Identification and characterization of functional regulatory variants in *Saccharomyces cerevisiae*

By Timothy Read
B.S., Northeastern University, 2006

A thesis submitted to the faculty of the graduate school in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Molecular, Cellular, and Developmental Biology, 2015

This thesis entitled: Identification and characterization of functional regulatory variants in *Saccharomyces cerevisiae*, written by Timothy Read, has been approved by the department of Molecular, Cellular, and Developmental Biology

_____

Robin Dowell

_____

Mark Winey

Date: _____

The final copy of this thesis has been examined by the signatories and will find that the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Read, Timothy J. (PhD, Molecular, Cellular, and Developmental Biology)
Identification and characterization of functional regulatory variants in *Saccharomyces cerevisiae*
Thesis directed by Dr. Robin D. Dowell

Abstract

Most genetic variants associated with disease occur within regulatory regions of the genome, underscoring the need to define the mechanisms that control differences in gene expression regulation between individuals. I discovered a pair of co-regulated, divergently oriented transcripts, *AQY2* and *ncFRE6*, that are expressed in one strain of *S.cerevisiae*, ∑1278b, but not in another, S288c. By combining classical genetics techniques with high-throughput sequencing, I identified a *trans*-acting single nucleotide polymorphism within the transcription factor *RIM101* that causes the background-dependent expression of both transcripts. Subsequent RNA-seq experiments revealed that deletion of *RIM101* in both backgrounds abrogated the majority of differential expression between S288c and ∑1278b and showed that *RIM101* regulates many more targets in S288c than in ∑1278b. However, expression profiling of both strains harboring either *RIM101* allele revealed that only three transcripts undergo a significant allele-dependent change in expression. Strikingly, hundreds of *RIM101*-dependent targets underwent a subtle but consistent shift in expression in the S288c *RIM101*-swapped strain, but not its ∑1278b counterpart. I conclude that ∑1278b may harbor a variant(s) that buffers against widespread transcriptional dysregulation upon introduction of a non-native *RIM101* allele, emphasizing the importance of accounting for genetic background when assessing the impact of a regulatory variant.

Acknowledgements

The work presented here would not have been possible without the support of many, whether involved in science or not. First and foremost, I would like to thank my advisor, Robin Dowell for her continued support and positivity throughout the years, and for letting me explore any question that was at all relevant to my project. I would like to thank my committee, Mark Winey, Rui Yi, Norm Pace, and Roy Parker for inspiring me to think critically about results. Thank you members of the Dowell lab, especially Dave Knox, Jess Vera, Amber Sorenson, Joe Rokicki, Joey Azofeifa, and Josephina Hendrix, whose friendship provided daily laughter and tremendous insight into my research. In particular I would like to thank Phil Richmond, who was critical to this work reaching its potential through help with high throughput data analysis. Personally, I would like to thank my mother, Cathy, my father, Stan, and my brother, Ben for their love and support through this very important phase of my life. Finally, I would like to thank my fiancé, Sam for being there for me every day and her incredible scientific reasoning skills, and my dog, Lucy, for her unconditional love and her sense of adventure.

Contents

Figures

Tables

Chapter I: Introduction

Since the completion of the first human genome more than a decade ago (Venter et al. 2001; Lander et al. 2001), the field of genomics has undergone a shift from describing the general content and architecture of the genome to understanding how it functions. However, even with thousands of genomes sequenced and a plethora of functional data available, the link between genotype and phenotype is unclear. One recently emerging theme is that phenotype is significantly affected by differences in gene expression regulation, especially transcriptional regulation. However, detailed mechanisms describing how genetic variation influences gene expression regulation and phenotype remain elusive.

1.1 Gene expression regulation contributes to the evolution of phenotypes

Well before the sequencing of a genome, some visionary researchers speculated that differences in gene expression regulation could contribute to the astounding phenotypic diversity observed in nature (Britten & Davidson 1969). In 1975, King and Wilson hypothesized that differences in gene expression regulation, rather than in the structure or function of the proteins being regulated, could drive the phenotypic differences between humans and chimpanzees (King & Wilson 1975). However, because no methods for obtaining genome sequences or measuring genome-wide expression profiles existed for the next 30 years or so, the hypothesis remained

untested. Today it appears that King and Wilson were correct, as humans and chimpanzees share approximately 96% genetic identity (The Chimpanzee Sequencing and Analysis Consortium 2005), with an even higher degree of similarity within coding regions, strongly implicating gene expression regulation in the phenotypic divergence between the species.

Recently, high throughput sequencing has enabled dozens of new assays geared at understanding how genomes are "read." For example, gene expression microarrays and RNA-seq allow for the quantification of transcripts throughout the genome (Mortazavi et al. 2008), while ChIP-seq catalogues the genomic occupancy of regulatory proteins such as transcription factors (TFs) and epigenetically modified histones (Barski et al. 2007). Together, these technologies have enabled researchers to make striking correlations between protein occupancy and gene expression that have fundamentally challenged traditional models of gene expression regulation.

In general, studies examining the role of gene expression regulation on phenotype take one of two approaches. One common approach asks the question, "How do differences in gene expression affect phenotype?" Alternatively, researchers can take advantage of the fact that expression patterns are simply intermediate phenotypes and ask, "How do differences in DNA sequence affect gene expression (Romero et al. 2012)?" Both approaches have yielded intriguing results, but strategies linking the two approaches to develop a unified model for how DNA affects phenotype remain elusive.

Transcription is regulated by the binding of transcription factors (TFs) to DNA. An early ChIP-seq study asked how genome-wide binding of two TFs is conserved among five vertebrate species (Schmidt et al. 2010). Surprisingly, the overwhelming majority of binding events are unique to a single species, and very few (only 35 out of about 30,000 total TF binding events) were conserved between all five species. Furthermore, a study comparing occupancy of RNA Polymerase II between immortalized lymphoblastoid cell lines revealed that ~32% of RNA PolII-occupied sites were different between humans and chimpanzees (Kasowski et al. 2010). These studies demonstrate that binding sites of protein factors known to affect gene expression are rapidly turning over throughout evolution and likely contribute to the phenotypic plasticity observed throughout evolution (Dowell et al. 2010). However, such results are correlations. In order to prove that differences in binding of various factors are important for gene expression and phenotype, more tractable systems have been exploited.

Model organisms are invaluable tools for understanding how differences in DNA sequence influence the evolution of gene expression and related phenotypes. One compelling example is that of pelvic fin reduction in the three-spine stickleback, which has been implicated in the invasion of freshwater habitats. Initial experiments suggested that the causal locus is a gene called Paired-like homeodomain transcription factor one (*Pitx1*) (Chan et al. 2010). Fine mapping of the locus implicated an enhancer element upstream of the gene in controlling pelvic fin development. Because the stickleback fish is a well-established model organism, researchers were able to show that deletion of the DNA element did have a large effect on *Pitx1* expression and pelvic fin development.

Furthermore, the phenotype could be rescued by the incorporation of a transgene harboring the enhancer element, proving that the enhancer DNA element, whose DNA sequence is under strong selective pressure, is the evolutionary driver of the phenotypic adaptation. Results such as this have led to an increase in research focused on identification and characterization of functional regulatory variants. However, detailed studies describing the mechanisms by which such variants function are rare.

1.2 Strategies for identification of functional regulatory variants

High throughput sequencing has enabled a number of strategies for identification of functional regulatory variants. Efforts to do so often focus on identifying DNA regions of high evolutionary conservation, often referred to as ultraconserved elements (Rands et al. 2014) as their function is likely conserved because they impart a fitness advantage. More recently, Genome-Wide Association Studies (GWAS) have identified thousands of SNPs linked to human traits, including disease. However, while GWAS studies are powerful tools for identifying variants, they rarely inform about the biological processes being affected.

Expression Quantitative Trait Loci (eQTL) studies present a unique opportunity to link DNA sequence to molecular function by correlating genomic features from large numbers of individuals with expression profiles. Many important genetic concepts pertinent to control of gene expression have been illuminated by eQTL studies (Cookson et al. 2009). For example, studies in yeast have revealed that differences in

cis regulatory elements (those existing near the gene being regulated) tend to be more commonly associated with inter- and intra-species differences in gene expression than *trans* factors and also tend to have a large effect on the gene being regulated (Schadt et al. 2003). Presumably, such *cis* elements can function by affecting the binding of TFs to DNA, resulting in altered transcription of the target gene. However, *trans* effects (DNA elements occurring distal to the gene being regulated) appear to be linked to more transcript levels than *cis* effects, though their detection is often limited by the need to perform multiple hypothesis testing and their effects can be subtle, yet widespread (Yvert et al. 2003). In any case, there is considerable evidence for co-evolution of *cis* and *trans* regulatory elements (Tirosh et al. 2009; Gordon & Ruvinsky 2012; Ahead 2005), further supporting a major role for gene expression regulation in phenotypic adaptation. Finally, trait-associated SNPs are enriched for eQTL (Nicolae et al. 2010), strongly supporting a role for expression-influencing variants in phenotypes including human disease.

While GWAS and eQTL studies make strong correlations about genetic variants and phenotypes, more direct, single locus studies are required to define the mechanistic basis of such events. For this reason, the brewer's yeast, *Saccharomyces cerevisiae* has been especially useful. *S.cerevisiae* can be easily genetically modified, allowing researchers to edit any non-essential DNA element (Storici et al. 2001). One early example of the utility of yeast as a model organism for the interrogation of the mechanisms governing gene expression came in 1986, when Rudolph and Hinnen generated a comprehensive set of promoter deletions upstream of the *PHO5* gene and

tested their effect on expression (Rudolph & Hinnen 1987). This approach led to the discovery of several important *cis*-regulatory DNA sequences required for appropriate expression of the *PHO5* gene under conditions of low inorganic phosphate.

More recently, the power of yeast genetic techniques has been combined with high throughput methodologies to uncover new concepts. For example, Xu et al used *S.cerevisiae* to define a role for antisense transcripts (Xu et al. 2011). Utilizing *S.cerevisiae*'s unique ability to be crossed, and phenotypes monitored in segregating populations of progeny, they first used microarrays to uncover correlations between antisense transcripts and expression of overlapping ORFs. They identified one particular antisense transcript whose expression was correlated with on/off expression of the overlapping gene, *SUR7*. By introducing mutations into the promoter of the antisense transcript they could completely abrogate its transcription. Loss of antisense transcription allowed for expression of *SUR7* in conditions where it was not previously expressed. Clearly, using high throughput methods to identify correlations, followed by direct genome editing techniques to prove their causality, is a useful approach, especially in *S.cerevisiae*. Today, with genome editing techniques becoming more feasible in mammalian cells, similar strategies will undoubtedly uncover principles of gene expression regulation in higher eukaryotes (Material et al. 2014).

1.3 Variants affecting gene expression regulation contribute to human disease

To date, GWAS studies have identified thousands of trait and disease-associated SNPs. Staggeringly, the vast majority (93%) of variants identified in these studies exist within non-coding regions of the genome, making it difficult to predict the biological processes they affect (Maurano et al. 2012). It is likely that a large portion of the non-coding SNPs are in some way involved in transcriptional regulation, potentially by altering enhancer or promoter sequences. Further support for this hypothesis comes from the observation that disease- and trait-associated SNPs are enriched for eQTL (Cookson et al. 2009; Pomerantz et al. 2009; Musunuru et al. 2010; Harismendy et al. 2011).

How are functional regulatory variants identified and classified? In addition to large-scale GWAS approaches, epigenetic modifications can be indicators of functional regulatory DNA elements (Fernandez & Miranda-Saavedra 2012; Papait et al. 2013; Hon et al. 2009). For example, H3K4 trimethylation is a staple of active promoters (Pekowska et al. 2011), and H3K36 is associated with actively transcribed chromatin (Krogan et al. 2003). Perhaps the most informative DNA mark is DNaseI hypersensitivity sites, as they often occur at sites of GWAS-identified *cis* regulatory elements (Maurano et al. 2012). This result strongly supports a model by which non-coding, *cis* regulatory variants are major drivers of human disease, and underscores the need to elucidate the mechanisms by which such variants exert their influence on the transcriptome.

Though the majority of disease and trait associated SNPs occur within non-coding regions of the genome, there are also a large number of disease-associated SNPs within transcription factors themselves, again supporting the hypothesis that

transcriptional regulation is vital to normal cellular function. Of particular interest are mutations within TFs common to human cancers (Lawrence et al. 2014; Kandoth et al. 2013). For example, mutations in the TF p53 are present in over 50% of cancers (Lawrence et al. 2013; Forbes et al. 2011). In addition to nonsense mutations, which usually cause complete loss of function phenotypes, many cancer-associated TF-linked SNPs are missense mutations, further complicating their mechanistic interpretation (Chang et al. 2013). One structure-based method showed that such missense mutations can alter DNA binding of the TF p53 in a manner that corresponds to gene expression, but it remains unclear how such structure-altering TF mutations result in varied disease states or outcomes (Ashworth et al. 2014). Clearly, mutations involved in transcriptional regulation are major contributors to phenotype, including human diseases such as cancer.

1.4 Genetic context influences gene expression phenotypes

Experiments designed to define the mechanisms by which regulatory variants function are usually carried out within a single genetic background. However, it is well established that genetic background effects are pervasive in nature for other phenotypes (Chandler et al. 2013). Only recently has it been appreciated that such effects also contribute to gene-expression regulation phenotypes (Dworkin et al. 2009).

S.cerevisiae represents an excellent model system for understanding how intra-species differences in DNA sequence affect various phenotypes. With hundreds of

sequenced strains available (Strope et al. 2015), *S.cerevisiae* is an attractive model for characterizing molecular phenotypes present within human populations. One recent study showed that divergence in the TF *Yrr1* between strains of *S.cerevisiae* not only contributes to a large-scale shift in a cellular phenotype (Drug resistance), but also in regulatory phenotype (DNA binding) (Gallagher et al. 2014). However, genome-wide binding profiles are highly background-dependent, as swapping *YRR1* alleles between diverse genetic backgrounds yields a diverse array of DNA binding patterns. While this study did not directly address the impact of TF binding on gene expression, it is implied that the differences in *Yrr1* binding also influence expression profiles among the assorted genetic backgrounds.

Differences in genetic background have also been linked to alternative transcriptional and phenotypic responses to drug treatment. For example, cancer cell lines respond dramatically differently to treatment with a commonly used MEK inhibitor (Litvin et al. 2015). Furthermore, the authors showed that the phenotypic differences were a result of discrepancies in the cell lines transcriptional response to treatment and proved that such discrepancies were due to mutations in the interferon pathway. As more genome sequences and corresponding transcriptomes become available, researchers will have the unprecedented opportunity to link DNA sequences to expression profiles and to various other phenotypic traits. Such an approach could be analogous to other well-established biomarkers used to predict disease states. With enough information, it may be possible for the field of “personalized transcriptomics” to become a reality (Montgomery & Dermitzakis 2011).

1.5 Summary

Incorporating genomic information into clinical practice is a major focus of personalized medicine. Despite the discovery of a large number of disease-associated genetic variants (Stranger et al. 2011; Welter et al. 2014), few clinical treatments have been developed that incorporate such information (Ramos et al. 2014). Furthermore, most disease-associated variants occur within regulatory regions of DNA (Maurano et al. 2012; Hindorff et al. 2009), making it particularly difficult to predict the biological processes they affect. Determining the mechanisms by which variants influence regulation, and hence phenotypic diversity among individuals, is paramount to a thorough understanding of genomics.

Uncovering the biological mechanisms underlying regulatory variants, as well as how variants interact with the myriad of genetic backgrounds present within a population, is a major focus of current research. It has become increasingly clear that genetic background contributes to phenotypes (Chandler et al. 2013; Chandler 2010; Matin & Nadeau 2001; Nadimpalli et al. 2015; Kim et al. 2009; Cubillos & Billi 2011) resulting in the seemingly infinite diversity observed, even within a species. Recent studies have suggested that transcript levels can be both powerful readouts for and determinants of disease states (Perou et al. 2000; Golub et al. 1999; Litvin et al. 2015; Calon et al. 2015). However, similar to other cellular phenotypes, expression differences among individuals are the product of an exceedingly complex genetic landscape. One reason for this complexity is that even subtle mutations can impart distinct regulatory

roles to transcription factors when placed into alternative genetic contexts (Dworkin et al. 2009; Gallagher et al. 2014).

Strategies linking genetic variants to gene-expression regulation often consist of one of two approaches. Expression quantitative trait loci (eQTL) studies combine genome-wide expression data with genome sequence information to uncover expression-influencing variants, including those linked to disease (Schadt et al. 2003; Brem & Clinton 2002; Jansen & Nap 2001; Nica & Dermitzakis 2013; Rockman & Kruglyak 2006; Westra et al. 2013; Westra & Franke 2014). In addition to variants themselves, eQTL studies have also uncovered many important genetic principles. For example, expression-influencing variants that occur near the gene being regulated, or in *cis*, tend to influence a single gene, whereas variants distal to the gene being regulated, or in *trans*, typically influence expression of many loci (Stranger et al. 2007; Göring et al. 2007; Montgomery & Dermitzakis 2011; Pickrell 2014; Battle et al. 2014). In contrast to the genome-wide eQTL approach, which correlates genetic variants with expression differences, much of our knowledge of the mechanisms underlying gene expression regulation comes from detailed, single-locus studies (Rudolph & Hinnen 1987; Houseley et al. 2008; Hainer et al. 2011; Hongay et al. 2006; Xu et al. 2011). However, such studies usually do not consider the effects of naturally occurring genetic variation on gene expression.

I sought to combine attributes of genome-wide and single-locus studies to understand the mechanistic basis by which genetic variant(s) result in altered gene expression regulation between two strains of *S.cerevisiae*. Here I describe principles

underlying the complexity of gene expression regulation and report evidence that genetic background strongly influences the extent to which a variant affects transcript levels throughout the genome. Specifically, I present studies aimed at understanding the molecular basis of transcriptional differences between two strains of yeast, focusing initially on the *AQY2/ncFRE6* locus. Chapter II focuses on efforts to map and characterize a variant responsible for differential expression between strains of S.cerevisiae, Chapter III discusses results from chapter II, and chapter IV describes my experience mentoring undergraduates in the International Genetically Engineered Machines (iGEM) program.

Chapter II: "A trans-acting variant within the transcription factor *RIM101* interacts with genetic background to determine its regulatory capacity"

The following results and discussion sections were adapted from "A *trans*-acting variant within the transcription factor *RIM101* interacts with genetic background to determine its regulatory capacity." Most bioinformatic analysis was performed by Phillip A. Richmond.

2.1 *Cis* variation controls background-specific co-regulation of *AQY2* and *ncFRE6*

Two strains of *S.cerevisiae*, S288c and ∑1278b, are a model system for how intraspecies genome sequence variation impacts phenotype (Dowell et al. 2010). We initially sought to identify a model locus where we could directly test how naturally occurring genetic variation impacts transcription factor (TF) binding and associated transcript levels (**Figure 2.1**). Specifically, we aimed to identify a genomic region that displays strain-specific binding of a TF that correlates with a nearby strain-specific expression difference. Ideally, our model locus would harbor at least one SNP within a predicted TF binding site, allowing us to directly test the effect of the SNP on TF binding. If TF binding can be modulated by the SNP, we could determine its influence on nearby expression, effectively defining a mechanism by which genome-sequence influences expression.

We performed strand-specific RNA-seq on S288c and Σ1278b. Not surprisingly given the sequence similarity (99.7%) of the strains, the majority of transcripts are expressed at similar levels between the strains (**Figure 2.2**). However, about 20% of genes are differentially expressed (DESeq, n=1207, Padj≤0.0005, minimum average expression ≥ 100 reads) (**Fig 2.2, Table 1**). Of the differentially expressed genes, gene ontology (GO) terms are enriched for categories such as transcription factor activity, mRNA binding, and oxidoreductase activity (Pval≤0.002) (**Table 6**). In addition to protein-coding genes, we identified 82 differentially expressed antisense transcripts (DESeq, n=82, Padj≤0.0005, minimum average expression ≥ 50 reads) (**Fig 2.2, Table 2**).

To better understand how expression patterns are gained or lost throughout intraspecies evolution, we initially focused on loci displaying an extreme differential expression phenotype between the strains (i.e. on in one strain and off in the other) **(n=62, Table 5)**. To identify a potentially regulatory SNP involved in the birth or death of a transcript, we examined the promoter regions of all 62 “extreme expressors’” and noted that one such region harbors a SNP located very near the transcription start site of a non-coding RNA, *ncFRE6*, which is transcribed in an antisense orientation to the *FRE6* ORF in ∑1278b, but not in S288c (**Figure 2.4)**. Closer inspection of the DNA sequence surrounding the *ncFRE6*-correlated SNP revealed that a consensus *Reb1* binding motif is interrupted in S288c relative to ∑1278b **(Figure 2.4, Red line)**. We reasoned that *Reb1* binding could activate *ncFRE6* expression in ∑1278b relative to S288c. Interestingly, expression of *ncFRE6* in ∑1278b correlated with approximately

50% reduction in *FRE6* mRNA levels specifically in ∑1278b, suggesting a possible transcriptional interference event (**Figure 2.4**).

We monitored occupancy of *Reb1* in S288c and ∑1278b by ChIP-seq. Because *REB1* shares 100% sequence identity between S288c and ∑1278b it is not surprising that most *Reb1* binding events are conserved between the backgrounds (83% of total binding events) (**Figure 2.3)**. However, there are a small number of strain-unique *Reb1* binding events, including a binding event in ∑1278b and not in S288c occurring at the position of the *ncFRE6*-asscoiated SNP (**Figure 2.4**). Because the SNP disrupts a preferred *Reb1* binding site in S288c relative to ∑1278b, we interconverted the SNP between backgrounds in an attempt to rescue binding in S288c and/or abolish binding in ∑1278b. We used the "delitto perfetto" method (Storici et al. 2001) to interconvert the SNP and observed that it was indeed necessary and sufficient to cause the *Reb1* binding discrepancy between the strains (**Figure 2.5**). However, we were surprised to observe that abolishing *Reb1* binding in ∑1278b did not reduce expression of *ncFRE6,* and rescuing binding in S288c did not increase expression of *ncFRE6* in S288c (**Figure 2.5**). This result highlights the importance of single locus studies for uncovering the causal variant(s) driving differences in transcript levels rather than assuming that correlations between TF binding and transcript levels are meaningful.

Since differential *Reb1* binding did not cause the differential expression of *ncFRE6* between S288c and ∑1278b, we next asked whether other *cis* elements influenced expression. *AQY2* is a divergently oriented gene that originates ~1kb upstream of the *ncFRE6* transcription start site and is also expressed specifically in

∑1278b. *AQY2* encodes a water channel that is disrupted by a premature stop codon in the vast majority of sequenced strains of *S.cerevisiae*, including S288c, but is functional in ∑1278b (Carbrey et al. 2001). The *AQY2/ncFRE6* promoter region has undergone significant genetic drift between S288c and ∑1278b. Harboring 21 SNPs, the *AQY2/ncFRE6* intergenic region is one of the most sequence-variable promoters between S288c and ∑1278b (**Figure 2.6**). Because a large number of SNPs within the region, both intergenic and within the body of each transcript, disrupt potential TF binding sites, we hypothesized that one or more of the SNPs drive the differential expression of *AQY2* and/or *ncFRE6*. Indeed, replacing all 30 SNPs in ∑1278b with those from S288c results in ~75% reduction of both *AQY2* and *ncFRE6* and replacing only the 15 *AQY2*-proximal SNPs results in ~50% reduction in the transcripts, indicating that DNA elements in both halves of the intergenic region contribute to expression levels of both *AQY2* and *ncFRE6* in Σ1278b **(Figure 2.4)**. Surprisingly, the expression levels of *both* transcripts were reduced to nearly identical levels in ∑1278b promoter-altered strains, implying that the two divergently oriented transcripts are co-regulated in *cis* (**Figure 2.7**). However, while introducing the S288c *cis* context into Σ1278b dramatically reduced expression of the transcripts, incorporation of 30 ∑1278b SNPs into S288c was completely ineffective at increasing *AQY2* and/or *ncFRE6* transcript levels **(Figure 2.7)**. Taken together, these results indicate that *AQY2* and *ncFRE6* are likely co-regulated and furthermore that a *trans*-acting factor(s) ultimately determines whether *AQY2* and/or *ncFRE6* are expressed.

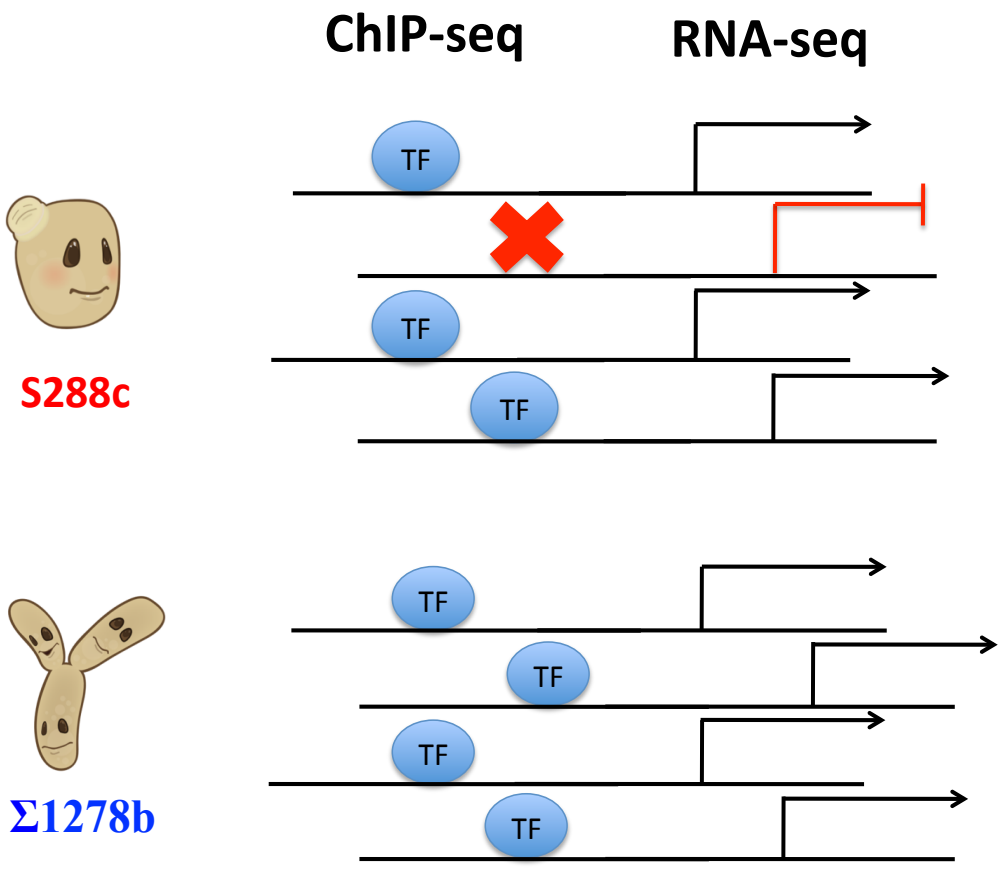


Figure 2.1: Strategy for identification of a functional regulatory variant. ChIP-seq measuring genome-wide occupancy of a TF can be combined with RNA-seq to identify sites where differential binding of the TF correlates with differential expression of the transcript between S288c and ∑1278b.

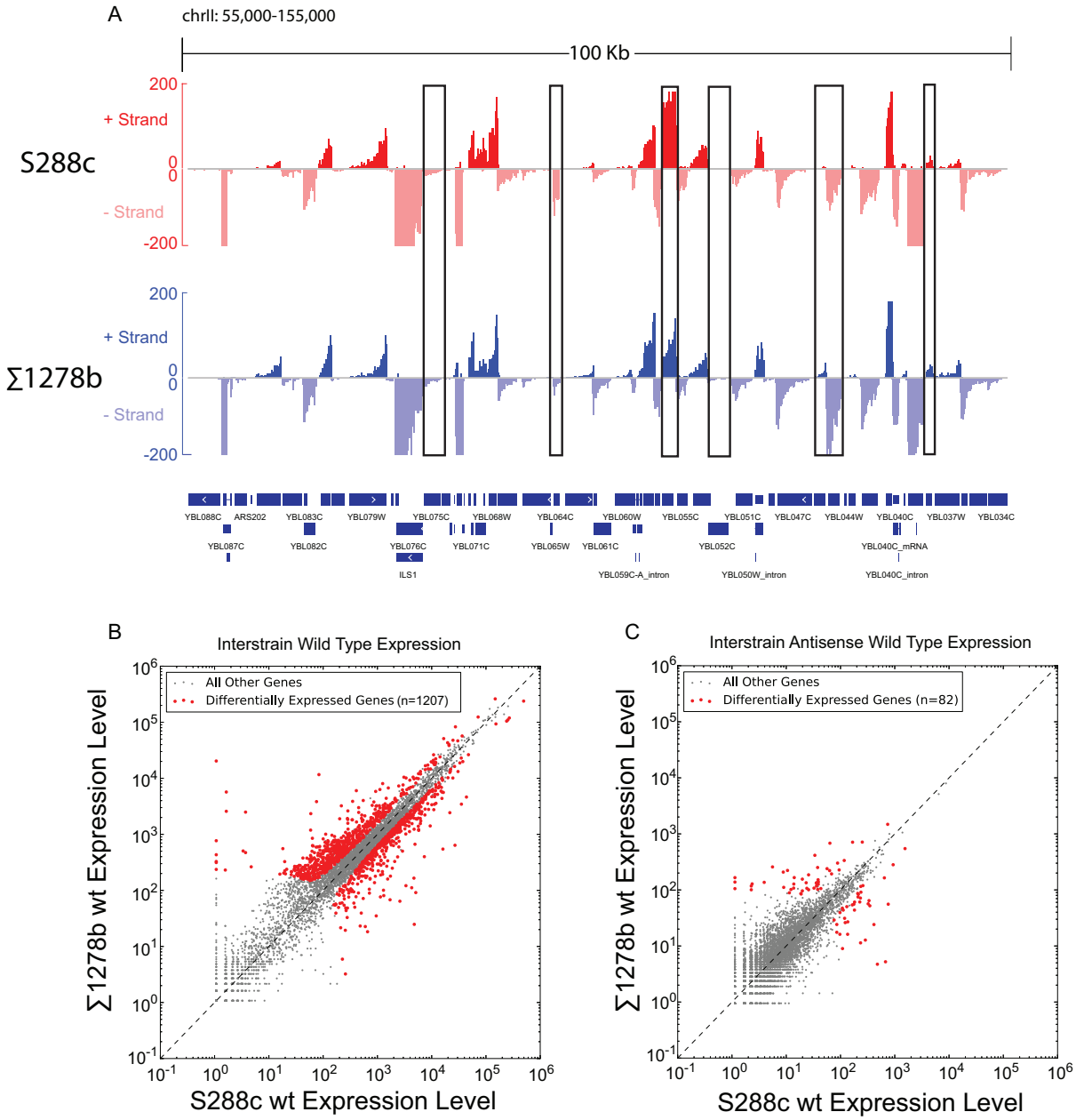


Figure 2.2: Most transcripts are expressed to similar levels between S288c and ∑1278b. (A) Integrated Genome Viewer (IGV) (https://www.broadinstitute.org/igv/) screenshot showing a representative region (100kb) of the genome between S288c (Red) and ∑1278b (Blue). Data are displayed as positive (Watson) strand above the axis and negative (Crick) strand below the axis. Black boxes = differentially expressed transcripts between S288c and ∑1278b. (B) Scatter plot displaying expression levels of 5682 genes in S288c relative to ∑1278b. Red dots (n=1207) are significantly differentially expressed between the wildtype strains (DESeq Padj ≤0.0005). (C) Scatter plot of antisense transcripts for 5682 genes between S288c and ∑1278b.

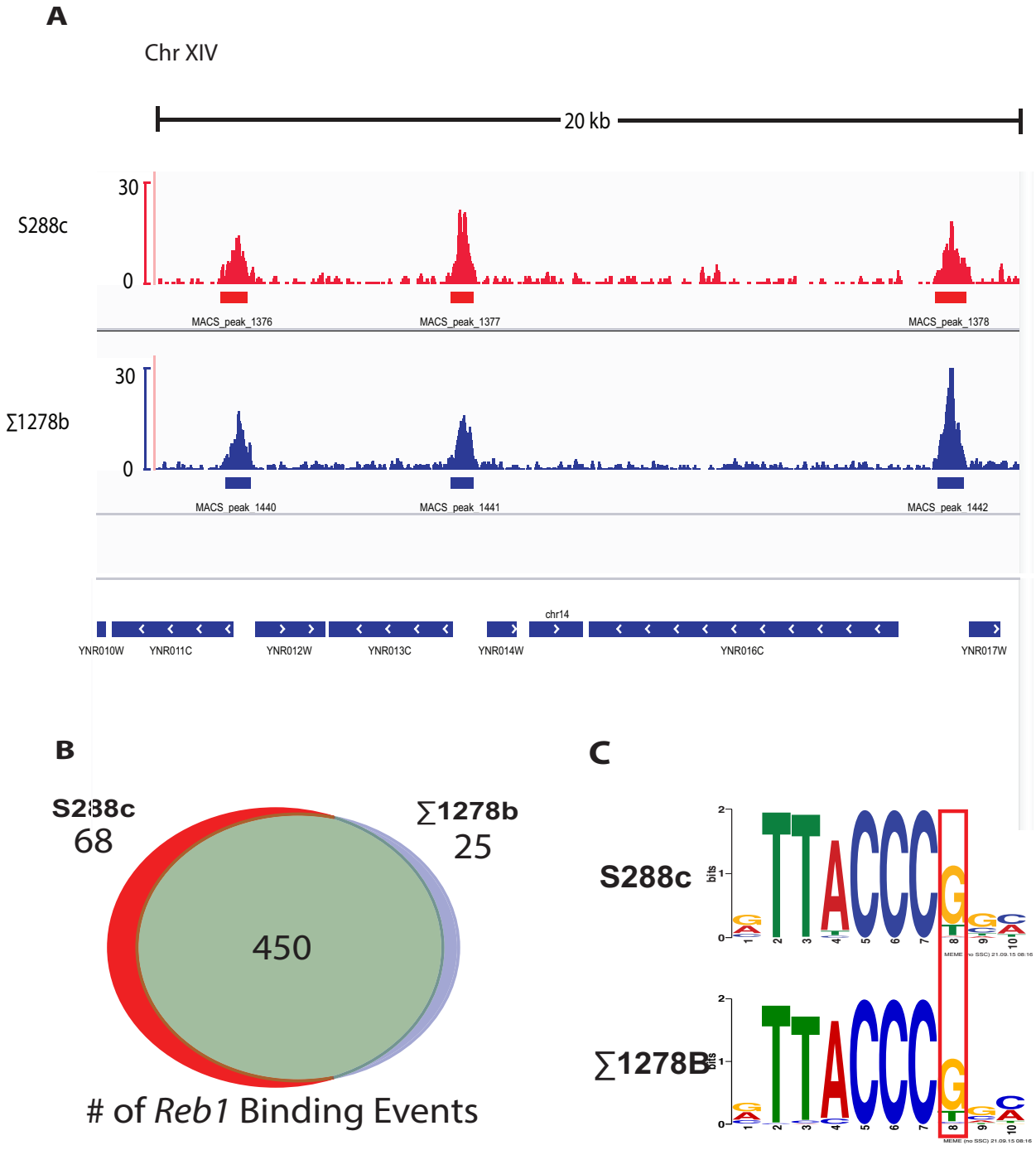


Figure 2.3: Most *Reb1* binding is conserved between S288c and ∑1278b. (A) IGV screenshot of a representative region of the genome (20kb) showing *Reb1* occupancy in S288c (Red) and ∑1278b (Blue). Data normalized to read depth. (B) Venn diagram displaying the number of *Reb1* binding sites occurring uniquely in S288c (Red), ∑1278b (Blue), or conserved between S288c and ∑1278b (Green). (C) Results of Position Specific Scoring Matrix (PSSM) analysis of *Reb1* ChIP-seq data from S288c and ∑1278b.

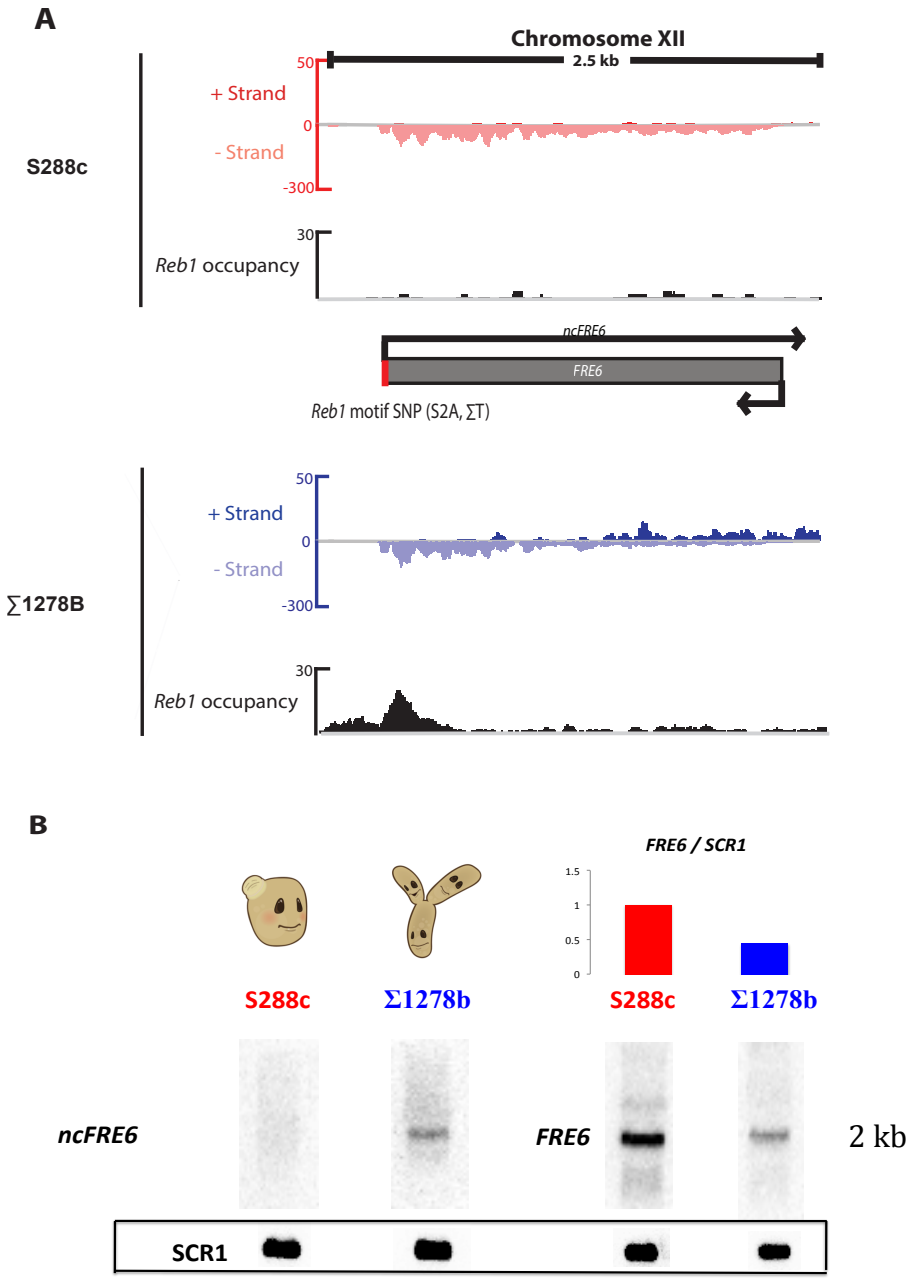


Figure 2.4: Does altered *Reb1* binding cause ∑1278b-specific expression of *ncFRE6*? (A) IGV screenshot displaying strand–specific RNA-seq of the *ncFRE6* region in S288c (Red) and ∑1278b (Blue). Data are displayed as positive (Watson) strand above the axis and negative (Crick) strand below the axis. *Reb1* ChIP-seq data in black for S288c and ∑1278b. Location of a single nucleotide polymorphism within a canonical *Reb1* binding site is represented by a red line. (B) Northern blot probing for expression of *ncFRE6* (Left) or *FRE6* mRNA (Right) in S288c and ∑1278b. Data normalized to loading control, *SCR1*. *FRE6* mRNA level quantified by densitometry.

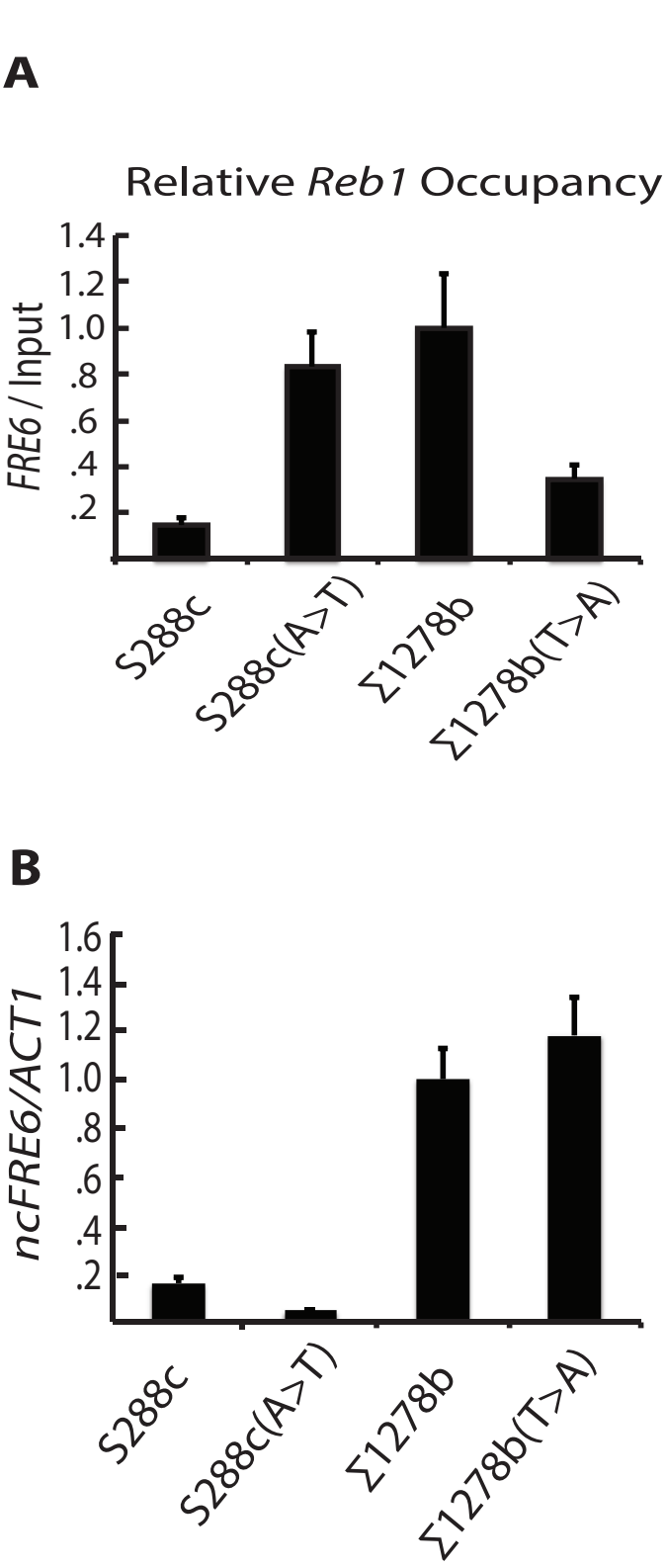


Figure 2.5: Differential binding of *Reb1* between S288c and ∑1278b is controlled by a SNP and does not cause differential expression of *ncFRE6*. (A) ChIP-qPCR displaying relative *Reb1* occupancy at the location of a *Reb1* binding site in S288c, ∑1278b and SNP-interconverted strains. Data normalized to input for each strain. (B) qRT-PCR showing levels of *ncFRE6* in S288c, ∑1278b and *Reb1* SNP-interconverted strains. All qRT-PCR data normalized with ∑1278b expression equal to one. (Throughout results qPCR data include error bars = SEM, n=3 biological replicates unless otherwise noted).

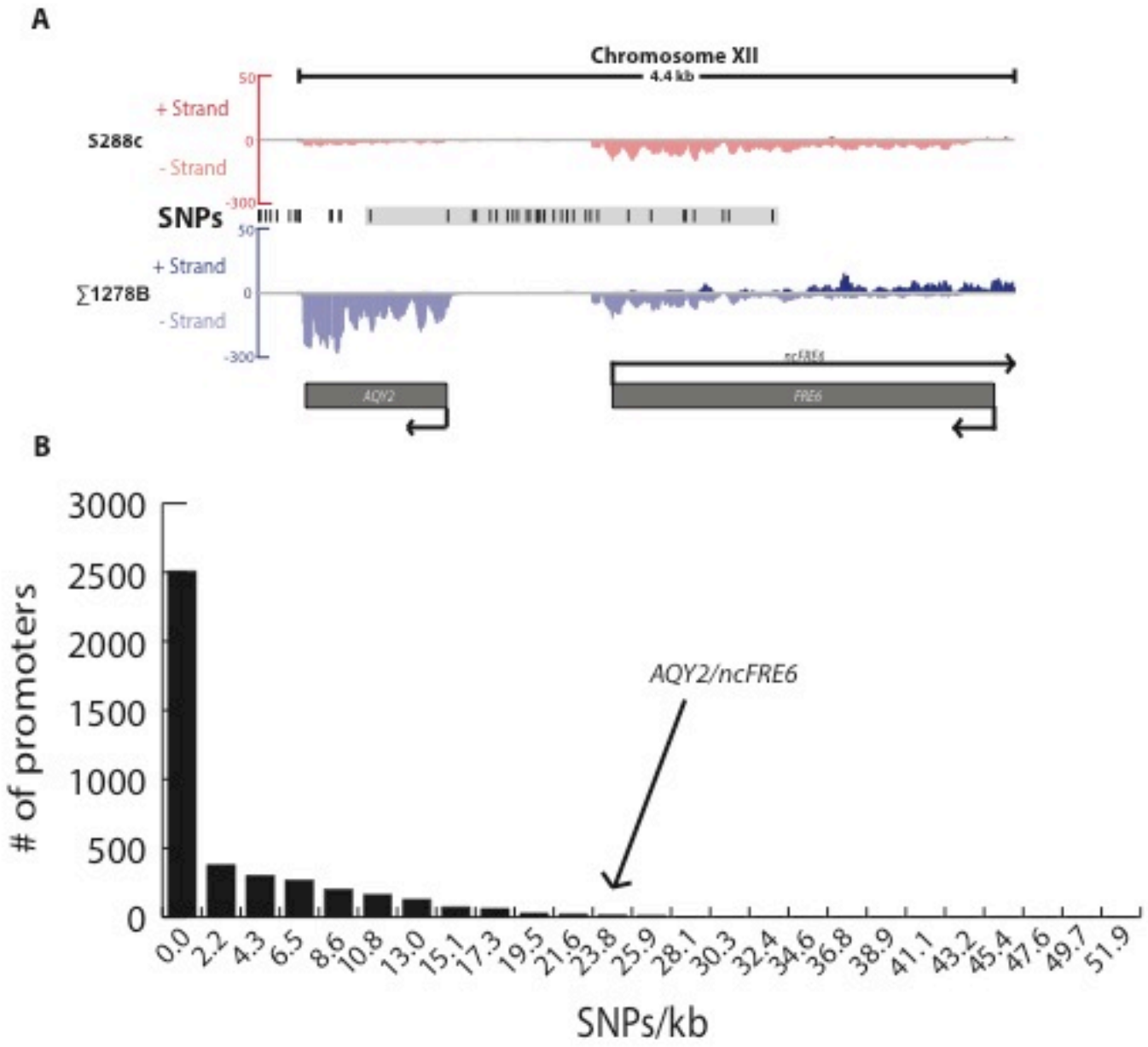


Figure 2.6: Is ∑1278-specific expression of *AQY2* and *ncFRE6* driven by *cis* variation? (A) IGV screenshot displaying strand–specific RNA-seq of the *AQY2/ncFRE6* region in S288c (Red) and ∑1278b (Blue). S288c chromosomal coordinates ChrXII: 35,200-39,570. Data are displayed as positive (Watson) strand above the axis and negative (Crick) strand below the axis. Single nucleotide variations between the strains are shown in black. The grey box highlights 30 interconverted SNPs. (B) Histogram displaying the number of promoters (Y axis, defined as the intergenic region upstream of each annotated ORF existing in both S288c and ∑1278b) relative to the SNP density of each promoter (X axis).

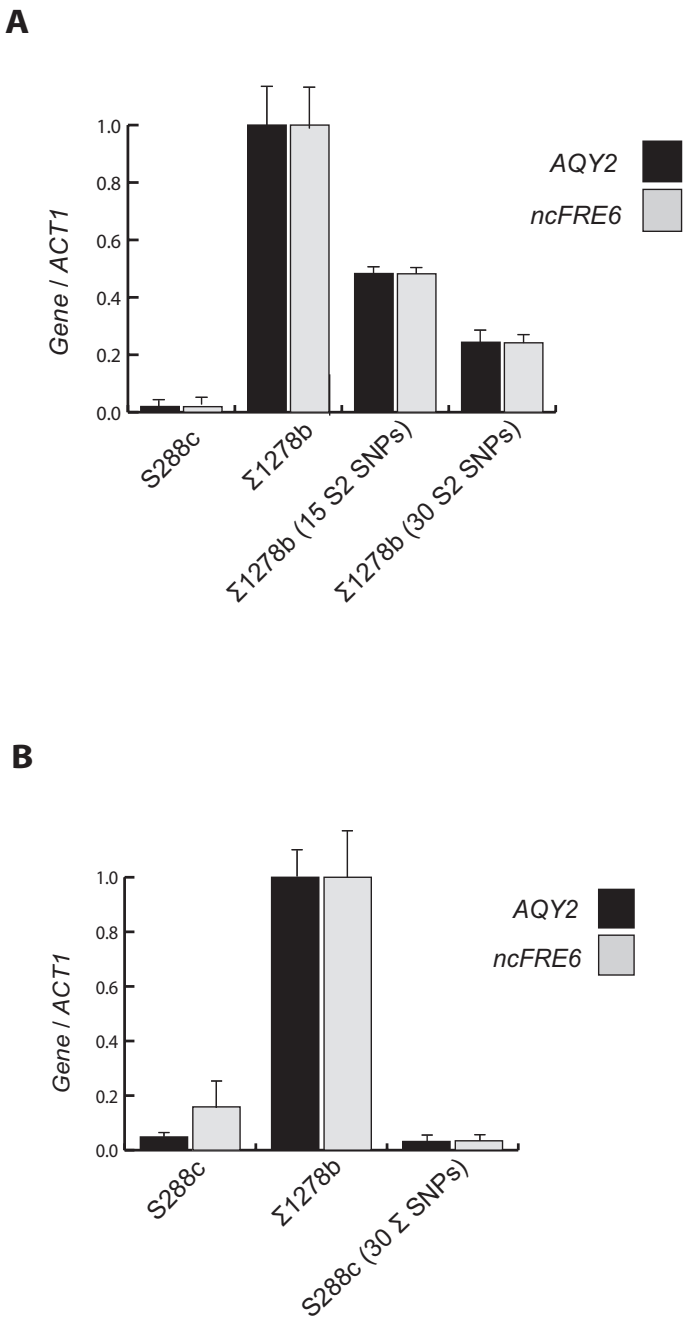


Figure 2.7: *AQY2* and *ncFRE6* are co-regulated by *cis* variation in ∑1278b, but a *trans* factor is epistatic to *cis* elements in S288c. (A) Relative expression of *AQY2* (black) and *ncFRE6* (grey) (Color scheme maintained throughout results) measured by qRT-PCR. (B) Relative expression of *AQY2* and *ncFRE6* measured by qRT-PCR. Data were normalized to *ACT1* levels. ∑1278b wildtype (wt) levels for both *AQY2* and *ncFRE6* were normalized to one throughout the paper. *AQY2* and *ncFRE6* levels in all

other strains are displayed relative to ∑1278b levels. RNA was prepped independently for each qRT-PCR experiment.

2.2 The transcription factor *RIM101* is epistatic to *cis*-linked variation with regards to expression of *AQY2* and *ncFRE6*

To learn about the genetic nature of the *trans* factor(s) that causes differential regulation of *AQY2* and *ncFRE6* between S288c and ∑1278b, we crossed the two strains and monitored expression of the transcripts in a heterozygous diploid. Neither *AQY2* nor *ncFRE6* expression is observed in the diploid strain **(Figure 2.8)**, implying that the S288c expression phenotype is dominant. To determine whether one or more *trans* factors control expression of *AQY2* and/or *ncFRE6*, we performed tetrad analysis assaying for expression of both *AQY2* and *ncFRE6*. We found that for each tetrad analyzed (Winge & Laustsen 1937), two haploid segregants express both *AQY2* and *ncFRE6* concurrently while the other two express neither transcript **(Figure 2.8)**. This 2:2 pattern of inheritance suggests that a single *trans* factor controls the on/off state of both transcripts, and that co-expression of the transcripts is conserved even in the unique genetic admixtures of the segregants.

We Sanger-sequenced the *AQY2/ncFRE6 cis* context within each haploid segregant to determine whether the S288c *cis* context exhibits a similar level of promoter activity as the ∑1278b cis context with regard to expression of *AQY2/ncFRE6* (**Figure 2.8, Red boxes harbor S288c *cis* context**). Indeed, both the S288c and Σ1278b *cis* contexts permit expression of *AQY2/ncFRE6,* but only in the absence of the

trans factor Furthermore, expression levels varied considerably between *AQY2/ncFRE6*-expressing segregants. Contrary to the results of the promoter swapped ∑1278b strain (**Figure 2.7**), the segregants that harbor the S288c *cis* context tend to express *higher* levels of *AQY2/ncFRE6* than those harboring the ∑1278b *cis* context (**Figure 2.8**). This result suggests that there are additional factors that alter the expression levels of *AQY2/ncFRE6*, but only within genetic backgrounds that lack the epistatic *trans*-factor. Furthermore, given that *aqy2* produces a non-functional protein in S288c, it is somewhat surprising that the S288c *cis* context possesses robust promoter activity.

In order to map the genetic location of the *trans* factor that causes differential expression of *AQY2/ncFRE6* between S288c and Σ1278b, we combined bulked segregant analysis (Kesseli 1991) with high throughput sequencing. A similar approach was developed previously, where microarrays were used to map complex phenotypes influenced by a large number of loci (Ehrenreich et al. 2010). We reasoned that the single dominant repressor would be present in all the non-expressing segregants of an S288c x ∑1278b cross. Therefore, the variant driving differential expression of *AQY2/ncFRE6* should always segregate according to the repression phenotype **(Figure 2.9)**. We isolated genomic DNA from 28 segregants: 14 that express *AQY2* and *ncFRE6* and 14 that do not. We pooled equal amounts of DNA from each strain in the two sets and performed high throughput sequencing of the pools. We then sought to identify regions of the genome inherited exclusively from S288c in the non-expressing strains and from Σ1278b in the expressing strains. Only one region fit this criteria: an

approximately 35kb region near the left arm of chromosome VIII **(Figure 2.10, 2.11)**. Because the heterozygous diploid does not express *AQY2* or *ncFRE6*, we reasoned that S288c likely harbors a repressor within this region.

To identify the locus within this region responsible for repression of *AQY2/ncFRE6* in S288c, we screened the S288c deletion library (Winzeler et al. 1999) for expression of *ncFRE6* in each of 12 gene deletions within the 35kb region. Of the deletions tested, only one, *rim101Δ*, de-repressed the transcripts **(Figure 2.11)**, strongly suggesting that the *RIM101* allele is the *trans* factor that represses *AQY2/ncFRE6* in S288c. In support of this hypothesis we note that *RIM101* is a well-characterized zinc finger transcriptional repressor and is one of the most sequence-variable transcription factors between S288c and ∑1278b**,** harboring 18 SNPs, 13 of which are non-synonymous **(Figure 2.12)**.

To confirm that the polymorphic *RIM101* allele controls expression of *AQY2/ncFRE6*, we interconverted the entire *RIM101* open reading frame (S288c: ChrVIII 51111-52988, Σ1278b: ChrVIII 49766-51655) between the strains and measured expression of *AQY2/ncFRE6*. Interconverting the *RIM101* allele is sufficient to repress expression in Σ1278b and to rescue expression in S288c, confirming that the *RIM101* alleles confer distinct *trans*-acting regulatory capacity with regards to *AQY2/ncFRE6* expression **(Figure 2.11)**. We concluded that one or more of the sequence variations between the strains are responsible for the difference in *RIM101* activity.

RIM101 is known to contribute to several phenotypes, and is required for haploid invasive growth in ∑1278b (Ryan et al. 2012). S288c cannot invade agar due to a loss

of function mutation in the TF *FLO8* and therefore is insensitive to null mutations in *RIM101*. We reasoned that the differences within the S288c and ∑1278b *RIM101* allele could affect the invasive growth phenotype in ∑1278b. However, the ∑1278b strain harboring the S288c *RIM101* allele, ∑1278b(S2*RIM101*), did not lose the ability to invade agar, implying that differences between the S288c and ∑1278b *RIM101* alleles do not affect the invasive growth phenotype (**Figure 2.13**).

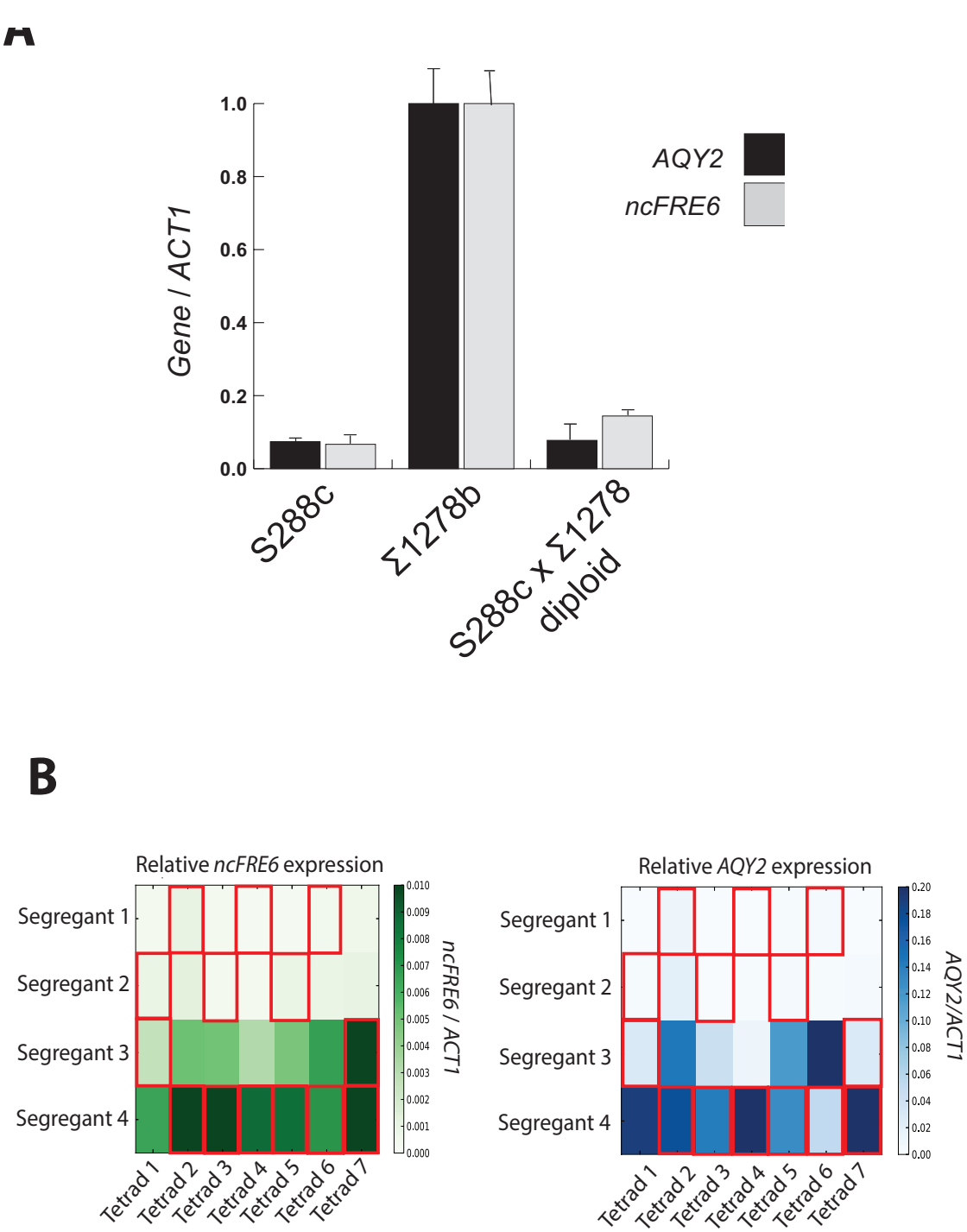


Figure 2.8: A single *trans* factor is epistatic to *cis*-linked variation with regards to expression of *AQY2* and *ncFRE6*(A) Relative expression of *AQY2* and *ncFRE6* in S288c and ∑1278b haploid strains and an S288c X ∑1278b diploid strain measured by qRT-PCR. (B) Heatmap displaying relative expression of *AQY2* (Blue) and *ncFRE6* (green) in 28 segregants of an S288c X ∑1278b heterozygous diploid as measured by qRT-PCR. Segregants were numbered according to expression level of *ncFRE6* (segregants 1 and 2 are non-expressing strains and 3 and 4 are expressing strains) for each of seven tetrads dissected.

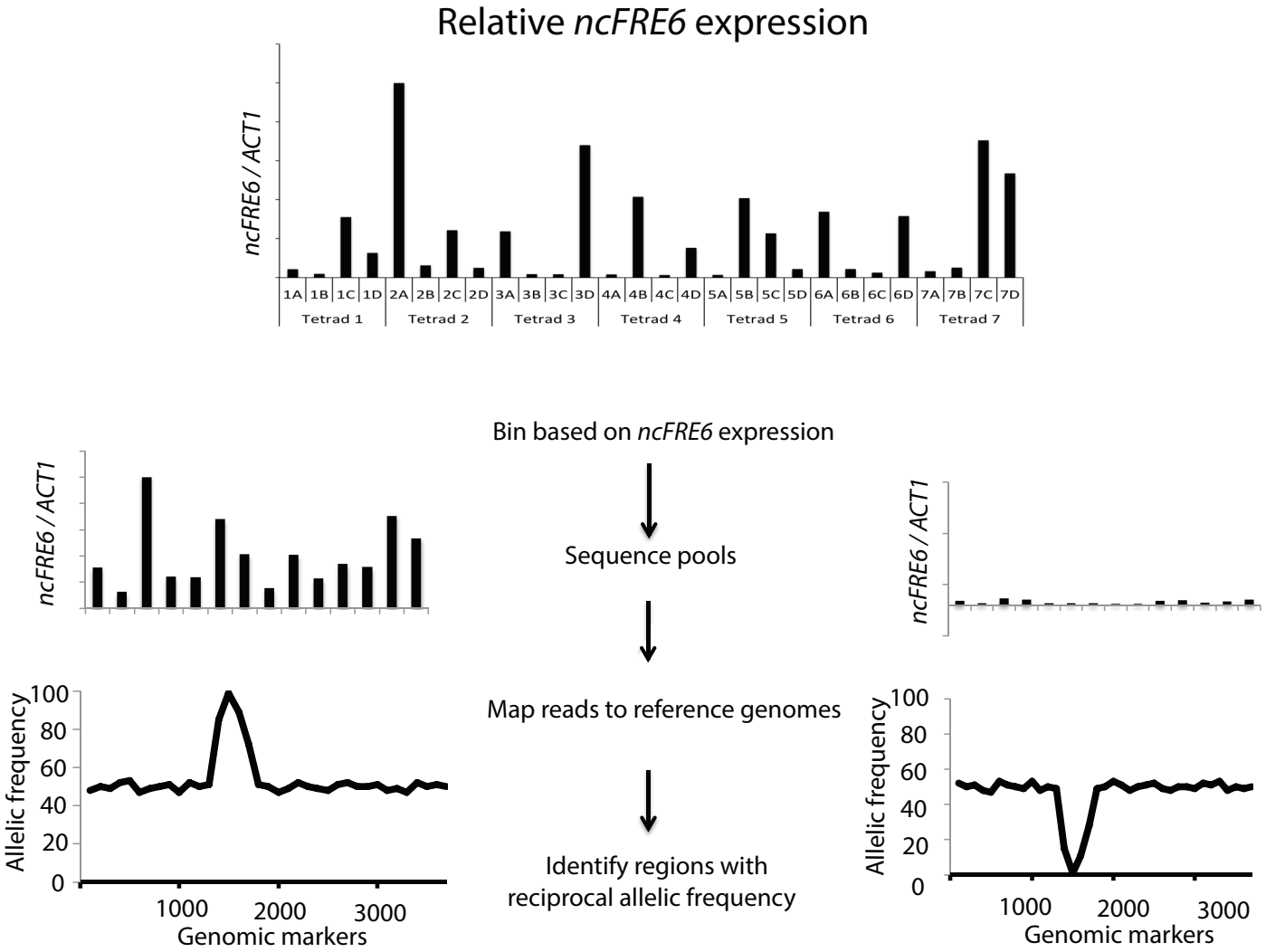


Figure 2.9: Schematic showing the workflow for expression-guided bulked segregant analysis (eBSA). Briefly, segregants were binned based on whether they express *AQY2/ncFRE6*. Genomic DNA from the expressing group was pooled separately from genomic DNA from the non-expressing pool. Pools were sequenced and reads mapped to both the S288c and ∑1278b genomes. The allelic frequency for each single nucleotide polymorphism is quantified based on read counts mapping to each genome for each pool. A region where only S288c alleles exist in one pool (i.e. expressors) and only ∑1278b alleles exist in the other (i.e. non-expressors) harbor the variant driving differential expression of the transcript being interrogated (i.e. *AQY2/ncFRE6*).

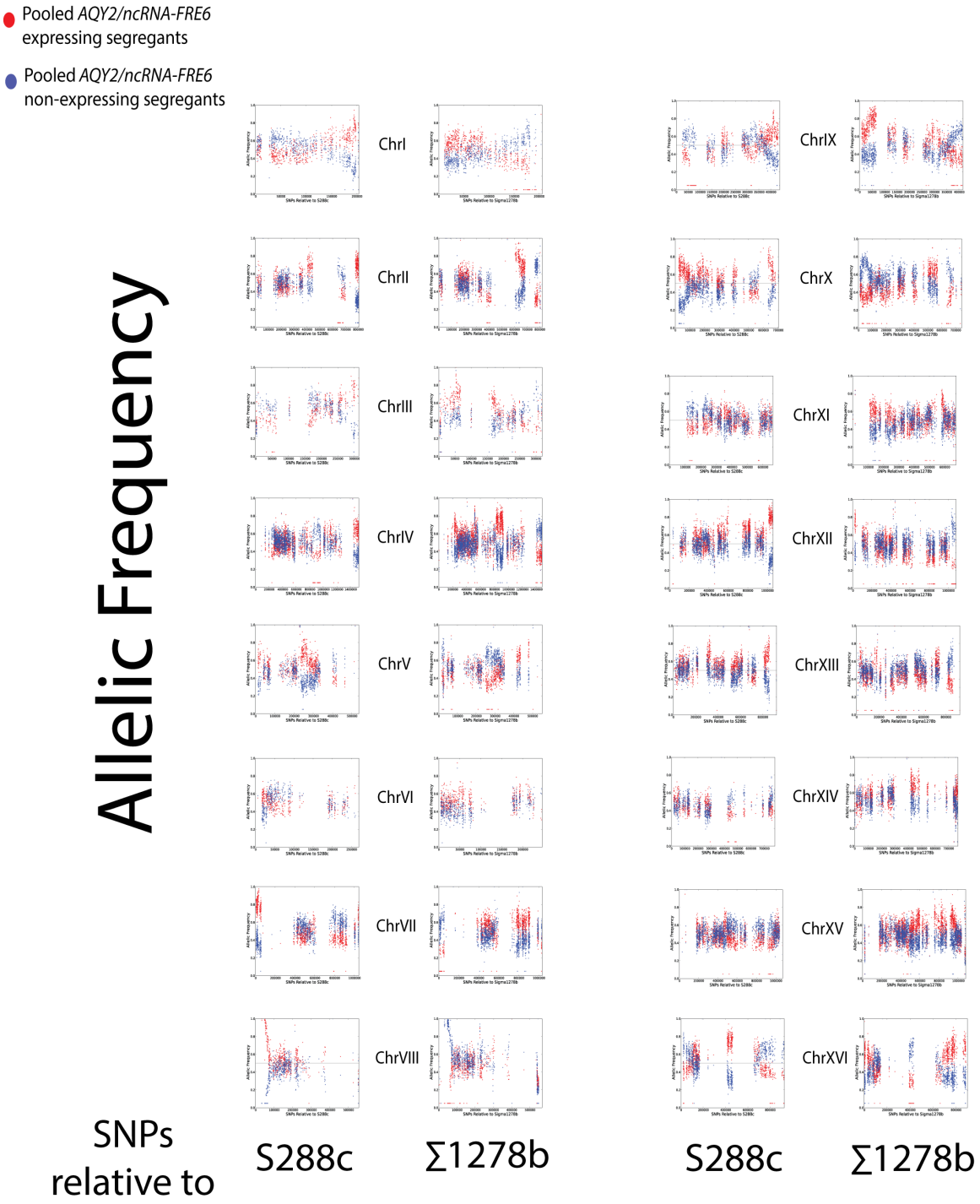


Figure 2.10: **Expression-guided bulked segregant analysis maps the *trans* factor to the left arm of chromosome VIII.** Scatter plots displaying the allelic frequencies of all SNPs between S288c and ∑1278b (dots) within pools of genomic DNA from either expressing or non-expressing segregants (Red = expressing, Blue = non-expressing). Plots are arranged by chromosome and genome mapped against. X-axis is position along the chromosome. Y-axis is allelic frequency. Red dots represent SNP frequency within the expressing pools.

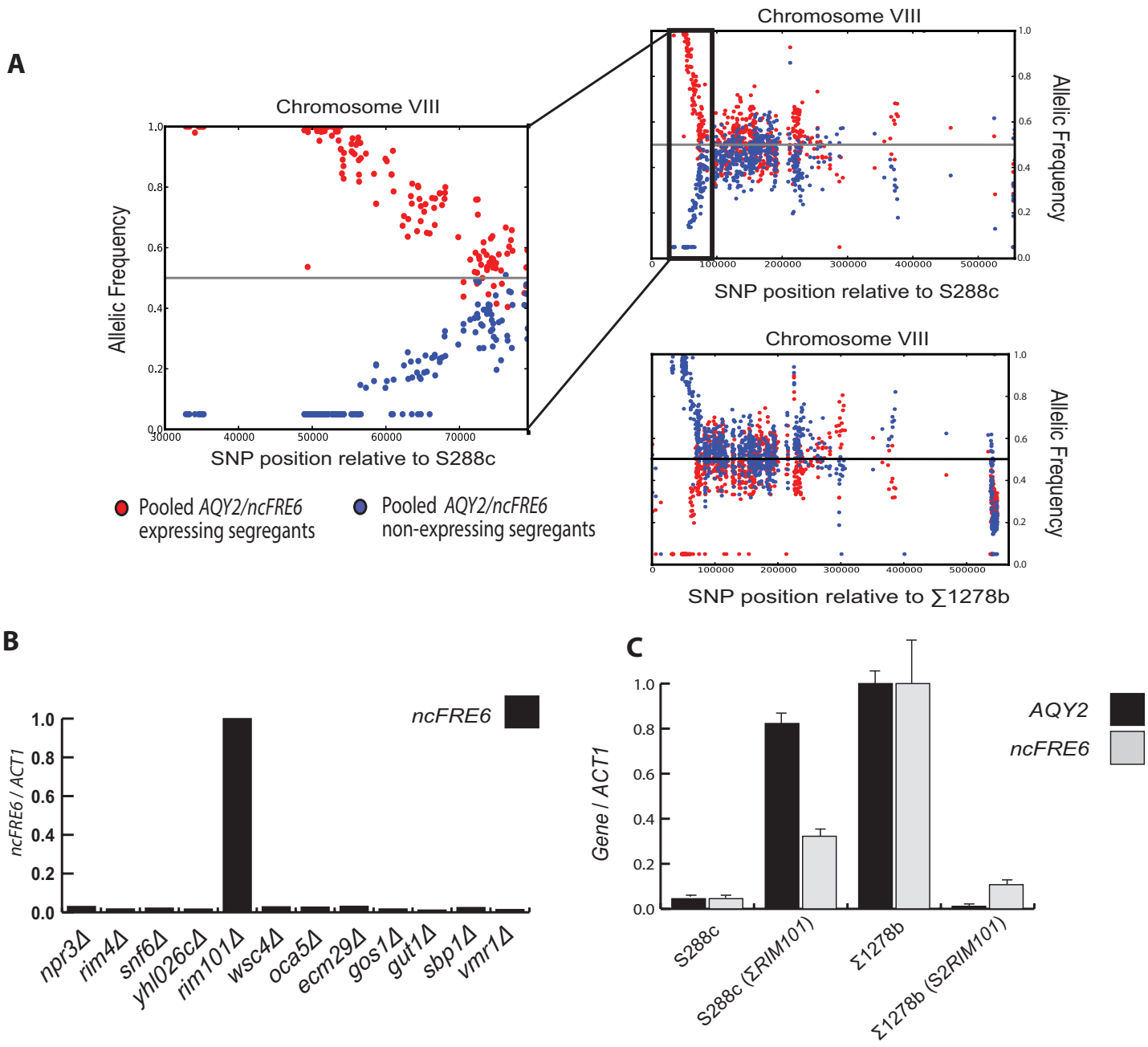


Figure 2.11: *RIM101* controls expression of both *AQY2* and *ncFRE6* in *trans*. (A) Scatter plot displaying the allelic frequency of every SNP (dots) between S288c and ∑1278b in 14 *AQY2/ ncFRE6* expressing strains (segregants 3 and 4 from Fig 2, red dots) and 14 non-expressing strains (segregants 1 and 2 from Fig 2, blue dots) across chromosome VIII (all chromosomes displayed in Fig S3). Strains were pooled based on expression, sequenced, and mapped to each reference genome. X-axis is position along chromosome eight. Y-axis is the allelic frequency of each SNP relative to the

genome being mapped to for each pool. Zoomed region represents the distal left arm of chromosome VIII where all SNPs segregate with either the expressing or non-expressing strains. (B) Relative expression of *ncFRE6* in gene deletions within the 35kb region identified in Fig 3A. (C) Relative expression of *AQY2* and *ncFRE6* in wildtype and *RIM101*-interconverted strains.

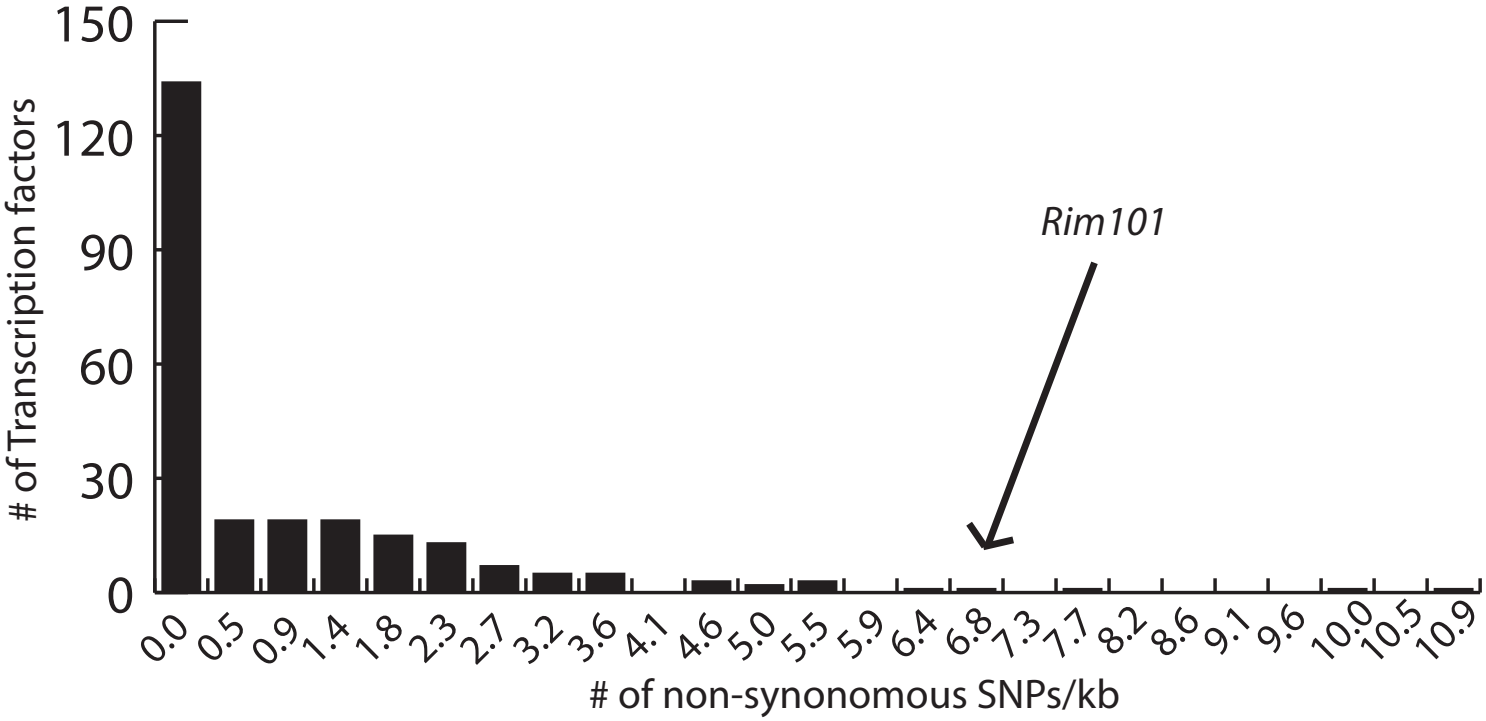


Figure 2.12: ***Rim101*** **is one of the most sequence-variable transcription factors between S288c and ∑1278b.** Histogram displaying the number of single nucleotide polymorphisms in 249 DNA-binding proteins. X-axis represents number of non-synonomous SNPs/kb. Y-axis is number of transcription factors.

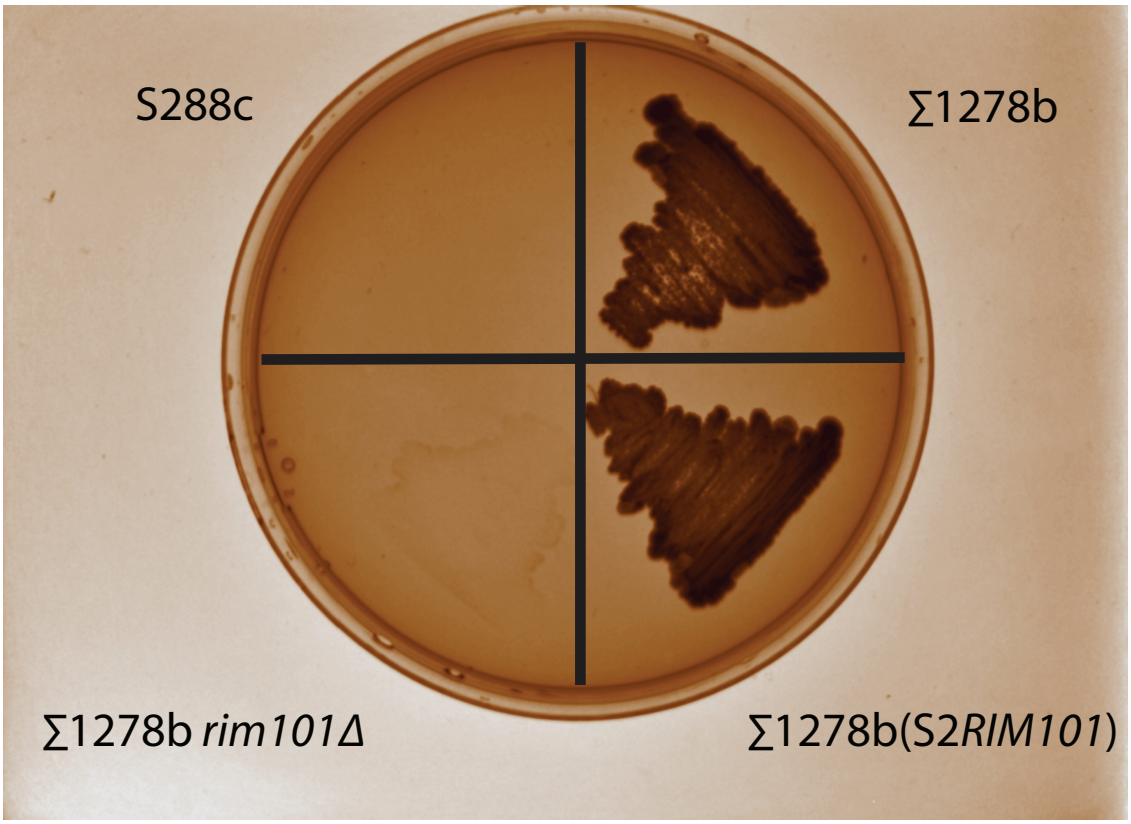


Figure 2.13: The S288c *RIM101* allele complements the ∑1278b *RIM101* allele in ∑1278b with regard to invasive growth. Strains were patched to YPD for two days and washed with gently running water before imaging.

2.3 Most differential expression between S288c and Σ1278b is *RIM101*-linked

To assess the impact of *RIM101* on genome-wide expression, we performed RNA-seq on *RIM101* deletion strains in both backgrounds. Consistent with *RIM101*'s role as a transcriptional repressor, the majority of the genes whose expression level changes upon deletion of *RIM101* in S288c became de-repressed (771 upregulated, 301 downregulated) **(Figure 2.14)**. Surprisingly, the effect of deleting *RIM101* in S288c was much larger than in Σ1278b **(Figure 2.14)**. While 1072 genes change expression levels in S288c *rim101Δ* relative to S288c wildtype, only 145 change in ∑1278b *rim101Δ* relative to ∑1278b wildtype. Furthermore, the ratio of de-repressed to repressed genes is opposite in the Σ1278b *RIM101* deletion (45 upregulated, 100 downregulated). This result suggests that *RIM101* is a stronger repressor in S288c than in ∑1278b. Consistent with a loss of repressive capacity in ∑1278b relative to S288c, we note that *AQY2/ncFRE6* levels do not change in the ∑1278b *rim101Δ* strain. Nevertheless, 145 genes do change expression in ∑1278b *rim101Δ*, implying that the disparate response to deletion of *RIM101* is not due to a complete loss of function of the ∑1278b *RIM101* allele.

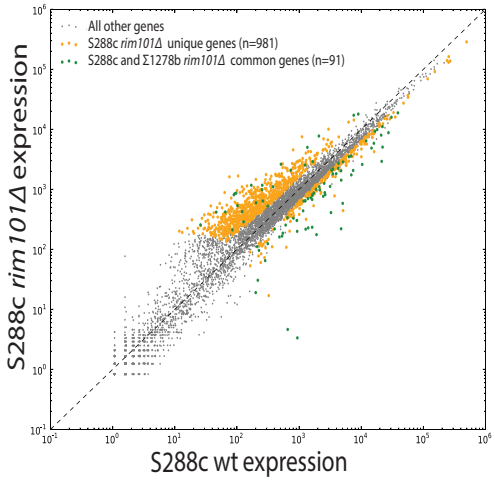
We next sought to determine the extent to which genome-wide differential expression between S288c and ∑1278b can be attributed to *RIM101*. We reasoned that genes that are differentially expressed between the wildtype strains but not the *RIM101* deletion strains are *RIM101*-dependent because removal of *RIM101* from the system eliminates the observed interstrain differential expression. Hence, these differences in

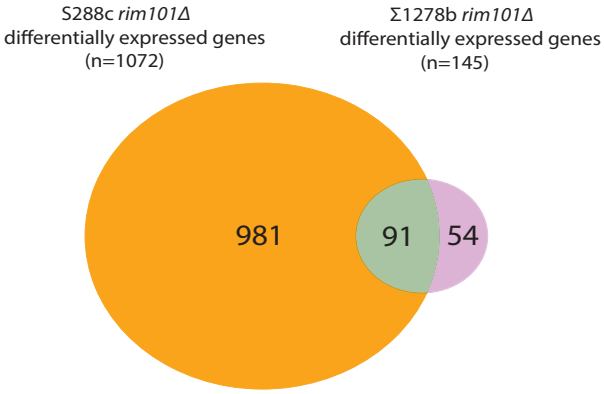
expression level between S288c and ∑1278b can be attributed to differences in *RIM101*-mediated regulation. Surprisingly, of 1207 differentially expressed genes between S288c and Σ1278b, over two-thirds (822) are in some way dependent on the presence of *RIM101* **(Figure 2.15 Red)**.

We next asked how expression of the 822 *RIM101*-dependent transcripts (as defined in Figure 2.15) changes upon loss of the *RIM101* allele in each strain background. Deleting *RIM101* in S288c resulted in a shift in *RIM101*-dependent gene expression toward ∑1278b wildtype expression levels **(Figure 2.16)**. However, deletion of *RIM101* in ∑1278b did not result in a shift toward S288c wildtype levels **(Figure 2.16)**. This asymmetric response to *RIM101* deletion is consistent with *RIM101* possessing augmented repressive capacity in S288c relative to ∑1278b.

Rim101 is regulated by several post-translational modifications that could be impacted by genetic background. For example, proteolytic cleavage of 70 C-terminal amino acids has been shown to activate the TF (Weishi & Mitchell 1997). We tested whether *RIM101* alleles were differentially cleaved in a background dependent manner. We N-terminally tagged each Rim101 allele with an HA tag and assessed cleavage by Western blot. Preliminary results indicate that genetic background does affect cleavage of the *Rim101* C-terminus (**Figure 2.17**). Both ∑1278b *Rim101* alleles undergo a similar cleavage pattern when in the S288c or ∑1278b backgrounds. However, the S288c allele shows a clear difference in cleavage when expressed in the ∑1278b background, rather than its native S288c background. It remains unclear whether this presumed difference

in cleavage pattern results in altered activity of the *Rim101* protein when placed in alternative genetic backgrounds.





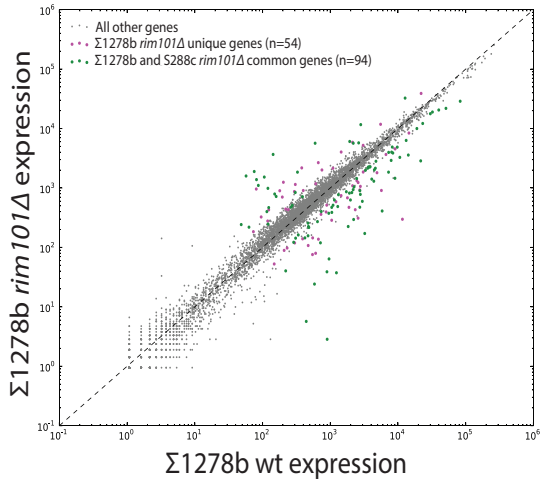
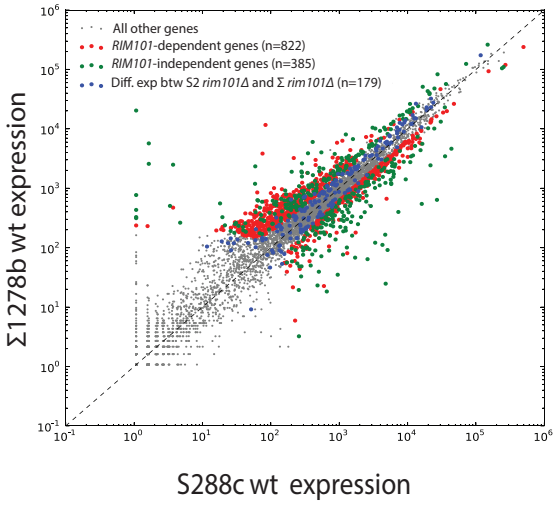
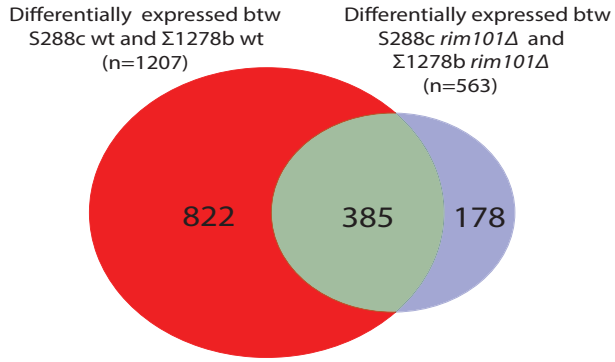


Figure 2.14 Deletion of *RIM101* affects the genome-wide expression pattern of S288c to a much greater extent than ∑1278b. (A) Scatter plots displaying expression levels of each gene (dots) in *rim101Δ* strains relative to wildtype. Venn diagram represents the number of genes differentially expressed in S288c *rim101Δ* relative to S288c wt (orange) or between ∑1278b *rim101Δ* and ∑1278b wt (magenta). The overlap (green) represents genes differentially expressed in both S288c *rim101Δ* and ∑1278b *rim101Δ* strains relative to respective wildtype strains. Dots on scatter plots are colored according to the Venn diagram.

A



B



C

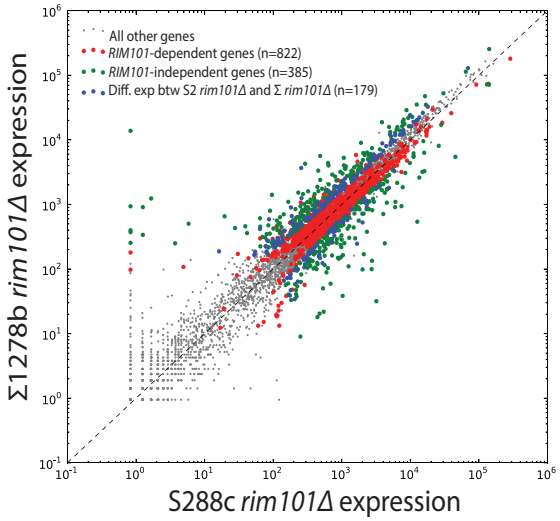


Figure 2.15: Most differential expression between S288c and Σ1278b is *RIM101*-linked Scatter plots displaying expression levels of each gene (dots) between *RIM101* wildtype S288c and ∑1278b strains (A) or between S288c *rim101Δ* and ∑1278b *rim101Δ* strains (C). Venn diagram (B) represents the number of genes differentially expressed between S288c and ∑1278b wt strains (Red) or between S288c *rim101Δ* and ∑1278b *rim101Δ* strains (blue). The overlap (green) represents genes differentially expressed in both comparisons. Dots on scatter plots are colored according to the Venn diagram. RNA-seq performed on 2 biological replicates for each strain. Differential expression called by DE-seq with Padj = 0.0005.

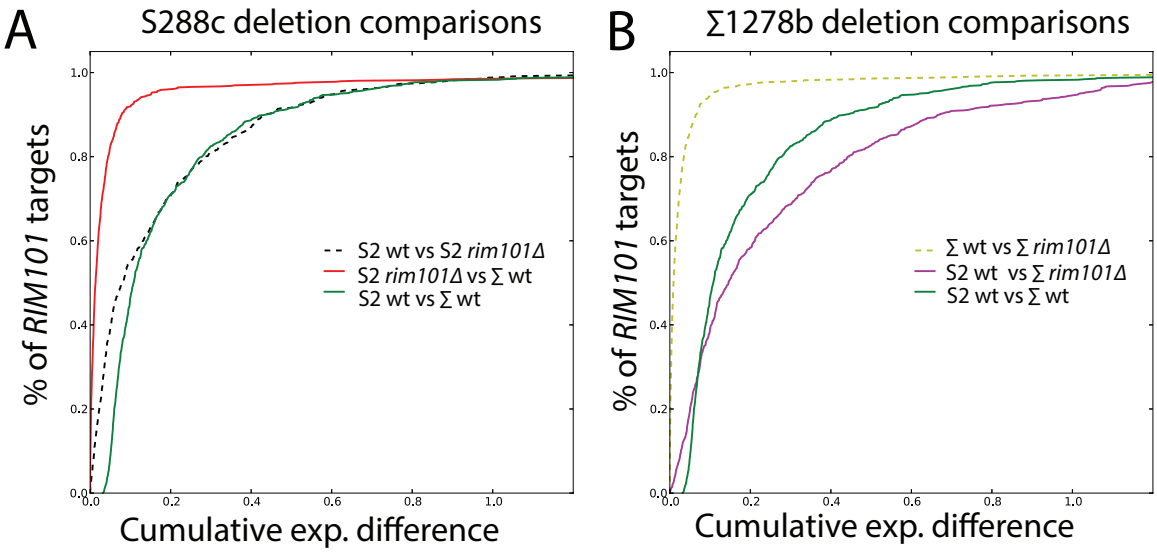


Figure 2.16: Deletion of *RIM101* results in an asymmetric transcriptional response between S288c and ∑1278b deletion and wildtype strains. CDF plot examining the impact of deleting *RIM101* on 822 *RIM101* targets in (A) S288c or (B) ∑1278b. Y-axis represents percentage of *RIM101* targets. X-axis represents cumulative differential expression for each comparison. Comparisons specified using S2 (S288c) and ∑ (∑1278b) as abbreviations.

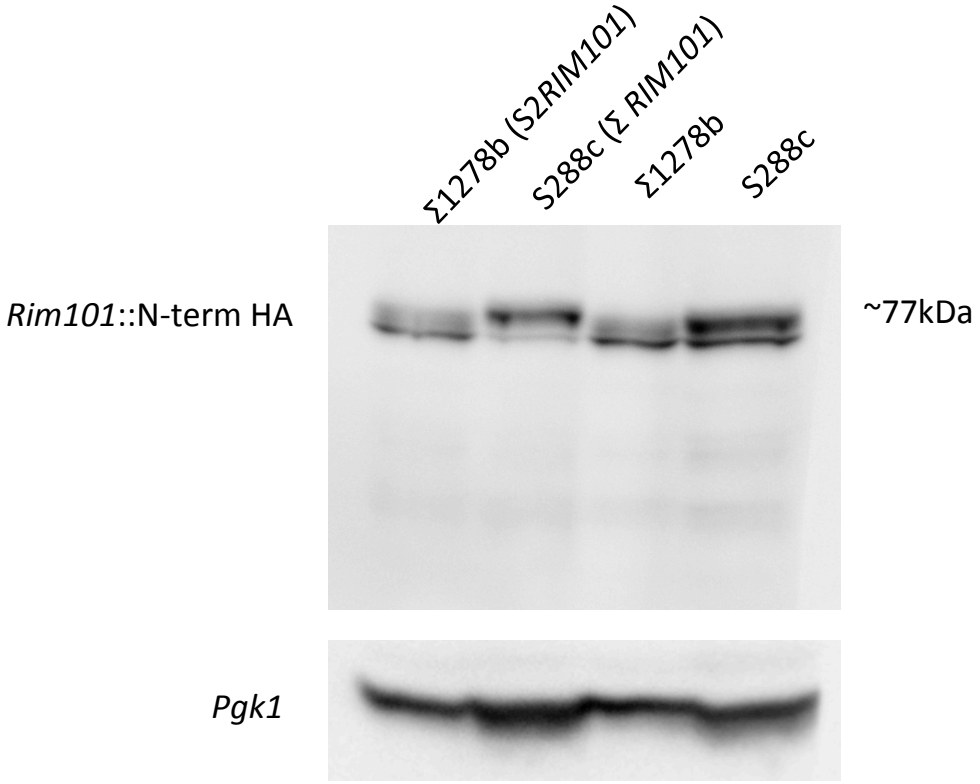


Figure 2.17: *Rim101* cleavage pattern is allele- and background- dependent. Western blot probing for *Rim101::HA* in S288c, ∑1278b, and *RIM101*-interconverted strains. *Pgk1* serves as a loading control.

2.4 The *RIM101* allele achieves remarkable specificity, but genetic background controls its regulatory capacity

We sought to distinguish whether the *RIM101* allele itself, or the *RIM101* pathway, imparts additional repressive capacity in S288c by swapping *RIM101* alleles between S288c and Σ1278b and assaying genome-wide expression by RNA-seq. Surprisingly, upon introduction of the non-native allele only three transcripts undergo statistically significant changes in expression in both backgrounds (Padj≤0.05) **(Figure 2.18)**. *AQY2, ncFRE6,* and *TIP1*—a cell surface mannoprotein—significantly change expression levels in response to incorporation of the non-native allele in both backgrounds. Such a focused, allele-dependent transcriptional response stands in stark contrast to other *trans*-regulators discovered in eQTL studies, which tend to affect expression of large numbers of genes. It remains unclear how only *AQY2*, *ncFRE6*, and *TIP1* are so dramatically influenced by interconversion of the *RIM101* alleles between backgrounds. Moreover, the allele-dependent expression level of *TIP1* is surprising given that the *TIP1* allele and promoter region are invariant between S288c and ∑1278b, and especially because no change in *AQY2/ncFRE6* expression was observed in the ∑1278b *RIM101* deletion strain, nor was *TIP1* expression changed in the S288c *RIM101* deletion strain. This high degree of allele-specificity suggests that unique combinations of factors can collaborate to effect specific sets of genes.

How *AQY2/ncFRE6* and *TIP1* are affected by interconversion of the *RIM101* allele, while hundreds of other *RIM101* targets remain largely unaffected, remains unclear. To better understand the mechanisms conferring such specificity we performed ChIP-qPCR for both alleles in either the S288c or ∑1278b backgrounds (**Figure 2.19**). Tiling of the *AQY2/ncFRE6* or *TIP1* promoters with qPCR primers revealed that *Rim101* binds to the *AQY2/ncFRE6* promoter region independently of genetic background or *RIM101* allele. This result suggests that the mechanism that allows *AQY2/ncFRE6* expression to be influenced by the *RIM101* allele does not impact *Rim101* binding. However, subtle differences were observed in the *TIP1* promoter. Whereas no binding of *Rim101* was observed in wildtype S288c, the ∑1278b allele appears to occupy the *TIP1* promoter when placed in the S288c background. Furthermore, the opposite trend is observed in the ∑1278b background, where the ∑1278b *Rim101* protein shows more occupancy than the S288c allele. Given the *TIP1* expression pattern, this result is consistent with the S288c allele being a stronger repressor than the ∑1278b allele. Furthermore, the mechanisms by which *RIM101* achieves target specificity at *TIP1* may involve binding directly to the promoter, as opposed to the *AQY2/ncFRE6* promoter, which appears to be affected independently of *Rim101* binding.

Although only three transcripts become significantly differentially expressed in both S288c and ∑1278b *RIM101* interconverted strains relative to their wildtype expression levels, many *RIM101*-dependent genes appear to be more highly expressed in S288c(∑*RIM101*) than in S288c wildtype, consistent with the S288c *RIM101* allele being a stronger repressor than the ∑1278b allele **(Figure 2.20)**. In fact, in

S288c(∑*RIM101*), expression levels of the 822 *RIM101*-dependent genes shift toward a pattern more similar to ∑1278b **(Figure 2.20)**, partially phenocopying the expression shift observed in S288c *rim101Δ* **(Figure 2.20).** However, incorporation of the strong S288c allele into ∑1278b does not result in a shift towards stronger repression of the same subset of genes **(Figure 2.21)** This asymmetry suggests that other background factors, and not solely the *RIM101* allele, are responsible for the gain of widespread *RIM101*-mediated repression in S288c, and that repression of *AQY2, ncFRE6* and *TIP1* are independent of the background effects. These results imply that the same transcription factor can display drastically altered activity depending on the background that it is present within, and that certain backgrounds, such as ∑1278b, buffer against widespread transcriptional dysregulation upon introduction of a new *RIM101* variant.

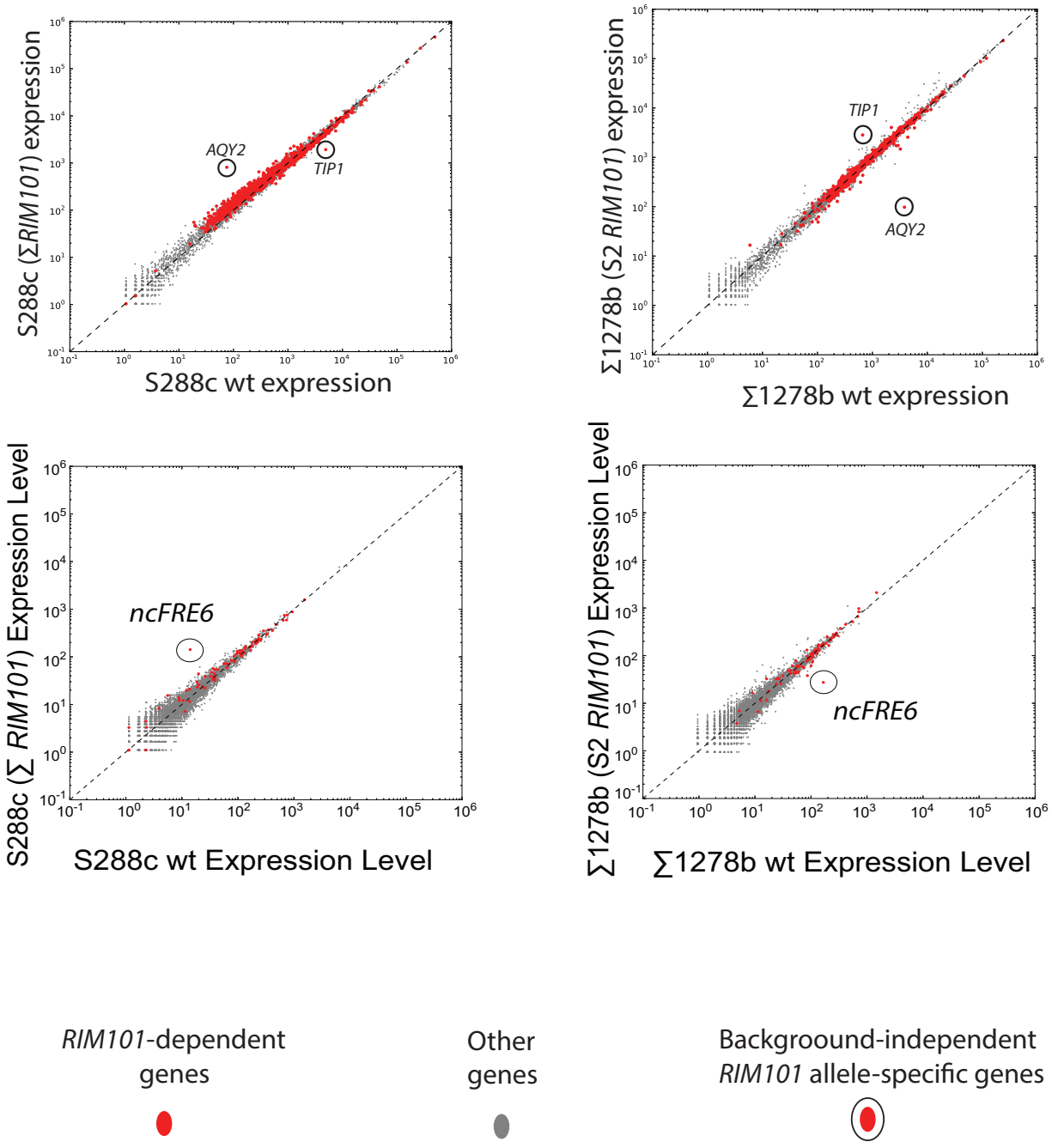


Figure 2.18: The *RIM101* allele achieves remarkable target specificity. (A) Scatter plots displaying expression levels for each gene (dots) in *RIM101* interconverted strains relative to their respective wildtype strains (Red = *RIM101*-dependent genes). Top = protein coding genes. Bottom = Antisense transcripts.

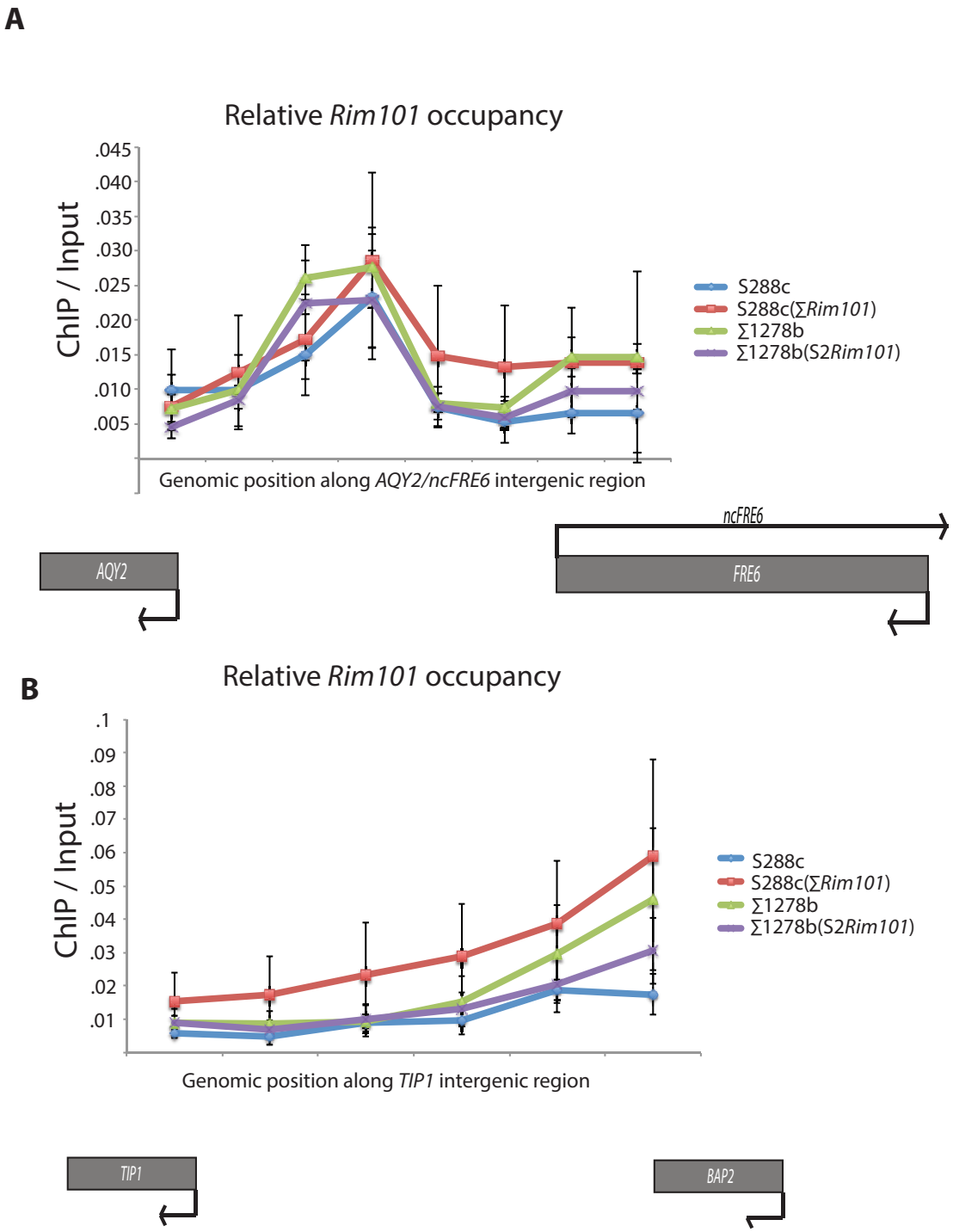
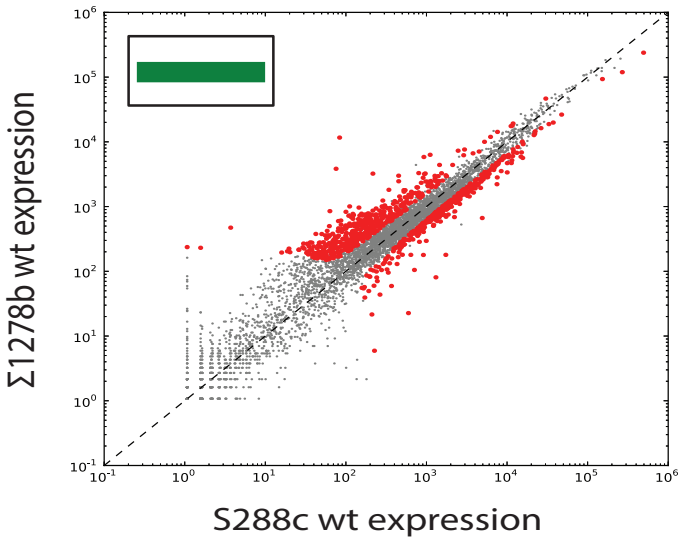
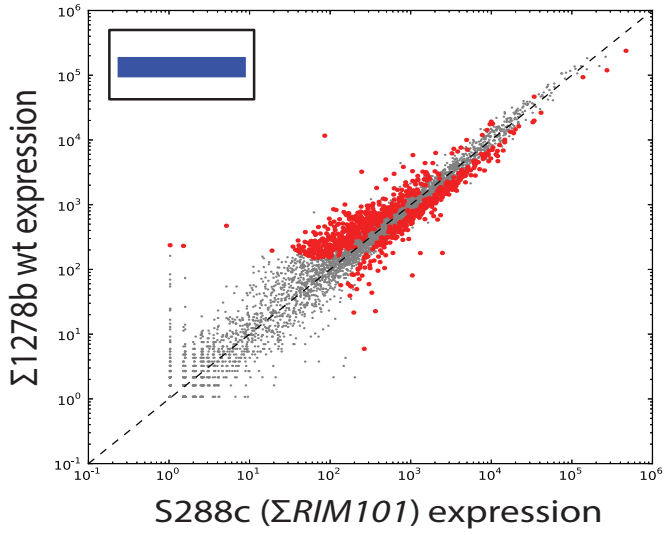


Figure 2.19: *Rim101* binding appears to be influenced by *RIM101* allele at *TIP1*, but not at *AQY2/ncFRE6*. (A) ChIP-qPCR displaying occupancy of *Rim101* within the intergenic region between *AQY2* and *ncFRE6* for wildtype and *RIM101* interconverted strains. (B) ChIP-qPCR displaying occupancy of *Rim101* within the intergenic region between *TIP1* and *BAP2* for wildtype and *RIM101* interconverted strains. Data displayed as ChIP relative to input DNA.

A



B



C

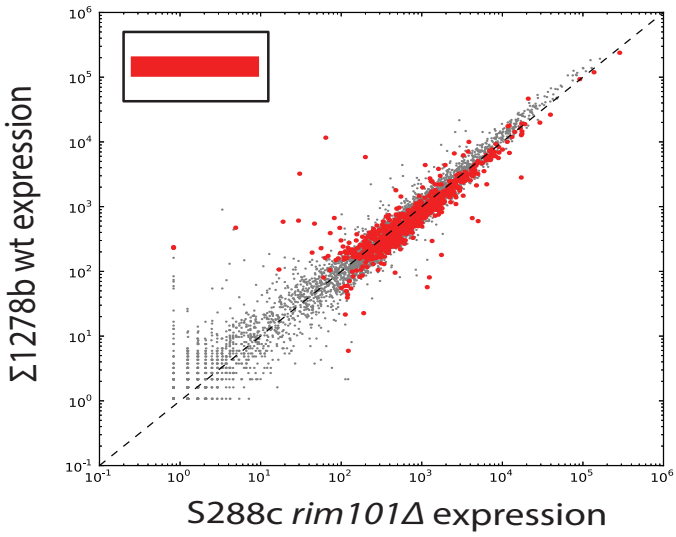
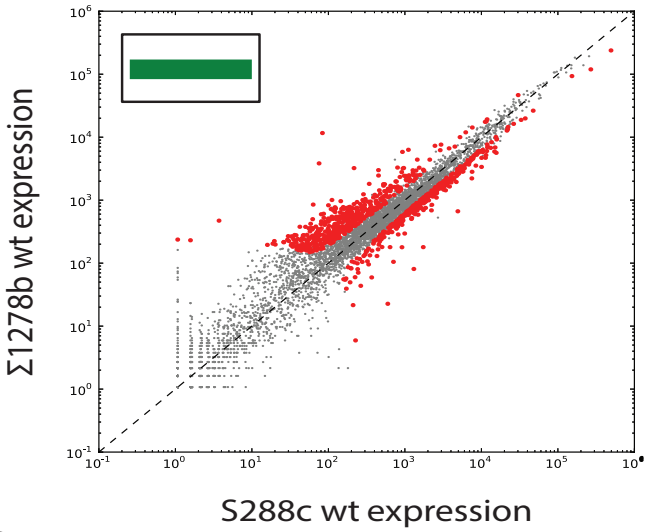
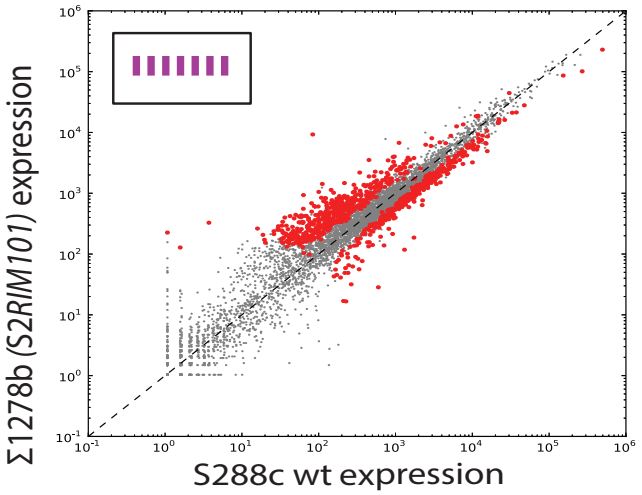


Figure 2.20: Introduction of the Σ1278b *RIM101* allele into S288c results in a large-scale shift in expression pattern that partially phenocopies the expression pattern observed in S288c *rim101Δ*. (A) Scatter plot displaying levels of each gene in Σ1278b relative to S288c wildtype strains (Red = *RIM101*-dependent genes). (B) Scatter plot displaying expression levels in the *RIM101*-interconverted strain, S288c(Σ1278b *RIM101*) relative to Σ1278b wildtype (Red = *RIM101*-dependent genes). (C) Scatter plot displaying expression levels in S288c *rim101Δ* relative to Σ1278b wildtype (Red = *RIM101*-dependent genes).

A



B



C

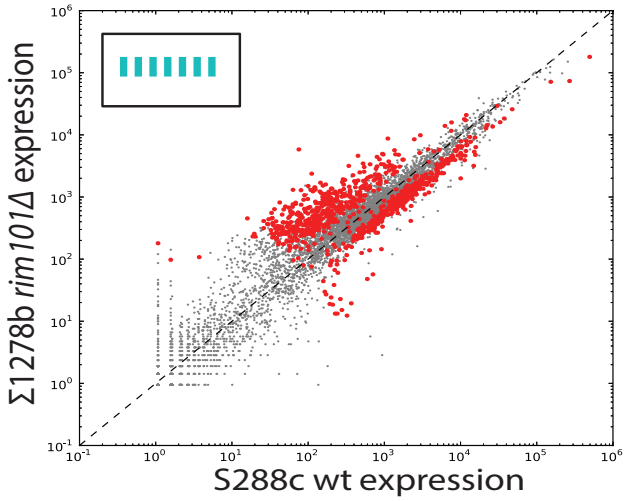


Figure 2.21: Introduction of the S288c *RIM101* allele into ∑1278b does not result in a shift in expression profile as it did in the S288c. (A) Scatter plot displaying levels of each gene in Σ1278b relative to S288c wildtype strains (Red = *RIM101*-dependent genes). (B) Scatter plot displaying expression levels in Σ1278b(S288c *RIM101*) relative to S288c wildtype (Red = *RIM101*-dependent genes). (C) Scatter plot displaying expression levels in Σ1278b *rim101Δ* relative to S288c wildtype (Red = *RIM101*-dependent genes).

A

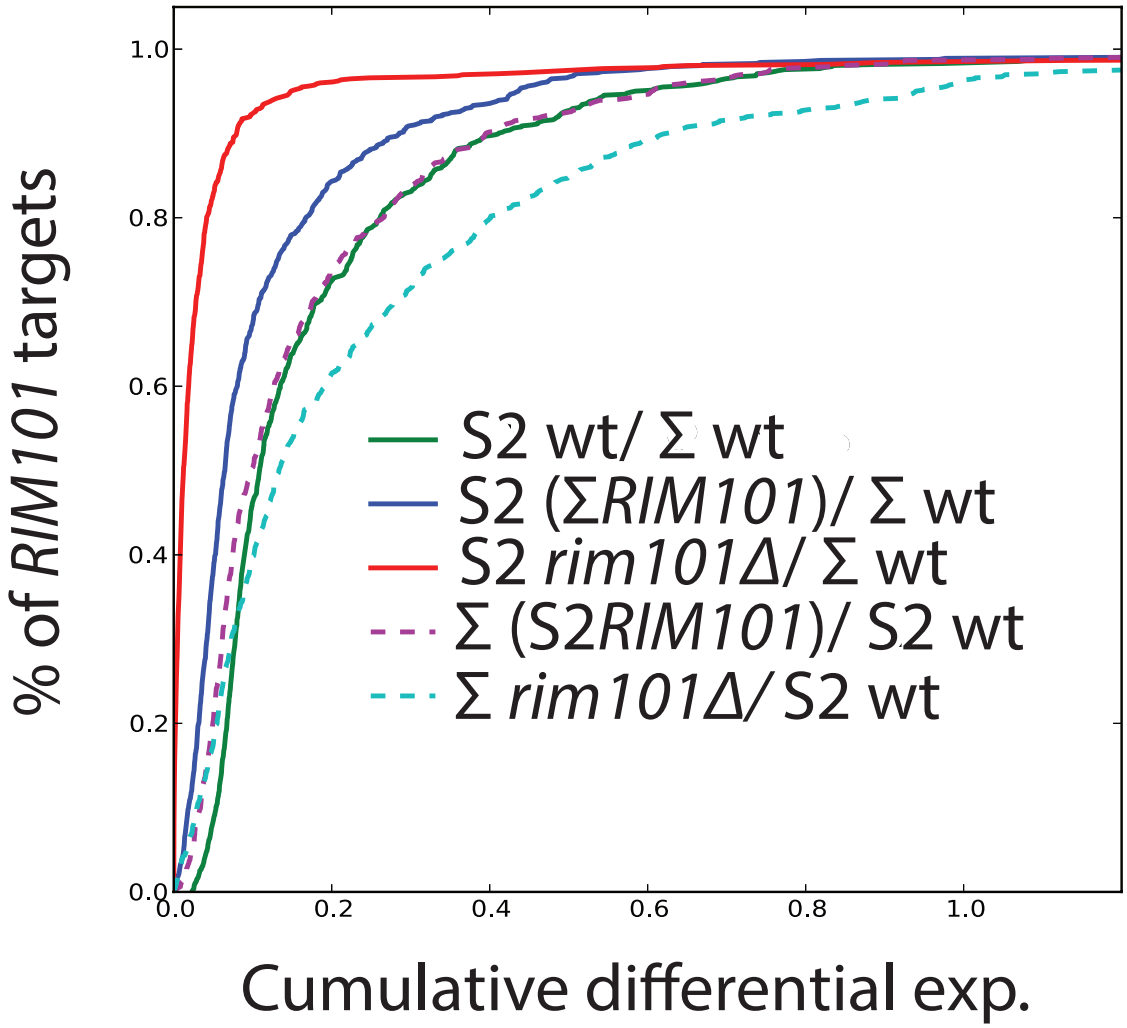


Figure 2.22: Genome-wide expression profiles are both *RIM101* and background-dependent. (A) Cumulative distribution function (CDF) plot showing the results of linear regression analysis of the distance of the 822 *RIM101*-dependent genes from a line of best fit for each strain comparison (Colored lines correspond to comparisons defined in legend*)*.

2.5 A single nucleotide polymorphism within *RIM101* is necessary and sufficient for expression of both *AQY2* and *ncFRE6*

Given that so few transcripts significantly change expression levels upon interconversion of *RIM101* alleles, we sought to better understand the molecular basis of such specificity. In order to infer the genomic feature or features within *RIM101* that contribute to repression of *AQY2* and *ncFRE6* in S288c, we screened two additional strains of *S.cerevisiae*, each with unique combinations of sequence variation within *RIM101*, for expression of *AQY2* and *ncFRE6* **(Figure 2.24)**. The *RIM101* DNA sequence includes 18 SNPs between S288c and Σ1278b, 13 of which alter the amino acid sequence of the *Rim101* protein **(Figure 2.23)**. In addition to the 13 non-synonymous SNPs, a poly-glutamine repeat stretch is expanded from four amino acids in S288c to eight in Σ1278b. We selected RM11-1a (Brem & Clinton 2002) and JAY291 (Argueso et al. 2009) for screening because they had distinct combinations of the sequence variations seen in S288c and Σ1278b. RM11-1a exhibits the *AQY2/ncFRE6*-repressed phenotype, suggesting that this strain harbors a *RIM101* allele capable of repressing the transcripts in a similar manner to S288c. However, the other strain, JAY291, expresses both transcripts, similar to Σ1278b. These results indicate that the two transcripts are expressed or repressed concurrently, implying the mechanism by which the transcripts are co-regulated is conserved across diverse *S.cerevisiae* strains.

We reasoned that the *RIM101* sequence necessary for repression must exist in both S288c and RM11-1a, but not in Σ1278b or JAY291. Alignment of the amino acid

sequences of *Rim101* across the strains revealed four non-synonymous SNPs and a truncated poly-glutamine stretch that exist solely in the repressive strains **(Figure 2.24)**. We sought to identify the one or more of these sequence variations between S288c and Σ1278b that control expression of *AQY2/ncFRE6*. Because variable length poly-glutamine tracks have been associated with altered protein structure and function, including altered protein-protein interactions (Schaefer et al. 2012), we first tested whether the altered poly-glutamine repeat length affected *RIM101*-mediated repression. After expanding the poly-glutamine tract in S288c and truncating it in Σ1278b, we tested for expression of *AQY2/ncFRE6* and detected no deviation from either wildtype strain, suggesting that the length of the poly-glutamine tract does not, by itself, affect *RIM101* activity at this locus **(Figure 2.25)**. Next we tested whether the four conserved amino acids were sufficient to affect *Rim101*-mediated repression of *AQY2/ncFRE6*. Indeed the collection of all four mutations is sufficient to rescue expression in S288c **(Figure 2.25)**. Replacing each amino acid individually revealed one critical amino acid residue with regards to regulation of *AQY2/ncFRE6*. In S288c, W249L is sufficient to de-repress *AQY2* and *ncFRE6* **(Figure 2.26)**. Furthermore, L249W is sufficient to repress *AQY2/ncFRE6* in Σ1278b. Hence, a single nucleotide polymorphism in the *RIM101* transcription factor determines whether *AQY2/ncFRE6* is expressed.

Finally, we sought to determine whether the amino acid present at position 249 is predictive of expression of *AQY2/ncFRE6* in other strains. We could predict expression of *AQY2/ncFRE6* in all five additional strains that we tested **(Figure 2.25)**. Strains with L249 all express *AQY2/ncFRE6*, and those with W249 do not. Clearly, position 249

within the *Rim101* protein is intimately linked to expression of *AQY2/ncFRE6* across a diverse array of strains. However, our ability to predict *AQY2/ncFRE6* expression was not conserved in a closely related species, *S.paradoxus* (**Figure 2.25**). Hence, the effect of the W249L *RIM101* mutation appears to be clade specific, indicating that it may be a recently evolved regulatory mechanism.

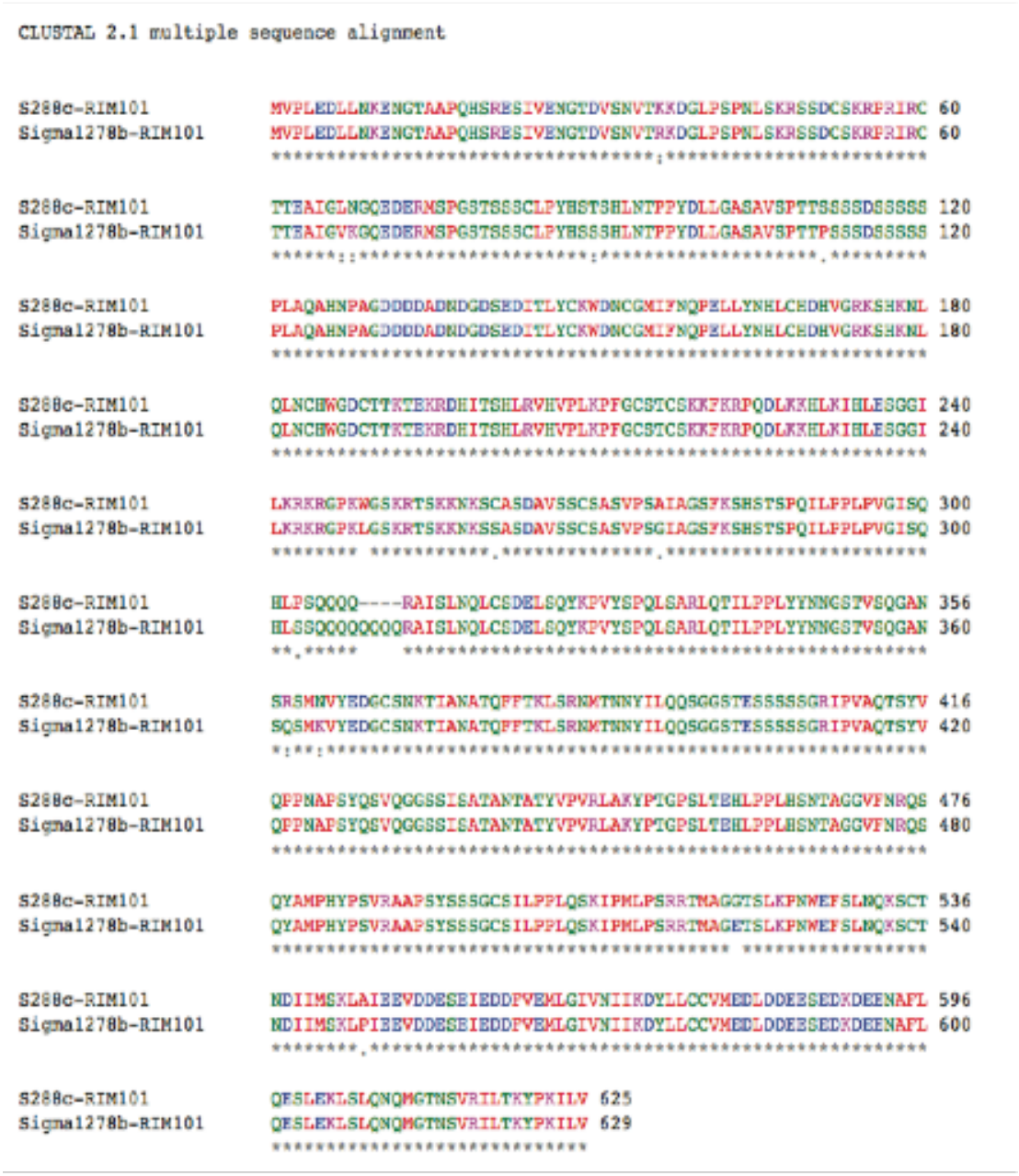


Figure 2.23: Alignment of the *Rim101* protein between S288c and ∑1278b. (A) *Rim101* protein sequence is highly polymorphic between S288c and ∑1278b. ClustalW protein alignment of *Rim101* showing 13 amino acid substitutions and a truncated poly glutamine tract in S288c relative to ∑1278b.

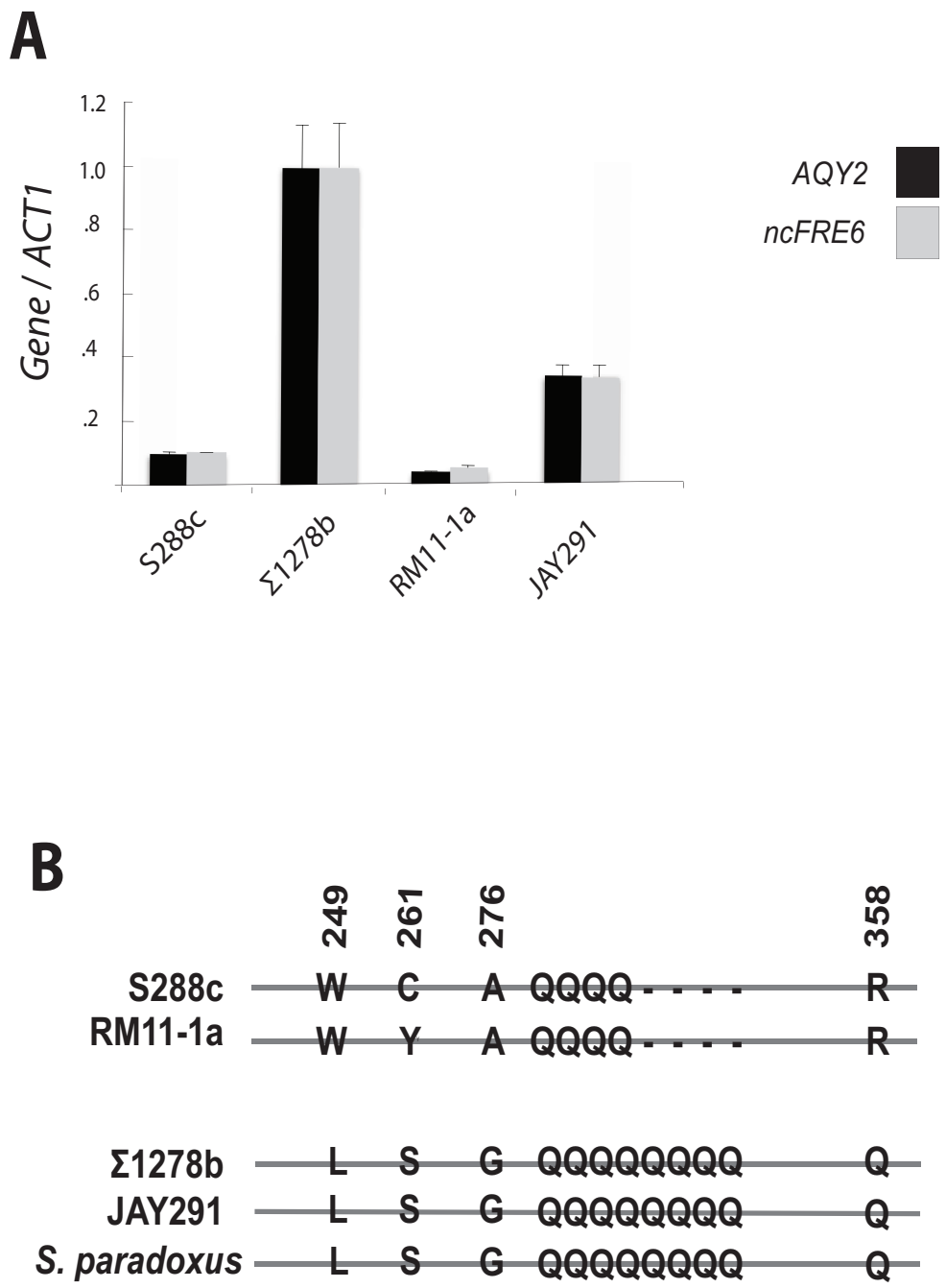


Figure 2.24: Expression of *AQY2/ncFRE6* in other *S.cerevisiae* strains could inform about amino acids necessary for expression. A) Relative expression of *AQY2* and *ncFRE6* in *S.cerevisiae* strains S288c, ∑1278b, RM11-1a, and JAY291 measured by qRT-PCR. (B) Sequence alignment of S288c, ∑1278b, RM11-1a, and JAY291 reveals features of *RIM101* conserved in *AQY2/ ncFRE6* expressing and non-expressing strains (Position numbers relative to S288c *Rim101* protein).

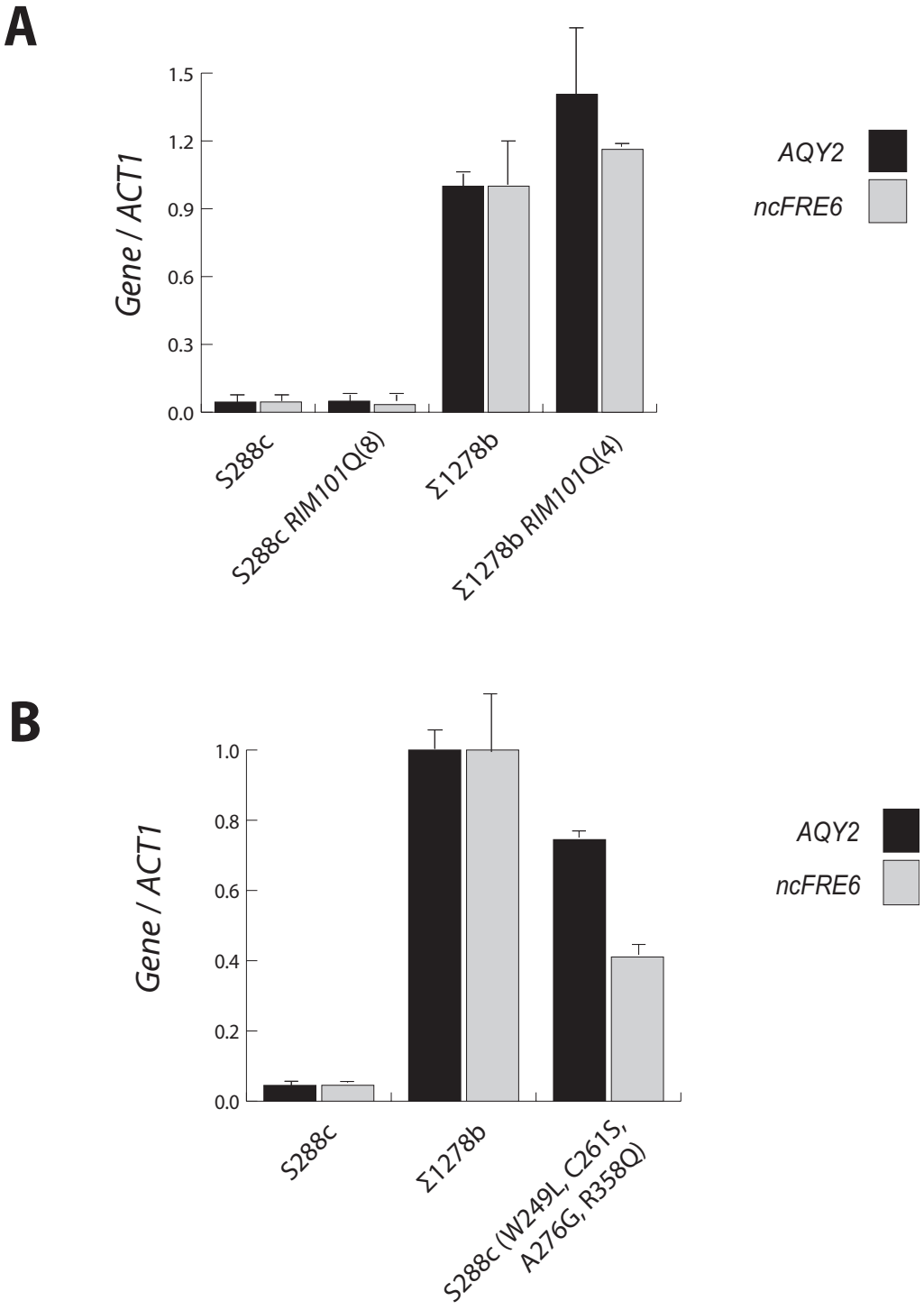


Figure 2.25: Poly-glutamine tract length does not influence *AQY2/ncFRE6* expression, but four conserved amino acids do. (A) Relative expression of *AQY2* and *ncFRE6* in a polyQ expanded S288c strain and a ∑1278b polyQ truncated strain relative to each wildtype strain measured by qRT-PCR. (B) Relative expression of *AQY2* and *ncFRE6* in an S288c strain harboring four ∑1278b SNPs measured by qRT-PCR.

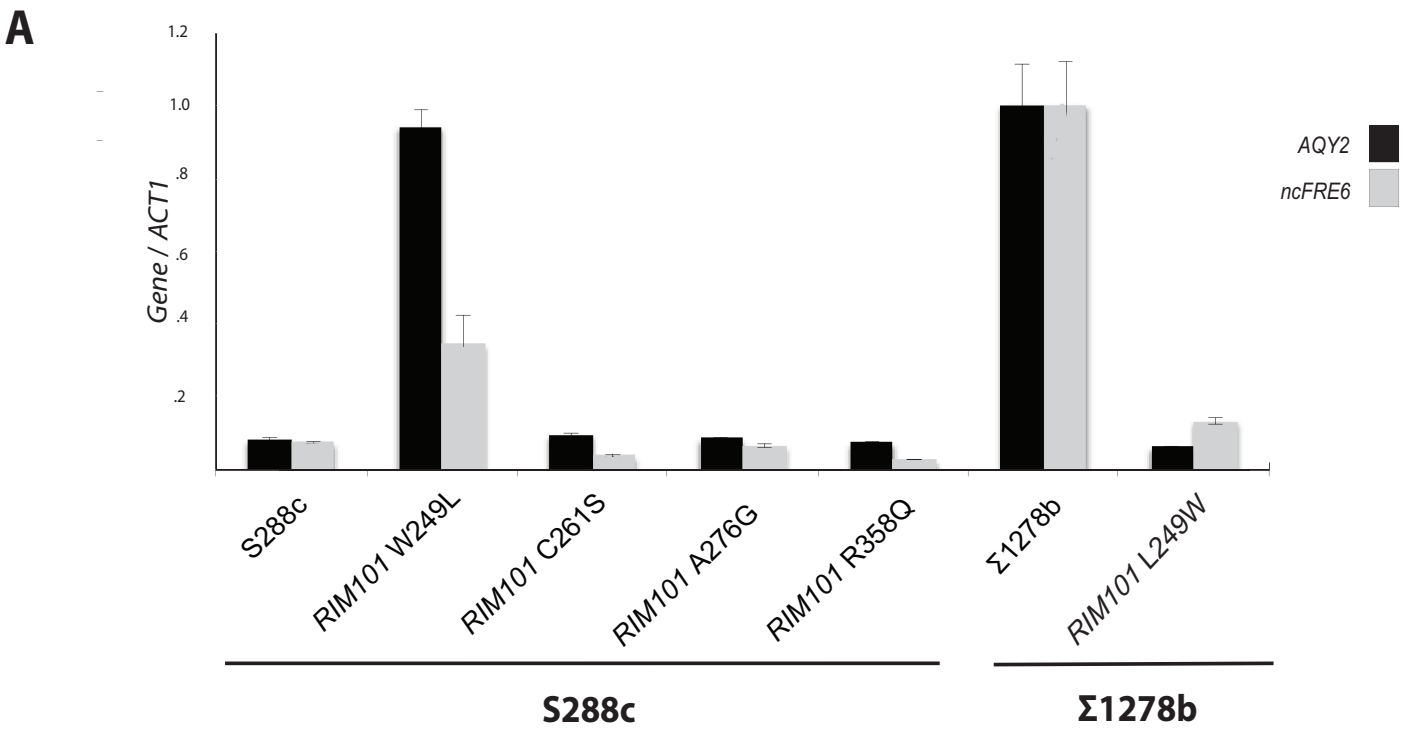
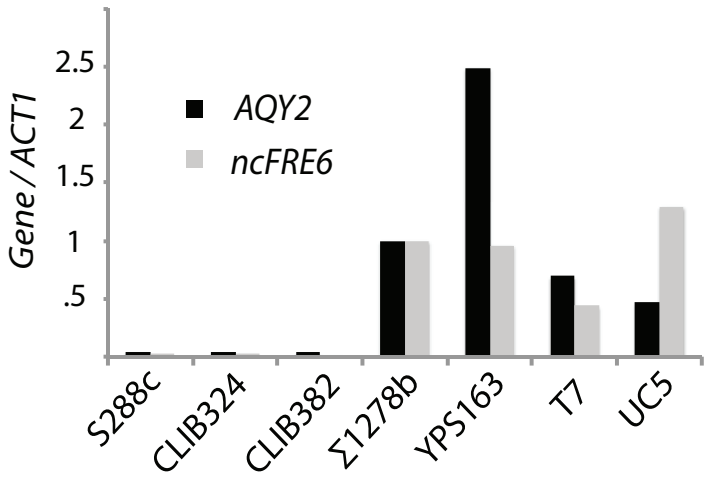
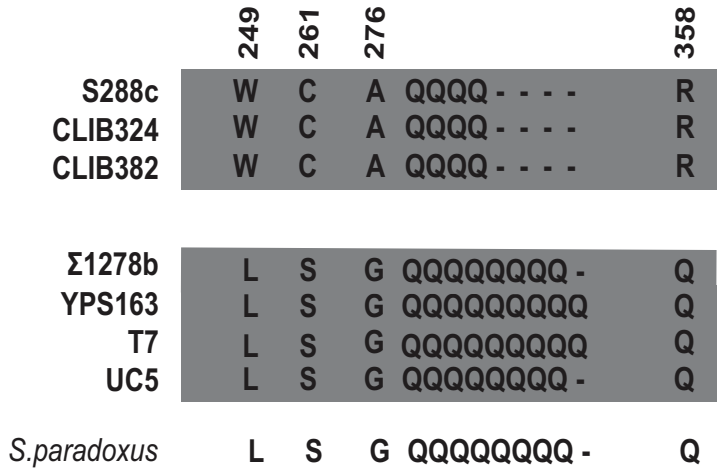


Figure 2.26: Amino acid position 249 is critical for controlling expression of *AQY2/ncFRE6*. (A) Relative expression of strains with individual *RIM101* point mutations measured by qRT-PCR.

A



B



C

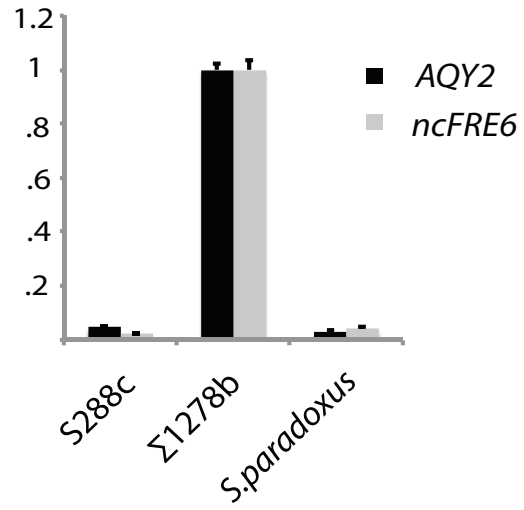


Figure 2.27: Position 249 within the *Rim101* protein determines the on/off state of *AQY2/ncFRE6* in five additional strains of *S.cerevisiae*, but not in *S.paradoxus*. (A) qRT-PCR of *AQY2* (Black) and *ncFRE6* (grey) in five additional strains of *S.cerevisiae*. (B) Alignment of the region of *Rim101* implicated in repression of *AQY2/ncFRE6*. (C) qRT-PCR of *AQY2* and *ncFRE6* in S288c, ∑1278b, and *S.paradoxus*.

Chapter III: Discussion

3.1 Discussion Summary

We discovered a *trans*-regulatory single nucleotide polymorphism within the transcription factor *RIM101* that causes strain-specific expression of a pair of co-regulated, divergently oriented transcripts, *AQY2* and *ncFRE6*. Subsequent RNA-seq analysis of *RIM101* deletion strains revealed that *RIM101* controls expression of many more targets in S288c than ∑1278b, and suggests that the majority of differential expression between the two strains is related to differences in the *RIM101* pathway. Swapping *RIM101* alleles between S288c and ∑1278b strongly affected expression of only three transcripts in both strains: *AQY2, ncFRE6,* and *TIP1*. However, consistent with results from *RIM101* deletion strains, hundreds of other *RIM101*-dependent genes underwent subtle changes in expression specifically in the S288c background, and not in ∑1278b.

3.2 Dissection of a regulatory circuit uncovers principles contributing to the complexity of gene-expression regulation

Our study highlights the complexity of transcriptional regulation, even at a single locus. For example, though *Reb1* binding is clearly regulated by a *cis* mutation between S288c and ∑1278b, and its binding pattern correlates with expression of *ncFRE6*, *Reb1* binding does not affect expression of *ncFRE6*. This result underscores the importance

of single locus studies for identifying the true sources of differential expression, rather than relying on correlations between TF binding and expression. Furthermore, our results provide a unique example of how *cis* and *trans*-linked DNA elements function in concert to affect gene expression. While either the S288c or ∑1278b *cis* context is capable of directing expression of *AQY2/ncFRE6*, differences in their promoter activities are only apparent in the absence of an epistatic *trans*-factor that we determined to be the transcription factor *RIM101*.

Although our study initially focused on transcriptional regulation at a single locus, much of the complexity governing the genome-wide regulatory capacity of *RIM101* arises from unknown background-dependent interactions that result in widespread differences in gene expression in *trans*. *RIM101* target genes undergo a widespread shift in expression pattern specifically in the S288c(∑*RIM101*) but not in ∑1278b(S2*RIM101*). While the physical mechanism underlying this asymmetric response is unknown, previous *RIM101*-based research could offer clues. In particular, *Rim101* is extensively post-translationally modified, including by phosphorylation (Nishizawa et al. 2010) and proteolytic processing (Weishi & Mitchell 1997). It is not known whether a background-specific, allele-dependent *RIM101* interaction influences either of these modifications. Also, W249L resides in close proximity to the C2H2 zinc finger DNA-binding domain of *Rim101*, raising the possibility that variation at this position could impact DNA binding in S288c, but not ∑1278b, perhaps endowing *Rim101* with altered regulatory capacity in certain genetic contexts.

How do alternate *RIM101* alleles achieve such remarkable specificity? Interconversion of the *RIM101* alleles between strain backgrounds strongly impacts only three transcripts, *AQY2*, *ncFRE6*, and *TIP1*, while other genes remain largely unaffected. How W249L, a mutation that has not been previously described, permits such specificity, remains unclear, though it is likely that such a phenomenon arises from allele-specific interactions with other genetic elements. However, the limited impact of the *RIM101* allele on expression of other genes implies that if this is the case, the interaction is specific to *AQY2/ncFRE6* and *TIP1*. Because the change in expression of *AQY2/ncFRE6* occurs in the opposite direction as *TIP1* (*AQY2/ncFRE6* higher in ∑*RIM101* strains, *TIP1* lower in ∑*RIM101* strains), it is possible that the mechanisms by which W249L elicits such a focused response are different between the two loci. Furthermore, expression of *AQY2/ncFRE6* or *TIP1* did not change in the ∑1278b *RIM101* deletion or the S288c *RIM101* deletion strains, respectively, further supporting a role for a W249L-specific interaction with other factors to influence *AQY2/ncFRE6* and *TIP1* expression specifically. Our results suggest that subtle mutations within TFs interact with genetic backgrounds to elicit unique combinations of gene expression patterns, likely expanding the phenotypic diversity observed within a population.

Such a focused, allele-dependent transcriptional response to a TF-linked variant stands in contrast to most known *trans*-regulators that strongly affect expression of large numbers of genes (Yvert et al. 2003). In order to understand the mechanisms by which such subtle mutations affect gene expression it may be necessary to undertake a systematic allele-swapping strategy. Such studies are likely to reveal concepts

important not only for understanding the biochemical nature of the variant itself, but also how the effect of the variant is propagated throughout alternate genetic backgrounds. Moreover, such an approach would afford researchers the ability to learn specifically about how variants within TFs, rather than other categories of genes typically discovered in eQTL studies (Yvert et al. 2003), affect gene expression. Our finding that a SNP within a TF that regulates hundreds of genes cause large-scale expression differences in so few transcripts supports a model in which specific TF alleles interact in a combinatorial manner to regulate specific sets of genes (Yvert et al. 2003).

One outstanding question is whether our findings regarding background or allele-dependent activities of a transcription factor will be generalizable to other complex biological systems, including those involved in disease. For instance, transcription factors, including zinc finger TFs, are frequently mutated in cancers (Ashworth et al. 2014) and other human diseases, yet little is known about how the mutations relate to disease progression or outcome. With an enormous amount of sequence and functional data now available through consortiums such as The Cancer Genome Atlas (TCGA) and the 1,000 Genomes Project (McVean et al. 2012),tools now exist to test whether different alleles of the same TF can lead to variable expressivity of disease-associated phenotypes by impacting transcriptional profiles.

3.3 Complex genetic interactions and evolution of the *RIM101* transcriptional regulatory network

RIM101-mediated regulation is affected not only by the *RIM101*-allele, but also the background that it is present within, suggesting that even in the relatively simple case of *RIM101*-mediated regulation of *AQY2/ncFRE6*, the regulatory pathways have diverged between S288c and ∑1278b. Furthermore, *S.paradoxus*, a species closely related to *S.cerevisiae*, does not conform to the same regulatory guidelines that govern the *S.cerevisiae* strains we tested. Although *S.paradoxus* harbors a *RIM101* allele that includes the *S.cerevisiae* expression-permissive *Rim101* L249 variant, *AQY2/ncFRE6* expression is absent, suggesting that the *RIM101*-dependent transcriptional regulatory circuit has been rewired between the species at this locus. Clearly, the regulatory pathways underlying even simple, binary expression patterns display extraordinary complexity that could contribute to the plasticity of gene expression regulation observed throughout evolution.

The *RIM101* allele-dependent interactions that we observed may contribute to the phenotypic diversity observed between S288c and ∑1278b. Because *AQY2* is non-functional in S288c, but functional in ∑1278b, the evolutionary pressures affecting expression of *AQY2* are likely different between the strains. Perhaps the subtle *RIM101* W249L variant, which strongly alters expression of only three transcripts, represents an example of genetic drift between the strains. *TIP1* is a cell-surface mannoprotein and *AQY2* is a cell surface water channel, raising the possibility that the focused *AQY2* and *TIP1* expression differences caused by W249L may result in an altered cell surface environment between the strains. Although we showed that the *RIM101* allele did not

affect haploid invasive growth, such a re-structuring of the cell surface could result in other *RIM101*-linked cell surface phenotypes.

Cryptic genetic variation (CGV) is genetic variation that influences a phenotype in certain environmental or genetic contexts, but not in others (Paaby & Rockman 2014). Although it is almost certain that CGV is common in nature, very few examples have been described in detail (Rutherford & Lindquist 1998; Milloz et al. 2008). Our study highlights a previously undescribed mechanism by which CGV can manifest. We propose that polymorphic transcription factors likely represent a source of CGV whereby certain genetic backgrounds buffer against widespread transcriptional dysregulation upon introduction of a non-native allele, while others are subject to a dramatic shift in gene expression. The regulatory capacity of *RIM101* is highly background-dependent and the interaction of *RIM101* with genetic background determines whether a cell will undergo widespread or localized changes in its transcriptional program upon introduction of an alternative *RIM101* allele.

Chapter IV: International Genetically Engineered Machines (iGEM)

4.1 Year one: Establishment of iGEM and a "Synthetic Biology" club at CU

The International Genetically Engineered Machines (iGEM) competition is an opportunity for undergraduates to get hands-on experience in the laboratory and compete at an international synthetic biology competition. Over the course of a three month period during the summer, teams are charged with the task of creating a new solution to a societal problem. Over the course of my PhD I mentored three iGEM teams, consisting of a total of about 30 undergraduate students from four departments. These three summers were by far some of the most rewarding moments of my graduate school career as I watched students with little to no experience take control of a project and think critically about experimental results.

The first year that I managed an iGEM team was the summer of 2012. This year turned out to be more of a learning experience about how to manage a team than obtaining results. Together, another graduate student, Joe Rokicki and I were able to establish iGEM within the CU-Boulder community, from recruitment of participants and the development of a "synthetic biology" club, to establishment of important contacts and funders. Today iGEM is thriving at CU.

Our first team consisted of five students, including one from Dartmouth, and focused on expressing an AHLase to combat quorum sensing. N-acyl Homoserine lactone (AHL) is produced by a bacterial cell and its concentration in an environment

can be sensed by other bacterium, a process called quorum sensing. This mechanism is important for processes such as biofilm formation, antibiotic resistance, and virulence. By expressing and purifying an enzyme capable of cleaving an AHL molecule, called an AHLase, we aimed to disrupt these processes. Although our team only made slight advances in the lab, we gave an outstanding presentation and poster at the iGEM jamboree at Stanford. Since graduating, all of the initial members of iGEM have gone on to start careers in science.

Although the first summer we spent most of the summer formulating an idea and teaching students basic cloning techniques, the experience was invaluable as we learned that to compete in iGEM we would need to make iGEM a more year round experience. For this reason we initiated CU-Boulders first “synthetic biology club” focused on narrowing down project ideas and designing experiments and controls before the summer begins. The club runs throughout spring semester, meeting once a week to discuss topics in synthetic biology and brainstorm for the following summer. Attendance usually consists of several graduate student and post-doctoral advisors as well as dozens of undergraduates. This club will be paramount to the success of future iGEM teams at CU.

4.2 Year two: “A calcium precipitable restriction enzyme”

After establishing iGEM in year one, we strived to be much more competitive the second year. We were able to recruit a larger team, consisting of six very creative full

time members as well as several part time participants. This years project was planned during the semester leading up to the summer of 2013 in synthetic biology club. Our idea was to come up with ways to make biological research easier and more cost effective. Our efforts during the summer were rewarded with the following publication in the Journal of the American Chemical Society's synthetic biology journal (JACS syn bio).

An Engineered Calcium-Precipitable Restriction Enzyme

Josephina Hendrix[1], Timothy Read[1], Jean-Francois Lalonde[1], Phillip K. Jensen[2], William Heymann[2], Elijah Lovelace[3], Sarah A. Zimmermann[1], Michael Brasino[2], Joseph Rokicki[1], & Robin D. Dowell[1,4*]

4.2.1 Abstract:

We have developed a simple system for tagging and purifying proteins. Recent experiments have demonstrated that RTX (Repeat in Toxin) motifs from the adenylate cyclase toxin gene (CyaA) of *B. pertussis* undergo a conformational change upon binding calcium, resulting in precipitation of fused proteins and making this method a viable alternative for bioseparation. We have designed an iGEM Biobrick comprised of an RTX tag that can be easily fused to any protein of interest. In this paper we detail the process of creating an RTX tagged version of the restriction enzyme EcoRI, and describe a method for expression and purification of the functional enzyme.

4.2.2 Introduction:

Commonly used methods for protein purification include high-performance liquid chromatography (HPLC) and other affinity based methods. While effective, these methods generate hazardous waste and require costly, limited-use materials. Recently developed methods for protein purification involve tagging the protein of interest and purifying with high heat or harsh chemical conditions, both of which can influence the activity of the protein[1].

In the last few years, the RTX motif of *B. pertussis* has been investigated as a possible alternative for tagging and purifying proteins[1]. The RTX motif consists of a nine amino acid sequence that repeats up to 40 times[2,3,1]. In the presence of calcium, these motifs undergo a conformational change resulting in precipitation of the polypeptide[4,2,3]. These motifs can theoretically be appended to any protein to allow for its precipitation and purification. The precipitation reaction occurs rapidly at room temperature and requires a lower salt concentration than other stimulus-induced tags, protecting a tagged protein from potential degradation.

The purification of restriction enzymes is of particular interest to the scientific community as they allow for site-specific cleavage of DNA to facilitate cloning. While not overly expensive, restriction enzymes do present a significant financial burden when used in bulk; therefore, we chose to develop a method of purifying EcoRI that was cheap, easy and effective for synthetic biology use.

4.2.3 Methods:

Standard Biobrick assembly protocols were followed to generate a plasmid that simultaneously expresses the EcoRI-RTX construct and EcoRI methylase. DH10B cells were cultured in 250mL flasks containing LB+100µg/mL Ampicillin at 37°C, shaking at 225 RPM until saturated. Cells were resuspended in 6 mL 50mM Tris-HCl pH 7.5 and lysed using a French press. Lysates were centrifuged at high speed for 20 minutes and supernatants were collected. $CaCl_2$ was added at increasing concentrations ranging from 0-100mM in a volume of 1 mL and incubated at room temperature for two minutes before centrifugation at 16,000 x g for two minutes. Supernatants were collected for SDS-PAGE analysis. Pellets were washed four times in 50mM tris-HCl, pH 7.5. Finally, pellets were resuspended in 50mM tris-HCl containing 50mM EGTA. EcoRI-RTX containing pellets were solubilized and centrifuged for 10 minutes at 16,000 x g.

To monitor purification of EcoRI-RTX, we performed SDS-PAGE analysis. Samples were boiled in SDS buffer containing DTT for 10 minutes and loaded into a 10% SDS-PAGE gel. To test the functionality of EcoRI, we digested an exogenous plasmid with 10 units of SpeI and 7.5 uL of EcoRI-RTX. Digestions were compared to exogenous plasmid digested with commercial SpeI and EcoRI.

4.2.4 Results:

To demonstrate EcoRI-RTX precipitation in response to calcium, we added increasing amounts of $CaCl_2$ to whole cell lysate from plasmid harboring *E. coli* and monitored accumulation of pellets after centrifugation. Indeed we did see pellets form

specifically in the calcium containing samples, with the optimal calcium concentration of 50mM. To verify that the pellet was in fact EcoRI-RTX, we performed SDS-PAGE analysis (Figure 1). A band is visible in the pellet at the predicted size.

Because EcoRI is an endonuclease, its expression in *E. coli* lacking EcoRI methylase is highly toxic[5]. We designed a Biobrick-compatible plasmid capable of expressing the EcoRI methylase and verified the plasmid was protected from cleavage by EcoRI (Figure 2A). After isolating the plasmid, we digested with commercially available EcoRI and PstI, whose sites flanked EcoRI methylase. The presence of a single band after the digest in addition to robust growth of the cells clearly shows that the methylase is being expressed and is protecting the EcoRI site from cleavage *in vivo*.

Finally, to test the activity of our engineered EcoRI, we used our EcoRI-RTX harboring resuspensions in a restriction digest (Figure 2B). For this experiment, our substrate was aplasmid containing the sequence of AmilCP, flanked by a SpeI site and an EcoRI site. Asindicated by the presence of two bands of the correct sizes in the gel, EcoRI-RTX was capable of digesting the plasmid, albeit with reduced efficiency compared to commercially available EcoRI. We tested the activity of the endonuclease activity after precipitation with 0, 50, and 100mM and found that 50mM was ideal.

4.2.5 Discussion:

We developed a simple system for tagging and purification of proteins, and demonstrated its effectiveness by purifying the EcoRI restriction enzyme. We designed

a plasmid containing an RTX tag sequence that can readily be fused to any protein of interest. We also verified that RTX-tagged EcoRI retained endonuclease activity after tagging and precipitation.

We note that there were specific aspects of the method where further optimization may be required. Unexpectedly, we observed reduced endonuclease activity in the sample precipitated in 100mM Ca^{2+} relative to 50mM Ca^{2+}. Because SpeI activity was also reduced in the 100mM sample, it is likely that residual Ca^{2+} in the solution may have generally inhibited restriction enzyme activity, a phenomenon that has been described previously[6]. Using lower concentrations of Ca^{2+} or more thorough washing of the pellet could help to alleviate this issue.

The RTX system is a useful protein purification tool that does not require harsh chemicals or extreme temperatures and significantly reduces costs associated with traditional protein purification approaches. Many proteins require specific temperature conditions for transportation and storage. RTX tagged proteins can be produced quickly on site, eliminating the need for long-term cold storage. This may ultimately allow extremely rural labs, even those with no, or only intermittently available electricity the ability to quickly purify proteins when they are needed. Overall, RTX precipitation offers a fast, inexpensive method for protein purification.

4.2.6 References:

1. Shur, O., Dooley, K., Blenner, M., Baltimore, M., & Banta, S. (2013). A designed, phase changing RTX-based peptide for efficient bioseparations. BioTechniques, 54, 197-206.

2. Chenal, A., Karst, J. C., Pérez, A. C. S., Wozniak, A. K., Baron, B., England, P., & Ladant, D. (2010). Calcium-induced folding and stabilization of the intrinsically disordered RTX domain of the CyaA toxin. Biophysical journal, 99(11), 3744-3753.

3. Pérez, A. C. S., Karst, J. C., Davi, M., Guijarro, J. I., Ladant, D., & Chenal, A. (2010). Characterization of the Regions Involved in the Calcium-Induced Folding of the Intrinsically Disordered RTX Motifs from the *Bordetella pertussis* Adenylate Cyclase Toxin. Journal of molecular biology, 397(2), 534-549.

4. Blenner, M. A., Shur, O., Szilvay, G. R., Cropek, D. M., & Banta, S. (2010). Calcium-induced folding of a beta roll motif requires C-terminal entropic stabilization. Journal of molecular biology, 400(2), 244-256.

5. Rubin, R. A., & Modrich, P. (1977). EcoRI methylase. Physical and catalytic properties of the homogeneous enzyme. Journal of Biological Chemistry, 252(20), 7265-7272.

6. Pingoud, Vera, Wende, Wolfgang, Friedhof, Peter, Reuter, Monika, Alves, Jurgen, Jeltshch, Albert, Mones, Letif, Fuxreiter, Monika, & Pingound, Alfred. (2009). On the divalent metal ion dependence of DNA cleavage by restriction endonucleases of the EcoRI family. Journal of molecular biology, 393, 140-160.

4.2.7 Figures

Figure 1) A) SDS-PAGE analysis of calcium precipitation of EcoRI-RTX. Calcium was added to EcoRI-RTX expressing whole cell lysate (supernatant S, pellet P) at increasing concentrations from 0mM to 100mM**,** run on a 10% SDS-PAGE gel and stained with coomassie blue.

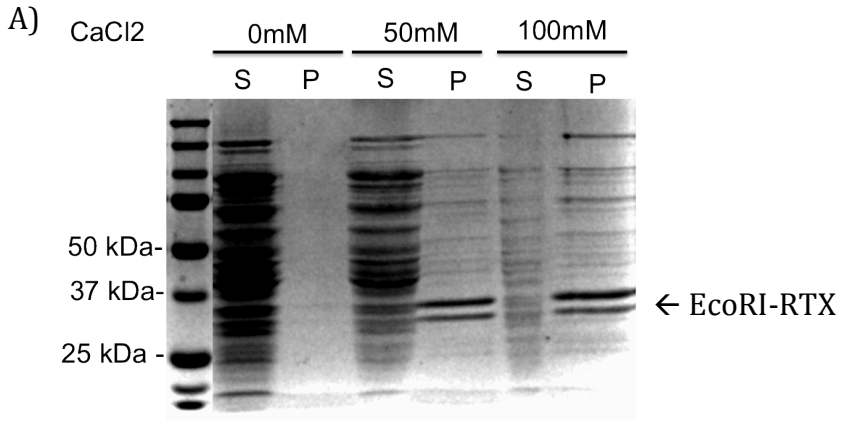
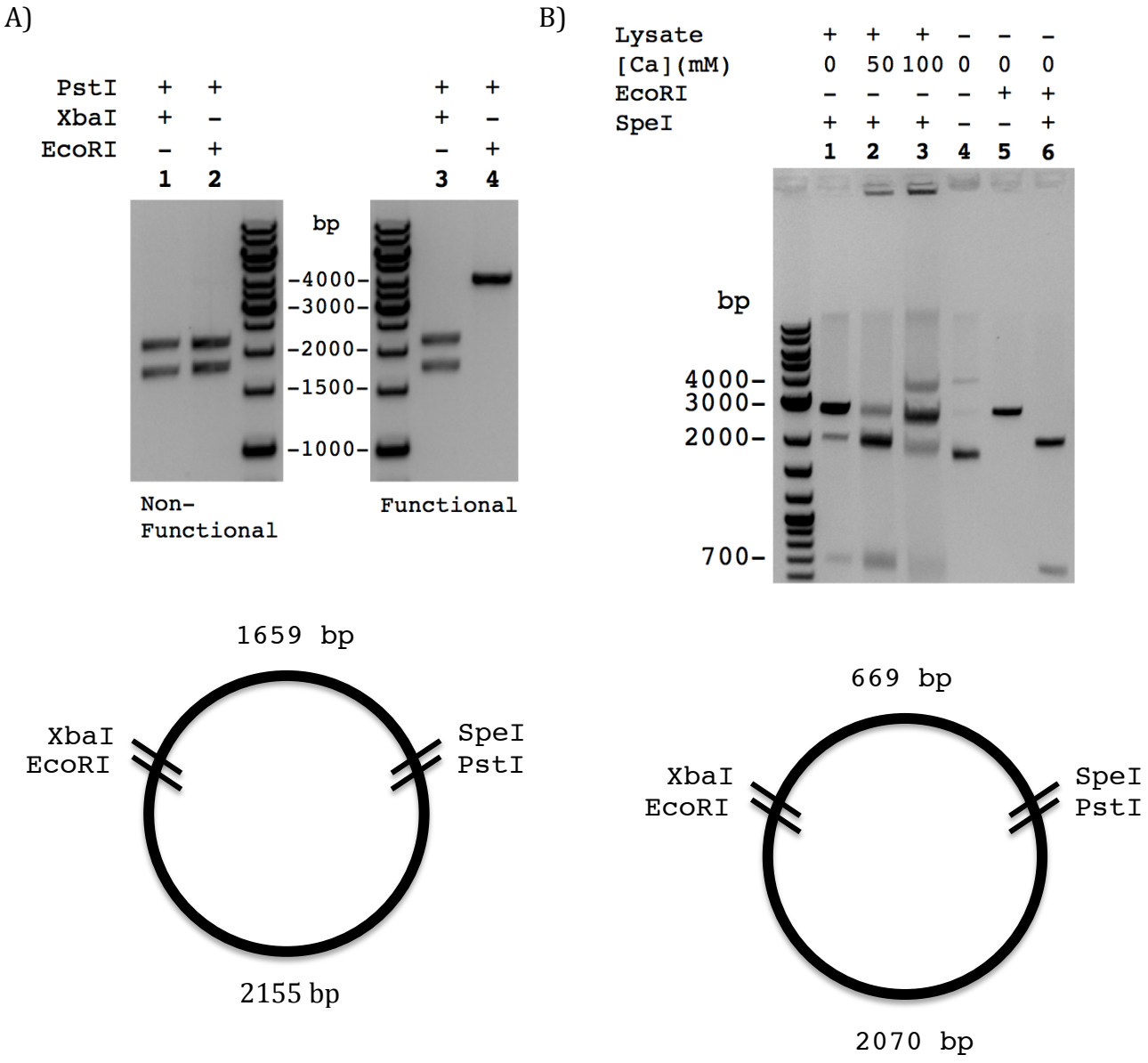


Figure 2) A. Methylase prevents EcoRI cleavage. A plasmid containing the sequence of a non-functional EcoRI methylase (interrupted by a premature stop codon) was digested by PstI and XbaI (Lane 1), or PstI and EcoRI (Lane 2). Plasmid expressing a functional EcoRI methylase was digested by PstI and XbaI (Lane 3) or PstI and EcoRI (Lane 4). Diagram of plasmid with sizes shown below. **B. EcoRI-RTX is functional**. An exogenous plasmid was run on a gel after digestion by lysate containing EcoRI-RTX (Lane 1), pellet after precipitation of EcoRI-RTX w/ 50mM $CaCl_2$ (Lane 2), or 100mM $CaCl_2$ (Lane 3). Exogenous plasmid prior to digestion (Lane 4), after digestion by commercially available EcoRI (Lane 5), or SpeI and EcoRI (Lane 6). Diagram of plasmid with sizes shown below.



4.3 Year three: A sequence specific alternative to antibiotics

With iGEM well established in the CU community, we were able to recruit about 20 members for the 2014 iGEM team. During synthetic biology club, students were tasked with coming up with an idea for the summer project. We heard proposals from 5 separate sub groups and settled on one project to focus on during the summer. This team was in a good position to succeed because we had several returning members from the previous years team.

The project that we undertook was an ambitious one. We chose to address the problem of antibiotic resistance by combining a well-established approach, bacteriophage, with a very modern molecular biology technique, CRISPR-cas. While bacteriophage have been considered for years as an alternative to antibiotics, significant safety concerns exist. For example, phage reproduce rapidly within their host and release new phage into the environment. Obviously it is not safe to treat human patients with a drug with an uncontrollable dose. Also, phage genomes can mutate, creating potentially harmful phage capable of damaging the host microbiome. For this reason, we needed a way to engineer a phage to continue to kill the bacteria that it infects while avoiding potentially harmful side effects.

To address safety concerns, we took advantage of a major breakthough in molecular biology. Clustered regularly interspaced short palindromic repeats (CRISPR) are sequences contained within the bacterial genome that encode a defense mechanism against invading phage. These elements are transcribed and the RNA

products physically associate with the cas9 endonuclease, which is guided to the DNA of the invading virus through complementary basepairing of the guide RNA with the DNA of the invading virus. One basepairing has been acheived, the cas9 endonuclease cleaves the DNA, resulting in a double stranded break. If the DNA is guided to a sequence of the bacterial genome, such cleavage results in cell death. While most groups have used the CRISPR-cas9 system for genome editing, we wanted to exploit the system to create a sequence-specific antibiotic.

Using a phagemid system we were able to generate large amounts of replication-deficient phage containing a plasmid harboring a minimal CRISPR array and a cas9 protein. We engineered the CRISPR array to be directed against a bacterial sequence. In doing so, we were able to successfully kill strains based on their genomic sequence. Once again our team gave an excellent presentation at the Giant Jamboree held in Cambridge, MA.

Over the three years that I mentored the iGEM team at CU, I found the experience to be an excellent outlet. Counter-intuitively, I felt that the times when I was mentoring were far more productive for my own research. Taking a few minutes throughout the day to think about something else often served to re-focus my attention on my own project. Hopefully the iGEM community will continue to grow at CU for years to come.

Materials and Methods

Strains, media, mirobiological techniques, and growth conditions.

S.cerevisiae strains used in this study were derived from BY4742 (S288c, his3Δ1, lys2Δ0, leu2Δ0, ura3Δ0) or L6441 (Σ1278b, ura3-52, leu2::hisG, his3::hisG). Other strains used in Fig 6A and Fig S8 including JAY291, RM11-1a, CLIB324, CLIB382, YPS163, T7, and UC5 are homothallic diploids generously donated by Justin Fay (Washington University). The *delitto perfetto* (Storici et al. 2001) method was used to edit genome sequences. For gene expression experiments, cells were grown in standard YPD media over night and saturated cultures were diluted and let grow to mid-log phase in YPD before washing pellets 1X with dH20 and snap freezing. Primers and plasmids used in this study are listed in supplementary materials and methods. Invasive growth phenotype assay was performed as described in (Gimeno et al. 1992) by patching cells onto a YPD plate for two days and washing the plate under gently running water before imaging.

qRT-PCR

RNA was extracted using a standard acid phenol chloroform extraction and DNased with RQ1 DNase (Promega) according to manufacturer's instructions. 1ug of RNA was reverse transcribed using Multiscribe reverse transcriptase (Life Technologies) with

random hexamers, except for *ncFRE6*, for which we used a gene specific RT primer due to the need to measure RNA levels strand-specifically. cDNA was measured using targeted qPCR primers and SYBR select (Life Technologies) on the Biorad CFX qPCR system.

Genome-wide expression profiling by RNA-seq

Strand specific RNA-seq libraries were made using the NEBNext Ultra Strand-specific RNA-seq library prep kit (NEB #E7420S/L) with manufacturers instructions. Briefly, RNA was isolated by standard acid phenol chloroform extraction and mRNA was purified with oligo (dT) dynabeads (Life Technologies). mRNA was fragmented and first strand synthesis performed with ProtoScript II reverse transcriptase and random hexamers. Second strand synthesis then incorporated Uridine residues into cDNA. cDNA was purified with AMPure beads (Agencourt). cDNA was then dA-tailed and NEBNext adaptors for Illumina were ligated before another AMPure purification. USER excision removed the second strand and libraries were amplified with NEBNext High Fidelity PCR master mix (NEB). NEBNext Multiplex oligos 1-12 (NEB #E7335) were incorporated during PCR. Libraries were quantified with the Qubit (Life technologies) before pooling at equimolar concentrations and sequencing on an Illumina Hiseq. Reads were mapped using bowtie2 and differential expression was assessed using DEseq (see Supp Methods for full description).

Expression guided bulked segregant analysis

S288c (BY4741) and Σ1278b (L6441) haploid strains were crossed to generate a heterozygous diploid. The diploid was sporulated on traditional sporulation media and haploid segregants were grown to mid log phase in YPD and genomic DNA was extracted with phenol chloroform and RNased (Ambion) according to manufacturer's recommendation. 28 haploid segregants of an S288c x Σ1278b cross were tested for expression of *AQY2/ncRNA-FRE6* by qRT-PCR. Genomic DNA was treated with RNAse (Ambion), and purified using Phenol chloroform. DNA concentrations were measured with the Qubit (Life Technologies) and pooled at 10nM separately for strains either expressing or not expressing *AQY2/ncFRE6*. DNA was sheared using the Covaris M220 ultrasonicator to an average size of 500 bp. DNA was blunted and dA-tailed before ligation of Illumina adapters. Libraries were amplified by Phusion polymerase with Illumina multiplex barcodes 1+2 for ten cycles before analysis on the Bioanalyzer (Agilent). Samples were sequenced on an Illumina HiSeq. Reads were mapped using bowtie2 and variants identified using GATK (see Supp Methods for full description).

Data availability

All raw data was submitted to the SRA under the accession number PRJNA285097. This includes the single-end 50bp Reb1::myc ChIP-seq in both S288c (BY4742) and ∑1278b (L6441) strain backgrounds, single-end 50bp pooled S288c:∑1278b tetrad

genome sequencing of expressors and non-expressors, and single-end 126bp RNA-seq libraries for S288c wildtype, S288c(∑*RIM101*), S288c-*rim101Δ,* ∑1278b wildytpe, ∑1278b(S2*RIM101*), and ∑1278b-*rim101Δ.*

Chromatin Immunoprecipitation

Briefly, *Reb1* C-terminal myc tagged strains SAV261 (S288c) and SAV273 (∑1278b) were derived from BY4742 (S288c) and L6441 (∑1278b) and generously provided by Gerald Fink (MIT). Alternatively Rim101 N-terminal 6x HA tagged strains were generated in BY4742, L6441, and corresponding RIM101-interconverted strains. 50mL cultures were grown to mid-log phase in YPD at 30^{o}C and fixed with 1% formaldehyde for 30 minutes and quenched with glycine for 10 minutes. Cells were pelleted, washed 1X with 10mL 1X TBS, and snap frozen. Cells were lysed in lysis buffer containing 50mM HEPES-KOH pH7.5, 140mM NaCl, 1mM EDTA, 1% Triton X-100, .1% sodium deoxycholate, and protease inhibitors by bead beating 5X for 4 minutes at 4^{o}C. Lysate was transferred to a 15mL conicle, centrifuged at 8500 rpm, and the pellet was washed 2X with lysis buffer. Chromatin was sheared using a Diagenode Bioruptor and immunuprecipitation was performed using anti-myc antibody (Sigma cat #: M4439) conjugated to protein G beads (Thermo Fisher Scientific cat #: 10004D) overnight at 4^{o}C. Beads were washed 2X with lysis buffer, 2X with lysis buffer + 500mM NaCl, and 2X with wash buffer containing 10mM Tris-HCl pH 8.0, 250mM LiCl, .5% NP40, .5% sodium deoxycholate, and 1mM EDTA. Chromatin was eluted from beads in TE + 1%

SDS for one hour at 65°C and cross-links were reversed in TE + .5% SDS for 8 hours at 65°C. DNA was purified using a QIagen PCR Purification kit (Qiagen cat #: 28104) before proceeding to qPCR or library prep. qPCR was performed using primers specified in supplemental materials and methods. Sequencing libraries were prepared by blunting the sheared DNA, A-tailing, and ligating Illumina Tru-seq adapters. Libraries were size selected and purified using a 2% agarose gel and a Qiagen Gel extraction kit. Adapter-ligated libraries were PCR amplified for 18 cycles and run on the Illumina Hi-seq 2000.

Northern analysis

RNA was purified by a standard acid phenol chloroform extraction and denatured for five minutes at 65°C. 20ug of RNA was run on a 1% agarose/MOPS/formaldehyde gel. Formaldehyde was rinsed from the gel 2X w/ DEPC H20 and RNA was transferred to a nitrocellulose membrane overnight. The membrane was washed 2X with 20X SSC and probed with a solution containing Church buffer and P32 labeled probe. For detection of *ncFRE6* and *FRE6* mRNA, the probe was in vitro transcribed using SP6 or T7 RNA polymerase to generate a strand-specific probe. For the *SCR1* control, a DNA probe was generated from a PCR amplified probe against *SCR1*. Blots were probed overnight at 65°C, washed 3X with 20X SSC, and visualized using the Typhoon scanner and analyzed by densitometry.

Western analysis

Cells were grown to mid log phase and normalized for cell number before protein isolation. Protein was isolated by a standard Trichloroacetic acid (TCA) extraction. Lysates were run on a 10% SDS/PAGE gel and transferred to a PVDF membrane overnight at 4^{o}C. The membrane was washed 3X with TBST, blocked with a 5% milk solution, and probed using an anti-HA antibody for one hour, followed by an HRP-conjugated secondary antibody for one hour.

Supplemental Materials and Methods

Identification of disrupted TF binding motifs in the *AQY2/ncFRE6* promoter

We used the Yetfasco (De Boer & Hughes 2012) "Scan sequences" tool to identify all TF binding motifs that exist within the S288c *AQY2/ncFRE6* SNP-dense region requiring a 95% match to a high quality motif. We cross-referenced the list to the list obtained through the same analysis in ∑1278b to obtain a list of motifs that exist in one strain but not the other due to a mutation(s) within the binding motif (Table S3).

Computational analysis of ChIP-seq data

Mapping

S288c *Reb1* ChIP-seq and ∑1278b *Reb1* ChIP-seq reads were mapped to the S288c reference genome (*S. cerevisiae* genome obtained on 06/26/2011, from from the *Saccharomyces* Genome Database, FTP SITE:

http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/ corresponding stable release from February 2011: http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/S288C_reference_genome_R64-1-1_20110203.tgz) using bowtie2 (Langmead & Salzberg 2012). They were then converted to BAM format using samtools (Li et al. 2009), and duplicate reads were removed via Picard's MarkDuplicates.jar (1.72). For visualization purposes, the duplicate removed reads were converted to pileup format using Bedtools (Quinlan & Hall 2010) genomeCoverageBed, and then normalized by read depth.

Peak Calling and Motif Enrichment

The peak caller MACS2 (v2.0.9) (Zhang et al. 2008) was run on the non-deduplicated mapped files using broad peaks and allowing for up to 5 duplicate reads at any position, resulting in ~1700 peaks for both S288c and ∑1278b. Those peaks were then subjected to a score cutoff (greater than or equal to 50) and compared (using bedtools intersectBed (v2.16.2) with any overlap). The unique peaks as determined from the intersection were then queried back to the original 1700 peaks for the other strain, to remove artifacts created by the score cutoff. This resulted in a conservative list of strain-unique peaks (68 in S288c and 25 in ∑1278b). Motif enrichment analysis was performed on all peaks with quality score greater than 50 using MEME (v4.10.4) (Bailey et al. 2009) looking for motifs of length 8-10.

Computational analysis of expression-guided bulked segregant analysis (eBSA)

Overall

The overall method of analysis for identifying the overrepresented alleles in the two pooled expression-guided bulked segregant analysis samples is as follows (Fig S4): First, the two samples were mapped to both the S288c and ∑1278b reference genomes. SNPs were called for each pool relative to either genome, with the expectation that SNPs will be heterozygous (i.e. having relatively equal allelic representation at all locations that did not affect expression of *AQY2/ncFRE6*). SNPs that were called as homozygous in both pools with reciprocal orientation (i.e. homozygous for one allele in the first pool and the other allele in the second pool), and were also called when reads were mapped to the opposite reference genome, were considered potentially linked to expression of *AQY2/ncFRE6*. Only one region of consistent homozygosity (multiple homozygous SNPs in succession) met these criteria: A ~35kb region on chromosome 8 containing 12 genes, including *RIM101* (Fig S5).

Raw data

The raw data, single-end 50bp reads, was obtained from the University of Colorado—Denver High Throughput Sequencing Core on the HiSeq2500. The two pooled samples (*AQY2/ncFRE6* expressors versus non-expressors), contained 19.5 and 25.1 million reads each, respectively. Raw reads were tested for adapter read-through or low quality using the FastQC tool (v0.11.2) (Leggett et al. 2013).

Mapping and variant Calling

High quality reads from each pooled sample were mapped to each of two *Saccharomyces cerevisiae* reference genomes. The reference sequence for the laboratory yeast strain S288c reference genome (*S. cerevisiae* genome obtained on 06/26/2011, from from the *Saccharomyces* Genome Database, FTP SITE: http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/ corresponding stable release from February 2011: http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/S288C_reference_genome_R64-1-1_20110203.tgz) as well as the reference sequence for the laboratory strain ∑1278b (reference genome obtained from Dowell 2010)(Dowell et al. 2010) . Reads were mapped using Bowtie2 in very-sensitive end-to-end mode with default score settings (v2.2.3)(Langmead & Salzberg 2012). After mapping, read information was converted into binary format for downstream analysis using Samtools view, sort, and index (v0.1.18)(Li et al. 2009). Variant calling was performed on the tailored read mappings using GATK UnifiedGenotyper (v2.4-9)(Schmidt 2009). Custom scripts were used to parse out and graph allelic frequencies on a per-SNP basis.

Identification of the region of the genome harboring RIM101 from Pooled Sequencing

In order to identify the single locus that segregated with expression of *AQY2/ncFRE6*, we parsed the allelic frequencies of every SNP called in the union of both groups when mapped against each genome. We searched for a genomic region matching the following criteria: 1) a region with high quality variants not representative of mapping artifacts due to differences between the genomes, 2) the region should be homozygous in both groups (i.e. all segregants within the pool had the either the S288c or the ∑1278b allele), 3) the orthologous region should also be homozygous when mapped against the other genome, and 4) the SNPs around the boundaries should gradually decrease in allelic frequency away from a binary homozygous region towards an even 1:1 ratio of alleles in each pooled set. Only one region fit this criteria, a roughly 35Kb region on the left arm of chromosome eight (v2.1.19)(Thorvaldsdóttir et al. 2013).

Analysis of non-synonomous SNPs in transcription factors

We sought to determine whether RIM101 was more or less polymorphic than all other transcription factors. Briefly, we split the genome of ∑1278b into 150mer reads, mapped them back to S288c using Bowtie2, and called SNPs using GATK UnifiedGenotyper to identify regions where the two genomes differ. Using custom scripts, we annotated SNPs over coding regions, including whether or not the SNPs cause amino acid changes. An entire list of proteins with annotated DNA binding domains (n=249) was retrieved from YetFasCo (De Boer & Hughes 2012) and was plotted as a histogram of non-synonymous mutations per kilobase for each gene (Fig S6).

RNA Sequencing

Overall

Single end, strand-specific RNA-seq was performed on six strains in biological duplicate. In order to discover the differentially expressed genes, the data was mapped back to its respective genome (S288c or ∑1278b), read counts over annotated genes were collected, each gene count was normalized for total depth over non-Ribosomal regions on a per-sample basis, and then the genes that exist in both genomes had their expression levels compared to identify those genes that are significantly different in expression.

Raw data

Raw data consisted of single-end 126bp reads, obtained from University of Colorado—Denver High Throughput Sequencing Core and was sequenced on the HiSeq2500. Raw reads were first converted into their reverse complement (due to the NEBNext Ultra-sensitive strand specific library prep) and sent through quality analysis to identify possible adapter read-through or quality-score biases using the FastQC tool (v0.11.2)(Leggett et al. 2013). We observed a large amount of adapter in the 3' end of reads. Hence, we hard-trimmed the reads to 50bp in the mapping process (below).

Mapping

Reads were mapped to their respective genomes (see Genome Sequencing) using Bowtie2 (Langmead & Salzberg 2012). We trimmed the reads to 50bp (bowtie2 option -5 76). We ran bowtie2 with --very-sensitive end-to-end alignment, and adjusted our scoring scheme to limit mismatches and insertions and deletions (-L,-20,0). They then underwent file format conversion into the binary format for downstream analysis using Samtools view, sort, and index (v0.1.18) (Li et al. 2009). Read mapping statistics are as follows:

Sample	Total Reads	Total Mapped	Replicate Pearson Correlation Coefficient
S288c_wt_rep1	20629107	19494132	.99622
S288c_wt_rep2	23279566	22241295	.99622
S288c_∑RIM101_rep1	21984130	20689010	.99832
S288c_∑RIM101_rep2	20697838	19804140	.99832
S288c_RIM101deleted_rep1	22891460	21029594	.99687
S288c_RIM101deleted	22651091	20403109	.99687

_rep2			
Sigma_wt_rep1	22893870	21484712	.99433
Sigma_wt_rep2	21994461	20801472	.99433
Sigma_S2RIM101_rep1	24951411	23444650	.98974
Sigma_S2RIM101_rep2	20918844	19767717	.98974
Sigma_RIM101deleted_rep1	21961295	20384668	.99485
Sigma_RIM101deleted_rep2	23072053	21499624	.99485

Quantification and Differential Expression

Per-gene read counts were attained using HTSeq over the coding regions present in the annotations for each genome (v0.6.1)(Anders et al. 2014). After acquiring read counts in each genome, the genes that exist in both genomes were placed into a count matrix (gene x sample), dubious ORFs were removed, Pearson correlation coefficients were calculated (table above), a LaPlace transformation of +1 was added to every gene (to remove divide-by-zero errors) and read into the R package for differential expression DESeq (v1.0)(Anders & Huber 2010). The output of the DESeq analysis identified differentially expressed genes with an adjusted p-value, as well as a Log2 fold change

for the comparison. For stringency, a cutoff for differential expression of p-adj $\leq$ 0.0005 and a minimum average expression between the comparisons of ≥ 100 reads was used. We observed roughly ~20% of genes as differentially expressed at this stringent cutoff. Since the data maintained strand information, HTSeq was run separately on antisense transcripts for each gene. The antisense gene counts were processed as above, but with a lower number of required reads (50 vs. 100) mapped, because antisense transcripts typically show lower expression than sense transcripts.

Statistical Analysis of Cumulative Differential Expression

In order to quantify the "cumulative differential expression"—or measure of the total difference between two samples' expression values—for a set of genes we used the residual sum of squares (RSS) in log-space for a set of genes relative to the linear regression fit to the background set of genes (entire set minus the gene set in question). We plot the RSS for individual comparisons as a CDF to highlight a reduction in the distribution of cumulative differential expression for different pairwise comparisons. In order to assess the significance of reduction of cumulative differential expression, we use the ANOVA one-tailed F-test to evaluate whether the variance between pairwise comparisons was equal (H_0: VarA = VarB,), or whether the variance was lower (H_l: VarA < VarB).

Plasmids used in this study

The pCORE plasmid was constructed by cloning a 1.5 kb *Bam*HI-*Hin*cII fragment harboring the *kanMX4* gene into the *Bam*HI-*Ssp*I sites of pFA6aKlURA3 (Storici et al. 2001).

Primers used in this study

Primers used to generate Reb1 binding motif mutants:

For amplification of pCORE to replace Reb1 SNP in both backgrounds:

Forward primer:

ACCAACACTGATATTCCTCGAAATACTCTATAATTCTCTCGAGCTCGTTTTCGACAC
TGG

Reverse primer:

TGTTAGAAACACCGTTTCTCAAAAACTCCTCGGTTACCCTCCTTACCATTAAGTTGA
TC

Primers to amplify genomic sequence of S288c or ∑1278b for replacement of Reb1 binding site SNP:

Forward primer:

GAAGGAGCCGGAGAGAAGAT

Reverse primer:

GGAGATTCATTAGCGGTCGT

Primers to test Reb1 occupancy by ChIP-qPCR:

Forward primer:

GGAGATTCATTAGCGGTCGT

Reverse primer:

GAAGGAGCCGGAGAGAAGAT

Primers used to generate *AQY2/ncFRE6 cis* context mutants:

For amplification of pCORE to replace 30 SNPs (i.e. S288c(30 ∑ SNPs) or ∑1278b(30 S2 SNPs)):

Forward primer:

CGGCTGTTCAGGTGGAATATAAGCATTGTCAACACCGGTGAGCTCGTTTTCGACAC TGG

Reverse primer:

TTGTTGGCAACACGTCAAAATTTTCAACGGTTGGAAAGATCCTTACCATTAAGTTGA TC

Primers for amplification of genomic DNA from S288c or ∑1278b to create template for transformation and counter selection. Includes 30 SNPs within the *AQY2/ncFRE6 cis* context:

Forward primer

TGGAATATAAGCATTGTCAACACC

Reverse primer

GCCCTTTTGTTCTTTTACTGTTG

For amplification of pCORE to replace 15 *AQY2* proximal SNPs in ∑1278b (i.e. ∑1278b(15 S2 SNPs)):

Forward primer:

AGGAACAAGAAAAAAGACATGCGCACACTAATAAGCTACGAGCTCGTTTTCGACAC TGG

Reverse primer:

GGAGGTGGCGCTGCAGTCCTTCTTTTCAGACCCAAGCAATCCTTACCATTAAGTTG ATC

For this strain a gBLOCK fragment (Integrated DNA Technologies) was synthesized to replace all 15 SNPs in ∑1278b with those from S288c.

Primers used to generate *RIM101* mutant strains:

Primers for amplification of pCORE for targeting to S288c *RIM101* ORF (S288c *rim101Δ* strains used in the study).

Forward primer

ACTGAAAACGGTAAAGTAGGTTTGTTTAAATTGACTTAAGGAGCTCGTTTTCGACAC TGG

Reverse primer

GCAAAGAAACAACTAAGAATAAAATATCCGACAATCCATATCCTTACCATTAAGTTG
ATC

To amplify pCORE for targeting to ∑1278b *RIM101* ORF (∑1278b *rim101Δ* strain used in this study).

Forward primer

ACTGAAAACGGTAAAGTAAGTTTGTTTAAATTGACTTAAGGAGCTCGTTTTCGACAC
TGG

Reverse primer

GCAAAGAAACAACTAAGAATAATATATCCAACAATTCATATCCTTACCATTAAGTTGA
TC

Primers for interconversion of *RIM101* allele between strains (after insertion of pCORE in place of RIM101).

Primers for amplification of *RIM101* allele from S288c for transformation into ∑1278b *rim101Δ*:

Forward primer

AACAAGTGCAAAGATAAAATACTGAAAACGGTAAAGTAAGTTTGTTTAAATTGACTT
AAG

Reverse primer

TACTATACAGCCGCAAAGAAACAACTAAGAATAATATATCCGACAATTCATATCATA

CCA

Primers for amplification of ∑1278b *RIM101* allele for transformation into S288c *rim101Δ*

Forward primer

AACAAGTGCAAAGATAAAATACTGAAAACGGTAAAGTAGGTTTGTTTAAATTGACTT

AAG

Reverse primer

TACTATACAGCCGCAAAGAAACAACTAAGAATAAAATATCCAACAATCCATATCATA

CCA

Primers for interconverting PolyQ repeat lengths between S288c and ∑1278b

To amplify pcore for replacement of S288c polyQ repeat:

Forward primer

CCCCCATTGCCCGTGGGTATATCTCAACATCTGCCTTCAGAGCTCGTTTTCGACAC

TGG

Reverse primer

TAGCTCGTCTGAGCATAGTTGGTTTAAGGAAATAGCCCGTCCTTACCATTAAGTTGA

TC

Primers to amplify pCORE for targeting to ∑1278b polyQ repeat:

Forward primer

CCCCCATTGCCCGTGGGTATATCTCAACATCTGTCTTCAGAGCTCGTTTTCGACACT

GG

Reverse primer

TAGCTCGTCTGAGCATAGTTGGTTTAAGGAAATAGCCCGTCCTTACCATTAAGTTGA

TC

Primers to amplify pCORE for replacement of four individual amino acid residues implicated in regulation of *AQY2/ncFRE6*.

Forward primer

GAAAGTGGCGGTATTTTGAAAAGAAAGAGGGGACCCAAATGAGCTCGTTTTCGACA

CTGG

Reverse primer

CGTTTGCTATGGTCTTATTAGAACAACCGTCCTCGTAGACTCCTTACCATTAAGTTG

ATC

Once pCORE inserted, primers used to check incorporation:

Forward primer

TCATCTGGAAAGTGGCGGTA

Reverse primer

GTGAAGAATTGGGTGGCGTT

gBLOCKs (IDT) were synthesized with each SNP, transformed, and counter selected for loss of the pCORE construct.

*All strains were confirmed by PCR and Sanger sequencing. Strains were initially selected for incorporation of the pCORE construct by growth on G418 and SC –ura and lack of growth on 5-FOA. After replacing the pCORE construct strains were tested for growth on 5-FOA and YPG and no growth on G418 or SC –URA. For polyQ repeat length strains and individual amino acid substitution strains, gBLOCK fragments were synthesized to incorporate altered polyQ lengths.

Primers used for qRT-PCR experiments:

Primer for gene-specific reverse transcription of *ncFRE6*:

CAGTGCTTTGCGTTCTACTA

qPCR primers:

Primers to measure *ncFRE6*

Forward primer

ATCGCTCGGAATAGTAAGGAAA

Reverse primer

CCCCAAATGAGCAAGGATAC

Primers to measure *ncFRE6* in Figures 4A, Figure S8.

Forward primer

TTTGAACACCAGCAACAACC

Reverse primer

ACAATATTGACCCGGTTTCG

Primers used to measure *AQY2*:

Forward primer

AACAGCCTAAACCCAAAGCA

Reverse primer

GCCGCTAGTGCTATGACTCC

Primer used for RT of ncFRE6 in *S.paradoxus*:

GGATCGTGCTGTCCTTGTTC

Primers to detect ncFRE6 in *S.paradoxus*:

Forward primer:

TTGAACACCAGCAACAACCC

Reverse primer:

GTCCCGGTTTTGAATGCCAT

Primers to detect *AQY2* in *S.paradoxus*:

Forward primer:

ACATTTCACACCTGGACCCA

Reverse primer:

TTTGTTTCCGGCTGTTCAGG

Primers used to detect relative occupancy at the *Reb1* SNP located near the start of *ncFRE6* by ChIP-qPCR:

Forward primer:

GGAGATTCATTAGCGGTCGT

Reverse primer:

GAAGGAGCCGGAGAGAAGAT

Tables

Table 1: 40 most differentially expressed genes between S288c and ∑1278b

id	Name	S288c wt	∑1278b wt	log2FoldChar	padj
YLR153C	ACS2	1.07218817	20335.0692	14.2111242	0
YBR115C	LYS2	1.64203189	2577.06044	10.6160285	1.22E-201
YBR296C	PHO89	5092.43527	91.375759	-5.8004004	2.23E-165
YGL028C	SCW11	3354.76114	121.254042	-4.7901053	9.31E-123
YLR411W	CTR3	59.2026968	2338.26879	5.30363408	9.68E-121
YLR163C	MAS1	1.07218817	769.121639	9.48650985	1.40E-118
YEL021W	URA3	1.64203189	5690.9904	11.7589819	4.32E-106
YOR390W	GO:0003674,	3.35156304	504.808078	7.23475712	3.73E-84
YHR136C	SPL2	2595.14042	64.0566273	-5.340321	6.05E-83
YBR302C	COS2	1.07218817	432.401013	8.65566797	6.21E-81
YLR285C-A	GO:0003674,	950.82841	35.067408	-4.7609823	1.71E-80
YML123C	PHO84	43715.9065	4628.67539	-3.239487	3.42E-79
YER124C	DSE1	4778.98688	24.7833605	-7.5911892	2.67E-78
YIR027C	DAL1	18.0922004	555.682731	4.94082162	8.13E-68
YLR154C	RNH203	1.07218817	327.542131	8.25497855	1.15E-67
YOR065W	CYT1	285.472051	2618.76103	3.197463	1.48E-66
YCL064C	CHA1	7325.52198	399.295298	-4.1974036	9.12E-64
YML132W	COS3	1.07218817	313.988458	8.19400959	3.26E-63
YDR281C	PHM6	658.284537	18.3005497	-5.1687524	2.90E-61
YER062C	HOR2	1543.97167	159.487602	-3.2751301	4.09E-60
YNR019W	ARE2	548.073703	4017.67574	2.87391931	5.83E-58
YPL019C	VTC3	21603.3567	2432.44869	-3.1507742	3.53E-55
YKR046C	PET10	408.974838	6016.90879	3.8789385	2.90E-54
YDR033W	MRH1	16104.7355	545.680882	-4.8832837	1.60E-53
YER044C	ERG28	1249.57546	8046.33564	2.68689389	1.48E-51
YOR153W	PDR5	35307.7804	3454.15926	-3.3535796	1.48E-51
YJR159W	SOR1	1.07218817	236.570994	7.78557126	2.13E-50
YBR093C	PHO5	4707.36807	165.349438	-4.8313306	4.28E-50
YGR049W	SCM4	280.051411	1918.40526	2.77614392	1.48E-49
YLR286C	CTS1	26879.0331	638.016254	-5.3967443	7.34E-49
YFL057C	AAD16	4.72359788	263.353811	5.80097237	1.58E-48
YJR048W	CYC1	349.300997	3878.99113	3.47313881	4.76E-47
YHR137W	ARO9	1136.36886	76.4906373	-3.8930042	2.00E-46
YNR067C	DSE4	2900.38705	501.594892	-2.5316509	1.34E-45
YDL124W	GO:0001950,	2358.24788	395.90809	-2.5744779	1.68E-45
YDR343C	HXT6	1.57453263	230.657421	7.19468425	1.39E-43
YOL151W	GRE2	1247.70243	193.854079	-2.6862309	1.13E-42
YDL037C	BSC1	257.654857	3.22729545	-6.3189703	1.41E-41
YDR380W	ARO10	481.660436	23.1461962	-4.3791694	1.57E-40
YJL107C	GO:0003674,	65.0958318	619.24067	3.24986315	3.53E-39

Table 2: 40 most differentially expressed antisense transcripts between S288c and ∑1278b

id	Name	S288c wt	∑1278b wt	log2FoldChar	padj
YLR343W	GAS2	675.087935	5.2263683	-7.0131228	5.39E-89
YBL005W	PDR3	478.484116	4.72437729	-6.662203	2.42E-75
YLR256W	HAP1	757.599093	55.6483244	-3.7670245	2.32E-49
YKR072C	SIS2	5.58510592	257.320958	5.52584051	1.37E-45
YJR103W	URA8	413.197977	24.1238775	-4.0982995	2.50E-45
YIL169C	HPF1'	287.140163	12.7410421	-4.4942	2.25E-43
YDR007W	TRP1	63.7813175	682.17916	3.41894488	9.49E-41
YNR055C	HOL1	234.221975	11.3643646	-4.3652875	9.14E-38
YFL033C	RIM15	10.560192	222.581386	4.39762498	1.22E-34
YJR160C	MPH3	1.13419094	138.063056	6.92751998	1.75E-34
YKR103W	NFT1	1.13419094	165.626988	7.19013042	8.69E-31
YHR071W	PCL5	8.81598114	208.143104	4.56131081	7.39E-27
YJL216C	IMA5	2.26838189	117.474866	5.69454477	1.16E-26
YDR107C	TMN2	33.6823331	414.610718	3.62169343	2.54E-25
YBR294W	SUL1	347.433059	52.1159166	-2.7369391	1.87E-24
YBR297W	MAL33	2.35423069	128.134061	5.76625452	5.03E-24
YEL022W	GEA2	2.26838189	100.603792	5.47087733	1.05E-23
YKL103C	LAP4	37.8757017	251.204712	2.729519	6.93E-21
YMR279C	GO:0016021,	323.501966	56.4860692	-2.5178075	9.47E-21
YLR342W-A	GO:0003674,	70.9298584	1.43208983	-5.6301992	5.89E-20
YPR194C	OPT2	108.977258	8.98370517	-3.6005727	6.92E-20
YNL053W	MSG5	25.8197671	185.897895	2.84796254	8.40E-20
YGL136C	MRM2	335.34989	64.4657924	-2.3790614	2.73E-19
YML066C	SMA2	166.897766	710.688381	2.09025247	5.71E-19
YDR452W	PPN1	19.8144954	155.676242	2.97392066	7.34E-19
YOL156W	HXT11	8.98767875	111.074999	3.62744176	7.34E-19
YNR056C	BIO5	135.348431	16.1626251	-3.0659447	7.07E-18
YCL039W	GID7	13.7910672	127.38867	3.20743097	8.82E-18
YLR278C	GO:0005634,	643.100277	164.139485	-1.9701214	4.18E-17
YLL051C	FRE6	14.0486136	165.987771	3.56257729	4.41E-17
YPR008W	HAA1	3.9267439	136.560363	5.12006148	5.25E-17
YOR202W	HIS3	13.7052184	107.727299	2.97458667	2.83E-14
YNL193W	GO:0003674,	5.23263246	73.6157347	3.81440533	3.44E-14
YPL017C	IRC15	933.6067	281.411173	-1.7301353	4.04E-14
YGR130C	GO:0003674,	56.5378496	239.560005	2.08309815	1.23E-13
YMR165C	PAH1	23.302917	134.268778	2.5265414	2.69E-13
YKL171W	NNK1	136.12107	26.9511155	-2.3364735	9.24E-13
YHR142W	CHS7	97.4727289	16.1626251	-2.5923371	9.77E-13
YEL011W	GLC3	180.878688	43.6526732	-2.0508805	5.85E-12
YHR031C	RRM3	11.4277583	88.9221442	2.96000032	6.65E-12

Table 3: 50 most differentially expressed genes in S288c rim101Δ

id	Name	S288c wt	S288c rim10	log2FoldChar	padj
YHL027W	RIM101	936.888614	3.3449476	-8.1297501	1.72E-138
YBR296C	PHO89	5092.43527	442.067228	-3.526018	9.15E-82
YER011W	TIR1	654.669427	4.62157164	-7.1462392	3.60E-61
YOL126C	MDH2	1621.59491	233.21592	-2.7976753	2.31E-50
YCL026C-B	HBN1	96.8642814	863.619358	3.15635891	2.37E-46
YMR319C	FET4	1397.95938	227.921567	-2.6167131	4.75E-42
YBR182C	SMP1	89.0591175	810.525738	3.18602278	5.46E-42
YEL060C	PRB1	509.498329	2618.41929	2.36154682	2.48E-39
YDR043C	NRG1	46.3740884	416.732767	3.16773171	2.44E-37
YOR389W	GO:0003674,	25.8973643	256.066451	3.30564107	1.46E-36
YJR004C	SAG1	22014.4824	5052.70038	-2.1233264	2.68E-36
YGL045W	RIM8	91.6759887	842.164446	3.19948615	3.86E-35
YEL040W	UTR2	10351.6241	2601.07168	-1.9926791	5.57E-32
YOL143C	RIB4	6919.02632	1740.25235	-1.9912725	1.72E-30
YKL216W	URA1	21282.0469	2922.84926	-2.8641896	1.70E-29
YPL088W	GO:0005575,	64.9309836	420.38449	2.69473047	9.20E-27
YNL274C	GOR1	255.067836	1169.47392	2.1969068	2.04E-25
YHL028W	WSC4	323.830677	16.9523355	-4.2556837	5.53E-24
YJR061W	GO:0003674,	169.967775	1169.26627	2.78227035	5.63E-24
YDR068W	DOS2	18.7970426	310.466425	4.04585966	1.06E-22
YNR044W	AGA1	3718.94452	777.423082	-2.2581214	1.08E-22
YDL241W	GO:0003674,	200.094193	19.0008646	-3.3965423	1.51E-22
YIL063C	YRB2	37.8939319	529.491137	3.80456779	7.20E-22
YJL196C	ELO1	5027.47443	583.890882	-3.1060632	2.46E-21
YNL065W	AQR1	2632.35936	812.020176	-1.696769	2.56E-21
YDR533C	HSP31	253.463454	986.847895	1.96104999	6.45E-21
YHL040C	ARN1	1358.45191	5007.46391	1.88211662	1.78E-20
YMR078C	CTF18	138.63417	625.910644	2.17467382	2.23E-20
YBR054W	YRO2	159.815737	19.9113305	-3.0047479	2.36E-20
YDL053C	PBP4	114.438537	499.178653	2.12498328	7.18E-20
YOL122C	SMF1	1272.91701	216.85734	-2.5533202	1.29E-19
YOR161C	PNS1	703.219092	201.062862	-1.8063276	1.49E-18
YOR010C	TIR2	104.099602	1669.86056	4.00369119	3.09E-18
YER001W	MNN1	2584.31609	879.918631	-1.5543405	4.02E-18
YKL096W	CWP1	1904.05424	175.579982	-3.4388743	9.79E-18
YDL002C	NHP10	55.2137908	275.938193	2.32124459	1.04E-17
YLR420W	URA4	7403.51257	2064.11859	-1.8426841	1.61E-17
YLR300W	EXG1	15969.1958	5848.04488	-1.4492654	1.62E-17
YNL225C	CNM67	65.6358259	305.923955	2.22061768	1.64E-17
YPL263C	KEL3	417.125298	1790.86227	2.10210167	1.88E-17

Table 4: 50 most differentially expressed genes in ∑1278b rim101Δ

id	Name	S288c wt	S288c rim10	log2FoldChar	padj
YHL027W	RIM101	936.888614	3.3449476	-8.1297501	1.72E-138
YBR296C	PHO89	5092.43527	442.067228	-3.526018	9.15E-82
YER011W	TIR1	654.669427	4.62157164	-7.1462392	3.60E-61
YOL126C	MDH2	1621.59491	233.21592	-2.7976753	2.31E-50
YCL026C-B	HBN1	96.8642814	863.619358	3.15635891	2.37E-46
YMR319C	FET4	1397.95938	227.921567	-2.6167131	4.75E-42
YBR182C	SMP1	89.0591175	810.525738	3.18602278	5.46E-42
YEL060C	PRB1	509.498329	2618.41929	2.36154682	2.48E-39
YDR043C	NRG1	46.3740884	416.732767	3.16773171	2.44E-37
YOR389W	GO:0003674,	25.8973643	256.066451	3.30564107	1.46E-36
YJR004C	SAG1	22014.4824	5052.70038	-2.1233264	2.68E-36
YGL045W	RIM8	91.6759887	842.164446	3.19948615	3.86E-35
YEL040W	UTR2	10351.6241	2601.07168	-1.9926791	5.57E-32
YOL143C	RIB4	6919.02632	1740.25235	-1.9912725	1.72E-30
YKL216W	URA1	21282.0469	2922.84926	-2.8641896	1.70E-29
YPL088W	GO:0005575,	64.9309836	420.38449	2.69473047	9.20E-27
YNL274C	GOR1	255.067836	1169.47392	2.1969068	2.04E-25
YHL028W	WSC4	323.830677	16.9523355	-4.2556837	5.53E-24
YJR061W	GO:0003674,	169.967775	1169.26627	2.78227035	5.63E-24
YDR068W	DOS2	18.7970426	310.466425	4.04585966	1.06E-22
YNR044W	AGA1	3718.94452	777.423082	-2.2581214	1.08E-22
YDL241W	GO:0003674,	200.094193	19.0008646	-3.3965423	1.51E-22
YIL063C	YRB2	37.8939319	529.491137	3.80456779	7.20E-22
YJL196C	ELO1	5027.47443	583.890882	-3.1060632	2.46E-21
YNL065W	AQR1	2632.35936	812.020176	-1.696769	2.56E-21
YDR533C	HSP31	253.463454	986.847895	1.96104999	6.45E-21
YHL040C	ARN1	1358.45191	5007.46391	1.88211662	1.78E-20
YMR078C	CTF18	138.63417	625.910644	2.17467382	2.23E-20
YBR054W	YRO2	159.815737	19.9113305	-3.0047479	2.36E-20
YDL053C	PBP4	114.438537	499.178653	2.12498328	7.18E-20
YOL122C	SMF1	1272.91701	216.85734	-2.5533202	1.29E-19
YOR161C	PNS1	703.219092	201.062862	-1.8063276	1.49E-18
YOR010C	TIR2	104.099602	1669.86056	4.00369119	3.09E-18
YER001W	MNN1	2584.31609	879.918631	-1.5543405	4.02E-18
YKL096W	CWP1	1904.05424	175.579982	-3.4388743	9.79E-18
YDL002C	NHP10	55.2137908	275.938193	2.32124459	1.04E-17
YLR420W	URA4	7403.51257	2064.11859	-1.8426841	1.61E-17
YLR300W	EXG1	15969.1958	5848.04488	-1.4492654	1.62E-17
YNL225C	CNM67	65.6358259	305.923955	2.22061768	1.64E-17
YPL263C	KEL3	417.125298	1790.86227	2.10210167	1.88E-17

Table 5: 62 genes with an on/off expression pattern between S288c and ∑1278b

SENSE: S2 Of Name					
id	Name	S288c wt	S288c rim10	log2FoldChar	padj
YBR115C	LYS2	1.64203189	2577.06044	10.6160285	1.22E-201
YBR294W	SUL1	9.07984983	148.629613	4.03290934	3.94E-23
YBR302C	COS2	1.07218817	432.401013	8.65566797	6.21E-81
YCL058W-A	ADF1	10.152038	108.829438	3.42222758	1.32E-05
YDR040C	ENA1	3.71890897	2491.82319	9.38810655	1.03E-09
YDR068W	DOS2	18.7970426	203.677065	3.43770593	5.15E-17
YDR312W	SSF2	16.8551641	101.782093	2.59422118	7.40E-06
YDR343C	HXT6	1.57453263	230.657421	7.19468425	1.39E-43
YEL021W	URA3	1.64203189	5690.9904	11.7589819	4.32E-106
YER029C	SMB1	19.1345389	127.732149	2.73887064	4.33E-10
YER053C-A	GO:0003674,	3.71890897	472.182238	6.98832053	1.78E-27
YFL057C	AAD16	4.72359788	263.353811	5.80097237	1.58E-48
YGL029W	CGR1	15.9478241	194.940819	3.61160471	1.33E-05
YGR142W	BTN2	19.666733	224.02867	3.50985415	7.55E-05
YGR213C	RTA1	13.7359484	120.021499	3.12726442	4.61E-14
YHR040W	BCD1	19.5693841	326.097802	4.05863446	6.83E-07
YIL161W	GO:0003674,	16.8175144	130.926521	2.96072097	1.51E-16
YIR027C	DAL1	18.0922004	555.682731	4.94082162	8.13E-68
YJL115W	ASF1	10.0845387	109.982024	3.44705074	4.76E-07
YJR159W	SOR1	1.07218817	236.570994	7.78557126	2.13E-50
YJR161C	COS5	1.07218817	162.00456	7.23933249	2.76E-14
YKR022C	NTR2	13.1661047	108.321343	3.04041705	7.55E-10
YLR153C	ACS2	1.07218817	20335.0692	14.2111242	0
YLR154C	RNH203	1.07218817	327.542131	8.25497855	1.15E-67
YLR163C	MAS1	1.07218817	769.121639	9.48650985	1.40E-118
YML062C	MFT1	11.7940699	157.836328	3.74229574	5.40E-11
YML132W	COS3	1.07218817	313.988458	8.19400959	3.26E-63
YMR227C	TAF7	14.9431351	129.915035	3.12001363	1.73E-05
YOR054C	VHS3	19.666733	119.179332	2.59930484	8.40E-07
YOR287C	RRP36	11.6590714	104.45426	3.16334654	4.67E-10
YOR390W	GO:0003674,	3.35156304	504.808078	7.23475712	3.73E-84
YPL056C	LCL1	18.6620441	106.204304	2.50866331	4.06E-09
SENSE: S2 On, ∑ off					
YDL037C		257.654857	3.22729545	-6.3189703	1.41E-41
YDR281C		658.284537	18.3005497	-5.1687524	2.90E-61
YER037W		226.440652	5.92768272	-5.2555209	4.11E-14
YHR033W		256.350322	19.4108058	-3.7231848	8.13E-19
YIL014C-A		137.831979	4.29992505	-5.0024553	1.78E-25
YKR103W		180.224962	2.15466584	-6.3861909	8.17E-38
YOR049C		100.680539	3.23199876	-4.9612144	1.40E-19

Antisense: S2 Off, ∑ on

YBR297W	2.35423069	128.134061	5.76625452	5.03E-24
YCL039W	13.7910672	127.38867	3.20743097	8.82E-18
YCL058C	16.4028443	104.881591	2.67674356	0.00204327
YDR452W	19.8144954	155.676242	2.97392066	7.34E-19
YEL021W	1.13419094	105.417244	6.53830352	2.37E-10
YEL022W	2.26838189	100.603792	5.47087733	1.05E-23
YFL033C	10.560192	222.581386	4.39762498	1.22E-34
YHR071W	8.81598114	208.143104	4.56131081	7.39E-27
YJL216C	2.26838189	117.474866	5.69454477	1.16E-26
YJR160C	1.13419094	138.063056	6.92751998	1.75E-34
YKR072C	5.58510592	257.320958	5.52584051	1.37E-45
YKR103W	1.13419094	165.626988	7.19013042	8.69E-31
YLL051C	14.0486136	165.987771	3.56257729	4.41E-17
YLR175W	12.218554	101.75882	3.05800838	1.01E-05
YOL156W	8.98767875	111.074999	3.62744176	7.34E-19
YOR202W	13.7052184	107.727299	2.97458667	2.83E-14
YPR008W	3.9267439	136.560363	5.12006148	5.25E-17

Antisense: S2 On, ∑ off

YBL005W	478.484116	4.72437729	-6.662203	2.42E-75
YIL169C	287.140163	12.7410421	-4.4942	2.25E-43
YLR343W	675.087935	5.2263683	-7.0131228	5.39E-89
YNR055C	234.221975	11.3643646	-4.3652875	9.14E-38
YNR056C	135.348431	16.1626251	-3.0659447	7.07E-18
YPR194C	108.977258	8.98370517	-3.6005727	6.92E-20

Table 6: Gene Ontology (GO) terms for genes differentially expressed between S288c and ∑1278b

GOID	GO_term	Cluster frequency
16491	oxidoreductase activity	94 out of 1207 genes, 7.8%
5515	protein binding	159 out of 1207 genes, 13.2%
988	protein binding transcription factor activit	43 out of 1207 genes, 3.6%
989	transcription factor binding transcription f	38 out of 1207 genes, 3.1%
3729	mRNA binding	53 out of 1207 genes, 4.4%
44822	poly(A) RNA binding	53 out of 1207 genes, 4.4%
8134	transcription factor binding	27 out of 1207 genes, 2.2%
9055	electron carrier activity	13 out of 1207 genes, 1.1%

References

Anders, S. & Huber, W., 2010. 'Differential expression analysis for sequence count data,' *Genome biology*, 11(10), p.R106.

Anders, S., Pyl, P.T. & Huber, W., 2014. 'HTSeq A Python framework to work with high-throughput sequencing data,' *bioRxiv*, 31(2), p.002824.

Argueso, J.J.L. et al., 2009. 'Genome structure of a Saccharomyces cerevisiae strain widely used in bioethanol production,' *Genome Research*, 19, pp.2258–2270.

Ashworth, J. et al., 2014, 'Structure-based predictions broadly link transcription factor mutations to gene expression changes in cancers,' *Nucleic Acids Research*, 42(21), pp.12973–12983.

Bailey, T.L. et al., 2009. 'MEME Suite: Tools for motif discovery and searching,' *Nucleic Acids Research*, 37, pp.202–208.

Barski, A. et al., 2007, 'High-Resolution Profiling of Histone Methylations in the Human Genome,' *Cell*, 129(4), pp.823–837.

Battle, A. et al., 2014. 'Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.' *Genome Research*, 24(1), pp.14–24.

De Boer, C.G. & Hughes, T.R., 2012. 'YeTFaSCo: A database of evaluated yeast transcription factor sequence specificities,' *Nucleic Acids Research*, 40(D1), pp.169–179.

Brem, R.B. & Clinton, R., 2002. 'Genetic Dissection of Transcriptional Regulation in Budding Yeast,' *Science*, 296(April), pp.752–756.

Britten, R.J. & Davidson, E.H., 1969. 'Gene Regulation for Higher Cells: A Theory,' *Science*, 165(3891), pp.349–357.

Calon, A. et al., 2015. 'Stromal gene expression defines poor-prognosis subtypes in colorectal cancer,' *Nature Genetics*, 47(February), pp.320–329.

Carbrey, J.M. et al., 2001. 'Aquaporins in Saccharomyces: Characterization of a second functional water channel protein,' *Proceedings of the National Academy of Sciences of the United States of America*, 98(3), pp.1000–5.

Chan, Y.F. et al., 2010. 'Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer.' *Science,* 327(5963), pp.302–5.

Chandler, C.H., 2010. 'Cryptic intraspecific variation in sex determination in Caenorhabditis elegans revealed by mutations,' *Heredity*, 105(5), pp.473–482.

Chandler, C.H., Chari, S. & Dworkin, I., 2013. 'Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution,' *Trends in Genetics*, 29(6), pp.358–66.

Chang, K. et al., 2013. 'The Cancer Genome Atlas Pan-Cancer analysis project,' *Nature Genetics*, 45(10), pp.1113–1120.

Cookson, W. et al., 2009. 'Mapping complex disease traits with global gene expression,' *Nature Reviews Genetics*, 10(3), pp.184–194.

Cubillos, F.A. & Billi, E., 2011. 'Assessing the complex architecture of polygenic traits in diverged yeast populations,' *Molecular Ecology*, 20, pp.1401–1413.

Dowell, R.D. et al., 2010. 'Genotype to Phenotype : A Complex Problem,' *Science* 328(April), p.2010.

Dowell, R.D., 2010. "Transcription factor binding in the evolution of gene regulation,' *Trends in Genetics*, 26(11), pp. 468-475

Dworkin, I. et al., 2009. 'Genomic consequences of background effects on scalloped mutant expressivity in the wing of drosophila melamogaster,' *Genetics*, 181(3), pp.1065–1076.

Ehrenreich, I.M. et al., 2010. 'Dissection of genetically complex traits with extremely large pools of yeast segregants,' *Nature*, 464(7291), pp.1039–42.

Fernandez, M. & Miranda-Saavedra, D., 2012. 'Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines,' *Nucleic Acids Research*, 40(10), pp.e77–e77.

Forbes, S. a et al., 2011. 'COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer,' *Nucleic acids research*, 39, pp.945–50.

Gallagher, J.E.G. et al., 2014. 'Divergence in a master variator generates distinct phenotypes and transcriptional responses,' *Genes and Development*, 28, pp.409–421.

Gimeno, C.J. et al., 1992. 'Unipolar cell divisions in the yeast S. cerevisiae lead to filamentous growth: Regulation by starvation and RAS,' *Cell*, 68(6), pp.1077–1090.

Golub, T.R. et al., 1999. 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,' *Science*, 286(5439), pp.531–537.

Gordon, K.L. & Ruvinsky, I., 2012. 'Tempo and Mode in Evolution of Transcriptional Regulation,' *PLoS Genetics*, 8(1), p.e1002432.

Göring, H.H.H. et al., 2007. 'Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes,' *Nature genetics*, 39(10), pp.1208–1216.

Hainer, S.J. et al., 2011. 'Intergenic transcription causes repression by directing nucleosome assembly,' *Genes & development*, 25(1), pp.29–40.

Harismendy, O. et al., 2011. '9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response,' *Nature*, 470(7333), pp.264–268.

Hindorff, L. a et al., 2009. 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits,' *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), pp.9362–9367.

Hon, G.C., Hawkins, R.D. & Ren, B., 2009. 'Predictive chromatin signatures in the mammalian genome,' *Human Molecular Genetics*, 18(R2), pp.R195–201.

Hongay, C.F. et al., 2006. 'Antisense Transcription Controls Cell Fate in Saccharomyces cerevisiae,' *Cell*, 127(4), pp.735–745.

Houseley, J. et al., 2008. 'A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster,' *Molecular Cell*, 32(5), pp.685–95.

Jansen, R.C. & Nap, J.P., 2001. 'Genetical genomics: The added value from segregation,' *Trends in Genetics*, 17(7), pp.388–391.

Kandoth, C. et al., 2013. 'Mutational landscape and significance across 12 major cancer types,' *Nature*, 502(7471), pp.333–339.

Kasowski, M. et al., 2010. 'Variation in transcription factor binding among humans,' *Science*, 328(5975), pp.232–235.

Kesseli, R. V, 1991. 'Identification of markers linked to disease-resistance genes by bulked segregant analysis : A rapid method to detect markers in specific genomic regions by using segregating populations,' 88(November), pp.9828–9832.

Kim, J., He, X. & Sinha, S., 2009. 'Evolution of regulatory sequences in 12 Drosophila species,' *PLoS Genetics*, 5(1).

King, M. & Wilson, A.C., 1975. 'Evolution at two levels in Humans and Chimpanzees,' *Science*,188(4184), pp.107-116.

Krogan, N.J. et al., 2003. 'Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II,' *Molecular and Cellular Biology*, 23(12), pp.4207–18.

Landry, C.R., 2005. 'Evolution and the Dysregulation of Gene Expression in Interspecific Hybrids of Drosophila,' *Genetics*, 171 pp. 1813-1822.

Langmead, B. & Salzberg, S.L., 2012. 'Fast gapped-read alignment with Bowtie 2,' *Nature Methods*, 9(4), pp.357–359.

Lawrence, M.S. et al., 2014. 'Discovery and saturation analysis of cancer genes across 21 tumour types,' *Nature*, 505(7484), pp.495–501.

Lawrence, M.S. et al., 2013. 'Mutational heterogeneity in cancer and the search for new cancer-associated genes' *Nature*, 499(7457), pp.214–8.

Leggett, R.M. et al., 2013. 'Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics,' *Frontiers in Genetics*, 4(DEC), pp.1–5.

Li, H. et al., 2009. 'The Sequence Alignment/Map format and SAMtools,' *Bioinformatics*, 25(16), pp.2078–2079.

Litvin, O. et al., 2015. 'Interferon α/β Enhances the Cytotoxic Response of MEK Inhibition in Melanoma,' *Molecular Cell*, 57, pp.784–796.

Mali, P. et al., 2014. 'RNA-Guided Human Genome Engineering,' *Science*, 823(2013), pp.823–827.

Matin, A. & Nadeau, J.H., 2001. 'Sensitized polygenic trait analysis,' *Trends in Genetics*, 17(12), pp.727–731.

Maurano, M.T. et al., 2012. 'Systematic Localization of Common Disease-Associated Variation in Regulatorty DNA,' *Science*, 337(September), pp.1190-1195.

McVean, G. a. et al., 2012. 'An integrated map of genetic variation from 1,092 human genomes,' *Nature*, 491(7422), pp.56–65.

Milloz, J. et al., 2008. 'Intraspecific evolution of the intercellular signaling network underlying a robust developmental system,' *Genes and Development*, 22(21), pp.3064–3075.

Montgomery, S.B. & Dermitzakis, E.T., 2011. 'From expression QTLs to personalized transcriptomics,' *Nature Reviews Genetics*, 12(4), pp.277–282.

Mortazavi, A. et al., 2008. 'Mapping and quantifying mammalian transcriptomes by RNA-Seq,' *Nature methods*, 5(7), pp.621–628.

Musunuru, K. et al., 2010. 'From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus,' *Nature*, 466(7307), pp.714–719.

Nadimpalli, S., Persikov, A. V. & Singh, M., 2015. 'Pervasive Variation of Transcription Factor Orthologs Contributes to Regulatory Network Evolution,' *PLOS Genetics*, 11(3), p.e1005011.

Nica, A. & Dermitzakis, E., 2013. 'Expression quantitative trait loci: present and future,' *Philosophical Transactions of the Royal Society B*, 368: 201203621.

Nicolae, D.L. et al., 2010. 'Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS,' *PLoS Genetics*, 6(4).

Nishizawa, M. et al., 2010. 'Pho85 kinase, a cyclin-dependent kinase, regulates nuclear accumulation of the Rim101 transcription factor in the stress response of Saccharomyces cerevisiae,' *Eukaryotic Cell*, 9(6), pp.943–951.

Paaby, A.B. & Rockman, M. V, 2014. 'Cryptic genetic variation: evolution's hidden substrate,' *Nature reviews. Genetics*, 15(4), pp.247–58.

Papait, R. et al., 2013. 'Genome-wide analysis of histone marks identifying an epigenetic signature of promoters and enhancers underlying cardiac hypertrophy,' *Proceedings of the National Academy of Sciences of the United States of America*, 110(50), pp.20164–9.

Pekowska, A. et al., 2011. 'H3K4 tri-methylation provides an epigenetic signature of active enhancers,' *The EMBO Journal*, 30(20), pp.4198–4210.

Perou, C.M. et al., 2000. 'Molecular portraits of human breast tumours,' *Nature*, 406(May), pp.747–752.

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

Pomerantz, M.M. et al., 2009. 'The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer,' *Nature genetics*, 41(8), pp.882–884.

Quinlan, A.R. & Hall, I.M., 2010. 'BEDTools: a flexible suite of utilities for comparing genomic features,' *Bioinformatics (Oxford, England)*, 26(6), pp.841–842.

Ramos, E.M. et al., 2014. 'Characterizing genetic variants for clinical action,' *American Journal of Medical Genetics, Part C: Seminars in Medical Genetics*, 166(1), pp.93–104.

Rands, C.M. et al., 2014. '8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage,' *PLoS genetics*, 10(7), p.e1004525.

Rockman, M. V & Kruglyak, L., 2006. 'Genetics of global gene expression,' *Nature Reviews Genetics*, 7(11), pp.862–72.

Romero, I.G., Ruvinsky, I. & Gilad, Y., 2012. 'Comparative studies of gene expression and the evolution of gene regulation' *Nature reviews. Genetics*, 13(7), pp.505–16.

Rudolph, H. & Hinnen, A, 1987. 'The yeast PHO5 promoter: phosphate-control elements and sequences mediating mRNA start-site selection. *Proceedings of the National Academy of Sciences of the United States of America*, 84(5), pp.1340–1344.

Rutherford, S.L. & Lindquist, S., 1998. 'Hsp90 as a capacitor for morphological evolution,' *Nature*, 396(6709), pp.336–342.

Ryan, O. et al., 2012. 'Global Gene Deletion Analysis Exploring Yeast Filamentous Growth,' *Science*, 337(6100), pp.1353–1356.

Schadt, E.E. et al., 2003. 'Genetics of gene expression surveyed in maize, mouse and man,' *Nature*, 205(October 2002), pp.1–6.

Schaefer, M.H., Wanker, E.E. & Andrade-Navarro, M. a., 2012. 'Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks,' *Nucleic Acids Research*, 40(10), pp.4273–4287.

Schmidt, D. et al., 2010. 'Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding,' *Science*, 328(5981), pp.1036–1040.

Storici, F., Lewis, L.K. & Resnick, M.A., 2001. 'In vivo site-directed mutagenesis using oligonucleotides,' *Nature Biotechnology*, 19, pp.773–776.

Stranger, B.E. et al., 2007. 'Population genomics of human gene expression,' *Nature Genetics*, 39(10), pp.1217–1224.

Stranger, B.E., Stahl, E. a. & Raj, T., 2011. 'Progress and promise of genome-wide association studies for human complex trait genetics,' *Genetics*, 187(2), pp.367–383.

Strope, P.K. et al., 2015. 'The 100-genomes strains , an S . cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen,' *Genome Research*, 25 pp.1–13.

Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P., 2013. 'Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration,' *Briefings in Bioinformatics*, 14(2), pp.178–192.

Tirosh, I. et al., 2009. 'A yeast hybrid provides insight into the evolution of gene expression regulation,' *Science,* 324(5927), pp.659–62.

The Chimpanzee Sequencing and Analysis Consortium, 2005. 'Initial sequence of the chimpanzee genome and comparison with the human genome,' *Nature*, 437(7055), pp.69–87.

Venter, J.C. et al., 2001. 'The sequence of the human genome,' *Science*, 291(5507), pp.1304–51.

Weishi, L. & Mitchell, A.P., 1997. 'Proteolytic Activation of Rimlp, a Positive Regulator of Yeast Sporulation and Invasive Growth,' *Genetics*, 145, pp.63–73.

Welter, D. et al., 2014. 'The NHGRI GWAS Catalog, a curated resource of SNP-trait associations,' *Nucleic Acids Research*, 42(D1), pp.1001–1006.

Westra, H.-J. et al., 2013. 'Systematic identification of trans eQTLs as putative drivers of known disease associations,' *Nature Genetics*, 45(10), pp.1238–1243.

Westra, H.-J. & Franke, L., 2014. 'From genome to function by studying eQTLs,' *Biochimica et biophysica acta*, 1842(10), pp.1896–1902.

Winge, O. & Laustsen, O., 1937. 'On two types of spore germination, and on genetic segregations in Saccharomyces demonstrated through single spore cultures,' *Cr Trav Lab Carlsberg Ser Physiol*, (24), pp.263–315.

Winzeler, E. a et al., 1999. 'Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis,' *Science* , 285(5429), pp.901–906.

Xu, Z. et al., 2011. 'Antisense expression increases gene expression variability and locus interdependency,' *Molecular systems biology*, 7(468), p.468.

Yvert, G. et al., 2003. 'Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors,' *Nature Genetics*. pp. 57–64.

Zhang, Y. et al., 2008. 'Model-based Analysis of ChIP-Seq (MACS),' *Genome Biology*, 9(9), p.R137.