Lines *S. cerevisiae*	Parental Lines *Saccharomyces*	Mitotype	Treatment	Fitness Pool
HY1	Sc[YPS128+DBVPG6765]	Sj	Sc	12°C	High
HY1	Sc[YPS128+DBVPG6765]	Sj	Sc	12°C	Low
HY1	Sc[YPS128+DBVPG6765]	Sj	Sc	Maltose	High
HY1	Sc[YPS128+DBVPG6765]	Sj	Sc	Maltose	Low
HY1	Sc[YPS128+DBVPG6765]	Sj	Sc	Acectic acid	High
HY1	Sc[YPS128+DBVPG6765]	Sj	Sc	Acectic acid	Low
HY2	Sc[YPS128+DBVPG6765]	Sj	Sj	12°C	High
HY2	Sc[YPS128+DBVPG6765]	Sj	Sj	12°C	Low
HY2	Sc[YPS128+DBVPG6765]	Sj	Sj	Maltose	High
HY2	Sc[YPS128+DBVPG6765]	Sj	Sj	Maltose	Low
HY2	Sc[YPS128+DBVPG6765]	Sj	Sj	Acectic acid	High
HY2	Sc[YPS128+DBVPG6765]	Sj	Sj	Acectic acid	Low
HY3	Sc[YPS128+Y12]	Sk	Sc	12°C	High
HY3	Sc[YPS128+Y12]	Sk	Sc	12°C	Low
HY3	Sc[YPS128+Y12]	Sk	Sc	Maltose	High
HY3	Sc[YPS128+Y12]	Sk	Sc	Maltose	Low
HY3	Sc[YPS128+Y12]	Sk	Sc	Acectic acid	High
HY3	Sc[YPS128+Y12]	Sk	Sc	Acectic acid	Low
HY4	Sc[YPS128+Y12]	Sk	Sk	12°C	Low
HY4	Sc[YPS128+Y12]	Sk	Sk	Maltose	High
HY4	Sc[YPS128+Y12]	Sk	Sk	Maltose	Low
HY4	Sc[YPS128+Y12]	Sk	Sk	Acectic acid	High
HY4	Sc[YPS128+Y12]	Sk	Sk	Acectic acid	Low
HY5	Sc[YPS128+Y12]	Seub	Sc	40°C	High

HY5	Sc[YPS128+Y12]	Seub	Sc	40°C	Low
HY5	Sc[YPS128+Y12]	Seub	Sc	4°C	High
HY5	Sc[YPS128+Y12]	Seub	Sc	4°C	Low
HY5	Sc[YPS128+Y12]	Seub	Sc	Ethanol	High
HY5	Sc[YPS128+Y12]	Seub	Sc	Ethanol	Low
HY6	Sc[YPS128+Y12]	Seub	Seub	40°C	High
HY6	Sc[YPS128+Y12]	Seub	Seub	40°C	Low
HY6	Sc[YPS128+Y12]	Seub	Seub	4°C	High
HY6	Sc[YPS128+Y12]	Seub	Seub	4°C	Low
HY6	Sc[YPS128+Y12]	Seub	Seub	Ethanol	High
HY6	Sc[YPS128+Y12]	Seub	Seub	Ethanol	Low

6.2.3 Sequencing, Mapping and Variant Calling

The hybrids were sequenced by Earlham Institute and GTCF in the University of Manchester. Paired-end raw Illumina sequence reads in fastq format were quality checked through FastQC 0.11.5 (Andrews, 2010) and trimmed through Trimmomatic (Bolger, et al., 2014). Filtered reads (Figure 6-2) were aligned to a combined reference genome with parental species (*S. cerevisiae* YP128 (Yue, et al., 2017) concatenated with *S. eubayanus* CBS12357 (Libkind, et al., 2011)) / *S. jurei* D5088 (Naseeb, et al., 2018) / *S. kudriavzevii* (assembled genome)) using bwa/0.7.16a (Li & Durbin, 2009). Local realignment was then performed to minimise the number of mismatching bases through RealignerTargetCreator and IndelRealigner through GATK 3.8. MarkDuplicates tool was then performed with picard/2.6.0 for removing the optical duplicates to control the alignments quality for variant analysis. Samtools were applied on step with raw bam files for sorting and indexing (Li, et al., 2009). Variant calling was then applied on the aligned reads using freebayes/1.0.2 (Garrison & Marth, 2012) with ploidy setting at 1 with the combined reference, --min-mapping-quality 30 --min-base-quality 20 --no-mnps. Variant calling outputs in vcd files were then filtered with only SNP results and transferred to csv files that included information of CHROM, POS, REF, ALT, AO, RO where CHROM is the chromosome where SNP located, POS is physical position in reference genome, REF is the base of reference genome ALT is the base of variants, AO is the allele depth of alternatives and RO is the allele depth that have same base with reference. Pool SNPs were further compared to the SNPs shown between founders through R script. Reads depths below 10 were excluded. For each pool, Matching allele files were stored separately by chromosome in txt for next stage allele frequency tests between high fitness and low fitness. Variant calling pipeline was illustrated in the Figure 6-2. The analysis for mapping and variant calling were performed on HPC service with qsub files. SNP filtering to genotype were analysed in r with scripts. R scripts, qsub files, raw vcf files were stored in GACT Rdrive.

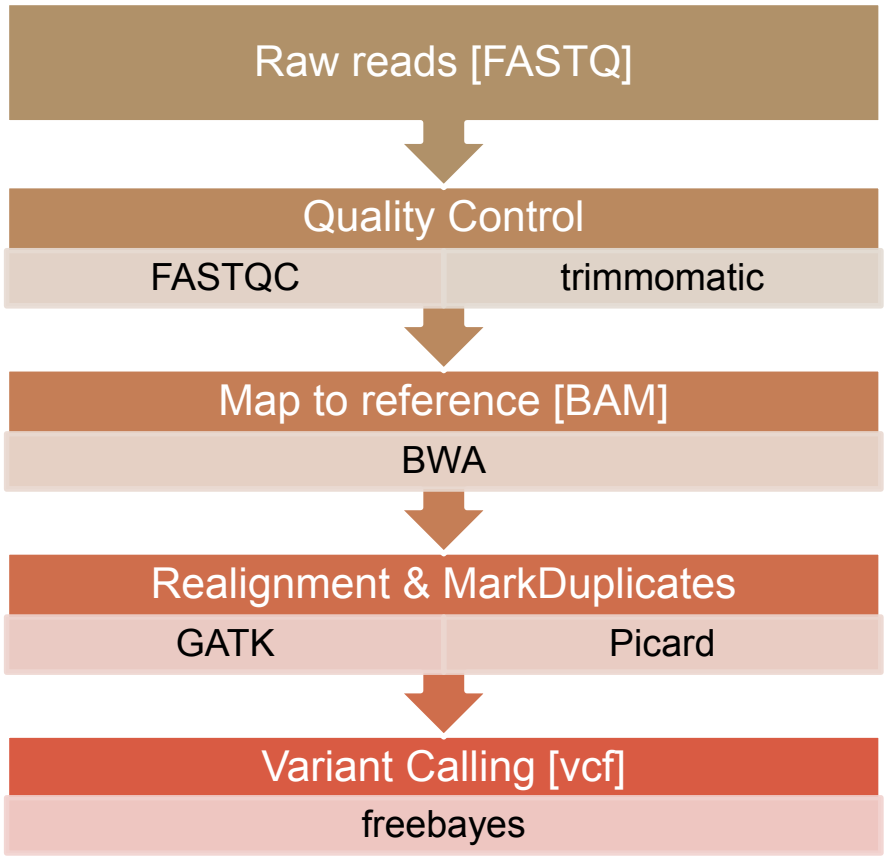


Figure 6-2 Variant Calling Workflow

6.2.4 *De novo* assembly

SPAdes assembler 3.9.0 was applied on the filtered reads of *S. kudriavzevii* IFO 1802 after quality check and trimming for assembly. Assembling quality was assessed through QUAST/4.3.

6.2.5 Allele frequency analysis

Every two pools that shown high fitness and low fitness with same cross founder and carried same mitochondria were further analysed through Multipool/0.10.2 programme for pool QTL analysis (Edwards & Gifford, 2012). Multipool analysis has shown the effectiveness in yeast QTL analysis. This method can identify the causal locus between 2 extreme pools which contains contrasting allele frequency enrichment intervals (Lee, et al., 2016). The pools with high resolution

segregants carried polymorphic loci. Rather than computing LOD score for each marker, Multipool compared the allele frequency difference under pool sequencing in bins in the chromosome through dynamic Bayesian network (DBN) model to estimate the locus location under the assumption for recombination rate in uniform distribution. Allele frequency for high fitness and low fitness with LOD score were displayed in the output figures generated through python 2.7 for each chromosome. Multipool were performed with setting: contrast mode, 3300 bp cM and 100 bp bins. The LOD score is calculated for the contrast degree in bins between high and low fitness pool. The QTL intervals from multipool results were located with span at least 20kb. Genetic features were annotated by 1-LOD drop interval from the peak marker bins.

6.3 Results

6.3.1 Assembly of *S. kudriavzevii* genome

Assembly of the available public genome data of *S. kudriavzevii* IFO1802 are in contig level with large amount fragmented genome which GCA_000167075.2 (Hittinger, et al., 2010) included 2054 scaffolds with total length 11,189,057 and UColDMed_2011_SRX055455 included 6671 scaffolds. In order to get high quality hybrid mapping, it is essential to improve the assembly of *S. kudriavzevii* genome sequence. In this study, an improved reference genome for *S. kudriavzevii* IFO1802 were supplied. Table 6-3 summarised the assessment of the assembly through quast. It clearly shows that from the number of contigs and the total length the assembly performed better. In the next stage, the analysis for *S. kudriavzevii* will use this assembly as reference genome. 5806 CDS genes were annotated by MAKER software using *S. cerevisiae* S288C genome coding sequences.

Table 6-3 Genome assembly report for scaffolding of *S. kudriavzevii* IFO1802

Assessment	Scaffold
Number of contigs (>= 0bp)	337
Number of contigs (>= 1000bp)	108
Number of contigs (>= 5000bp)	67
Number of contigs (>= 10000bp)	59
Number of contigs (>= 25000bp)	49
Number of contigs (>= 50000bp)	44
Total length (>= 0bp)	11713061
Total length (>= 1000bp)	11639197
Total length (>= 5000bp)	11545229
Total length (>= 10000bp)	11486219
Total length (>= 25000bp)	11315714
Total length (>= 50000bp)	11137286
Largest contig	653594
GC (%)	39.61
N50	319497
N75	201332
L50	13
L75	24
Coverage	378

Table 6-4 Lists of QTL numbers for *S. cerevisiae* × *S. jurei* hybrids

Hybrid Group	Selection	*S. cerevisiae* QTLs	*S. jurei* QTLs
HY1	12°C	4	13
HY1	Maltose	8	7
HY1	Acetic acid	7	8
HY2	12°C	17	20
HY2	Maltose	12	9
HY2	Acetic acid	8	6

6.3.2 Identification of QTLs in hybrid *S. cerevisiae* × *S. jurei*

When examining the difference in allele frequency among hybrids pools, a large number of intervals with contrasting allele frequency structure were identified under high fitness pool and poor fitness pool. Different QTL intervals with large allele frequency difference were identified through Multipool analysis (
Table 6-4). 17 prominent QTL intervals were identified in HY1 group under the cold temperature 12°C. Four QTL intervals were identified in *S. cerevisiae*. Noticeably, 2 of the QTL intervals in chromosome V (99300 - 138200) and chromosome XIV (326500 - 344500) linked to three genes (GIM4, ALF1, GIM3) are involved in microtubule biogenesis (Figure 6-3). GIM3 was validated in the previous study as cold-sensitive genes (Geissler, et al., 1998). In addition, 5 genes (VPS73, YEA6, GGC1, MRPL11, RMD9) were clustered with relatedness in mitochondrial protein or transporter. Different QTL intervals were identified corresponding *S. cerevisiae* alleles and *S. jurei* alleles for HY1 (Figure 6-4). In chromosome IV, one QTL interval range from 95900 to 107100 was located from *S. cerevisiae* alleles where no effect was shown as *S. jurei* alleles. In contrast, four QTL intervals contained in range from 382600 to 1453300 were located from *S. jurei* alleles but not identified in *S. cerevisiae* QTLs. Moreover, 5 genes (GAL7, GAL10, GAL1, SIM1, GAL3) identified in *S. jurei* allele that involved in carbohydrate metabolism pathway. Different genes were identified from HY2 group under the same condition. However, function clusters were shown similarity with HY1. Among the genes localised though the QTL intervals, 7 genes (ASK1, BIK1, TIM19, KAR9, SPC34, DYN2, TUB4) are also involved in microtubule. Furthermore, 5 genes (FUS1, FUS3, KSS1, PKC1, MSG5) were involved in MAPK signaling pathway.

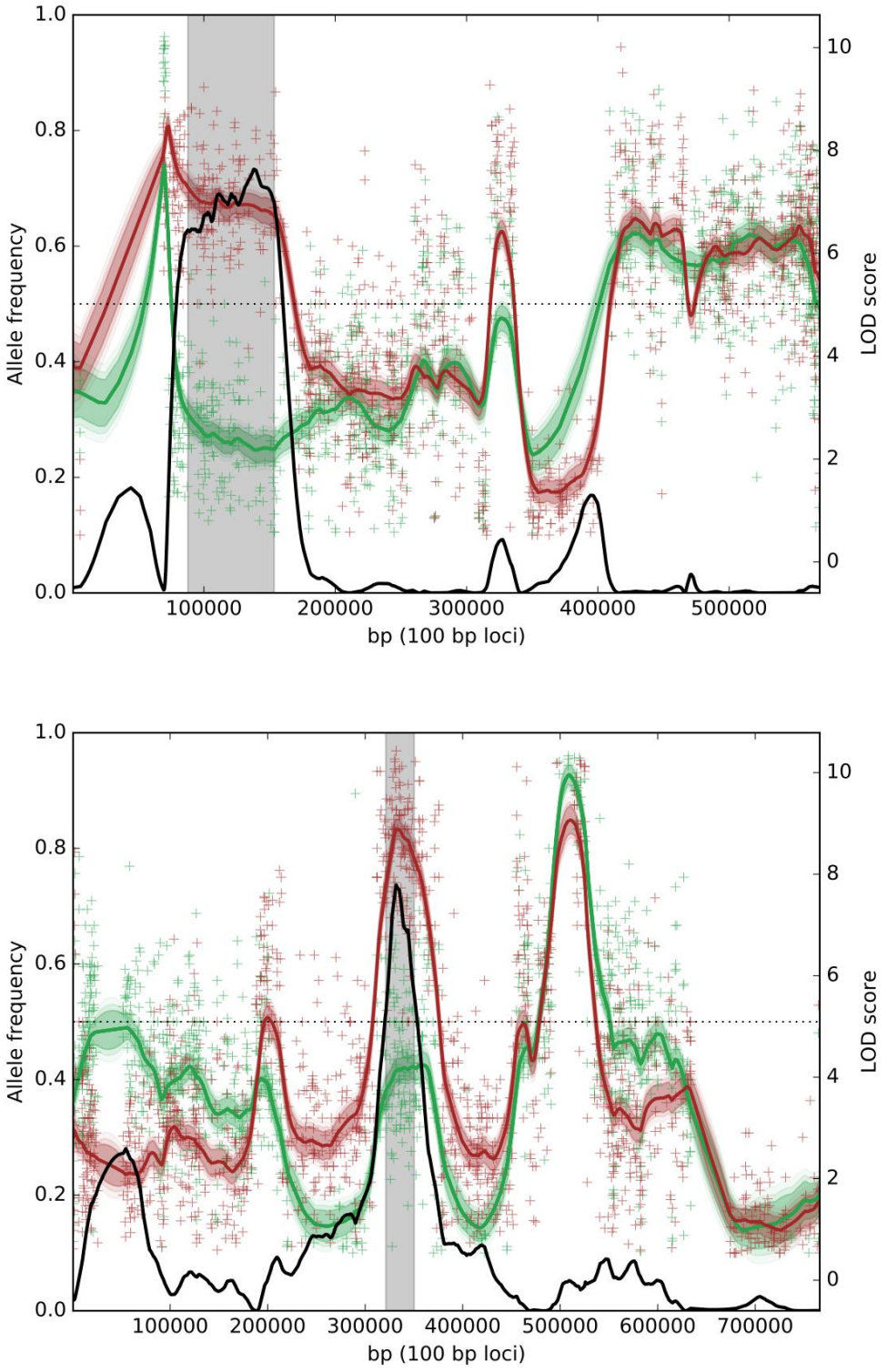


Figure 6-3 Multipool output for HY1 chromosome V and XIV

Allele frequencies were displayed with causative locus region. Red lines and dots represent the allele frequency of variants in the high fitness pool and green for poor fitness. Black lines represent LOD scores. Peaks were shown distinct separation of allele frequencies.

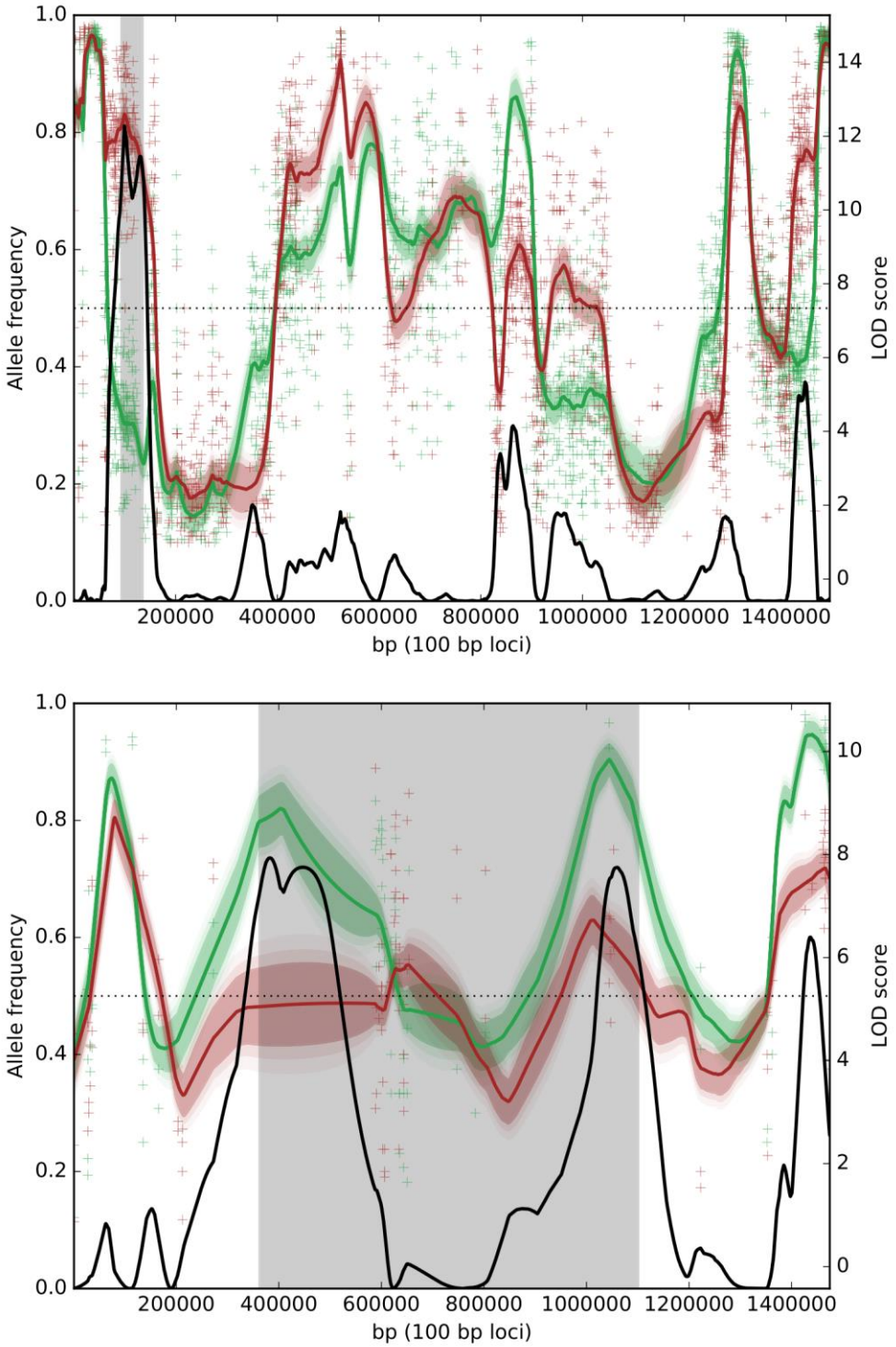


Figure 6-4 Different QTLs were identified for HY1 on chromosome IV

The figure on the top illustrated the allele frequency and LOD scores of the comparison in S. cerevisiae. The figure on the bottom illustrated the allele frequency and LOD scores of the comparison in S. jurei. Red lines and dots represent the allele frequency of variants in the high fitness pool and green for poor fitness.

Apart from selection under temperature, maltose tolerance is another major feature in brewing strain assessment. Several peaks were shown extreme difference of allele frequency in both of HY1 and HY2 under high maltose. Interestingly, both of the analysis for HY1 and HY2 shown the largest difference on the tail of chromosome II located the peak at 790700 included in the gene SUL1 which control the sulfate uptake (Figure 6-5). In addition, these peak intervals were close to the subtelomere region of MAL gene families where MAL31 which was the high-affinity maltose transporter, MAL33 and MAL32 involved in maltose catabolism (Louis, et al., 2014). In the analysis for HY1, several genes were shown relatedness with glycoprotein were also shown significance under maltose selection (GAS1, YOR1, DDR2, GTB1, INA1, IRC18, PUN1, VBA2, PRC1, BGL2, FKS3, SUL1). Two genes (ADH2, ADH3) located on chromosome XIII which are ADH members that involved in the pathway of degradation of aromatic compounds.

In addition, acetic acid is also an important feature in industrial brewing. However less QTLs shown than the other two selection. 7 QTL intervals were linked to genes (LEU2, MXR2, LPD1, GRX1, HBN1, HIS4, FRM2) that involved in oxidation-reduction process which is expecting under acetic acid as well as ADH3.

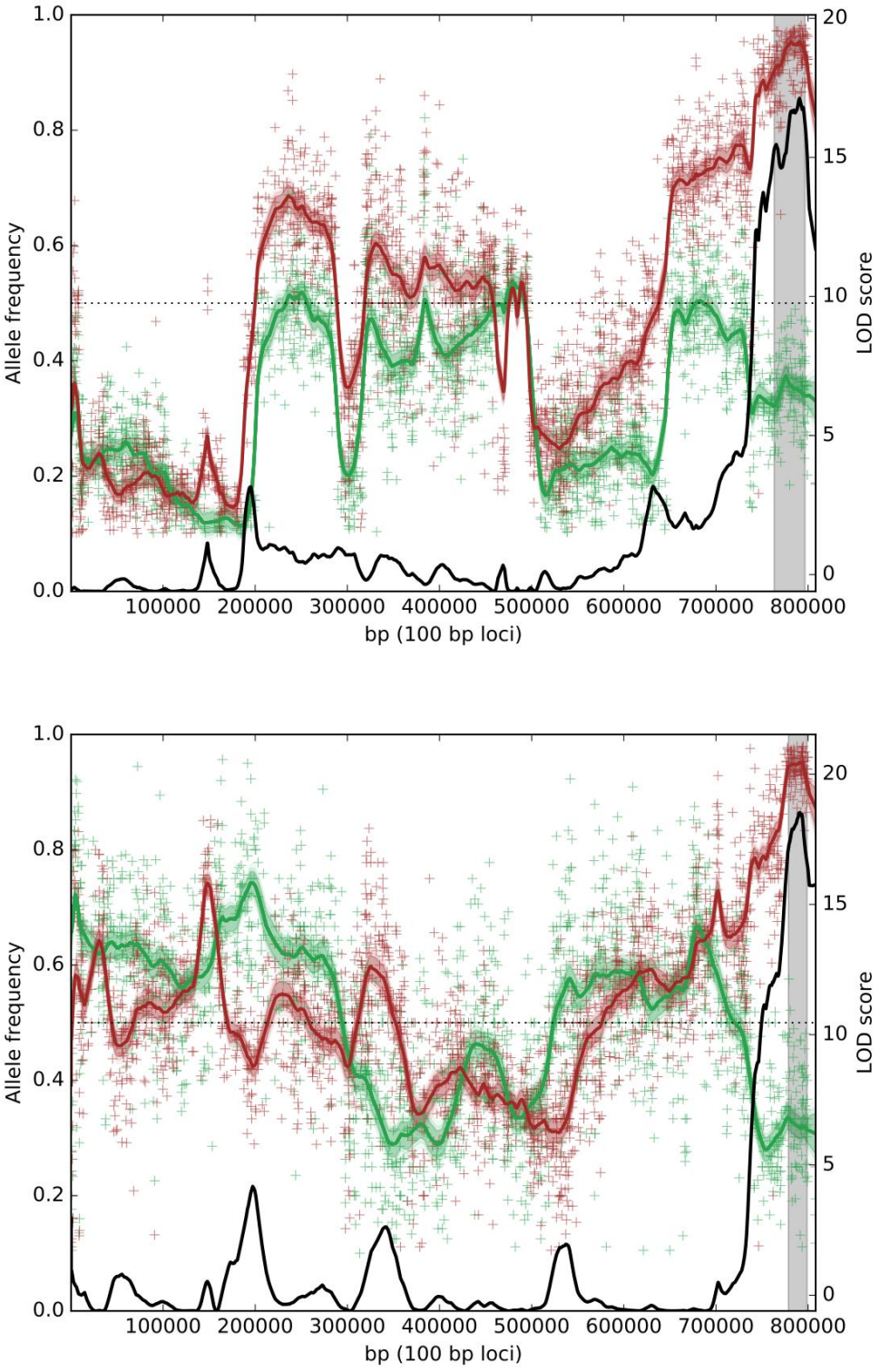


Figure 6-5 Multipool output for HY1 and HY2 chromosome II under maltose

Allele frequencies were displayed with causative locus region. Red lines and dots represent the allele frequency of variants in the high fitness pool and green for poor fitness. Black lines represent LOD scores. Peaks were shown distinct separation of allele frequencies.

6.3.3 Identification of QTLs in hybrid *S. cerevisiae* × *S. kudriavzevii*

For the hybrids group HY3 and HY4 with *S. cerevisiae* × *S. kudriavzevii*, a large amount of QTL intervals were identified with distinct allele distribution between high and low fitness (Table 6-5). Under Low temperature, 7 QTL intervals were identified in HY3 as significance under the *S. kudriavzevii* allele comparison where linking to genes (ERP5, VAC7, MNT2, PMT6, PLB3, MNT4, YCR061W) related to Glycoprotein. Among these genes, MNT2 and MNT4 are the obvious genes that are known as mannosyltransferase involved in O-linked glucosylation (Romero, et al., 1999). In addition, many genes identified in *S. cerevisiae* comparison of HY3 were related to mitochondrion and transmembrane.

Table 6-5 Lists of QTL numbers for *S. cerevisiae* × *S. kudriavzevii* hybrids

Hybrid Group	Selection	*S. cerevisiae* QTLs	*S. kudriavzevii* QTLs
HY3	12°C	32	13
HY3	Maltose	38	32
HY3	Acetic acid	24	28
HY4	12°C	20	25
HY4	Maltose	24	24
HY4	Acetic acid	20	10

Under the selection of maltose, QTL intervals were located close to the subtelomere region of chromosome II where shown the obvious difference under high fitness pool and low fitness pool both in HY3 and HY4 (Figure 6-6). Apart from overlaps on chromosome II, three more QTL intervals were overlapped across the genome between HY3 and HY4 under maltose. These overlapping regions were included 15 genes that are located on chromosome XIII, XIV, XV. 6 genes (ENV9, ERG24, DGA1, YTA12, PRM1) involves in transmembrane. Several QTL intervals were identified for the difference of high and poor fitness under acetic acid. Among them, two QTL intervals in HY3 group, located on chromosome VIII, were linked to three genes (STB5, SKN7, SCH9) that responses to oxidative stress. Gene STB5 is an essential for regulating pentose

phosphate pathway (Larochelle, et al., 2006) and acetaldehyde tolerance (Matsufuji, et al., 2010)

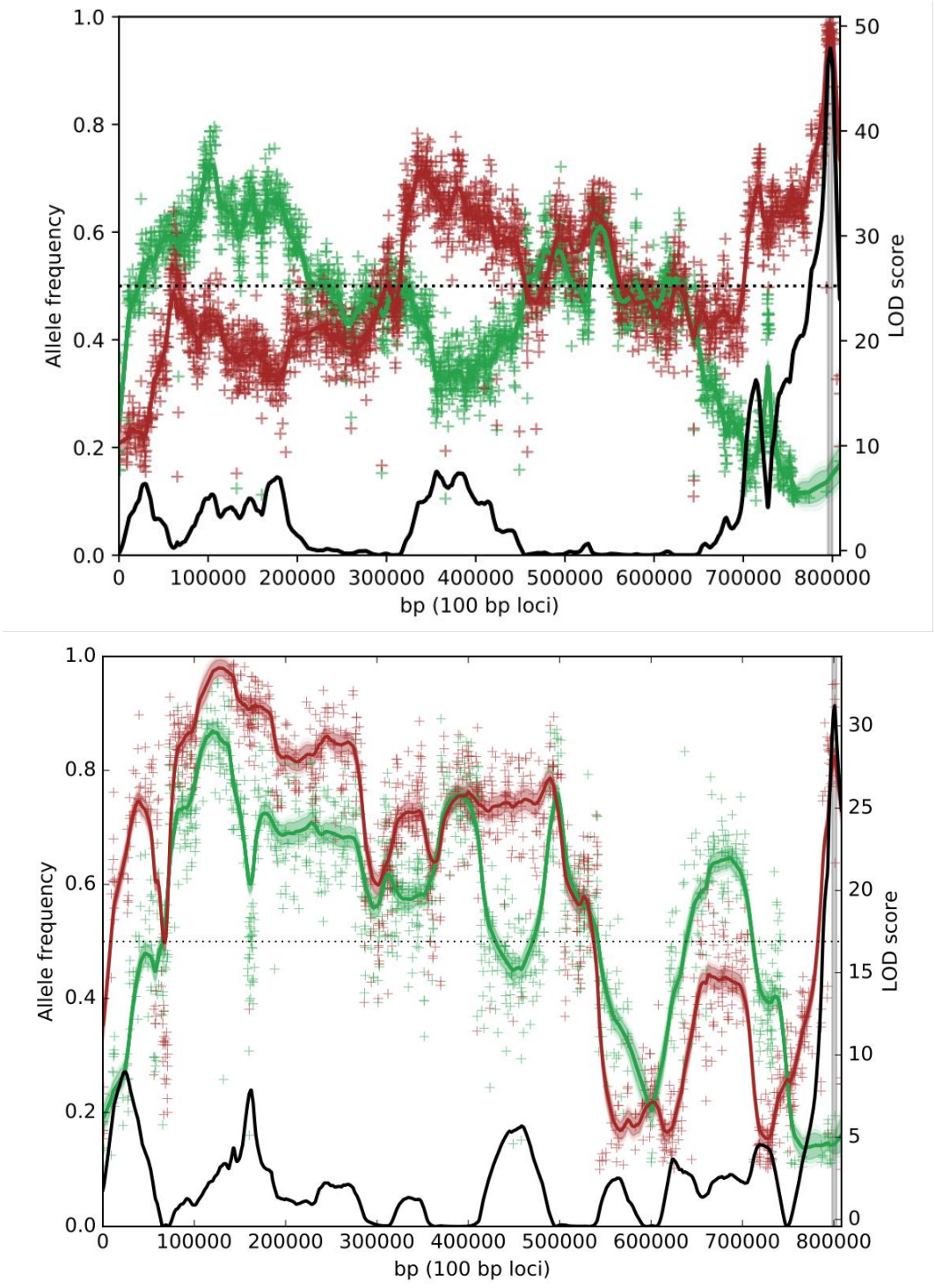


Figure 6-6 Multipool output for HY3 and HY4 chromosome II under maltose selection

6.3.4 Assessment of QTL in hybrid *S. cerevisiae* × *S. eubayanus*

Different selection conditions were performed on the hybrid *S. cerevisiae* × *S. eubayanus* that are high temperature in 40 °C, cold temperature in 4 °C and ethanol. The hybrids carrying different mitochondria identified different QTLs sets, especially under high temperature and ethanol conditions. HY5 carries mitochondria of *S. cerevisiae*, and HY6 carries mitochondria of *S. eubayanus* (Table 6-6). For selection of high temperature, two QTL intervals were overlapped between HY5 and HY6 that linked to 11 genes. The first overlapping region were located at chromosome II linked to four adjacent genes (APL3, YBL036C, POL12, STU1). The other overlapping region were located at chromosome XV linked to 7 genes (SFG1, DGK1, SLY41, SNU66, SPS4M NOP58). SLY41 involved in ER to Golgi transport and DGK1 involved in diacylglycerol kinase activity and suppress of SLY1 which has temperature sensitive mutation (Kosodo, et al., 2001). Interestingly, SLY1 were not shown in the peak or drop in the QTL intervals but both HY5 and HY6 identified causal regions close to SLY1 which were on chromosome IV 792100 - 820600 in HY5 and 827200 - 849000 in HY6 (Figure 6-7). These two QTL intervals are both adjacent to SLY1 but not covered within this gene. This might be caused by that the high effect locus were located on the highest peak causing a narrow 1-LOD drop range, excluding locus which shown lower effect but still over threshold. Hence, a further verification on the hybrids is required to localise the causal locus.

Table 6-6 Lists of QTL numbers for *S. cerevisiae* × *S. eubayanus* hybrids

Hybrid Group	Selection	*S. cerevisiae* QTLs	*S. eubayanus* QTLs
HY5	40°C	18	10
HY5	4°C	13	8
HY5	Eth	15	8
HY6	40°C	26	13
HY6	4°C	13	11
HY6	Eth	26	14

Freezing temperature condition was applied on these hybrids at 4 Celsius, which is a very stressful condition for yeast growth. One QTL interval was located at subtelomere region in chromosome I under for *S. cerevisiae* in HY5. In this region, the well-known gene FLO1 was occurred in both QTL intervals. Several studies have already shown the evidence of FLO1 in contribution of stress tolerance during fermentation under cold temperature (Deed, et al., 2017). Several QTL intervals for HY5 linked to genes for DNA/RNA helicase that affected bring cold stress including DBP9, HAS2, SUB2, MSS116, MOT1, MCM5, RVB1. In HY6, a number of QTL intervals linked to genes for telomere maintenance (PIF1, RRM3, STM1, SWD3, GBP2) and DNA repair (PCD1, OGG1, PMS1, HRQ1, PIF1, HUG1, MKT1, REV3, RRM3, NSE3, MLH3). Ethanol tolerance is one of the most desired features for industrial application of yeasts.

Several QTL intervals were overlapped in *S. cerevisiae* comparison and *S. eubayanus* comparison. Among the overlaps, Gene MPD1 were occurred in HY5 and HY6 under comparison of *S. eubayanus*. MPD1 involved in protein folding and recent studies have showed the correlation between ethanol stress response with unfolded protein response (Navarro-Tapia, et al., 2018). In addition, GAL1, GAL7 and GAL10 involved in carbohydrate metabolism pathway were included in QTL intervals of HY5 under *S. cerevisiae*. In HY6, 4 genes (REV3, ACO2, LIP5, TYW1) were included in QTL intervals for HY6 *S. cerevisiae* comparison that shown relatedness on iron-sulfur cluster binding. Ethanol is known to be associated with loss of iron homeostasis. One previous study showed that the consequence of increasing free Fe^{2+} could be affecting iron-sulfur clusters and potentially affect mitochondrial function (Gomez, et al., 2014).

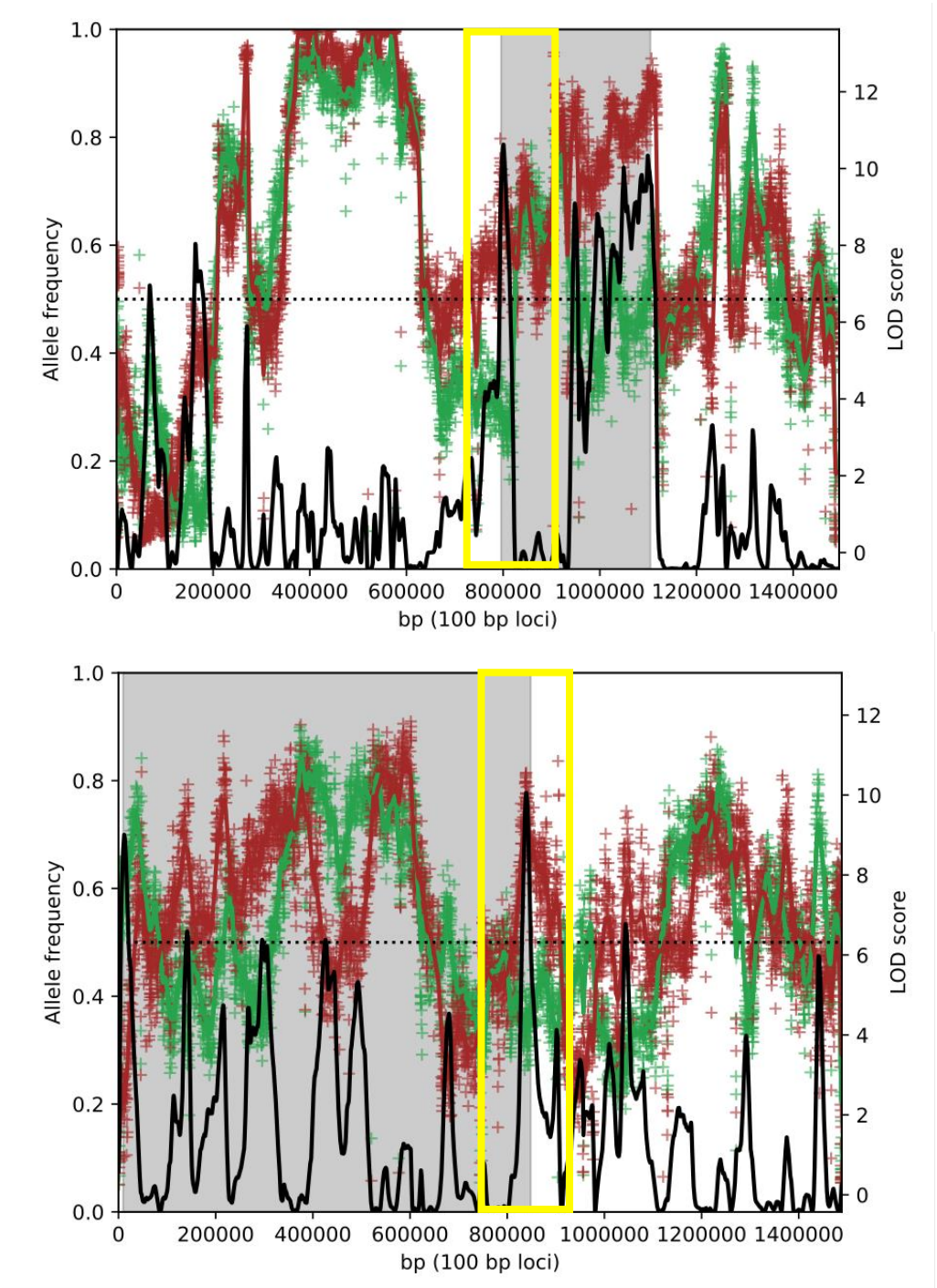


Figure 6-7 Multipool output for HY5 and HY6 chromosome IV under high temperature selection.

The peak QTLs that adjacent to SLY1 were highlighted in yellow box.

6.4 Discussion

In this chapter, the genetic features of de novo interspecies hybrids *S. cerevisiae* × *S. eubayanus*, *S. cerevisiae* × *S. kudriavzevii*, *S. cerevisiae* × *S. jurei* under different environment stress conditions were assessed using pool QTL mapping approach. The fertile yeast hybrids opened the door for generating high resolution hybrid segregants. After 12 rounds of mating and recombination, the phenotype of the 12th generation hybrid segregants achieved a high diversity that allowed pool selection for extreme performance in high fitness and low fitness. Many genetic variants with polymorphic markers were identified through whole genome sequencing and variant calling. Under the QTL analysis through Multipool, a large number of QTL intervals were identified under the comparison between high fitness and low fitness pools for 6 hybrid groups with 3 different conditions. A lot of candidate genes and regions were identified in different hybrid groups. Among these genes, multiple factors contribute to the fitness performance of yeast hybrids. The number of *S. cerevisiae* variants of all six hybrid groups are much larger than the number of the variants in other *Saccharomyces* species. However, the number of QTLs identified through Multipool varies among different hybrid groups. For the hybrids in crosses *S. cerevisiae* × *S. jurei* and *S. cerevisiae* × *S. eubayanus*, more QTLs of *S. cerevisiae* alleles were identified in the hybrids who inherited non *S. cerevisiae* mitochondria comparing to the hybrids who inherited *S. cerevisiae* mitochondria. Moreover, the number of QTLs detected in *S. cerevisiae* alleles is much larger than the QTLs identified in *S. jurei* and *S. eubayanus* alleles. In the other hand, the largest number of QTLs were identified from hybrid group of *S. cerevisiae* × *S. kudriavzevii* among all conditions. The largest set of markers were identified as QTLs in both *S. cerevisiae* alleles and *S. kudriavzevii* alleles under maltose condition. Subtelomeric regions were identified in almost every experiment for the contribution of maintenance under stress condition associated with the brewing environment. MAL gene families at the end of chromosome II exhibited distinct allele frequency differences between high fitness and low fitness with the selection of maltose stress. In addition, adhesion gene FLO1 were identified under freezing temperature in hybrid of *S.*

cerevisiae and *S. eubayanus*. Furthermore, under the selection of acetic acid, genes involved in oxidation-reduction process were identified. An improved genome sequence of *S. kudriavzevii* IFO1802 were assembled and annotated. This genome presents a longer assembly at chromosome level and smaller gap size than then the public reference genome. 5806 coding genes were annotated. With the advent of long read DNA sequencing technic such as Nanopore sequencing and Pacbio, repetitive elements and subtelomere region could be further assembled.

Chapter 7 Concluding Remarks and Future Work

7.1 Conclusion

Quantitative traits of yeast cell growth were studied for different purposes under different agents and different perspectives. A large volume of genotype data and experimental datasets for yeast were collected and analysed for understanding the role of the genetic diversity and its relationship to the responses of growth. Different yeast strains exhibit a high diversity of phenotypes.

The first part of this project is using yeast as model organism to identify candidate genes that are related to growth under chemical agents for human study. Different causative markers were identified with different treatments. Some of these markers are well known for the corresponding agents, which validates the approach. In Chapter 2, the genotype data of six bi-parental cross F1 segregants, generated with four founders NA, WE, SA, WA in SGRP4, that have clean lineages and their growth phenotypes under DOX treatment were examined. Phenotypic diversity was observed for F1 segregants under DOX treatment. QTL analysis captured the association between the genotypes and the changes of phenotype which can be used for genetic feature selection. By comparing overlaps of QTL intervals with different F1 crosses, 9 candidate genes (SPF1, SLN1, AAC1, PMR1, CCC1, HFA1, UPS1, CCP1) were localised. In addition, two-dimensional QTL analysis identified epistasis effects involved in degradation pathway. These QTLs were linked to genes involved in metal ion binding and mitochondrial processes. The analysis result suggests that these F1 yeast segregants are useful in identifying candidate genes for various treatments. However, the F1 genotype data contains limited markers leading to only a few QTLs identified within a large genomic span.

In Chapter 3, various agents were applied to the high resolution 12th generation 4-way cross segregants, SGRP4-X, that displayed larger phenotypic variation than the F1 segregants. The distinct response to agents and the dense markers leads to a large number of loci identified by QTL analysis. Narrower QTL intervals were identified which targets a finer range down to a single gene in some cases. Different complex trait landscapes come out for yeast growth under different agents. No master QTL or genes occurred in every QTL analysis. However, in most of genotoxic agents, shared QTL underlying responses to DNA damage were observed. The frequently occurring genes among these agents were further analysed and QTL interactions were computed.

In Chapter 3, a massive QTL report for the analysis of the F12 generation data provides a lot of interesting sites for downstream analysis, including gene candidate nucleotide polymorphism and functional clustering for further experimental validation. From the QTL plot, it is apparent that a number of significant markers are located in the vicinity of adjacent regions. This might potentially cause the risk that adjacent neutral variants are identified as QTLs linked to genes emerging in one cluster as the causal QTLs were shown large effect so that adjacent regions were included from the statistical analysis. To solve this complexity and locate candidate genes in a cluster, genetic linkages between F12 populations were assessed and genetic linkages were assessed among F12 population and fine mapping were applied to find the causative genes with the modelling in effect changing detection. By identifying QTL clusters rather than focusing on the top-ranking significance, the selected gene candidates were able to be included with large and small effect.

In Chapter 5, as time is an important factor correlated with the dynamic changes of yeast growth, temporal QTL analysis was performed to analyse F12 genetic variants under an expanded phenotype dimension assessing growth rate in each hour. Large-scale analysis has produced a large number of genetic findings indicating the importance of tracking growth traits that develop over time. In the temporal QTL analysis of X-rays, many genes were identified in stages rather

than showing an effect throughout the growth period. The growth of the segregants were disturbed by short-term responses and long-term responses QTLs that showed accumulation effect. This provides a high throughput way for tracking dynamic responses of natural genetic variation to stress conditions, in addition to the advantages for studying complex traits using the SGRP4X population.

Apart from the analysis for F12 multi-parental lines of *S. cerevisiae*, a de novo hybrid yeast QTL analysis was performed in Chapter 6. Whole genome sequences of interspecies hybrids *S. cerevisiae* × *S. eubayanus*, *S. cerevisiae* × *S. kudriavzevii*, *S. cerevisiae* × *S. jurei* were used for identifying genetic variants with responses to different fermenting tolerance conditions. Each of these hybrids were used to generate high resolution hybrid segregants in 12^{th} generation populations as with the SGRP-4 F12 population. These hybrids have a large number of genetic variants inherited from their founders which were identified by variant calling after mapping to a reference for each parental species. The diverse phenotypic range in each hybrid gave the chance for selecting samples into two extreme pools based on the fitness performance. The multipool QTL analysis revealed multiple alleles were associated with the performance under biotechnological and fermentation related traits. Different QTLs emerged from the segregants generated in same founders with different mitochondria. These results indicate that mitochondria might interact with the nuclear genome and control the responses to the stress conditions. Also, subtelomeric regions were found to contribute to responses to stress conditions, especially significant in the tolerance of high concentrations of maltose.

Overall, these results suggested that the formation of complex traits may involve complex mechanisms with interactions of multiple factors under dynamics of environment and time. From the QTL analysis under other chemotherapy agents, known involved genes are identified from the analysis such as genes for responses of DNA damage, TOR1 gene detected as expected in the gene sets for Rapamycin. Several novel genes with indirect role were also identified. For

DOX treatment, genes related to metal ion binding and mitochondrion were observed. Functional mostly unknown gene MNN14 was identified through the analysis under multiple agents and validation on PQ has shown positive (Georghiou, 2017). For fermentation targets, identified genes showed direct and indirect mechanisms of the maintenance and metabolism. Moreover, Epistasis are further looked for DOX treatment, interactions were observed related to degradation pathways that were missed in the single QTL scan. The approaches for looking QTL analysis in temporal dynamic under X-ray radiation observed gene RAD4. Through the modelling and analysing of yeast complex trait landscapes, many candidate genes and regions were identified that can be applied for further validation and screening of industrial strains.

7.2 Future Work

De novo hybrid analysis was performed for pooling samples with robust multipool QTL analysis. In the future, a high throughput arrays with individual sequencing segregants could be used in QTL analysis. This can help to further identify causative makers within a finer range and locate involved causative genes with better statistical power. The limitation of multipool analysis is that although non additive markers can be mapped, it is not possible to identify the interaction patterns. With the development of analysis method and further data supplied, epistasis effect (i.e gene-gene interaction) could be detected for yeast *de novo* hybrids. In addition, the current analysis considered allele effect separately under comparison with each founder due to the fact that multipool only provides analysis for haplotype. Hence, a dedicated QTL method for hybrid yeast can be further developed.

Apart from the QTL analysis for polymorphism markers, it is possible that the structural variation of hybrid can give the additional perspective for yeast variation and evolution. Further experiments and corresponding data are required for that analysis. In addition, in this study, only diploid hybrid analysis was summarised. Additionally, ploidy is an important feature of yeast hybrid genome that affects the hybrid's functional attributes. Besides of diploid hybrid yeast, triploid and allotriploid yeast are widely found in the nature and used in industry. Fertile tetraploidy has been successfully generated through a haploid yeast and a triploid yeast to produce high diversity hybrid progeny. Further analysis will focus on this interesting point as well.

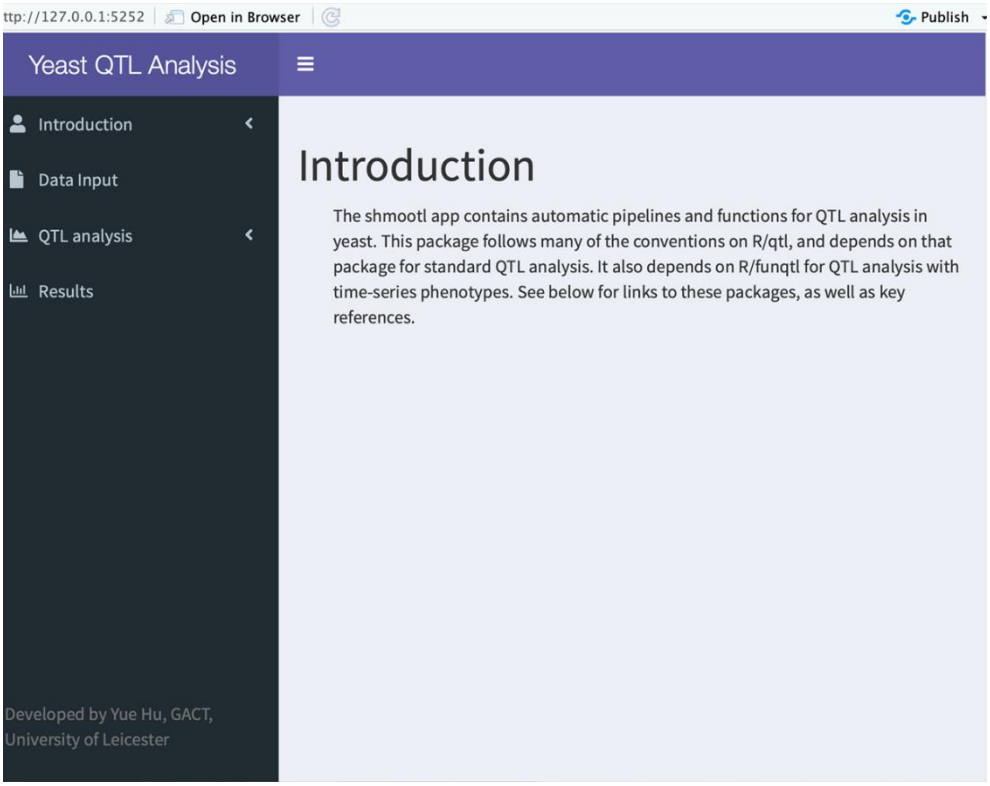


Figure 7-1 Overview of the application

During this PhD project, I was involved in the development of r package shmootl for yeast QTL analysis. In order to help researchers easily access the data analysis pipeline, an interactive web application was implemented by the author using the R Shiny package. The application integrated functions from the Shmootl/R package and provided an easy to use graphical user interface which can either run on a server or a personal PC with R installation of version higher

than 3.5.3. The graphical interface is user friendly, providing a navigation panel which allows the user to quickly go to different stages (Figure 7-1).

This application could be used simultaneously by multiple users, improving the analyse speed for different projects. This application takes several input files required by different process stages, then output a HDF5 file containing detailed computation of LOD scores for each data point and the threshold of permutation. This application contains four sections. The first section introduces the purpose of the application. This section also explained the data analysis workflow from data input to generating results. The second section is the data input section. A user can upload five different files which will be used in the process. These files include:

1. QTL input file including phenotype data and genotype data.
2. Genetic map file
3. Physical map file
4. Covariate file including mating type information
5. Annotation file

The input files will be uploaded to the server so that the user's original data file wouldn't be changed. After uploading the files, the user can easily follow all the steps by selecting the desired settings with the easy to use user interface. All steps can be completed within a few clicks on the buttons of the application. A default setting is provided for the user as a guideline on how to setup each stage. There are three stages of data analysis implemented in this application. The first stage is data preparation. During this stage input data will be normalised according to the user's choice, then a genetic map will be generated based on the normalised data. The genetic map will then be stored to the output report. The second stage is single QTL analysis. This stage will take the output genetic map from the previous stage as input as perform scan one QTL analysis based on user setting. The user can assign a LOD threshold, a significance level, and a number of permutations times to the application (Figure 7-2).

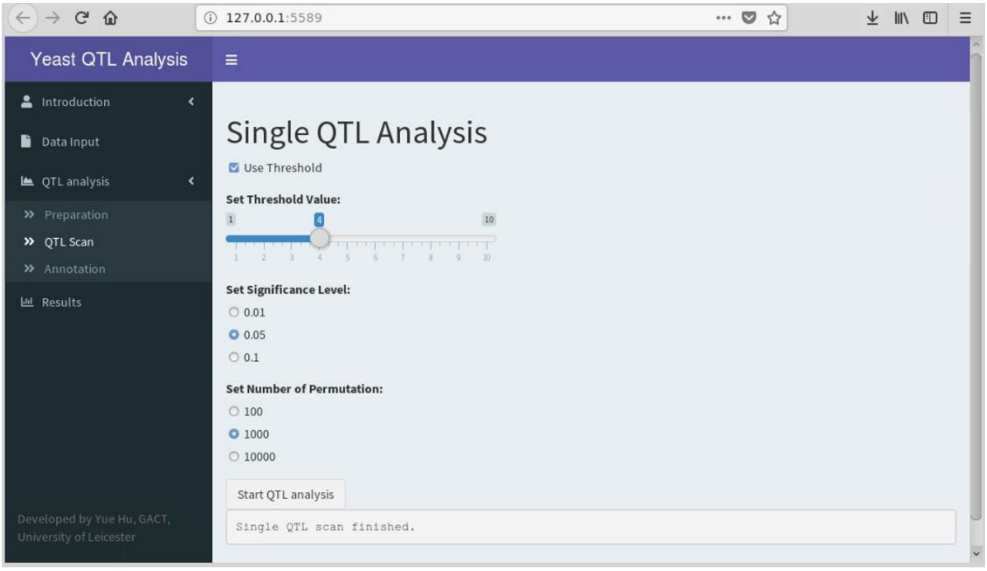


Figure 7-2 Single QTL Analysis page

The output of this stage will then be annotated during the final stage with the gene annotation file based on the reference genome. After the data analysis pipeline, the user can choose to download the report in excel, pdf, or hdf5 format. The hdf5 format contained details about the user settings, which is useful for adjusting parameters in a finer grade. The application offered basic functionalities at this stage. More integration of the Shmootl/R library functions could be done in the future, such as scan two functions. Further improvement on the application could be to allow the usage of high-performance computing cluster, so that the analysis could be speed up. Another improvement could be implementing real-time data visualisation as a part of output with graphics and tables. This would give the researchers a direct view and assessment of the experimental data. Further integrations like Intermine could also be done, so that the user can convert the analysis result directly into ready to use data such as getting the gene function and build complex queries to find homologues.

Bibliography

Almayouf, S., 2018. *Personal Communication.* Leicester: University of Leicester.

Andrews, S., 2010. *FastQC: a quality control tool for high throughput sequence data.* Cambridge: Babraham Bioinformatics, Babraham Institute.

Ansorge, W. J., 2009. Next-generation DNA sequencing techniques. *New biotechnology,* 25(4), pp. 195-203.

Antony, E., Khubchandani, S., Chen, S. & Hingorani, M. M., 2006. Contribution of Msh2 and Msh6 subunits to the asymmetric ATPase and DNA mismatch binding activities of Saccharomyces cerevisiae Msh2--Msh6 mismatch repair protein. *DNA repair,* Volume 5, pp. 153-162.

Arnez, J. G. & Moras, D., 1997. Structural and functional considerations of the aminoacylation reaction. *Trends in Biochemical Sciences,* 22(6), pp. 211-216.

Barton, D. B. H. et al., 2018. PHENOS: a high-throughput and flexible tool for microorganism growth phenotyping on solid media. *BMC microbiology,* Volume 18, p. 9.

Bergström, A. et al., 2014. A high-definition view of functional genetic variation from natural yeast genomes. *Molecular biology and evolution,* Volume 31, pp. 872-888.

Blackburn, E. H., 1991. Structure and function of telomeres. *Nature,* 350(6319), p. 569.

Bloom, J. S. et al., 2013. Finding the sources of missing heritability in a yeast cross. *Nature,* Volume 494, pp. 234-237.

Bokulich, N. A. & Bamforth, C. W., 2013. The microbiology of malting and brewing. *Microbiol. Mol. Biol. Rev.,* Volume 77, pp. 157-172.

Bolger, A. M., Lohse, M. & Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* Volume 30, pp. 2114-2120.

Borneman, A. R. et al., 2012. The genome sequence of the wine yeast VIN7 reveals an allotriploid hybrid genome with Saccharomyces cerevisiae and Saccharomyces kudriavzevii origins. *FEMS yeast research,* Volume 12, pp. 88-96.

Borneman, A. R. & Pretorius, I. S., 2015. Genomic insights into the Saccharomyces sensu stricto complex. *Genetics,* Volume 199, pp. 281-291.

Brem, R. B. & Kruglyak, L., 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences,* 102(5), pp. 1572-1577.

Brion, C. et al., 2013. Differential adaptation to multi-stressed conditions of wine fermentation revealed by variations in yeast regulatory networks. *BMC genomics,* Volume 14, p. 681.

Broman, K. W. & Sen, S., 2009. *A Guide to QTL Mapping with R/qtl.* London: Springer.

Broman, K. W., Wu, H., Sen, Ś. & Churchill, G. A., 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics,* Volume 19, pp. 889-890.

Buijs, N. A., Siewers, V. & Nielsen, J., 2013. Advanced biofuel production by the yeast Saccharomyces cerevisiae. *Current opinion in chemical biology,* Volume 17, pp. 480-488.

Burkhardt, R. W., 2013. Lamarck, evolution, and the inheritance of acquired characters. *Genetics,* 194(4), pp. 793-805.

Buschini, A., Poli, P. & Rossi, C., 2003. Saccharomyces cerevisiae as an eukaryotic cell model to assess cytotoxicity and genotoxicity of three anticancer anthraquinones. *Mutagenesis,* 18(1), pp. 25-36.

Cariaso, M. & Lennon, G., 2012. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic acids research,* Volume 40, pp. D1308--D1312.

Chan, K. et al., 2012. Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. *PLoS genetics,* Volume 8, p. e1003149.

Chatterjee, K., Zhang, J., Honbo, N. & Karliner, J. S., 2010. Doxorubicin cardiomyopathy. *Cardiology,* Volume 115, pp. 155-162.

Chong, J. X. et al., 2015. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *The American Journal of Human Genetics,* 97(2), pp. 199-215.

Cortés-Ledesma, F. & Aguilera, A., 2006. Double-strand breaks arising by replication through a nick are repaired by cohesin-dependent sister-chromatid exchange. *EMBO reports,* Volume 7, pp. 919-926.

Cubillos, F. A. et al., 2011. Assessing the complex architecture of polygenic traits in diverged yeast populations. *Molecular ecology,* Volume 20, pp. 1401-1413.

Cubillos, F. A. et al., 2017. Identification of nitrogen consumption genetic variants in yeast through QTL mapping and Bulk segregant RNA-seq analyses. *G3: Genes, Genomes, Genetics,* Volume 7, pp. 1693-1705.

Cubillos, F. A., Louis, E. J. & Liti, G., 2009. Generation of a large set of genetically tractable haploid and diploid Saccharomyces strains. *FEMS yeast research,* Volume 9, pp. 1217-1225.

Cubillos, F. A. et al., 2013. High-resolution mapping of complex traits with a four-parent advanced intercross yeast population. *Genetics,* Volume 195, pp. 1141-1155.

Darwin, C., 1859. *On the origin of species by means of natural selection.* London: John Murray.

Deed, R. C., Fedrizzi, B. & Gardner, R. C., 2017. Saccharomyces cerevisiae FLO1 gene demonstrates genetic linkage to increased fermentation rate at low temperatures. *G3: Genes, Genomes, Genetics,* Volume 7, pp. 1039-1048.

Demae, M. et al., 2007. Overexpression of two transcriptional factors, Kin28 and Pog1, suppresses the stress sensitivity caused by the rsp5 mutation in Saccharomyces cerevisiae. *FEMS microbiology letters,* Volume 277, pp. 70-78.

Dennis, G. et al., 2003. DAVID: database for annotation, visualization, and integrated discovery. *Genome biology,* Volume 4, p. R60.

Dotiwala, F. et al., 2013. DNA damage checkpoint triggers autophagy to regulate the initiation of anaphase. *Proceedings of the National Academy of Sciences,* Volume 110, pp. E41--E49.

Duina, A. A., Miller, M. E. & Keeney, J. B., 2014. Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. *Genetics,* 197(1), pp. 33-48.

Dujon, B., 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *TRENDS in Genetics,* Volume 22, pp. 375-387.

Dujon, B. A. & Louis, E. J., 2017. Genome diversity and evolution in the budding yeasts (Saccharomycotina). *Genetics,* Volume 206, pp. 717-750.

Dunn, C. D., Lee, M. S., Spencer, F. A. & Jensen, R. E., 2006. A genomewide screen for petite-negative yeast strains yields a new subunit of the i-AAA protease complex. *Molecular biology of the cell,* Volume 17, pp. 213-226.

Edwards, M. D. & Gifford, D. K., 2012. High-resolution genetic mapping with pooled sequencing. *BMC Bioinformatics,* Volume 13, p. S8.

Ehrenreich, I. M., Gerke, J. P. & Kruglyak, L., 2009. Genetic dissection of complex traits in yeast: insights from studies of gene expression and other phenotypes in the BY× RM cross. *Cold Spring Harbor Symposia on Quantitative Biology,* Volume 74, pp. 145-153.

Ehrenreich, I. M. et al., 2010. Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature,* Volume 464, p. 1039.

Eriksson, P. R., Ganguli, D., Nagarajavel, V. & Clark, D. J., 2012. Regulation of histone gene expression in budding yeast. *Genetics,* Volume 191, pp. 7-20.

Evans, L. M. et al., 2018. Narrow-sense heritability estimation of complex traits using identity-by-descent information. *Heredity,* Volume 121, pp. 616-630.

Fahrenkrog, B., 2016. Histone modifications as regulators of life and death in Saccharomyces cerevisiae. *Microbial Cell,* Volume 3, p. 1.

Falconer, D. S., 1996. *Introduction to Quantitative Genetics.* 4th ed. London: Longman.

Fedorov, D. V. et al., 2013. HSM6 gene is identical to PSY4 gene in Saccharomyces cerevisiae yeasts. *Russian journal of genetics,* Volume 49, pp. 286-293.

Finsterbusch, T. & Mankertz, A., 2009. Porcine circoviruses - small but powerful. *Virus research,* 143(2), pp. 177-183.

Fisher, R. A., 1930. The evolution of dominance in certain polymorphic species. *The American Naturalist,* 64(694), pp. 385-406.

Frezza, C., 2018. *Histidine metabolism boosts cancer therapy.* s.l.:Nature Publishing Group.

Garg, M. B. et al., 2012. Predicting 5-fluorouracil toxicity in colorectal cancer patients from peripheral blood cell telomere length: a multivariate analysis. *British journal of cancer,* Volume 107, p. 1525.

Garrison, E. & Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907.*

Geissler, S., Siegers, K. & Schiebel, E., 1998. A novel protein complex promoting formation of functional α-and γ-tubulin. *The EMBO Journal,* 17(4), pp. 952-966.

Georghiou, D., 2017. *Perosonal Communication.* Leicester: University of Leicester.

Gibson, B. & Liti, G., 2015. Saccharomyces pastorianus: genomic insights inspiring innovation for industry. *Yeast,* Volume 32, pp. 17-27.

Glazier, A. M., Nadeau, J. H. & Aitman, T. J., 2002. Finding genes that underlie complex traits. *Science,* Volume 298, pp. 2345-2349.

Goffeau, A. et al., 1996. Life with 6000 genes. *Science,* 274(5287), pp. 546-567.

Gomez, M. et al., 2014. Malfunctioning of the iron-sulfur cluster assembly machinery in Saccharomyces cerevisiae produces oxidative stress via an iron-dependent mechanism, causing dysfunction in respiratory complexes. *PLoS One,* 9(10), p. e111585.

Han, F. F., Guo, C. L. & Liu, L. H., 2013. The effect of CHEK2 variant I157T on cancer susceptibility: evidence from a meta-analysis. *DNA and cell biology,* 32(6), pp. 329-335.

Han, J. et al., 2007. Rtt109 acetylates histone H3 lysine 56 and functions in DNA replication. *Science,* Volume 315, pp. 653-655.

Harvey, L. et al., 2000. *Molecular Cell Biology.* 4th ed. New York: W. H. Freeman.

Hebly, M. et al., 2015. S. cerevisiae× S. eubayanus interspecific hybrid, the best of both worlds and beyond. *FEMS Yeast Research,* Volume 15.

Hinks Roberts, A., 2019. *Saccharomyces Hybrids: Generation And Analysis,* Leicester: Department of Genetics, University of Leicester.

Hittinger, C. T. et al., 2010. Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature,* Volume 464, p. 54.

Horigome, C. et al., 2016. PolySUMOylation by Siz2 and Mms21 triggers relocation of DNA breaks to nuclear pores through the Slx5/Slx8 STUbL. *Genes & development,* Volume 30, pp. 931-945.

J Mayhew, A. & Meyre, D., 2017. Assessing the heritability of complex traits in humans: methodological challenges and opportunities. *Current genomics,* Volume 18, pp. 332-340.

Jordan, P. W., Klein, F. & Leach, D. R. F., 2007. Novel roles for selected genes in meiotic DNA processing. *PLoS genetics,* Volume 3, p. e222.

Kaiser, B. K. et al., 2007. Disulphide-isomerase-enabled shedding of tumour-associated NKG2D ligands. *Nature,* 447(7143), p. 482.

Killick, R. & Eckley, I. A., 2014. changepoint: An R Package for Changepoint Analysis. *Journal of Statistical Software,* Volume 58, pp. 1-19.

Kim, H., Kim, A. & Cunningham, K. W., 2012. Vacuolar H+-ATPase (V-ATPase) promotes vacuolar membrane permeabilization and nonapoptotic death in stressed yeast. *Journal of Biological Chemistry,* Volume 287, pp. 19029-19039.

Klukovich, R. & Courchesne, W. E., 2016. Functions of Saccharomyces cerevisiae Ecm27p, a putative Na+/Ca2+ exchanger, in calcium homeostasis, carbohydrate storage and cell cycle reentry from the quiescent phase. *Microbiological research,* Volume 186, pp. 81-89.

Kosodo, Y. et al., 2001. Multicopy suppressors of the sly1 temperature-sensitive mutation in the ER--Golgi vesicular transport in Saccharomyces cerevisiae. *Yeast,* Volume 18, pp. 1003-1014.

Kristofich, J. et al., 2018. Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme. *PLoS genetics,* 14(8), p. e1007615.

Krogerus, K., Holmström, S. & Gibson, B., 2018. Enhanced wort fermentation with de novo lager hybrids adapted to high-ethanol environments. *Appl. Environ. Microbiol.,* Volume 84, pp. e02302--17.

Kroymann, J. & Mitchell-Olds, T., 2005. Epistasis and balanced polymorphism influencing complex trait variation. *Nature,* 435(7038), p. 95.

Kuilman, T. et al., 2015. Identification of Cdk targets that control cytokinesis. *The EMBO journal,* Volume 34, pp. 81-96.

Laco, J. et al., 2010. Adenine nucleotide transport via Sal1 carrier compensates for the essential function of the mitochondrial ADP/ATP carrier. *FEMS yeast research,* Volume 10, pp. 290-296.

Larochelle, M., Drouin, S., Robert, F. & Turcotte, B., 2006. Oxidative stress-activated zinc cluster protein Stb5 has dual activator/repressor functions required for pentose phosphate pathway regulation and NADPH production. *Molecular and cellular biology,* Volume 26, pp. 6690-6701.

Lassmann, T. & Sonnhammer, E. L., 2005. Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics,* 6(1), p. 298.

Lee, D.-H.et al., 2010. A PP4 phosphatase complex dephosphorylates RPA2 to facilitate DNA repair via homologous recombination. *Nature structural & molecular biology,* Volume 17, p. 365.

Lee, J. T., Taylor, M. B., Shen, A. & Ehrenreich, I. M., 2016. Multi-locus genotypes underlying temperature sensitivity in a mutationally induced trait. *PLoS genetics,* Volume 12, p. e1005929.

Leenders, E. K. S. M. et al., 2018. Cancer prevention by aspirin in children with Constitutional Mismatch Repair Deficiency (CMMRD). *European Journal of Human Genetics,* Volume 26, p. 1417.

Lempiäinen, H. & Shore, D., 2009. Growth control and ribosome biogenesis. *Current opinion in cell biology,* Volume 21, pp. 855-863.

Lengronne, A. et al., 2006. Establishment of sister chromatid cohesion at the S. cerevisiae replication fork. *Molecular cell,* Volume 23, pp. 787-799.

Libkind, D. et al., 2011. Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proceedings of the National Academy of Sciences,* Volume 108, pp. 14539-14544.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows--Wheeler transform. *bioinformatics,* Volume 25, pp. 1754-1760.

Li, H. et al., 2009. The sequence alignment/map format and SAMtools. *Bioinformatics,* Volume 25, pp. 2078-2079.

Liti, G. et al., 2009. Population genomics of domestic and wild yeasts. *Nature,* Volume 458, p. 337.

Liti, G. & Louis, E. J., 2012. Advances in quantitative trait analysis in yeast. *PLoS genetics,* 8(8), p. e1002912.

Lodish, H. F., 2016. *Molecular cell biology.* 8th ed. New York: W.H. Freeman/Macmillan Learning.

Louis, E. J., Becker, M. M. & Marion, M., 2014. *Subtelomeres.* Berlin: Springer.

Lynch, M. & Walsh, B., 1998. *Genetics and analysis of quantitative traits.* Sunderland, MA: Sinauer.

Lynch, M. & Walsh, B., 1998. Genetics and analysis of quantitative traits. *Sunderland,* Volume 1, pp. 535-557.

Massey, T. H. & Jones, L., 2018. The central role of DNA damage and repair in CAG repeat diseases. *Disease models & mechanisms,* Volume 11, p. dmm031930.

Matsufuji, Y. et al., 2010. Transcription factor Stb5p is essential for acetaldehyde tolerance in Saccharomyces cerevisiae. *Journal of basic microbiology,* Volume 50, pp. 494-498.

Maulina, T. et al., 2019. The Usage of Curcumin as Chemopreventive Agent for Oral Squamous Cell Carcinoma: An Experimental Study on Sprague-Dawley Rat. *Integrative cancer therapies,* Volume 18, p. 1534735418822094.

Mayhew, A. J. & Meyre, D., 2017. Assessing the heritability of complex traits in humans: methodological challenges and opportunities. *Current genomics,* 18(4), pp. 332-340.

Mendel, G., 1865. Versuche uber pflanzen-hybriden. *Vorgelegt in den Sitzungen.*

Metzker, M. L., 2010. Sequencing technologies - the next generation. *Nature reviews genetics,* 11(1), p. 31.

Miles, C. & Wayne, M., 2008. Quantitative trait locus (QTL) analysis. *Nature Education,* 1(1), p. 208.

Mirkes, E. M., Walsh, T., Louis, E. J. & Gorban, A. N., 2015. Long and short range multi-locus QTL interactions in a complex trait of yeast. *arXiv preprint arXiv:1503.05869.*

Morgan, T. H., 1910. Sex limited inheritance in Drosophila. *Science,* 32(812), pp. 120-122.

Nadeem, M. A. et al., 2018. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnology & Biotechnological Equipment,* Volume 32, pp. 261-285.

Naseeb, S. et al., 2018. Whole genome sequencing, de novo assembly and phenotypic profiling for the new budding yeast species Saccharomyces jurei. *G3: Genes, Genomes, Genetics,* Volume 8, pp. 2967-2977.

Naumov, G. I. et al., 2000. Three new species in the Saccharomyces sensu stricto complex: Saccharomyces cariocanus, Saccharomyces kudriavzevii and Saccharomyces mikatae.. *International journal of systematic and evolutionary microbiology,* Volume 50, pp. 1931-1942.

Navarro-Tapia, E., Querol, A. & Pérez-Torrado, R., 2018. Membrane fluidification by ethanol stress activates unfolded protein response in yeasts. *Microbial biotechnology,* Volume 11, pp. 465-475.

Nguyen, H.-V., Legras, J.-L., Neuvéglise, C. & Gaillardin, C., 2011. Deciphering the hybridisation history leading to the lager lineage based on the mosaic genomes of Saccharomyces bayanus strains NBRC1948 and CBS380T. *PLoS One,* Volume 6, p. e25821.

Nguyen, T. T. T. et al., 2015. Fitness profiling links topoisomerase II regulation of centromeric integrity to doxorubicin resistance in fission yeast. *Scientific reports,* Volume 5, p. 8400.

Ojuederie, O. & Babalola, O., 2017. Microbial and plant-assisted bioremediation of heavy metal polluted environments: a review. *International journal of environmental research and public health,* Volume 14, p. 1504.

Pellicer, J., Fay, M. F. & Leitch, I. J., 2010. The largest eukaryotic genome of them all?. *Botanical Journal of the Linnean Society,* 164(1), pp. 10-15.

Penzo, M., Montanaro, L., Treré, D. & Derenzini, M., 2019. The Ribosome Biogenesis - Cancer Connection. *Cells,* 8(1), p. 55.

Peris, D. et al., 2014. Population structure and reticulate evolution of S accharomyces eubayanus and its lager‐brewing hybrids. *Molecular Ecology,* Volume 23, pp. 2031-2045.

Peter, J. et al., 2018. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature,* 556(7701), p. 339.

Pilzecker, B., Buoninfante, O. A. & Jacobs, H., 2019. DNA damage tolerance in stem cells, ageing, mutagenesis, disease and cancer therapy. *Nucleic acids research,* Volume 47, pp. 7163-7181.

Radovic, S., Rapisarda, V. A., Tosato, V. & Bruschi, C. V., 2007. Functional and comparative characterization of Saccharomyces cerevisiae RVB1 and RVB2 genes with bacterial Ruv homologues. *FEMS yeast research,* Volume 7, pp. 527-539.

Rainey, M. D., Charlton, M. E., Stanton, R. V. & Kastan, M. B., 2008. Transient inhibition of ATM kinase is sufficient to enhance cellular sensitivity to ionizing radiation. *Cancer research,* Volume 68, pp. 7466-7474.

Romero, P. A. et al., 1999. Mnt2p and Mnt3p of Saccharomyces cerevisiae are members of the Mnn1p family of α-1, 3-mannosyltransferases responsible for adding the terminal mannose residues of O-linked oligosaccharides. *Glycobiology,* Volume 9, pp. 1045-1051.

Salinas, F. et al., 2012. The genetic basis of natural variation in oenological traits in Saccharomyces cerevisiae. *PLoS One,* Volume 7, p. e49640.

Scannell, D. R. et al., 2011. The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the Saccharomyces sensu stricto genus. *G3: Genes, Genomes, Genetics,* Volume 1, pp. 11-25.

Schaschl, H., Aitman, T. J. & Vyse, T. J., 2009. Copy number variation in the human genome and its implication in autoimmunity. *Clinical & Experimental Immunology,* 156(1), pp. 12-16.

Shalem, O. et al., 2011. Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA Pol II. *PLoS genetics,* Volume 7, p. e1002273.

Shen, H. et al., 2013. Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. *PloS one,* Volume 8.

Simpson-Lavy, K., Xu, T., Johnston, M. & Kupiec, M., 2017. The Std1 activator of the Snf1/AMPK kinase controls glucose response in yeast by a regulated protein aggregation. *Molecular cell,* Volume 68, pp. 1120-1133.

Sirugo, G., Williams, S. M. & Tishkoff, S. A., 2019. The missing diversity in human genetic studies. *Cell,* Volume 177, pp. 26-31.

Smith, S., 1960. Lamarck and modern genetics. *The Eugenics review,* 52(1), p. 47.

Snoek, T. et al., 2015. Large-scale robot-assisted genome shuffling yields industrial Saccharomyces cerevisiae yeasts with increased ethanol tolerance. *Biotechnology for biofuels,* Volume 8, p. 32.

Speicher, M., Antonarakis, S. E. & Motulsky, A. G., 2010. *Vogel and Motulsky's Human Genetics: Problems and Approaches.* 4th ed. Berlin: Springer.

Steyer, D. et al., 2012. QTL mapping of the production of wine aroma compounds by yeast. *BMC genomics,* Volume 13, p. 573.

Strober, B. J. et al., 2019. Dynamic genetic regulation of gene expression during cellular differentiation. *Science,* Volume 364, pp. 1287-1290.

Suh, Y. & Vijg, J., 2005. SNP discovery in associating genetic variation with human disease phenotypes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* 573(1-2), pp. 41-53.

Taylor, M. B. et al., 2016. Diverse genetic architectures lead to the same cryptic phenotype in a yeast cross. *Nature communications,* Volume 7, p. 11669.

Thorn, C. F. et al., 2011. Doxorubicin pathways: pharmacodynamics and adverse effects. *Pharmacogenetics and genomics,* Volume 21, p. 440.

Tkach, J. M. et al., 2012. Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nature cell biology,* Volume 14, p. 966.

Tora, L., 2002. A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription. *Genes & development,* Volume 16, pp. 673-675.

Tsabar, M. et al., 2016. Asf1 facilitates dephosphorylation of Rad53 after DNA double-strand break repair. *Genes & development,* Volume 30, pp. 1211-1224.

Van den Broek, M. et al., 2015. Chromosomal copy number variation in Saccharomyces pastorianus is evidence for extensive genome dynamics in industrial lager brewing strains. *Applied and Environmental Microbiology,* 81(18), pp. 6253-6267.

Wang, Q.-M.et al., 2012. Surprisingly diverged populations of S accharomyces cerevisiae in natural environments remote from human activity. *Molecular ecology,* Volume 21, pp. 5404-5417.

Wang, X. & Kruglyak, L., 2014. Genetic basis of haloperidol resistance in Saccharomyces cerevisiae is complex and dose dependent. *PLoS genetics,* 10(12), p. e1004894.

Warde-Farley, D. et al., 2010. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research,* Volume 38, pp. W214--W220.

Ware, J. S., Roberts, A. M. & Cook, S. A., 2012. Next generation sequencing for clinical diagnostics and personalised medicine: implications for the next generation cardiologist. *Heart ,* 98(4), pp. 276-281.

Watson, J. D. & Crick, F. H. C., 1953. Molecular structure of nucleic acids. *Nature,* 171(4356), pp. 737-738.

Weiss, C. V. et al., 2018. Genetic dissection of interspecific differences in yeast thermotolerance. *Nature genetics,* Volume 50, p. 1501.

Wei, W.-H., Hemani, G. & Haley, C. S., 2014. Detecting epistasis in human complex traits. *Nature Reviews. Genetics,* Volume 15, p. 722.

Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.

Wilkening, S. et al., 2014. An evaluation of high-throughput approaches to QTL mapping in Saccharomyces cerevisiae. *Genetics,* Volume 196, pp. 853-865.

Xu, P. et al., 2013. AtMMS21, an SMC5/6 complex subunit, is involved in stem cell niche maintenance and DNA damage responses in Arabidopsis roots. *Plant physiology,* Volume 161, pp. 1755-1768.

Yue, J.-X.et al., 2017. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature genetics,* Volume 49, p. 913.

Zhang, P. et al., 2017. Aspirin suppresses TNF-α-induced MMP-9 expression via NF-κB and MAPK signaling pathways in RAW264. 7 cells. *Experimental and therapeutic medicine,* Volume 14, pp. 5597-5604.

Zhang, W. et al., 2016. Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nature communications,* Volume 7, p. 12619.

Zhang, Y. et al., 2013. The SWI/SNF chromatin remodeling complex influences transcription by RNA polymerase I in Saccharomyces cerevisiae. *PloS one,* Volume 8, p. e56793.

Zhang, Y. et al., 2013. Protein chemical synthesis by serine and threonine ligation. *Proceedings of the National Academy of Sciences,* Volume 110, pp. 6657-6662.

Zuo, X. et al., 2015. Whole-exome SNP array identifies 15 new susceptibility loci for psoriasis. *Nature communications,* Volume 6, p. 6793.

Appendix A. Scripts and Pipelines

A.1 QTL Analysis in F1 and F12

```
#################################################################
#PBS -N QTL_PREP
#PBS -l walltime=02:00:00
#PBS -l vmem=10gb
#PBS -l nodes=1:ppn=1
#PBS -t 1-3
#################################################################
cd "${PBS_O_WORKDIR}"
#################################################################
# Load dependencies.
module load java/1.8
module load R/3.5.1
# Get name of cross input file.
CROSSFILE=$(awk "NR == ${PBS_ARRAYID}" crossfiles.txt)
# Prep cross file for QTL analysis.
Rscript -e 'library(shmootl)' -e 'run()' prep   --datafile
"${CROSSFILE}"
--normseq
# Push genetic map into cross file.
Rscript -e 'library(shmootl)' -e 'run()' pushmap \
    --mapfile AESW12fc.gmap.csv  --datafile "${CROSSFILE}"
#################################################################
```

```
#############################################################
#PBS -N QTL_SCAN
#PBS -l walltime=2:00:00
#PBS -l vmem=8gb
#PBS -l nodes=1:ppn=8
#PBS -t 1-3
#############################################################
cd "${PBS_O_WORKDIR}"
#############################################################
# Load dependencies.
module load java/1.8
module load R/3.5.1
# Get name of cross input file and set prefix.
CROSSFILE=$(awk "NR == ${PBS_ARRAYID}" crossfiles.txt)
if [[ "${CROSSFILE}" =~ (^|\/)(.+)[.]csv$ ]]
then PREFIX="${BASH_REMATCH[2]}"
else die "unknown cross filename pattern '${CROSSFILE}'"
fi
# Run scanone pipeline
Rscript -e 'library(shmootl)' -e 'run()' scanone \
  --n.perm 1000 --alpha 0.05 --n.cluster 8 --method mr \
  --infile "${CROSSFILE}" --h5file "${PREFIX}.hdf5"
# Create scanone report
Rscript -e 'library(shmootl)' -e 'run()' report \
  --h5file "${PREFIX}.hdf5" --report "${PREFIX}.pdf"
# Create scanone report
Rscript -e 'library(shmootl)' -e 'run()' report \
  --h5file "${PREFIX}.hdf5" --report "${PREFIX}.xlsx"
#############################################################
```

```
#################################################################
#PBS -N QTL_ANNO
#PBS -l walltime=02:00:00
#PBS -l vmem=10gb
#PBS -l nodes=1:ppn=1
#PBS -t 1-3
#################################################################
cd "${PBS_O_WORKDIR}"
#################################################################
# Load dependencies.
module load java/1.8
module load R/3.5.1
# Get name of scanfiles and get annofile.
SCANFILE=$(awk "NR == ${PBS_ARRAYID}" scanfiles.txt)
ANNOFILE="/scratch/gact/shared/data/genomes/SGD_S288C_R64-2-
1/SGD_S288C_R64-2-1.gff"
# Push physical map into scan file.
Rscript -e 'library(shmootl)' -e 'run()' pushmap --mapfile
AESW12fc.pmap.csv --datafile "${SCANFILE}"
# Run annotation pipeline
Rscript -e 'library(shmootl)' -e 'run()' annoqtl --infile
"${SCANFILE}" \
  --annofile "${ANNOFILE}" --outfile "${SCANFILE}"
#################################################################
```

A.2 Overlap analysis for F12 QTL analysis under multiple agents

```
#################################################################
overlap.r
#################################################################
overlap<- function(Features,filenumber){
    gene <- Reduce(union, Features)
    test.list <- Features
    gene_count <- matrix(0, nrow = length(gene), ncol = 1)
    for (i in 1:length(gene)){
        for(j in 1:length(test.list)){
            if (gene[i] %in% test.list[[j]]){
                gene_count[i] <- gene_count[i] +1
            } }}
    result <- gene[gene_count >= filenumber]
    result <- result[which(result != "NA")]
    return(result)}
#################################################################
```

A.3 QTL Analysis app on shiny

```
#
# This is the server logic of a Shiny web application. You
can run the
# application by clicking 'Run App' above.
#
# Find out more about building applications with Shiny here:
#
#    http://shiny.rstudio.com/

library(shiny)
```

```
library(shmootl)
options(shiny.maxRequestSize=30*1024^2)
# Define server logic required to draw a histogram
shinyServer(function(input, output, session) {
  annoAnalysis <- reactive({
    #run pushmap
    withProgress(message = "Scanone in process", detail =
"This might take a while", {
      pmap = input$pmapfile
      anno = input$annofile
      scanfile <- session$userData$scanonefilepath
      print(scanfile)
      setProgress(value = 0.1, detail = "Running pushmap...")
      run_pushmap(mapfile = pmap$datapath, datafile =
scanfile)
      setProgress(value = 0.3, detail = "Running QTL
annotation...")
      run_annoqtl(infile = scanfile, annofile =
anno$datapath, outfile = scanfile)

      output$annoresult <- renderText("QTL Annotation
finished.")
      })
  })

  scanAnalysis <- reactive({
    withProgress(message = "Scanone in process", detail =
"This might take a while", {
      crossFile <- input$crossfile
      prefix <- as.character(strsplit(crossFile$name,
'.csv'))
```

```
      #TODO: remove hard coding
      session$userData$fileprefix <- strsplit(prefix, '
')[[1]][2]
      scanfile <- paste0(session$userData$fileprefix,
'.hdf5')
      session$userData$scanonefilepath <- scanfile
      incProgress(0.1, detail = "Scanone running...")
      #set threshold and significant level

      if(input$usethreshold){
        run_scanone(threshold = input$thresvalue, alpha =
0.05, method = 'mr', n.cluster = 4,
                    infile = crossFile$datapath, h5file =
scanfile)
      }

      else{
        run_scanone(n.perm = input$permutation, alpha =
input$significance, method = 'mr', n.cluster = 4,
                    infile = crossFile$datapath, h5file =
scanfile)
      }

      setProgress(1)
      output$scanoneanalysis <- renderText("Single QTL scan
finished.")
    })
  })

  prepareData <- reactive({
    withProgress(message = "Data preparation in process", {
```

```
      crossinFile <- input$crossfile
      gmapinFile <- input$gmapfile
      print(gmapinFile$datapath)
      incProgress(0.1, detail = "Files loaded")
      run_prep(crossinFile$datapath, normseq = TRUE)
      setProgress(0.5, detail = "Preprocessing...")
      run_pushmap(mapfile = gmapinFile$datapath, datafile =
crossinFile$datapath)
      output$dataprep     <-renderText("Data     preparation
finished.")
      setProgress(1)
    })
  })

  output$crossfileres <- renderText({
    inFile <- input$crossfile
    if (is.null(inFile)) {
      print("Please upload cross file.")
    } else {
      print(paste0("Cross    file    '",    inFile$name    ,"'
uploaded!"))
      #print(paste0("Path: ", inFile$datapath))
    }
  })

  output$gmapfileres <- renderText({
    inFile <- input$gmapfile
    if (is.null(inFile)) {
      print("Please upload gmap file.")
    } else {
```

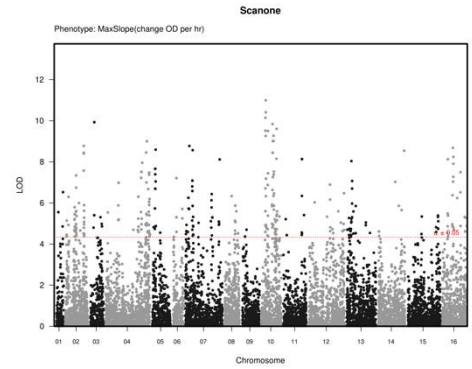
```
      print(paste0("GMAP    file    '",    inFile$name    ,"'
uploaded!"))
      #print(paste0("Path: ", inFile$datapath))
    }
  })

  output$pmapfileres <- renderText({
    inFile <- input$pmapfile
    if (is.null(inFile)) {
      print("Please upload pmap file.")
    } else {
      print(paste0("PMAP    file    '",    inFile$name    ,"'
uploaded!"))
      #print(paste0("Path: ", inFile$datapath))
    }
  })

  output$covfileres <- renderText({
    inFile <- input$covfile
    if (is.null(inFile)) {
      print("Please upload covariate file.")
    } else {
      print(paste0("Covariate   file   '",   inFile$name   ,"'
uploaded!"))
      #print(paste0("Path: ", inFile$datapath))
    }
  })

  output$annofileres <- renderText({
    inFile <- input$annofile
    if (is.null(inFile)) {
```

```
    print("Please upload .gff file.")
  } else {
    print(paste0("GFF file '", inFile$name ,"' uploaded!"))
    #print(paste0("Path: ", inFile$datapath))
  }
})

observeEvent(input$btnprep, {
  isolate(prepareData())
})

observeEvent(input$btnscan, {
  isolate(scanAnalysis())
})

observeEvent(input$btnanno, {
  isolate(annoAnalysis())
})

output$downloadPdf <- downloadHandler(
  filename = function() {
    paste0(session$userData$fileprefix, ".pdf")
  },
  content = function(file) {
    run_report(h5file = session$userData$scanonefilepath,
               report = file)
  }
)

output$downloadXlx <- downloadHandler(
  filename = function() {
```

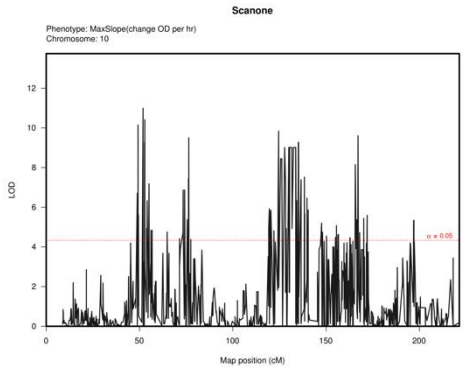
```
      paste0(session$userData$fileprefix, ".xlsx")
    },
    content = function(file) {
      run_report(h5file = session$userData$scanonefilepath,
                 report = file)
    }
  )
})
```

Appendix B. Additional Figures

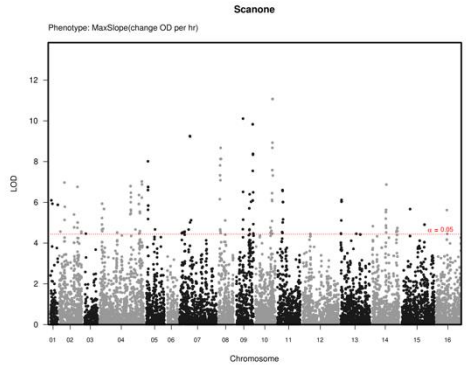
B.1 QTL analysis output for Chapter 3



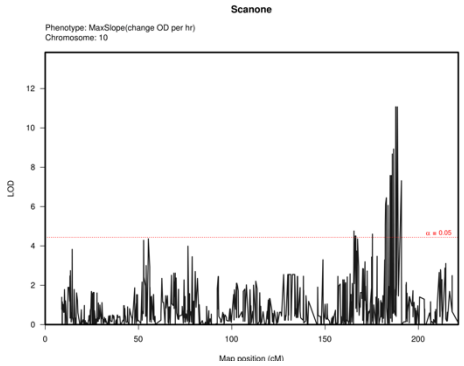
FU 2.5mM



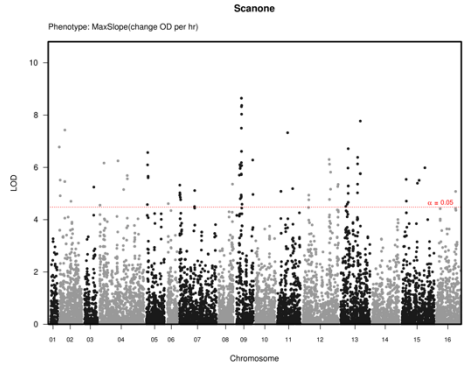
FU 2.5mM peak marker



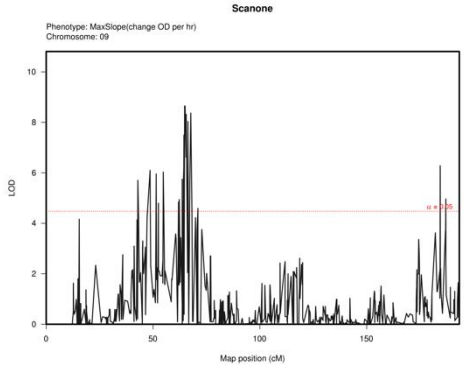
FU 5mM



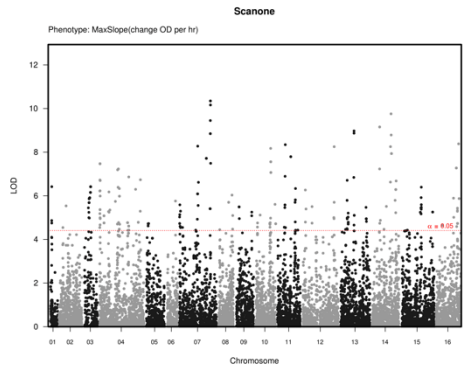
FU 5mM peak marker



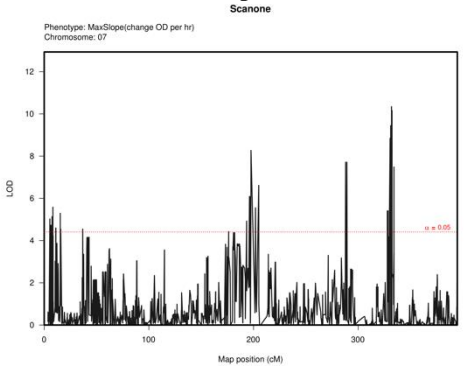
ASP 10mM



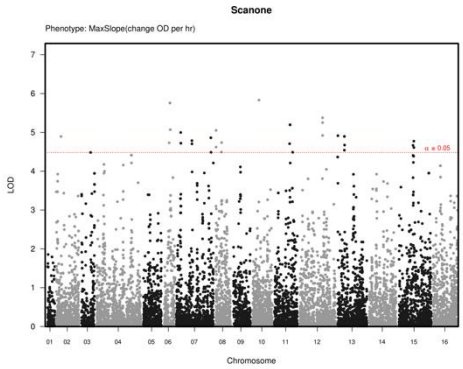
ASP 10mM peak marker



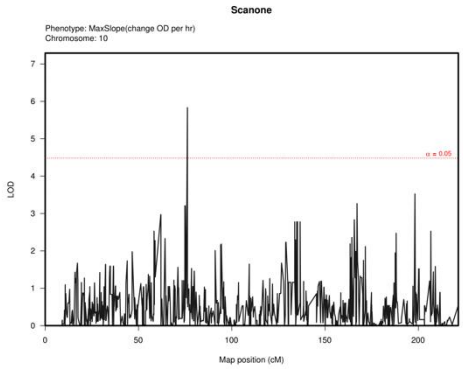
ASP 35mM



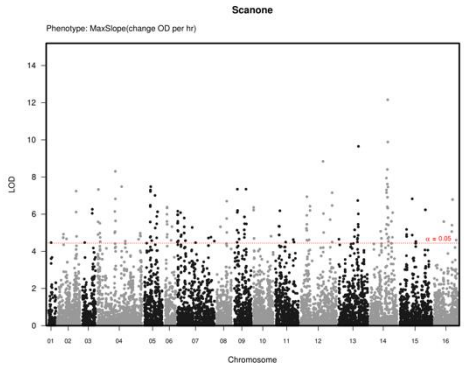
ASP 35mM peak marker



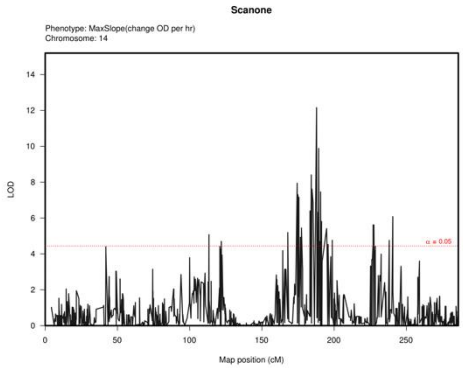
CIS 1mM



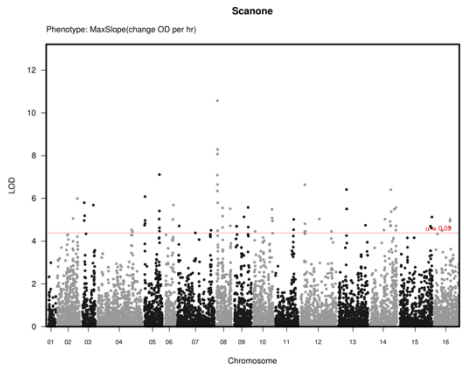
CIS 1mM peak marker



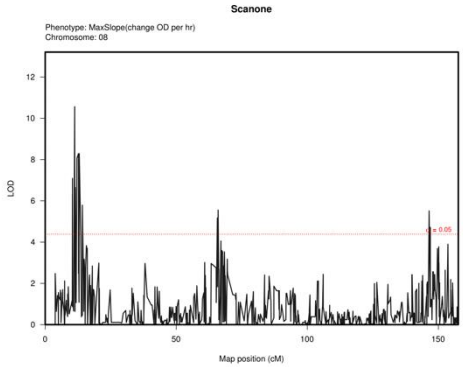
CCM 200uM



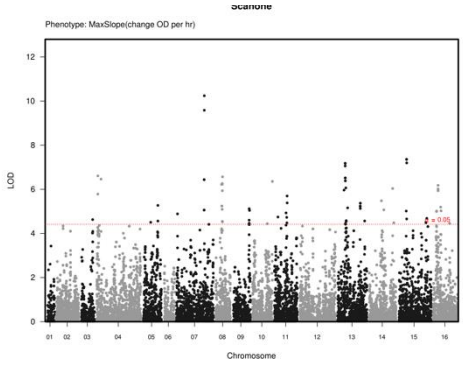
CCM 200uM peak marker



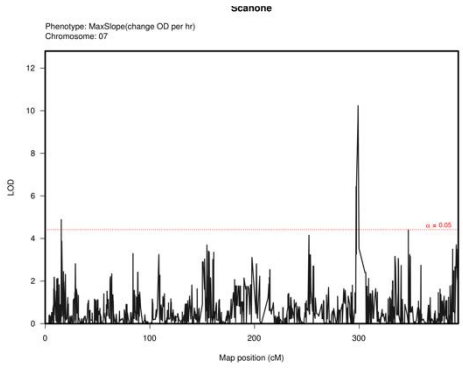
CCM 500uM



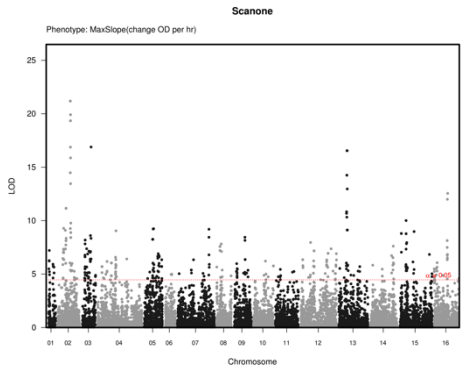
CMM 500uM peak marker



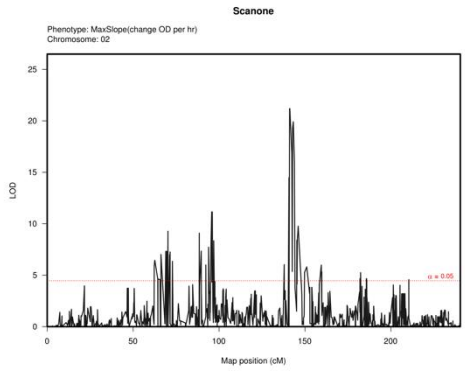
DSF 70uM



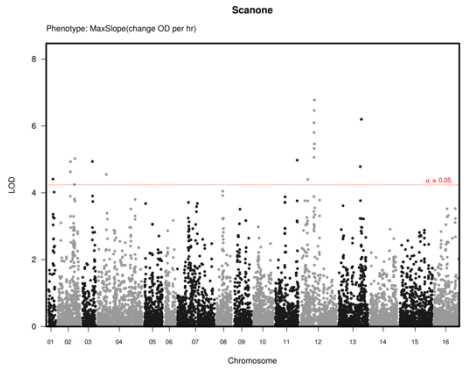
DSF 70uM peak marker



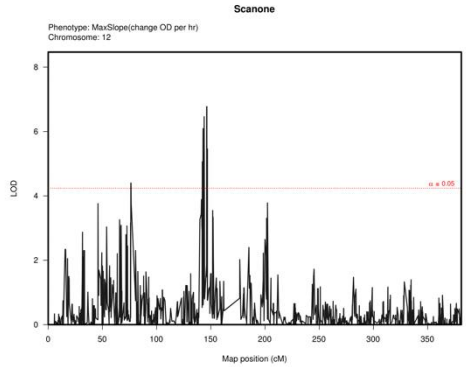
DOX



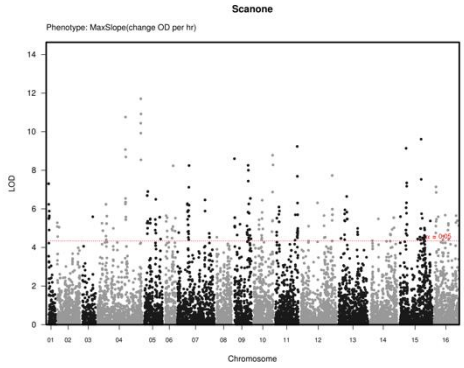
DOX peak marker



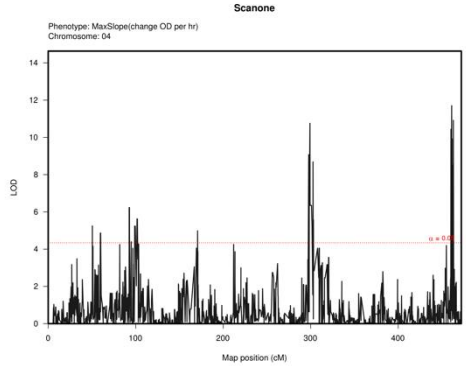
EPA



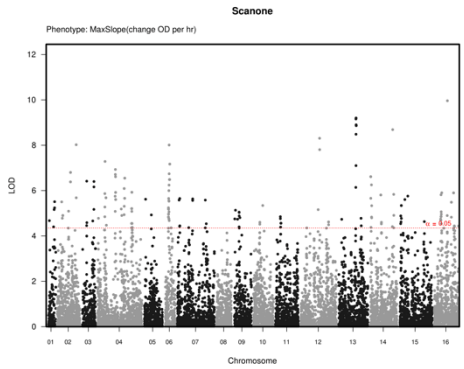
EPA peak marker



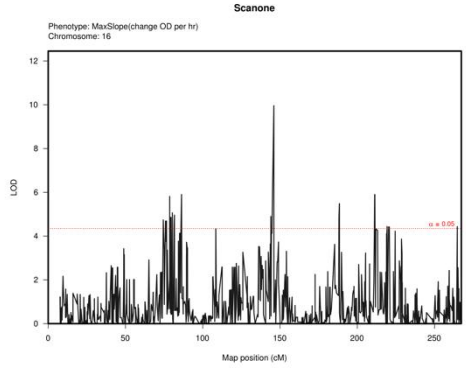
HU



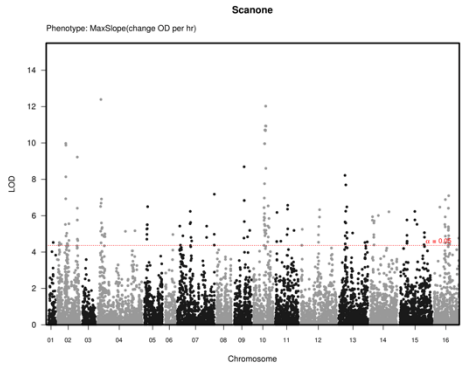
HU peak marker



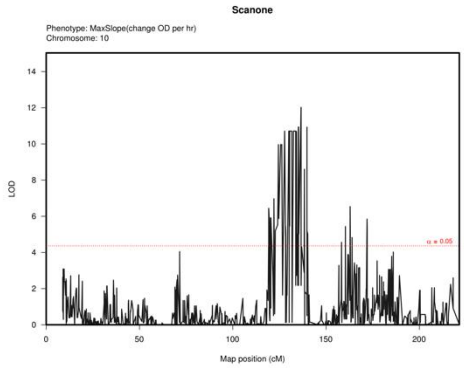
MT



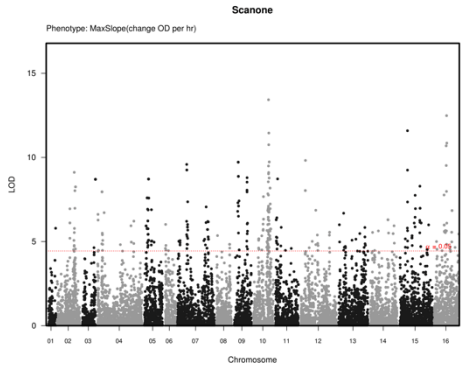
MT peak marker



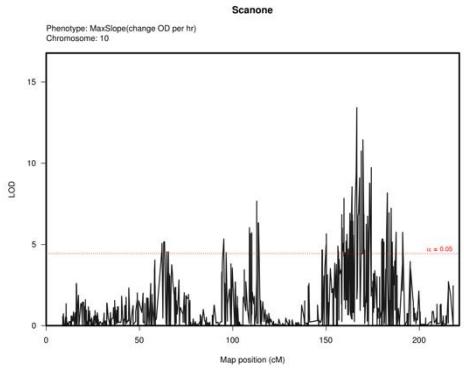
MT + DOX



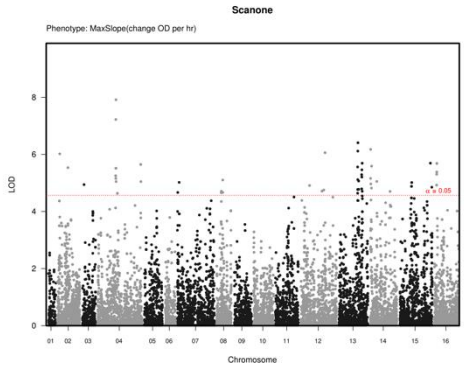
MT + DOX peak marker



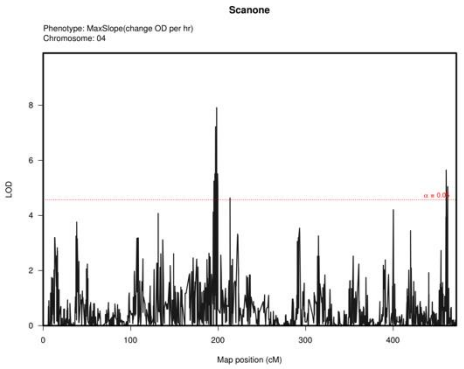
MET



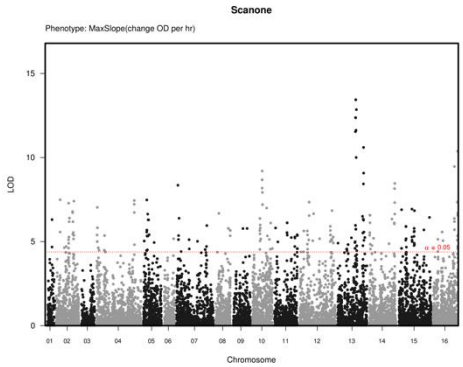
MET peak marker



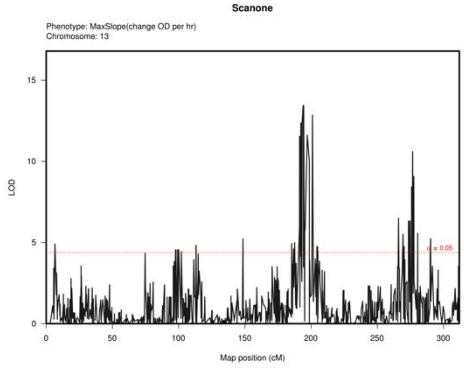
MMS



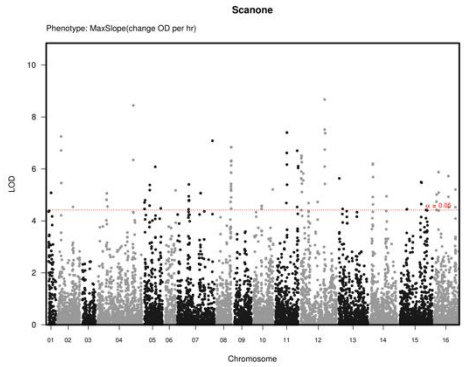
MMS peak marker



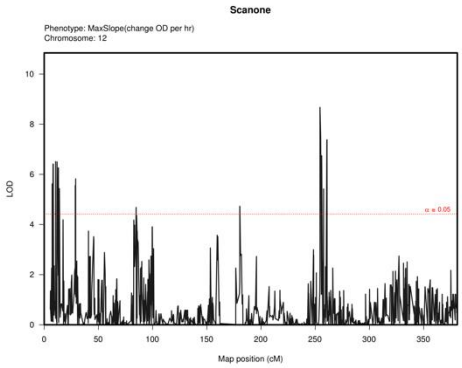
PQ



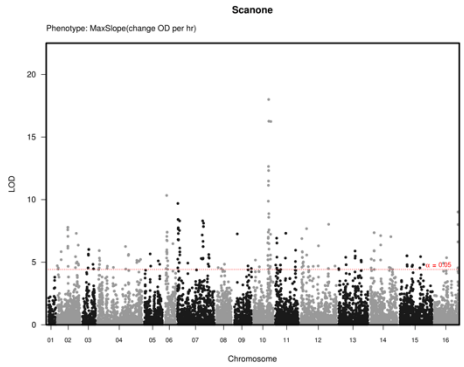
PQ peak marker



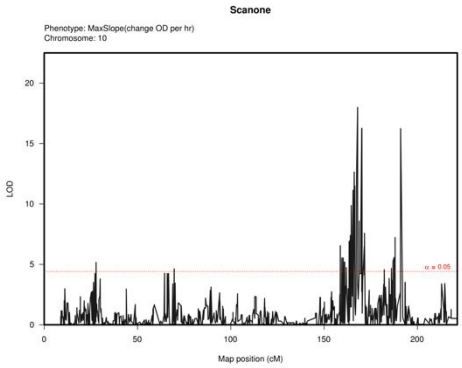
Phleo



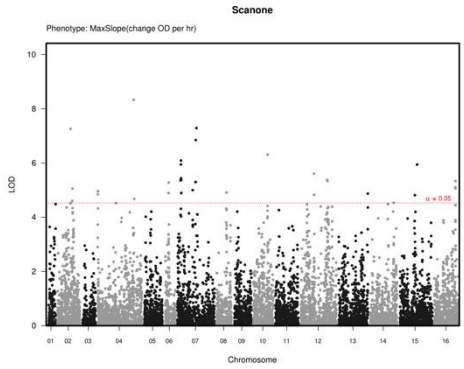
Phleo peak marker



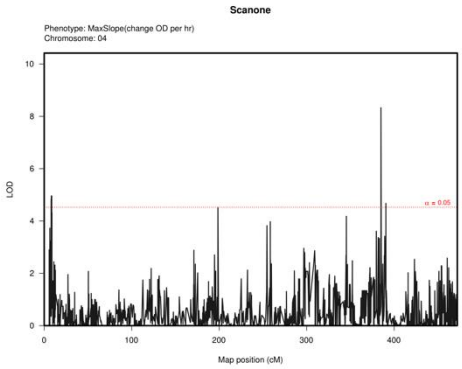
RAP



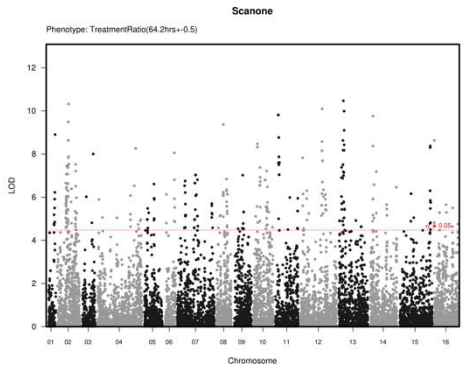
RAP peak marker



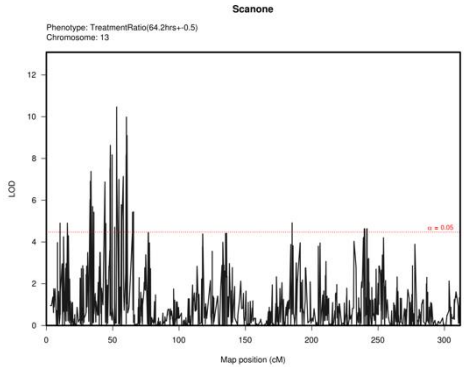
SLA



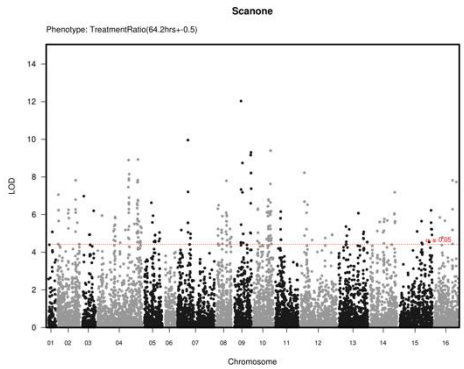
SLA peak marker



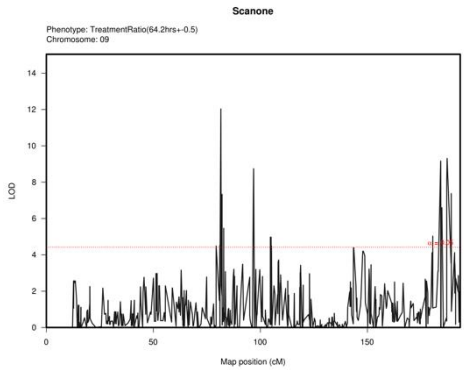
FU 2.5mM (TR)



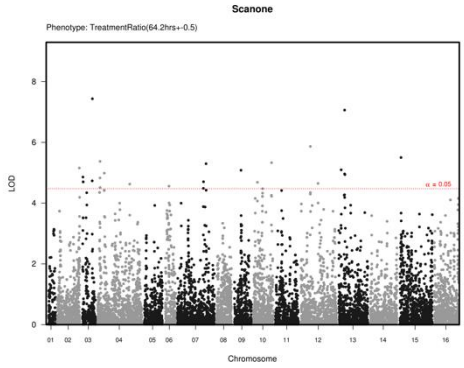
FU 2.5mM peak marker (TR)



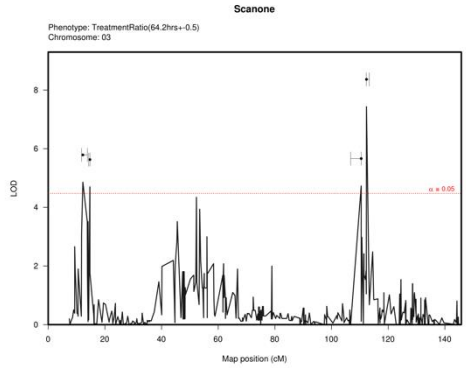
FU 5mM (TR)



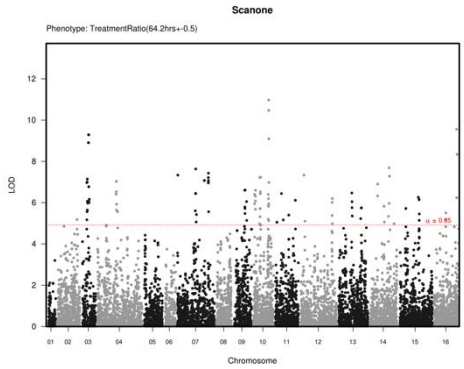
FU 5mM peak marker (TR)



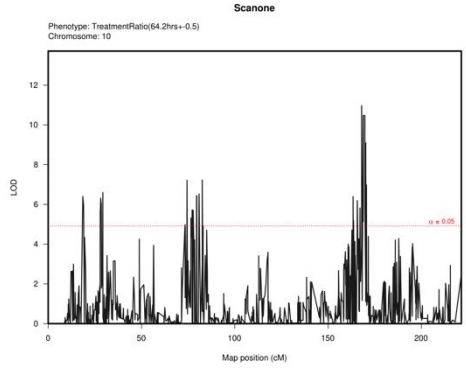
ASP 10mM (TR)



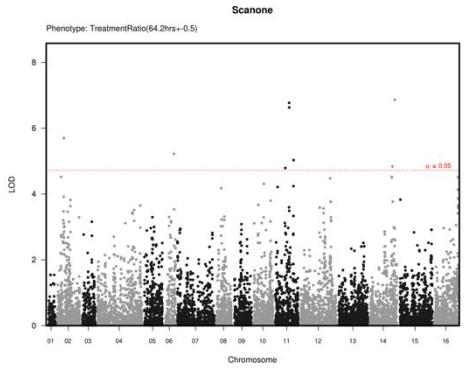
ASP 10mM peak marker (TR)



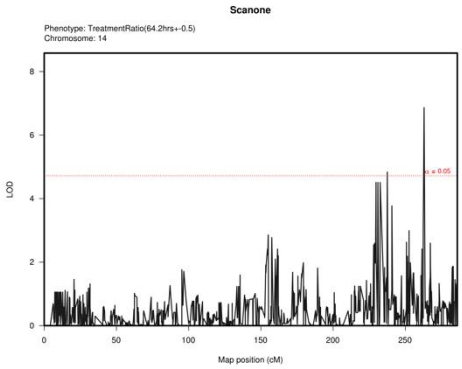
ASP 35mM (TR)



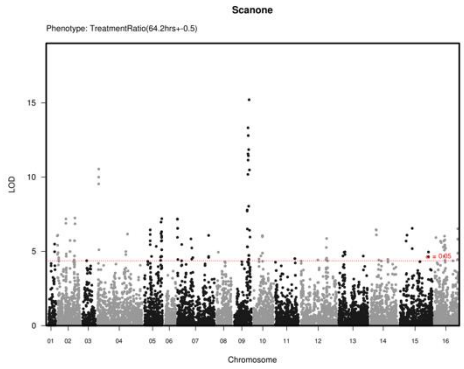
ASP 35mM peak marker (TR)



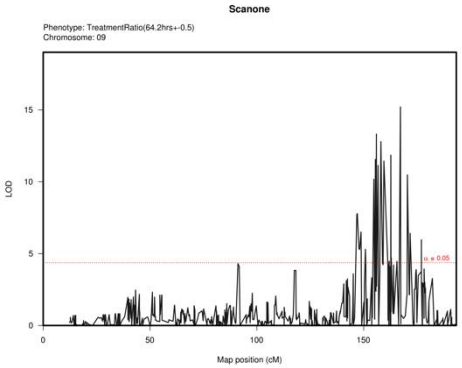
CIS 1mM (TR)



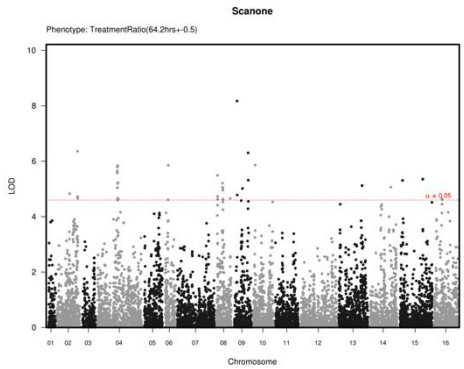
CIS 1mM peak marker (TR)



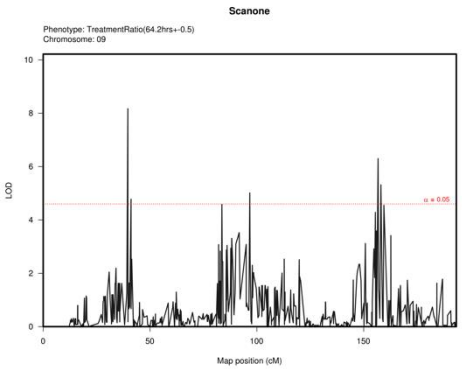
CMM 200uM (TR)



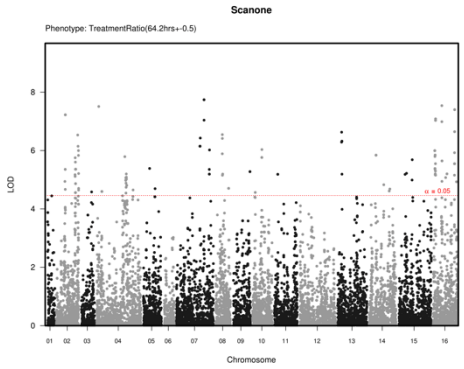
CMM 200uM peak marker (TR)



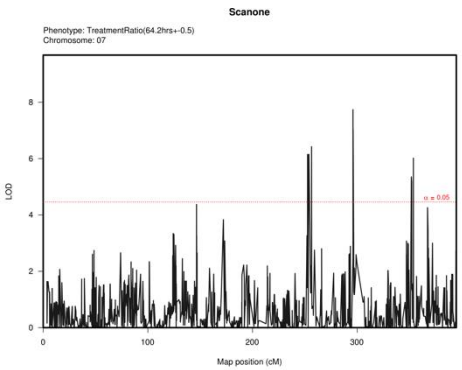
CMM 500uM (TR)



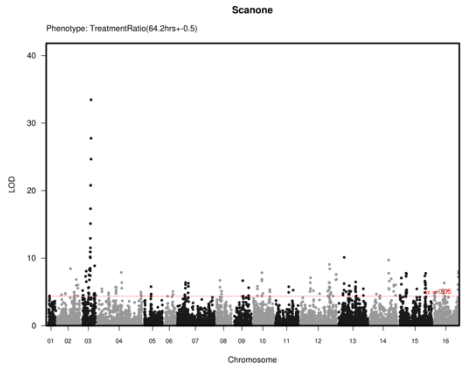
CMM 500uM peak marker (TR)



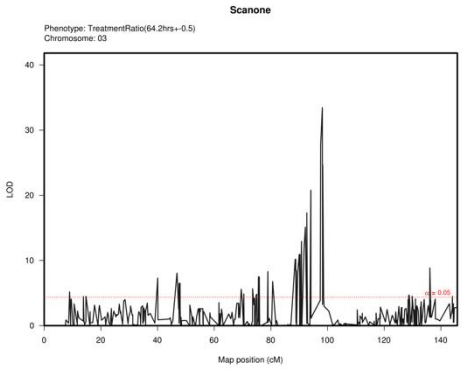
DSF 70uM (TR)



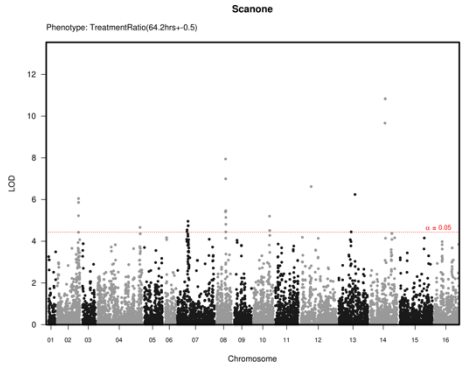
DSF 70uM peak marker (TR)



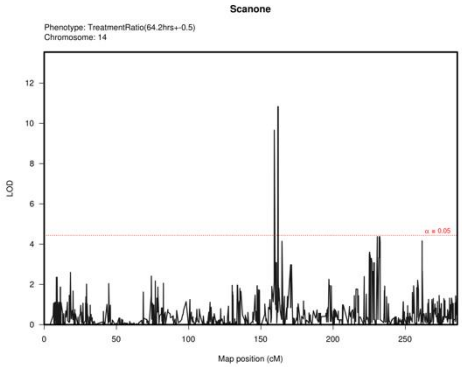
DOX (TR)



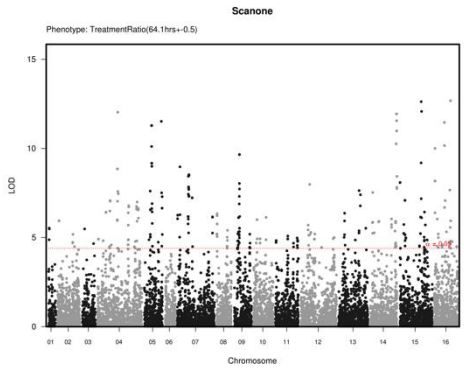
DOX peak marker (TR)



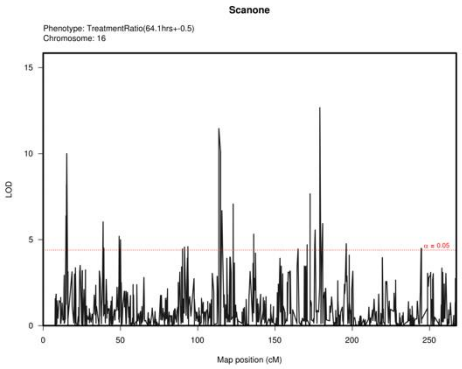
EPA (TR)



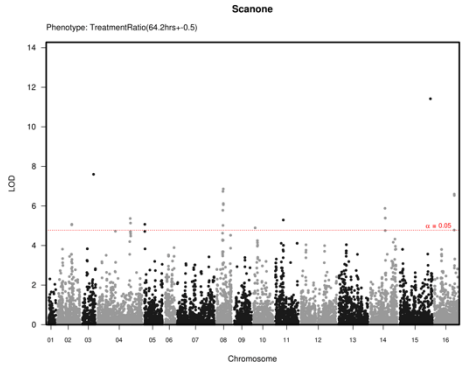
EPA peak marker (TR)



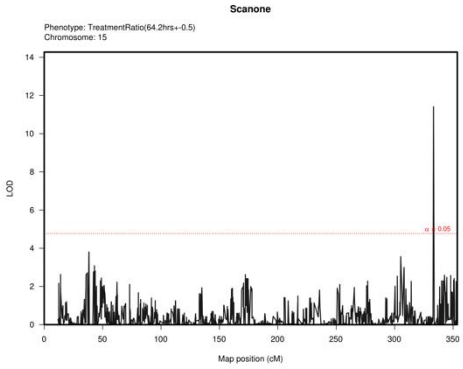
HU (TR)



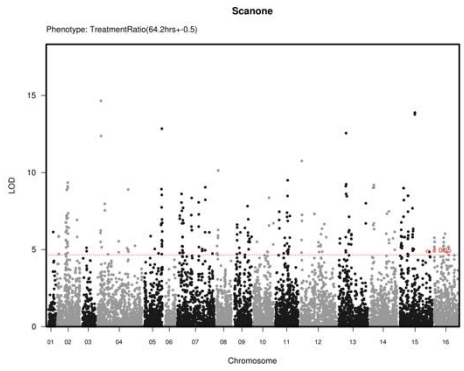
HU peak marker (TR)



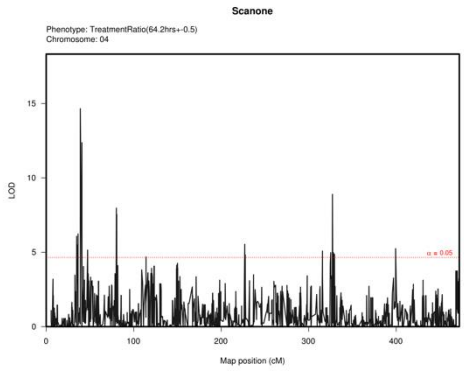
MT (TR)



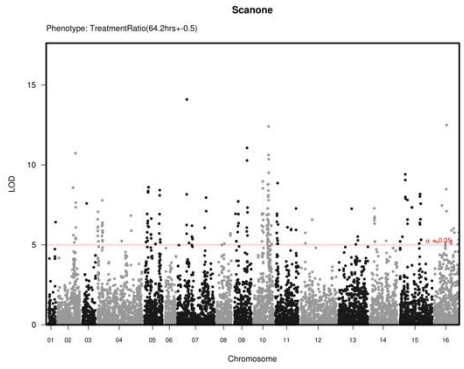
MT peak marker (TR)



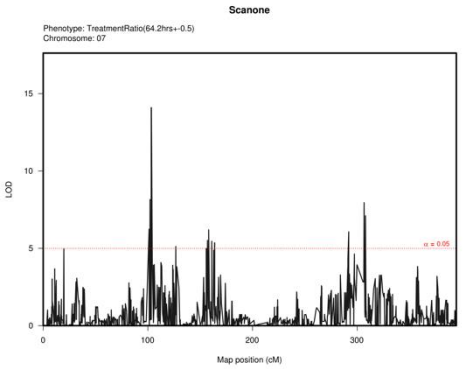
MT + DOX (TR)



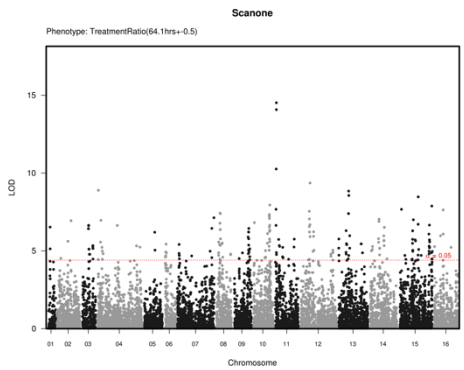
MT + DOX peak marker (TR)



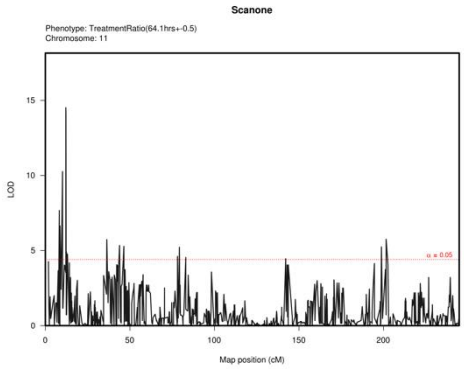
MET (TR)



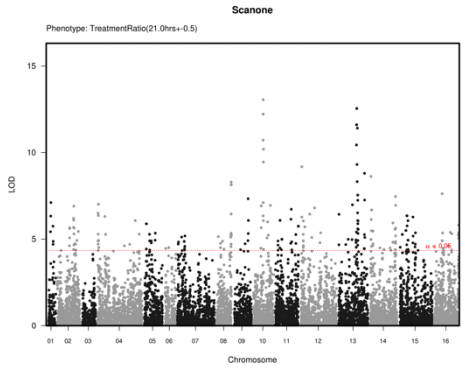
MET peak marker (TR)



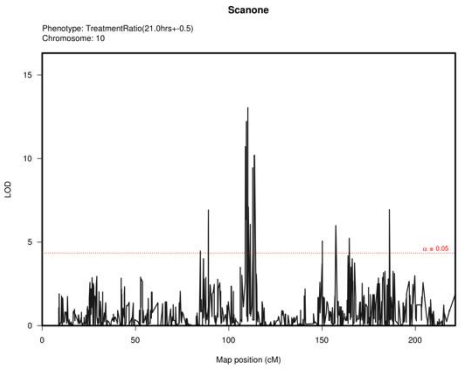
MMS (TR)



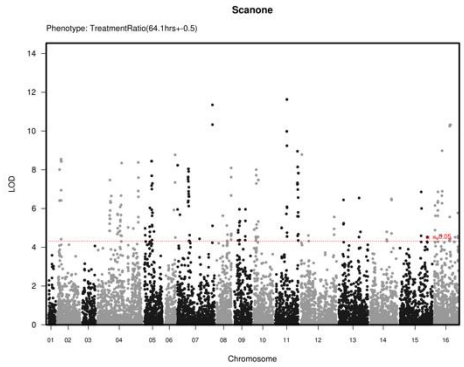
MMS peak marker (TR)



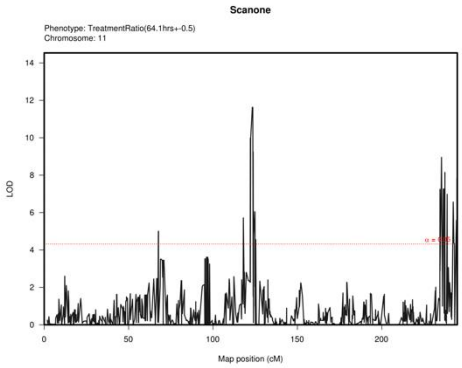
PQ (TR)



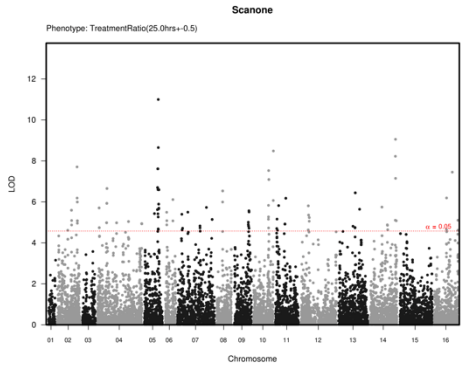
PQ peak marker (TR)



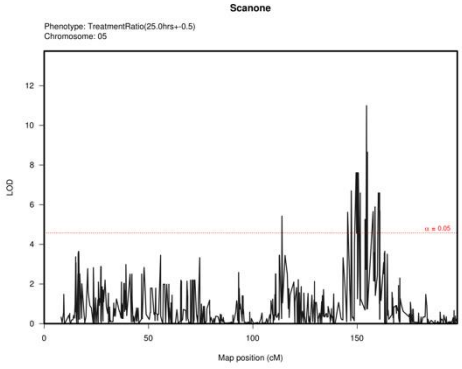
Phleo (TR)



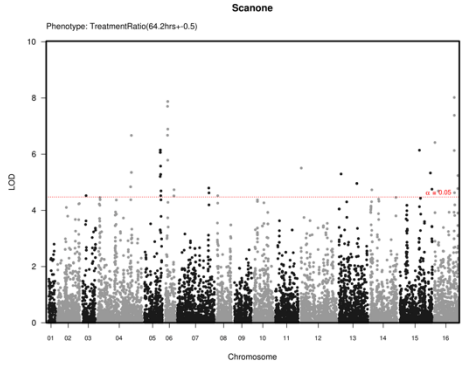
Phleo peak marker (TR)



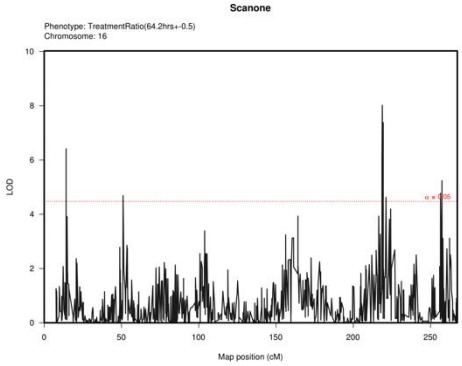
RAP (TR)



RAP peak marker (TR)



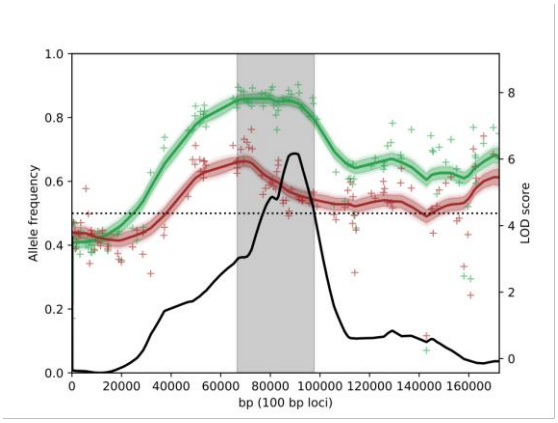
SLA (TR)



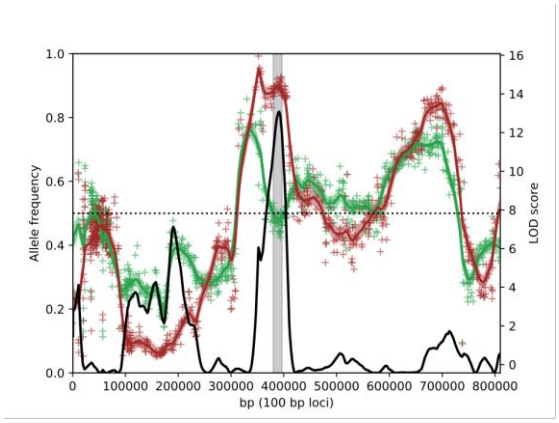
SLA peak marker (TR)

B.2 Multipool analysis output

B.2.1 H179 Low Temperature *S.eubayanus* QTL output

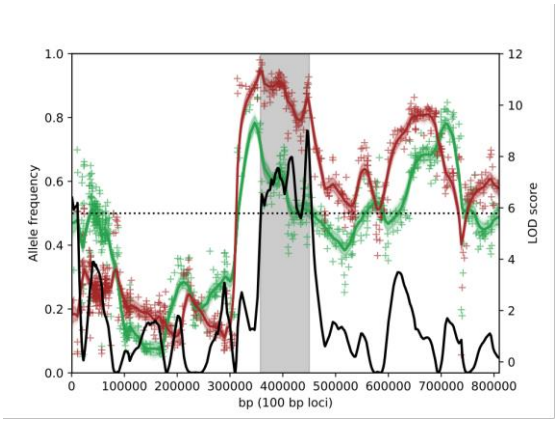


Chromosome I

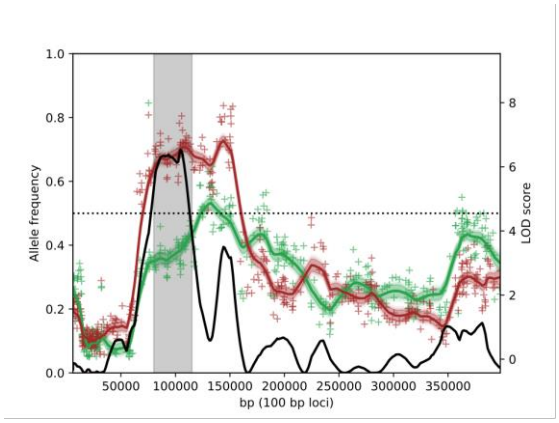


Chromosome VIII

B.2.2 H179 High Temperature *S.eubayanus* QTL output

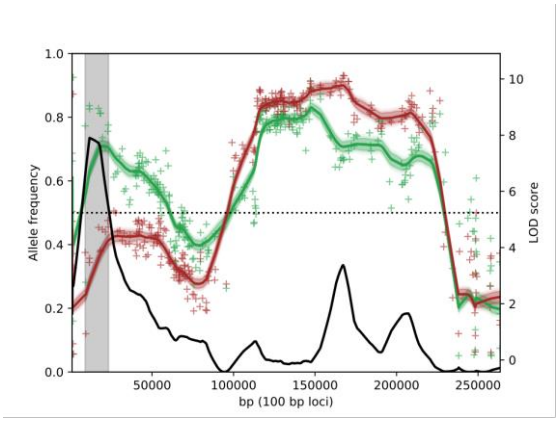


Chromosome VIII

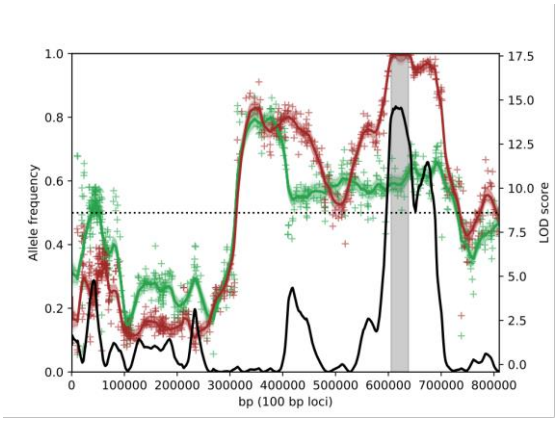


Chromosome IX

B.2.3 H179 High Ethanol *S.eubayanus* QTL output

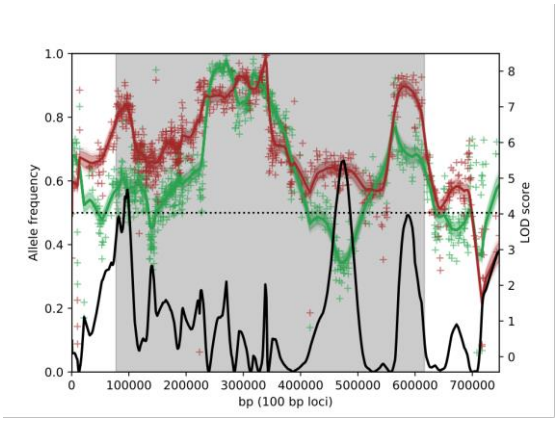


Chromosome VI

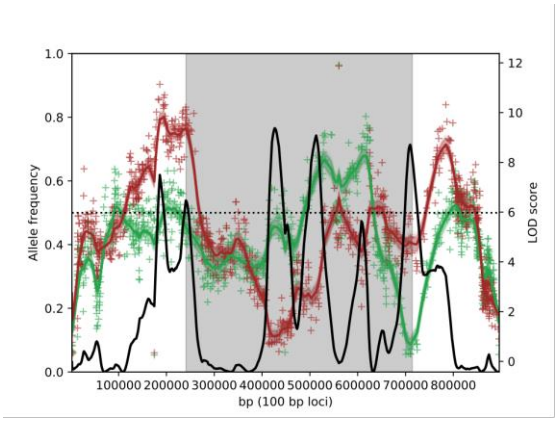


Chromosome VIII

B.2.4 H188 Low Temperature *S.eubayanus* QTL output

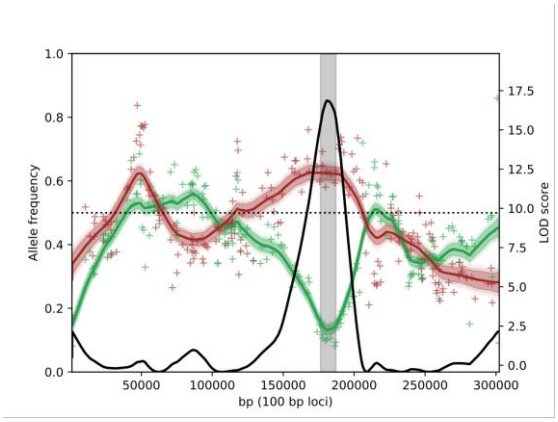


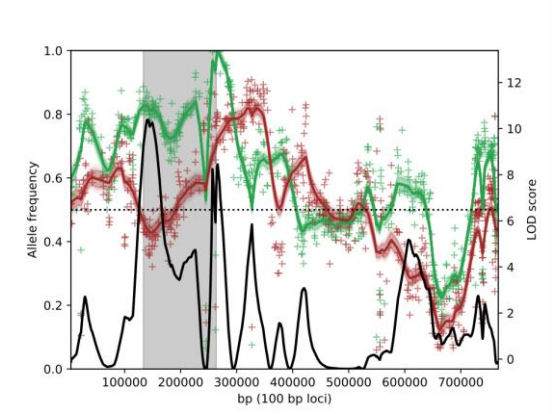
Chromosome X



Chromosome XVI

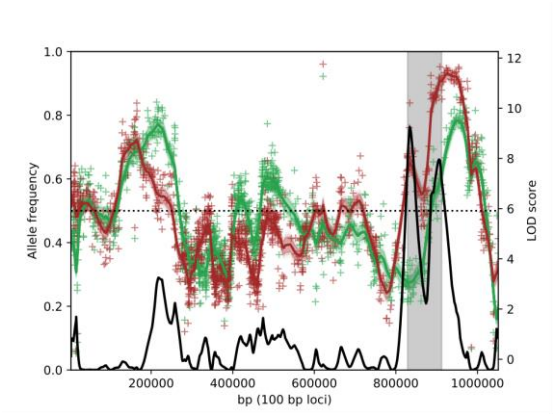
B.2.5 H188 High Temperature *S.eubayanus* QTL output



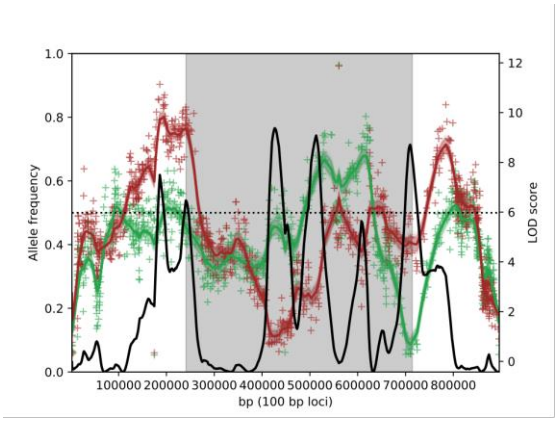


Chromosome III

B.2.6 H188 High Ethanol *S.eubayanus* QTL output



Chromosome VII



Chromosome XVI

Appendix C. Additional Tables

C1. F1 Map Information:

Marker ID	Chromosome	Position (bp)	Genetic Map (cM)
c01:0038000 (CNE1)	1	38000	0.00
c01:0064000 (CDC24)	1	64000	10.36
c01:0078000 (FUN12)	1	78000	44.99
c01:0095000 (SAW1)	1	95000	59.36
c01:0114000 (ATS1)	1	114000	67.23
c01:0158000 (RFA1)	1	158000	82.96
c01:0170000 (ADE1)	1	170000	88.43
c01:0191000 (YAT1)	1	191000	119.08
c02:0045000 (ROX3)	2	45000	0.00
c02:0116000 (YBL055C)	2	116000	38.99
c02:0205000 (SCT1)	2	205000	69.63
c02:0308000 (TLC1)	2	308000	91.47
c02:0383000 (YBR072C-A)	2	383000	118.41
c02:0472000 (LYS2)	2	472000	147.17
c02:0547000 (SPP381)	2	547000	190.92
c02:0639000 (YBR208C)	2	639000	214.39
c02:0696000 (YBR238C)	2	696000	255.70
c02:0770000 (SSH1)	2	770000	292.46
c03:0048000 (MGR1)	3	48000	0.00
c03:0099000 (BUD3)	3	99000	36.78
c03:0135000 (ADP1)	3	135000	52.54
c03:0188000 (SNT1)	3	188000	65.53
c03:0200000 (MAT)	3	200000	66.58
c03:0209000 (YCR045C)	3	209000	68.67
c03:0219000 (CTR86)	3	219000	85.86
c03:0232000 (RAD18)	3	232000	90.19
c03:0251000 (PAT1)	3	251000	108.92
c03:0283000 (CDC39)	3	283000	143.57
c04:0024000 (LRG1)	4	24000	0.00
c04:0052000 (GCS1)	4	52000	28.75
c04:0102000 (YDL199C)	4	102000	61.35
c04:0152000 (GLT1)	4	152000	116.26
c04:0189000 (NOP14)	4	189000	145.01
c04:0279000 (POL3)	4	279000	168.49
c04:0375000 (MTF2)	4	375000	217.52
c04:0464000 (GAL3)	4	464000	244.40

c04:0496000 (VPS54)	4	496000	248.66
c04:0547000 (HEM13)	4	547000	260.27
c04:0628000 (RLI1)	4	628000	290.89
c04:0707000 (ARO1)	4	707000	312.73
c04:0814000 (RSM24)	4	814000	337.92
c04:0902000 (RAD9)	4	902000	374.70
c04:0999000 (PEX10)	4	999000	393.41
c04:1080000 (GIC2)	4	1080000	407.77
c04:1146000 (YDR336W)	4	1146000	432.92
c04:1170000 (MRP1)	4	1170000	438.40
c04:1254000 (SAC7)	4	1254000	457.12
c04:1333000 (PPM1)	4	1333000	485.87
c04:1422000 (000RE2)	4	1422000	514.61
c04:1507000 (YDR535C)	4	1507000	566.49
c05:0026000 (YEL068C)	5	26000	0.00
c05:0088000 (MCM3)	5	88000	61.59
c05:0161000 (YND1)	5	161000	75.96
c05:0236000 (SAH1)	5	236000	96.22
c05:0329000 (ILV1)	5	329000	137.54
c05:0395000 (SPR6)	5	395000	154.77
c05:0464000 (SCC4)	5	464000	179.96
c05:0546000 (BMH1)	5	546000	201.80
c06:0038000 (SWP82)	6	38000	0.00
c06:0067000 (YFL034W)	6	67000	7.87
c06:0077000 (AGX1)	6	77000	18.22
c06:0086000 (BST1)	6	86000	26.05
c06:0094000 (FRS2)	6	94000	32.69
c06:0154000 (RPN11)	6	154000	67.32
c06:0198000 (ROG3)	6	198000	87.58
c06:0241000 (DUG1)	6	241000	111.07
c07:0055000 (MTO1)	7	55000	0.00
c07:0143000 (IME4)	7	143000	43.76
c07:0199000 (SUT1)	7	199000	72.51
c07:0268000 (RSM23)	7	268000	101.25
c07:0333000 (PAN2)	7	333000	130.00
c07:0432000 (MIG1)	7	432000	153.47
c07:0515000 (SNU71)	7	515000	173.72
c07:0614000 (ADE6)	7	614000	197.22
c07:0674000 (VAS1)	7	674000	211.59
c07:0757000 (PEX4)	7	757000	246.24

c07:0848000 (ERG1)	7	848000	257.90
c07:0916000 (ZPR1)	7	916000	292.53
c07:1009000 (RAD2)	7	1009000	329.31
c07:1061000 (ERV29)	7	1061000	354.50
c08:0037000 (GUT1)	8	37000	0.00
c08:0119000 (STP2)	8	119000	43.75
c08:0198000 (INM1)	8	198000	147.68
c08:0222000 (GIC1)	8	222000	155.55
c08:0244000 (OSH3)	8	244000	168.55
c08:0273000 (SAM35)	8	273000	181.54
c08:0337000 (YHR113W)	8	337000	203.40
c08:0433000 (PRP8)	8	433000	207.75
c08:0498000 (AIM46)	8	498000	242.40
c09:0038000 (SUC2)	9	38000	0.00
c09:0072000 (SLN1)	9	72000	39.00
c09:0131000 (POG1)	9	131000	57.71
c09:0192000 (YIL091C)	9	192000	84.63
c09:0276000 (P000P1)	9	276000	149.92
c09:0347000 (EPS1)	9	347000	182.51
c09:0380000 (GAT4)	9	380000	192.87
c10:0049000 (YJL206C)	10	49000	0.00
c10:0106000 (ERG20)	10	106000	23.48
c10:0179000 (PBS2)	10	179000	54.13
c10:0249000 (BC0001)	10	249000	69.91
c10:0312000 (UTP18)	10	312000	91.76
c10:0390000 (VPS53)	10	390000	126.40
c10:0465000 (ILV3)	10	465000	161.03
c10:0552000 (YJR061W)	10	552000	193.63
c10:0612000 (YJR096W)	10	612000	215.49
c10:0671000 (NMD5)	10	671000	227.16
c10:0726000 (YJR154W)	10	726000	266.16
c11:0037000 (TRP3)	11	37000	0.00
c11:0104000 (FAS1)	11	104000	54.88
c11:0134000 (MRP49)	11	134000	60.36
c11:0173000 (AVT3)	11	173000	70.74
c11:0256000 (UTP11)	11	256000	94.24
c11:0324000 (MSN4)	11	324000	124.88
c11:0419000 (PRP40)	11	419000	221.15
c11:0504000 (SPO14)	11	504000	239.88
c11:0567000 (CCP1)	11	567000	274.53

c11:0578000 (Y000R073C)	11	578000	283.64
c11:0598000 (HBS1)	11	598000	290.31
c12:0028000 (YLL056C)	12	28000	0.00
c12:0126000 (YEH1)	12	126000	69.30
c12:0197000 (AAT2)	12	197000	101.89
c12:0291000 (GAL2)	12	291000	143.20
c12:0372000 (HOG1)	12	372000	166.68
c12:0391000 (YLR122C)	12	391000	186.95
c12:0439000 (YLR149C)	12	439000	207.21
c12:0533000 (ATG26)	12	533000	250.96
c12:0606000 (BNA5)	12	606000	285.60
c12:0702000 (YLR278C)	12	702000	309.09
c12:0780000 (PEX30)	12	780000	336.03
c12:0873000 (PSY3)	12	873000	414.46
c12:0917000 (S000I2)	12	917000	423.57
c12:0969000 (YLR422W)	12	969000	440.80
c12:1037000 (LEU3)	12	1037000	484.57
c13:0053000 (ZDS2)	13	53000	0.00
c13:0144000 (ORC1)	13	144000	25.17
c13:0215000 (NDC1)	13	215000	59.80
c13:0308000 (SPO20)	13	308000	84.96
c13:0382000 (STB2)	13	382000	143.09
c13:0467000 (MUB1)	13	467000	186.84
c13:0526000 (ECM16)	13	526000	204.06
c13:0588000 (MSS11)	13	588000	231.00
c13:0659000 (VTI1)	13	659000	254.49
c13:0756000 (ZRC1)	13	756000	287.08
c13:0849000 (ABZ2)	13	849000	308.91
c14:0047000 (YNL313C)	14	47000	0.00
c14:0128000 (SEC2)	14	128000	41.32
c14:0208000 (SIN4)	14	208000	71.95
c14:0281000 (DUG3)	14	281000	90.66
c14:0374000 (FYV6)	14	374000	136.97
c14:0446000 (YNL095C)	14	446000	202.27
c14:0539000 (ALG11)	14	539000	234.87
c14:0622000 (MRP7)	14	622000	255.12
c14:0720000 (POP2)	14	720000	324.41
c15:0059000 (ARG8)	15	59000	0.00
c15:0143000 (TRM10)	15	143000	26.93
c15:0205000 (INP54)	15	205000	50.41

c15:0276000 (LAG2)	15	276000	89.40
c15:0369000 (YOR019W)	15	369000	114.59
c15:0442000 (C000A2)	15	442000	130.37
c15:0504000 (R000I1)	15	504000	162.95
c15:0604000 (ELG1)	15	604000	175.94
c15:0687000 (MSB1)	15	687000	199.43
c15:0759000 (ODC2)	15	759000	216.67
c15:0834000 (YTM1)	15	834000	247.32
c15:0920000 (LDB19)	15	920000	296.34
c15:0991000 (CIN1)	15	991000	351.25
c15:1047000 (ATF1)	15	1047000	378.18
c16:0037000 (PLC1)	16	37000	0.00
c16:0104000 (RVB2)	16	104000	34.64
c16:0201000 (RTT10)	16	201000	78.40
c16:0302000 (COX11)	16	302000	127.43
c16:0390000 (SEC16)	16	390000	162.07
c16:0480000 (YPL039W)	16	480000	192.71
c16:0541000 (CHL1)	16	541000	216.19
c16:0630000 (CSR2)	16	630000	236.44
c16:0658000 (MSF1)	16	658000	249.43
c16:0727000 (YPR097W)	16	727000	286.20
c16:0793000 (SCD6)	16	793000	320.84
c16:0899000 (SEC23)	16	899000	346.01

C2. Sample sequences information for Chapter 6

Sample Name	#Reads	Mean Q30 R1	Mean Q30 R2	Lane	Coverage
H179_Eth_max	29780159	150	144	4	184
H179_4c_max	23684265	150	144	4	146
H179_40c_max	28604033	150	144	4	176
H179_Eth_min	27049359	150	144	4	167
H179_4cmin	31927533	150	144	4	197
H179_40cmin	22364617	150	144	4	138
H188Eth_max	17510300	150	144	4	108
H1884c_max	20155855	150	144	4	124
H188_40c_max	12571291	150	144	4	78
H188Eth_min	24335455	150	144	4	150
H1884cmin	17913417	150	149	4	111
H18840cmin	26800614	150	144	4	165
H179_fermentation	31370218	150	144	4	194
H188_fermentation	23572349	150	139	4	145
JRY9185	18465108	249	219	2	378