Modeling Distributions of Chromosomal Modifications Using Chromosomal Features

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Joshua Adam Baller

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Chad Myers, Daniel Voytas

February 2012

Acknowledgements

J. Gao played an integral role in the production of the Ty5 and Ty1 datasets.

H. Wang, D. Mayhew and R. Mitra made additional Ty5 data available prior to publication.

M. Hickman made K. lactis ORC data available prior to publication

Also, thanks to J. Carlis for advice on data storage as well as to R. Bushman and N. Milani for advice on data processing and statistical approaches.

This dissertation is dedicated to:

My mother, father and sister

> You have given me a great deal of love and support. This dissertation is the culmination of years of work, long nights and little free time. I wouldn't have made it to this point without your patience and understanding.

My thesis advisors: Dan, Chad and (honorarily) Judy

> These discoveries were not made in a vacuum. You have been my sounding boards, my strongest allies and my harshest critics. I hope in my own career I can be equal to your examples.

And to all my mentors and advisors throughout the years

> Your encouragement and advice inspired me to push my limits and discover new frontiers.

Abstract

Chromatin plays a major role in the regulation and evolution of genomic DNA. The advent of high-throughput sequencing, and the subsequently increasing availability of sequencing data from chromatin immunoprecipitation experiments, is leading to a comprehensive view of the chromatin landscape in key model organisms such as *S. cerevisiae*. To date, little has been done to exploit the availability of such data. My work develops a logistic regression based framework capable of dissecting the observed distribution of a particular chromosomal modification. This framework models the observed distribution in terms of other known chromosomal features in the organism. I have applied this approach to the distributions of Ty5 and Ty1 retrotransposons, identifying previously unknown integration patterns. For Ty5, I identified integration, independent of the canonical mechanism, at sites of open DNA. For Ty1, I identified precise integration events on a single surface of nucleosomes found near Polymerase III transcribed genes. Additionally, a similar logistic regression approach was developed to predict origins of replication in terms of nucleosome patterning. This resulted in a 200-fold enrichment for origin sites and over 7000-fold enrichment when ORC occupancy data was considered. Together these studies present a general model capable of utilizing the available chromosomal data to provide either mechanistic models or site predictions in a variety of organisms.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Bp – basepairs

ChIP – Chromatin Immuno-Precipitation

ddNTPs - dideoxynucleotide triphosphates

DNA – Deoxyribonucleic Acid

FISH - Fluorescence *in situ* Hybridization

HIV - Human Immunodeficiency Virus

HML/HMR – *unclear, abbreviation used exclusively since at least 1970*. Possibly stands for Homothallic Left and Right from early screens (similar to HS gene names from heat shock strains)(Takamu and Oshima 1970).

IN – Integrase

LEDGF - Lens Epithelial-Derived Growth Factor

LTR – Long Terminal Repeat

MAT Loci – Mating Type Loci

MNase – Micrococcal Nuclease

NPS – Nucleosome Positioning Sequence

PCA – Principle Component Analysis

PCR – Polymerase Chain Reaction

POL – Polymerase

RNA – Ribonucleic Acid

RSS – Recombination Signal Sequences

SNP – Single Nucleotide Polymorphism

SVM – Support Vector Machine

TERT - Telomerase Reverse Transcriptase

Ty – Transposon in Yeast

1. Introduction

Chromosome biology in eukaryotes has historically been dissected from both the small-scale, DNA sequence (Lander, Linton et al. 2001) and protein occupancy, and from the very large-scale, chromosome karyotyping via dye banding and Fluorescence *in situ* Hybridization (FISH) (Rudkin and Stollar 1977). Both of these vantage points have promoted the concept of the chromosome as a linear, static entity. However, further dissection has shown that beneath the condensed shape of the chromosome and above the DNA sequence, there is a hierarchy of dynamic interactions regulating the inner workings of the organism (Misteli 2007).

Precisely targeted chromosomal modifications are an essential part of this dynamic character. Consider, for instance: the role of transcriptional activators which, after binding to specific DNA sequences, recruit a series of proteins driving nucleosome remodeling at transcription start sites. This has the effect of recruiting RNA polymerase and removing heterochromatin, all ultimately increasing gene expression (Kadonaga 2004). Consider that in mammals, precise V(D)J recombination is necessary for proper immune system function. V(D)J recombination assembles the antigen receptor by stitching together series of disparate loci. The recombination event is driven by the VDJ recombinase which is targeted to the site of action by a conserved Recombination Signal Sequences (RSSs) in the DNA at the recombination point. (Jung and Alt 2004). Finally, consider the selection of integration sites for retrotransposons and retroviruses. For both types of retroelements, continued survival requires transcription of the integrated element. Retrotransposons, which lack genes to infect new host cells, have an additional selective pressure to avoid integration sites that may lower the fitness of their host (Bushman 2003). In all the cases discussed, regardless of whether the chromosomal modifications affects the DNA or associated proteins, the modifications can have a significant and long-lasting impact on the expression of surrounding genes and local genome evolution. While transcriptional activation and

V(D)J recombination are well understood, there are many chromosomal modifications where it is unclear how the modifying activity is targeted to the modified sites. Chapters 1 and 2 of this thesis concern the targeting mechanisms of particular retrotransposons to specific regions of the genome and/or to specific positions within nucleosomes.

In part to address the question of how chromosomal modifications are targeted, the body of my work focuses on analyzing the distribution of chromosomal modifications with respect to other known features of the genome. I hypothesize that a single computational framework can be created to interrogate the distribution of any specific chromosomal modification and identify a succinct subset of chromosomal features associated with the modification. This hypothesis is supported by evidence in the following three chapters where my computational framework, consisting of a classifier, a database of chromosomal features and a selection of mathematical transforms, is applied to three distinct chromosomal modifications. This genome-wide investigation of chromosomal modifications is enabled by improvements in high-throughput sequencing technology. These improvements have led to the increasing availability of datasets describing features of the chromosome; a fact particularly true in model organisms. Machine learning techniques were applied to the available data in order to build a sparse, multidimensional model describing the observed pattern of chromosomal modifications.

The success of the computational framework is predicated on a rich array of interactions between known features of the chromosome. It is assumed that a subset of the interactions will be associated with the modification pattern under investigation, and thus will be detected by the framework. Without the associations there is nothing to detect. Furthermore, the value of my framework is based on a need within the scientific community for better means by which to analyze these interactions. The current literature justifies this predicate as it details interactions

between DNA sequence, DNA modifications, nucleosomes, DNA bound proteins and other chromosomal features.

Chromosomes and Chromatin in the Three Domains of Life

All three domains of life, Eukaryota, Bacteria and Archaea, use DNA, a double helix (Watson and Crick 1953) made up of a linear sequence of base-paired nucleotides, Adenine with Thymine and Guanine with Cytosine in anti-parallel strands, to encode the majority of their heritable information (Avery, Macleod et al. 1944). A cell's DNA is organized into chromosomes, packaged DNA helices that are passed to progeny cells. Chromosomes are a fundamental feature of living organisms and are found in all three domains.

Large-scale chromosomal structure varies between the three domains. The three major large scale differences concern the telomeres, the centromere and supercoiling. The telomeres are the terminal ends of linear chromosomes (Blackburn and Szostak 1984). Since, Archaea and Bacteria (Prokaryotes) typically have a single circular chromosome, while Eukaryotes have multiple linear chromosomes (Kates, Kushner et al. 1993), telomeres are generally only found in Eukaryotes.

Centromeres are the regions of the genome, often centrally located, used to tether the chromosome segregation machinery (Blackburn and Szostak 1984). The mechanism of centromere activity is best understood in Eukaryotes. In Eukaryotes centromeres serve as the attachment point for microtubules, protein filaments that pull chromosomes to opposites poles of a dividing cell during cell division (Dorn and Maddox 2011). An analogues filamentous structure, tied to the replication origin, has been identified for Bacteria (Livny, Yamaichi et al. 2007). Also, recent work in Archaea has identified a complex, with a single subunit sharing homology with the Bacterial segregation complex, that binds chromosome and forms filaments (Kalliomaa-Sanford, Rodriguez-Castaneda et al. 2012).

Lastly, circular genomes allow for the introduction of persistent supercoiling into the genome by topoisomerase enzymes. Supercoils are additional twists or writhes in the chromosome beyond that of the DNA helix. Positive supercoiling inhibits melting of the DNA helix while negative supercoiling promotes melting. Both positive and negative supercoils can cause writhe, the twisting of separate parts of the helix together, resulting in genome compaction (Koster, Crut et al. 2010). While supercoiling is particularly evident in circular genomes due to that lack of unbound chromosomal ends, supercoiling is also observed in linear genomes. As the DNA helix is unwound ahead of replication machinery, positive supercoils develop. Similarly, negative supercoils develop as the DNA is rewound behind the machinery (Liu and Wang 1987). Theoretically, these supercoils will eventually dissipate down the length of the chromosome due to the presence of unbound chromosomal ends. However, in practice the length of the chromosomes, as well as the presence of bound proteins and intra-chromosomal loops, prevents this dissipation (Koster, Crut et al. 2010).

The presence and absence of these large scale structures is associated with the types of proteins bound to the chromosomes in the different domains of life. The chromosomes of all three domains are bound in protein, referred to as chromatin, which is responsible for protecting, compacting and regulating transcriptional or repair activity on nearby DNA. The primary component of chromatin in Eukaryotes is the nucleosome. The nucleosome is a highly-conserved essential octomer made up of histone proteins, generally two copies each of H3, H4, H2A and H2B, which spools approximately 147bp of DNA (Clapier and Cairns 2009). Much of the nucleosomes regulatory activity is derived from the unstructured N-terminal ‘tail’ regions of the histones. These tails serve as a binding site for other proteins, many of which recognize specific covalent modification made to tail amino acids (Campos and Reinberg 2009). A structurally similar tetramer forms the nucleosomes in studied Archaea (Pereira, Grayling et al. 1997). In

contrast, in Bacteria, a disparate set of proteins, referred to as nucleoid proteins fulfill this role (Sandman, Pereira et al. 1998).

Another component of chromatin is the transcriptional machinery. Eukaryotic and Archaeal RNA polymerases contain up to 15 subunits (Huet, Schnabel et al. 1983) and share greater sequence similarity with one another in the large subunits than either does with Bacteria (Zillig, Palm et al. 1988). In contrast, the RNA polymerase of Bacteria is relatively simple, consisting of four subunits plus the σ initiation factor, which is responsible for targeting the polymerase to the transcription site (Sweetser, Nonet et al. 1987). Despite their greater similarity, the transcription machinery of Archaea and Eukaryotes show considerable divergence. In an analysis of 280 predicted transcription related proteins in Archaea, 168 had homologs only in Prokaryotes, 51 in Eukaryotes and 61 in both (Kyrpides and Ouzounis 1999). In all, these facts illustrate the distinct differences between the three domains of life.

The chromosomal modifications to which I have applied my computational model all occurred and were detected in Eukaryotes. As such, I will confine further discussion of chromosomal features to the Eukaryota Domain.

Chromosomal Features of Eukaryotes

Chromosomal features of a Eukaryotic cell range in size from a single base pair to a significant fraction of a chromosome. The smallest features, DNA modifications and paired bases are 1 bp in size. DNA sequence motifs can range in magnitude from 10^0 to 10^2 bp. Nucleosomes cover 1.47*10^2 bp and gene annotations range from 10^2 bp (yeast small ORF) (Kastenmayer, Ni et al. 2006) to at least 10^6 bp (large human gene) (Kent, Sugnet et al. 2002). For purposes of discussion, small-scale features encompass DNA sequence motifs, transcription factor binding footprints (Hesselberth, Chen et al. 2009) and DNA modifications. Intermediate-

scale describes nucleosomes, small nucleosome arrays and small genes. Finally, large-scale features encompass large genes, telomeres/sub-telomeres and broad features of the genome.

DNA Sequence Properties and Modifications

Different sequences of base pairs impart different physical properties to the DNA helix. The simplest of these properties is the melting temperature, the ease with which the two stands of DNA separate. This property is largely governed by the ratio of A/T base pairs in comparison to G/C base pairs. This effect derives from the strength of the three hydrogen bonds of a G/C pair relative to the two hydrogen bonds of the A/T pair. The DNA sequence also has numerous other effects on DNA properties, for instance minor and major groove size, stiffness and stacking energy (Friedel, Nikolajewa et al. 2009).

In most Eukaryotes, *S. cerevisiae* being an exception, direct methylation of cytosine in CpG di-nucleotides can influence histone modification and local chromatin packaging (Cedar and Bergman 2009). In particular, CpG methylation has been associated with repression of transcription and the formation of heterochromatin (Zemach, McDaniel et al. 2010). Thus it is clear that DNA sequence and modifications are tightly linked to the distribution of small and intermediate scale features of the chromosome.

Repetitive DNA Motifs

Particular repeating DNA sequences have been associated with other features of the genome. Consider, for instance, the telomeres. All vertebrates have a conserved telomere repeat (TTAGGG)n (Meyne, Ratliff et al. 1989). The ciliate Tetrahymena, in which telomeres were originally investigated, have the repeat (TTGGGG)n (Blackburn and Gall 1978). Most Eukaryotes have similar sequence, though a few standouts do exist. In the case of yeasts in the genus *Saccharomyces* such as *bayanus*, *paradoxus*, *cerevisiae*, *mikatae* and *exiguus*, the repeats are degenerate with the motif (T[G]2-3(TG)1-6) (Shampay, Szostak et al. 1984; Cohn,

McEachern et al. 1998; Teixeira and Gilson 2005). All of these sequences serve as a recognition site for Telomerase and its subunit Telomerase Reverse Transcriptase (TERT). As the replication fork approaches the end of the chromosome it is eventually unable to prime the lagging strand resulting in the loss of a fragment of terminal DNA. Telomerase along with its template RNA compensates for this loss by providing a template for telomere extension. (Yu, Bradley et al. 1990)

Interestingly, in the well-studied fruit fly, *Drosophila melanogaster*, no short repeat or telomerase is present. Instead the terminal sequence is made up of tandem copies of the retrotransposons HeT-A and TART and the telomeric end is maintained through targeted retrotransposition (Louis 2002). A similar process occurs in *S. cerevisiae* if the functioning of the telomerase is impaired. When telomerase is impaired an endogenous subtelomeric sequence, partly non-repetitive DNA found adjacent to the telomeres, known as the Y' element, is mobilized with the help of Ty1 retrotransposons. The resulting Y'-Ty1 hybrid recombines with and extends the existing telomeric sequence (Maxwell, Coombes et al. 2004).

Sites of repetitive DNA can also develop distinct chromatin. The subtelomeric regions of *S.cerevisiae* consist of heterochromatin, tightly packaged DNA, spreading from the telomere. Similar structures can be found spreading from the centromeres of *Schizosaccharomyces pombe.* These region of heterochromatin can also act in *trans* to nucleate heterochromatin at remote sites (Talbert and Henikoff 2006).

Repetitive DNA motifs do more than simply affect the intermediate-scale, positioning local proteins such as telomerase or nucleosomes. Through mechanisms like chromatin spreading, intermediate-scale effects can expand to a larger-scale, defining entire regions of the chromosome.

Nucleosomes and Histone Modifications

Nucleosomes are of intermediate size covering hundreds of basepairs of DNA. Crystal structures of nucleosomes bound to DNA have shown that 146-147 bp of DNA are wrapped around the histone core in two loops (Richmond and Davey 2003). Micrococcal Nuclease (MNase) digestion studies, where single stranded or bare DNA is preferentially digested, have shown that the nucleosome protects a similarly sized region from degradation (Kaplan, Moore et al. 2009). The protective property of the nucleosome also excludes other regulatory proteins from accessing the DNA. This is evidenced by the need for nucleosome clearance before the Pol II RNA polymerase can bind to a transcription start site (Fuda, Ardehali et al. 2009). Similarly, the positioning of a single nucleosome within an origin of replication can block origin functioning (Simpson 1990; Lipford and Bell 2001). Conversely, the deformation of DNA as it winds around the nucleosome has been implicated in improving access to DNAs major groove for integrating viruses like HIV (Pruss, Bushman et al. 1994; Wang, Ciuffi et al. 2007).

Nucleosomes are positioned on the chromosome by DNA sequence motifs and specific proteins, known as chromatin remodelers. Genome-wide, high-throughput studies of nucleosome positioning have used MNase nuclease digestion and gel-purification to specifically isolate DNA wrapped around nucleosomes. The isolated DNA has then been sequenced to identify sites with nucleosome occupancy (Lee, Tillo et al. 2007; Valouev, Ichikawa et al. 2008; Tsankov, Thompson et al. 2010). These studies have identified distinct regions of the genome, those that contain uniformly localized nucleosome and those that show significant delocalization (Mavrich, Ioshikhes et al. 2008).

As discussed previously, different DNA sequence patterns can affect the physical properties of the DNA helix (Friedel, Nikolajewa et al. 2009). Changes in DNA flexibility and curvature enhance nucleosome formation and minimize sliding (Satchwell, Drew et al. 1986;

Salih, Salih et al. 2007; Ioshikhes, Hosid et al. 2011; Wang, Bryant et al. 2011). The effects of particular sequences have been identified through a combination of *in vitro* and *in vivo* studies. These studies have identified two distinct Nucleosome Positioning Sequence (NPS) classes, the RR/YY class and the WW/SS class (R=A or G, Y=C or T, W=A or T, S=G or C) (Ioshikhes, Hosid et al. 2011). The RR/YY class frequently consists of AA and TT motifs alternating every 5bp with the R nucleotides being proximal to the nucleosome core and the Y nucleotide distal. The RR/YY class of NPSs has some innate curvature, causing the DNA to fit to the nucleosome, but also retains considerable flexibility. Due to the flexibility of the RR/YY NPSs they have been associated with more mobile nucleosomes (Salih, Salih et al. 2007; Ioshikhes, Hosid et al. 2011). The WW/SS NPS class is also characterized by a periodic motif every 5bp. However, the WW and SS sequences are associated with sites where the major groove of the DNA helix faces towards or away from the histone. The WW/SS NPS class exhibits innate curvature similar to that of the RR/YY but forms a stiffer structure. The stiffness of the WW/SS NPS class has led to its association with sites of well-positioned nucleosomes (Trifonov 2010; Ioshikhes, Hosid et al. 2011).

NPSs only describe ~50% of genome-wide nucleosome positioning. A significant percentage of the remaining positioning is thought to be driven by chromatin remodelers and barrier proteins (Valouev, Ichikawa et al. 2008). The effect of nucleosome remodeling is particularly evident at sites of gene transcription, where nucleosomes are found to be similarly localized across a population of cells. In *S. cerevisiae,* a particular chromatin remodeler is known to bind at the UASg, a quick acting regulatory locus. The remodeler, known as RSC, positions a nucleosome at the UASg in nearly every cell in the analyzed population (Wang, Bryant et al. 2011). Evidence suggests that given a firmly placed barrier, such as those found at the UASg, chromatin remodelers will than pack nucleosomes into a tight array against the barrier. This

remodeling activity will often override sequence based positioning. (Mavrich, Ioshikhes et al. 2008; Sadeh and Allis 2011).

Nucleosomes have roles beyond that of a physical block to other proteins' activities. Nucleosomes also play a central role in the condensation and organization of DNA. For human cells, it is estimated that chromosomal DNA without packaging would require ~4 * 10^7 μm^3 of space within the cell. However, given that the human cell nucleus has a volume of ~100 μm^3, a compaction of 5 orders of magnitude is needed simply for the chromosomes to fit within the cell. Nucleosomes provide the first level of packaging, providing sevenfold compaction simply by acting as a spool for DNA (Bloom and Joglekar 2010). The role of nucleosomes in the next level of compaction, the 30nm fiber, is unclear. Studies have suggested the presence of a nucleosome solenoid, however electron microscopy has been unable to valid that hypothesis. In fact, a more recent study using cryo-electron microscopy, where the sample is first vitrified to preserve native structure, suggests a lack of any consistent higher-order structure. (Woodcock and Ghosh 2010)

Nucleosomes act in an organizational capacity by serving as recruitment sites for other proteins. This additional function is facilitated by covalent changes to the histone tails. Two such modifications are acetylation or methylation. Acetylation modifications are added to Lysine residues by Acetyltransferases (HAT) and removed by Histone Deacetylases (HDAC). Methylation modifications are added to either Lysines or Arginines by Histone Methyltranserases (HMT). Until recently it was unclear if histone demethylation occurred *in vitro*; however both Lysine and Arginine Demethylases have been identified, all with a common JmjC domain. Lysine demethylases have been identified in yeast, human and fruit fly while arginine demethylases have, currently, only been found in mice (Chang, Chen et al. 2007; Metzger and Schule 2007). Other, equally important modifications include ubiquitinilation, sumoylation and phosphorylation.

Histone acetylation has often been associated with increased transcription while methylation has been associated with decreased transcription. It should be understood, however, that with multiple potential modification sites and the possibility for di- or tri- modifications at each site, the literature on the histone code is still incomplete. (Kouzarides 2007). The protein responsible for binding the TATA sequence motif in human, the TATA Binding Protein (TBP), and two associated proteins, TBP-associated factors, (TAF1 and TAF3), bind directly to histone tails that are both diacetylated and trimethylated (Gardner, Allis et al. 2011). Interruption of this interaction resulted in decreased transcription, implicating histone modifications in transcriptional regulation. Similar interactions have been identified between histone modifications and components of the DNA repair pathway. One such example is the interaction of Crb2 with phosphorylation and methylation sites on histones in *Schizosaccharomyces pombe* (Du, Nakamura et al. 2006).

Nucleosomes protecting and package DNA, as well as acting as organizational sites. Additionally, when small-scale features, such as DNA motifs or chromatin remodelers, fix a nucleosome in place, it can result in changes to the nucleosome patterning over a considerably larger scale. The role of nucleosomes as an intermediary between chromosomal features makes them a particularly interesting part of the computational framework.

Genes and gene organization

A primary role of DNA sequence is to encode protein sequences, as encapsulated in the central dogma of molecular biology. The central dogma state that, in general, information flows from DNA to RNA to protein. This occurs by the transcription of DNA by an RNA polymerase, forming RNA, and the translation of RNA by the ribosome, forming a protein. (Crick 1970). The DNA sequence coding for a protein is generally referred to as a gene or an open reading frame (ORF). However, as already discussed, DNA sequence has roles beyond coding for genes,

including effecting chromatin binding and positioning. One effect on chromatin, specific to genes, is the binding of transcription factors. Transcription factors bind to particular DNA sequences, or motifs, regulating the transcription of nearby genes. This activity is often derived from their role in recruiting chromatin remodelers or directly recruiting the polymerase machinery (Fuda, Ardehali et al. 2009). For example, the well characterized TATA binding motif, found 25bp just upstream of many Eukaryotic transcription start sites, promotes DNA melting and the binding of the RNA Pol II pre-initiation complex (Smale and Kadonaga 2003).

Genes have been identified chromosome wide based on specific initiation and termination sequences as well as protein occupancy data. More recently, it has become possible to detect genes by sequencing all mRNA in the cell and mapping it back to genomic sites (Wang, Gerstein et al. 2009). While genes have a primary function, the production of RNA transcripts, they also define a collection of distinct chromatin domains. For instance, there are the regulatory region, the protein coding region and the site of termination. While the make-up of these regions does vary between genes, they are generally more similar than dissimilar. As such gene annotations describe not only genomic spans with a particular activity but also the arrangement of a series of sub-features within the gene (Fuda, Ardehali et al. 2009).

Chromosome Conformation

The largest feature that can be considered in the chromosomal context is the three-dimensional structures of the chromosomes themselves. Recent analyses have identified sites of intra- and inter- chromosomal interaction. These interactions were then used to build a constraint system describing the shape of the chromosomes within the nuclear envelope. Specific studies include, the three-dimensional conformation of *S. cerevisiae* (Duan, Andronescu et al. 2010) and *H. sapiens* (Lieberman-Aiden, van Berkum et al. 2009) genomes. These studies show the close packing of centromeres toward the center of the chromosomal bundle, with telomeres on the

outside. Information about inter- and intra- chromosomal interactions can be used to better understand patterns occurring over large scales.

Detection of DNA Modifications

The detection of whole-genome patterns of DNA and chromatin modification, like those discussed above, requires whole-genome high-throughput techniques. For DNA modifications there are a few options, some, such as microarrays, involve detecting differences from a reference sequence, others, sequencing sites of interest. While sequencing provides more information than measures of difference, only recently has sequencing reached sufficient capacity to be of use for whole genome analyses (t Hoen, Ariyurek et al. 2008).

The most common technique for detecting DNA sequence differences has been the microarray (Schena, Shalon et al. 1995). Microarrays are slides spotted with a grid of various DNA strands, half duplexes, referred to as probes. A second set of half-duplexes, referred to as targets, are then washed over the plate. Because of the tendency for DNA strands to form duplexes, complementary or near-complementary strands will anneal to one another. The abundance of hybridized DNA is then quantified, often through the covalent attachment of fluorophores to targets. Using fluorophores, quantification is based the fluorescent intensity of each spot (Schena, Shalon et al. 1995). To distinguish differences in DNA sequence from differences in hybridization strength, which varies considerably by sequence (SantaLucia and Hicks 2004), usually a control target is mixed with a case target under investigation. Different fluorophores distinguish the two targets so that the relative abundance of the two fluorophores provides the relative abundance of a particular sequence in the case set while controlling for hybridization strength (Shalon, Smith et al. 1996). Different selections of probes can be used to address different questions. A microarray with probes selected to represent evenly spaced regions of a genome detects duplications or deletions of regions of the chromosome (Pinkel, Segraves et

al. 1998). Conversely, arrays to detect single nucleotide polymorphisms (SNPs), single basepair differences between genomes, include multiple spots representing different sequence permutations of specific single loci in the genome (Syvanen 2001) to detect sequence variations at a given locus. However, due to limitation in the number of spots on a single microarray slide, it has been difficult to explore both large-scale and small-scale variation simultaneously. Furthermore it is only possible to explore small-scale variation in a comprehensive, genome-wide manner at loci that have been selected for analysis. This leaves out loci that may differ but have not been previously identified as important for a given study.

DNA sequencing returns the exact sequence of a locus and as such provides more detailed data than microarrays which report the similarity of the sequence relative to the probes on the array. However, DNA sequencing has been slow to overtake microarrays as the method of choice for genome-wide studies. This has been due to difficulties in scaling DNA sequencing to levels needed to efficiently sequence a genome's worth of sequence data. Historically, the first technique for directly sequencing purified DNA was the Maxam-Gilbert method published in 1977. The Maxam-Gilbert method required radioactive labeling of one end of the DNA strand. A series of ionizing chemicals were then used to break the DNA strand at specific bases, with concentrations selected to only cause one break per strand. The resulting fragments were then run in lanes on a gel and visualized via autoradiography. The nucleotides at each position in the sequence could be determined by a combination of the distance each fragment traveled and the lane it was in. This method was capable of sequencing 100bp of DNA. The cost per basepair is unclear but given the need for radioactive isotopes and toxic chemical the method was not particularly scalable (Maxam and Gilbert 1992). The Maxam-Gilbert method was quickly replaced by chain termination methods, also known as Sanger sequencing, which are still used today for most low throughput applications. The Sanger method relies on dideoxynucleotide

triphosphates (ddNTPs), rather than harsh chemicals, to produce DNA fragments. Four different sequencing reactions are prepared and each is spiked with a different ddNTP (ddATP, ddGTP, ddCTP or ddTTP). The ddNTPs are nucleotide analogues which, if incorporated into the DNA strand, prevent further stand elongation. Due to the presence of both dNTPs and ddNTPs in each reaction mixture, only a fraction of strands terminate at each site. As a result the sequence can be read by gel electrophoresis, with a method similar to that of Maxam-Gilbert. The maximum length sequenced in the original paper was 110bp (Sanger, Nicklen et al. 1977) . However, even with the Sanger method improving the sequencing process, by the time of the next major advancement in DNA sequencing technology in 1986, only 4,000,000 bp of DNA had been sequenced. In 1986, fluorescent markers were attached to the different ddNTPs allowing for the detection and differentiation of bands by an automated computer system. The maximum sequencing length at that point was approximately 500bp (Smith, Sanders et al. 1986). The final development that set the stage for modern Sanger sequencing was capillary electrophoresis. Capillary electrophoresis allowed size separation of DNA fragment in small quantities and without the need for gels (Monnig and Kennedy 1994). However, even with the development of sophisticated parallelized Sanger sequencers capable of sequencing millions of bases a week (Lander 2011), the cost of sequencing a megabase of DNA remained above $500 until mid 2007 (Wetterstrand 2011).

There are additional limitations, beyond cost, in using traditional sequencing. Traditional sequencing can only detect the most prevalent sequence in a sample. The presence of sequence variation will lead to the detection of multiple fluorescent dyes and thus, multiple nucleotides at a particular position. As such, if variation in the sequence is suspected, the amplicons are sub-cloned into a series of plasmids. Then, either the amplicons or a sample of the resulting plasmids are sequenced (Shendure and Ji 2008). Genome-wide sequencing using this method has been

quite difficult given that PCR, sub-cloning and sequencing are limited to 10^3 to 10^4 basepair per colony while even the *S. cerevisiae* genome is approximately 1.2*10^7 bp. The human genome project overcame the barriers of traditional sequencing through concerted effort and large sums of money (Lander, Linton et al. 2001). The use of similar techniques for re-sequencing or detection of *de novo* DNA modifications genome-wide is simply not cost or resource effective. Today's electrophoretic sequencers may be able to sequence hundreds of distinct, sub-cloned sequences simultaneously but they still cannot match the ability of high-throughput sequencers to read millions of distinct sequences without a separate isolation step. However, electrophoretic sequencers maintain one edge, they still consistently produce longer sequence reads (~700bp) than high-throughput techniques (Lander 2011) and have a faster turn-around time.

Recent technological advances have provided a series of new cost effective genome-wide sequencing options. Collectively, these new sequencing options are referred to as high-throughput sequencing. Since the dawn of the high-throughput sequencing era, the cost of DNA sequencing has dropped dramatically from a cost of $1000 per megabase (10^6 bp) in July of 2007 to $10 per megabase in July 2008. Today, sequencing costs are less than $0.10 per megabase and continue to drop (Wetterstrand 2011). With this rapid decrease in price has come an explosion in the total volume of sequencing, evidenced by four orders of magnitude increase in total basepairs submitted to the EMBL/NCBI Sequence Read Archive (SRA) since 2007, putting the current total at 166 trillion bases (Leinonen, Akhtar et al. 2012). Furthermore, the size of the SRA is on track to double within the next year. Additionally the NCBI Gene Expression Omnibus (GEO) archive has been expanded to accept any and all published high-throughput functional genomic data (Barrett, Troup et al. 2009).

High throughput sequencing encompasses a variety of new sequencing technologies, each with distinct benefits and shortcomings. Perhaps the best known offerings are 454 Life Sciences'

pyrosequencing method (Margulies, Egholm et al. 2005) and Illumina's reversible dye terminator method (Bentley, Balasubramanian et al. 2008; Mardis 2008). The 454 sequencing method uses a small fiberoptic slide containing millions of individual wells. Small beads, each bound with a distinct template sequence to be analyzed, are loaded into these wells along with reaction enzymes, in particular a high-fidelity polymerase (Margulies, Egholm et al. 2005). Each of the four deoxynucleotide triphosphates (dATP, dTTP, dCTP, dGTP) are cycled into the wells and, if not incorporated into the growing DNA strand, are degraded before the next dNTP is introduced. Concurrently, the polymerase molecule synthesizes a new strand based on the supplied template molecule. If the introduced dNTP matches the nucleotide needed to continue synthesis it will be incorporated into the strand, resulting in the release of a molecule of pyrophosphate. A separate bead is included in each well, containing ATP sulfurylase and luciferase. These enzymes jointly convert the pyrophosphate molecule into photons detectable by optical sensors (Mardis 2008). Illumina's reversible dye terminator method is based on a similar sequence-by-synthesis concept. However, in the case of Illumina sequencing is performed by binding template sequences to a slide and first amplifying them into a large cluster on the plate. From there 3'-OH blocked dideoxynucleotide triphosphates (ddNTPs) are incorporated by DNA polymerase (Bentley, Balasubramanian et al. 2008). The blocked nucleotides prevent DNA stand extension, ensuring only a single nucleotide is added. Each nucleotide is fluorescently marked so it can be distinguished and read by laser excitation. The 3'-OH blocking and fluorescent labels are then reversed allowing extension to continue (Mardis 2008).

In general the 454 pyrosequencing method produces longer sequence reads (150 to 500bp) but fewer reads overall. Additionally, 454 exhibits a higher insertion/deletion rate when the same nucleotide occurs multiple time consecutively within a sequence (Mardis 2008). Conversely, the Illumina sequencers average reads of about 75bp. However, paired end reads, where both ends of

the read are sequenced 75bp deep, effectively double the read length. With shorter reads, certain sections of the genome become indistinguishable due to regions of repetitive or duplicated sequence. Thus as the length of sequence reads becomes longer, more of the sequences becomes uniquely mapable and more of the genome is covered. With more reads, there is a greater chance that a low frequency modification will be sequenced, thus increasing the sensitivity of the experiment. Lastly, the discrete nature of sequence based detection also minimizes oversaturation, due to an increased linear detection range (t Hoen, Ariyurek et al. 2008).

Detection of Chromatin Modifications

The detection of Chromatin modifications is a more involved procedure. The primary means of detecting protein occupancy at a genomic location is Chromatin Immuno-Precipitation (ChIP). ChIP uses an antibody against the chromatin feature of interest to separate the feature of interest from the rest of the genome. Since chromatin is bound with ionic bonds either directly or via an intermediate to a segment of DNA, the isolated chromatin is still bound to its target DNA. The DNA is then separated from the proteins and sequenced to identify the distribution of the protein of interest on the DNA (Thorne, Myers et al. 2004). In most cases, the ionic bonds holding the chromatin to the DNA are not strong enough to survive the pull down (Orlando 2000). As a result, formaldehyde is used as a reversible cross-linker, producing temporary covalent bonds between amino groups stabilizing the DNA-chromatin interaction. Cross-links are removed after the pulldown using high salt and high temperature incubation. The result is DNA sequences enriched for sites of protein occupancy. The identity of these sequences is then determined by either high-throughput sequencing (Barrett, Troup et al. 2009) or microarray analysis (Schena, Shalon et al. 1995).

Importantly, sequence reads from many different studies of chromosomal modification are available from sites such as the SRA and GEO. The reads submitted to these sites include studies

of chromatin modification and DNA accessibility, as well as DNA modifications. The availability of these studies greatly facilitates the analysis of other chromosomal features.

Algorithms in the Literature

While increased sequencing capacity has been a boon for the biological community, there has been a bottleneck in obtaining insights into the large amount of data. The central problem is a paucity of novel tools for analyzing the newly produced dataset in the context of other, publicly available, datasets. Those tool that do exist narrowly focus on specific type of data. To date, most studies regarding the targeting of chromosomal modifications have focused on DNA sequence based mechanisms. As a result, an impressive array of tools are available for detecting and analyzing sequence based features. Perhaps the best known set of sequence analysis tools is the MEME suite (Bailey, Boden et al. 2009), consisting of the MEME algorithm (Bailey and Elkan 1994), MAST (Bailey and Gribskov 1998), GLAM (Frith, Saunders et al. 2008) and an assortment of other algorithms. The MEME suite contains tools for sequence motif discovery, search and comparison and is generally used to investigate sequence-based targeting, such as that of transcription factors. Other sequence-based techniques have been applied such as k-mer frequency analyses (Zhu, Byers et al. 2009), where the prevalence of fixed length sequences is evaluated. These techniques are highly specialized for DNA sequence analysis and would have to be substantially reworked to be applied to other data types. The focus on DNA sequence analysis was understandable given the prevalence of whole genome sequences and few whole genome chromatin studies. However, the SRA now contains a fairly comprehensive picture of the chromatin state in model organisms such as *S. cerevisiae* and tools need to be developed to take advantage of this data. *S. cerevisiae* now has multiple whole genome nucleosome maps (Lee, Tillo et al. 2007), ChIP studies of all transcription factor binding sites (Zhu, Byers et al. 2009), ChIP studies of histone acetylation and methylation status (Kurdistani, Tavazoie et al. 2004;

Pokholok, Harbison et al. 2005), as well as binding site distributions for proteins such as ORC and MCM.

A challenge with the diverse range of studies is the need for a framework capable of handling diverse data types. Importantly, a series of papers from Frederic Bushman's lab at University of Pennsylvania provided solutions to this challenge. They applied logistic regression analysis to a modification targeting problem, the analysis of the distribution of retroviral integrations (Mitchell, Beitzel et al. 2004). Specific retroelement distributions included HIV, MLV, ASLV, L1, AAV and Sleeping Beauty (a retrotransposon). This study however, focused primarily on sequence-based features and not chromatin features. The only chromatin based feature was DNAase sensitivity, a general measure of chromatin occupancy at a site on the chromosome. Other features included whether the sites were in a gene or an exon, the gene density in varying ranges from the integration sites, the level of expression of nearby genes and the presence of CpG islands. Additionally, they considered the sequence composition in the 20bp window centered on the integration sites (Berry, Hannenhalli et al. 2006).

The consideration of chromatin in targeting is important, particularly given a number of well-known targeting mechanisms that rely on tethering to chromatin. Some cases to consider are the tethering of the HIV virus to the LEDGF transcription factor (Shun, Botbol et al. 2008) and the targeting of Ty3 and Ty5 retrotransposons to the TFIIIB subunit of Pol III and Sir4 protein, respectively, in S. cerevisiae (Xie, Gai et al. 2001; Yieh, Hatzis et al. 2002). Additionally, chromatin affects the accessibility and conformation of DNA, promoting or inhibiting chromosome modifications. This is evident in HIV, where a secondary determinant of targeting is points of DNA distortion found on DNA strands wrapped around histones (Pruss, Bushman et al. 1994).

Some techniques have been established explicitly for chromatin analysis, in particular to analyze the 'histone code' (Strahl and Allis 2000). The histone code is a series of acetylation, phosphorylation and methylation marks occurring on particular amino acids on particular histones. There is significant evidence that the presence or absence of certain marks plays in role in the regulation of local modifications either by providing a binding site for a particular chromatin element or by adjusting the binding strength of the histone to DNA (Hecht, Laroche et al. 1995; Dhalluin, Carlson et al. 1999). Many of the techniques applied have used an unsupervised approach, looking for clusters of similar histone marks and determining if they are associated with particular effects (Kurdistani, Tavazoie et al. 2004). Two examples of supervised approaches applied a Naïve Bayes Classifier (Liu, Kaplan et al. 2005) and associative rule mining (Wang, Dai et al. 2010), to try and identify histone marks associated with increased transcription.

Model Selection

There are many models theoretically capable of modeling a response variable (the modification of interest) in terms of a set of independent variables (the features). Only a subset of these has received attention in the current biological literature, including naïve Bayes classifiers (Liu, Kaplan et al. 2005), associative rule mining, hidden Markov models and others. There is not a single set of criteria that allow a model to be applied to biological application, but rather specific criteria based on specific applications. However, given that biological data can be encoded on a variety of scales, for instance: continuous, discrete, categorical or binary. It is often convenient if a model can handle diverse data types.

The field of possible models can be narrowed by articulating the goals and assumptions of the model. In particular, information is available distinguishing sites of modification from sites without modifications; this suggests the use of a supervised learning approach, which excludes pattern identification techniques such as Principal Component Analysis (PCA) (Jolliffe 2002),

hierarchical clustering (Johnson 1967) or associative rule mining (Wang, Dai et al. 2010) which primarily group features within the feature set. Nonetheless, as the set of available features becomes more complete, these grouping techniques may be useful in order to catalog the full range of feature patterns found within an organism.

An additional question to consider is the nature of the supervised learning task. Is the chromosomal modification under investigation a discrete event, perhaps binary in occurrence or non-occurrence? Or, does it exhibit a continuous spectrum of severity? Classification models are applied to discrete events while regression models are applied to continuous events. Classification is generally an easier modeling problem; if a continuous distribution can be reasonably coerced to a discrete distribution, classification is desirable. In order to discretize data a threshold is chosen and sites above the threshold are determined to be modified sites while sites below the threshold are determined to be non-modified sites. One biological justification for this approach is to assume that the rate of modification across the genome is roughly bimodal; some sites form a baseline while others are significantly elevated. There may not always be a clear division, but assuming a small number of factors are driving the targeting of the modification, the division is likely a fair one. However, biological events are rarely truly discrete. For example, a DNA mutation, while discrete when observed as a single instance, may occur at different rates at different positions. As most occurrences are investigated, the rate of occurrence becomes a continuous distribution over the genome.

Lastly, there exist two distinct groups of models: discriminative and generative. Discriminative models simply describe the chromosomal modification in terms of the conditional distribution based on the feature set. Generative models describe the full joint distribution. Only the conditional distribution is necessary for classification and regression (Bouchard and Triggs 2004). Proper training of a joint distribution can require significantly more data, particularly for

high dimension problems. For the problem under consideration there is a clear upper bound on the number of data points available and, in a sense, it has already been reached. In most genome-wide studies a single base pair of the chromosome equates to a single data point and the chromosome has a roughly fixed size. Multiple organisms could be considered simultaneously to increase the number of data points available. However, this approach would complicate inferring biological mechanism, particularly if the biological mechanism under investigation differs between the organisms. Since, the machine learning task at hand is of high dimensionality a discriminative model is more practical.

Of the discriminative classifiers, logistic regression, decision trees and support vector machines (SVMs) are particularly well known and widely applied. Logistic regression is mathematically simple. The least-squared error of the equation $f(z) = \frac{1}{1+e^{-z}}$ is minimized by adjusting the vector of regression coefficients z. The vector of regression is, $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$, where n is the number of features in the model, x_i is the value of a particular feature, i, and the coefficient $\beta_i$is the contribution of feature i to the model's prediction (Dreiseitl and Ohno-Machado 2002). There are multiple varieties of decision trees, however at their core they all rely on a similar mechanism. The algorithm starts with the full dataset and finds a subset that, with its complement, minimizes the total impurity of the two subsets. The two subsets are defined as child nodes of the original subset and the algorithm is applied recursively to each node. This process continues until a stopping criterion is met, such as the existence of no subset capable producing a minimum increase in node purity (Loh 2011). The basic SVM model identifies a hyperplane that perfectly separates two classes while maintaining a maximum margin from each. In practice a 'soft margin' SVM is used, which identifies a hyperplane that divides the two classes as cleanly as possible, as measured by a penalty function. Through the use of kernel

transforms the SVM decision space can be made non-linear allowing for more complex classification tasks (Cortes and Vapnik 1995).

From this list, logistic regression was chosen as the primary analyses tool. While logistic regression often exhibits weaker predictive power than kernel-based SVMs, it is generally easier to infer the features used in prediction when using the logistic regression model. SVMs complicate inference in that they return the support vectors of the separating hyperplane rather than the separating features. (Dreiseitl and Ohno-Machado 2002). For logistic regression, the features used in prediction indicate chromosomal features associated with the analyzed distribution, providing candidate biological mechanisms. Additionally, the choice of logistic regression was influenced by its use and successful application in current biological literature (Berry, Hannenhalli et al. 2006). On the other hand, decision trees are highly flexible and provide easy inference, though they often exhibit lower predictive power than logistic regression models (Long, Griffith et al. 1993; Bensic, Sarlija et al. 2005). Furthermore, decision trees tend to be unstable; small perturbations to the training data can have large effects on the resulting model. A close relation to decision trees, random forests show better prediction and greater stability than decision trees but, again, complicate inference (Dietterich 2000).

Sparse Modeling

The focus of the described analysis is to predict the primary biological mechanisms behind the observed distribution of chromosomal modification. Identification of important chromosomal features is based on the assumption that only a small number of the features used in the models truly drive targeting. Other features may have a minor effect, or, due to random chance, may show a spurious association with the modification in question. The greater the number of features provided for model training, the greater the probability of spurious associations, eventually leading to model over-fitting. To simplify the trained model and reduce over-fitting, a

regularization parameter is applied to the logistic regression model in the form of the Least Absolute Shrinkage and Selection Operator (LASSO) (Friedman, Hastie et al. 2010). LASSO penalizes models proportionally to the value of the L1 norm of the β coefficients. Norms describe ways of measuring a vector's length and must apply to all vectors in a given vector space. The L2 norm, otherwise known as the Euclidean norm, is perhaps the most familiar to non-mathematicians as it is the n-dimensional generalization of the Pythagoras theorem. In practice, LASSO is implemented by setting a series of L1 constraints, referred to as lambdas and determining the optimal model for each constraint.

Other regularization methods include Ridge regression and elastic net, which simply use the L2 norm or a combination of L1 and L2 norms respectively. LASSO is preferable to ridge regression because, while both reduce over-fitting, ridge regression does not promote a sparse model (Hoerl and Kennard 1970; Le Cessie and Van Houwelingen 1992). The elastic net provides both the sparsity of LASSO and retention of redundant features found in ridge regression however it also adds an additional parameter that must be estimated (Friedman, Hastie et al. 2010). Since estimating additional parameters can increase the likelihood of model over-fitting, LASSO regularization was chosen.

Applications of the Model

The logistic regression model was applied to two classes of biological problems: modeling of retrotransposon targeting and prediction of origin of replication sites. Retrotransposons are mobile genetic elements: fragments of DNA capable of replicating and moving within a host genome as reviewed in (Beauregard, Curcio et al. 2008). They replicate through a multistep process: transcription of an integrated element by a host polymerase, reverse transcription of the transcript by an encoded reverse-transcriptase and integration of the resulting cDNA by an encoded integrase. Retrotransposons differ from retroviruses in that they lack the proteins

necessary to efficiently exit a cell and infect a new host. Random integration of retrotransposon copies would inevitably lead to integration into, and disruption of, an essential gene leading to death of the host and subsequent 'death' of the retrotransposon. To avoid this fate retrotransposons have evolved precise targeting mechanisms with which they avoid sensitive loci (Lesage and Todeschini 2005). Understanding the mechanisms of retrotransposon targeting has potential benefits beyond simple scientific advancement. In particular, retrotransposons can be made to target particular loci via tethering of a targeting determinant (Wang, Johnston et al. 2007) or modification of the targeting domain.

Origins of replication are the initiation sites for the replication of genomic DNA (Masai, Matsumoto et al. 2010). Origin sites affect local genome evolution due to proofreading biases (Marsolier-Kergoat and Goldar 2012) and regulation of nearby genes (Omberg, Meyerson et al. 2009). The proof reading bias is particularly noticeable in well studied bacterial genomes where the same origin sites are used every S phase (McLean, Wolfe et al. 1998). Eukaryotic origins fire in a staggered fashion rather than all at once (Gilbert 2002). Furthermore only a subset of the apparently competent sites actually fire (Xu, Aparicio et al. 2006). While there are high-throughput techniques for identifying origin sites that can fire *ex vivo* (Liachko, Bhaskar et al. 2010), the technique used for direct analysis of *in vivo* firing is quite laborious, requiring separation of replication intermediates on a 2D gel (Brewer and Fangman 1987; Nawotka and Huberman 1988). Whole genome approaches exist and rely on *de novo* DNA synthesis. One approach uses the chemical bromodeoxyuridine (BrdU), a thymidine analogue that is incorporated into the new DNA strand during replication. To investigate a particular segment of time within the replication cycle BrdU is pulsed into the experiment to begin replication tracking; hydroxyurea (HU) is used to arrest DNA replication and prevent further BrdU incorporation. Fragments containing BrdU are isolated with a BrdU specific antibody (Lengronne, Pasero et al.

2001). A similar technique identifies replication origins based on the presence of Okazaki fragment, a product of lagging strand synthesis (Gerbi and Bielinsky 1997). The resolution of the BrdU based method depends on the pulse/chase timescale of the experiment. However, only origins that replicate within the pulse/chase window can be detected necessitating larger timescales. This results in a tradeoff between the resolution of site detection and the rate of site detection. The Okazaki based method exhibits a similar tradeoff but tends toward high-resolution with low detection. A computational method for predicting active *in vivo* origins by examining the status of nearby chromatin would predict a large number of sites with higher resolution. Such a method would promote a better understanding of origins in model organisms such as *S. cerevisiae,* as well as faster identification of origins in other species.

The following three chapters address the application of logistic regression to three specific applications. The Ty5 and Ty1 retrotransposon studies, chapters 1 and 2 respectively, validate that the basic logistic regression model is capable of identifying biologically relevant phenomena. In both of these chapters the logistic regression model highlights key features that play a role in the fine scale targeting of the transposons, with minimal modification to the modeling pipeline between studies. However, despite the success of these applications, a key relationship between Ty1 and nucleosomes was largely ignored by the logistic regression model. The third chapter, covering prediction of origins of replication, aims to address shortcomings in the model as applied to Ty5 and Ty1 while simultaneously applying the model to a new class of problems. Where the Ty5 and Ty1 problems focused on describing the observed distribution, the origin problem focuses on predicting the distribution. Additional background on the particular biological phenomena can be found at the beginning of each chapter.

2. Chapter 1: Ty5 Integration Site Selection

General Overview of Retrotransposon Biology

As discussed, chromosomal modifications can have a profound impact on genome structure and evolution. This is particular evident when the chromosomal modification in question is a mobile genetic element. For many mobile elements, integration sites selection is distinctly non-uniform. These target site biases are particularly well-documented for the long-terminal repeat (LTR) retrotransposons and retroviruses (Bushman 2003; Sandmeyer 2003; Ciuffi and Bushman 2006). In most eukaryotes, retrotransposons constitute a large fraction of the genetic material, comprising, for example, up to half of the human genome (Goodier and Kazazian 2008). Retrotransposons attain such high copy numbers by reverse transcribing their mRNA into cDNA, which becomes inserted into new genomic sites through the action of the retrotransposon-encoded integrase (IN) protein (Beauregard, Curcio et al. 2008). cDNA integration has genetic consequences for the host: it can create mutations, and genome rearrangements, and deletions can result due to recombination between repetitive retrotransposon sequences scattered throughout the genome. In addition to genetic consequences of transposition, retrotransposons are often epigenetically modified and define distinct chromatin domains (Slotkin and Martienssen 2007). The combined genetic and epigenetic consequences of retrotransposition on host genomes are significant, and this impact is determined by the final step in retrotransposition, namely the choice of where cDNA inserts into the genome. Understanding mechanisms of retroelement target site choice, therefore, has value for both basic and applied research.

In the best studied cases, retroelement target site choice is dictated by interactions between IN and specific DNA-bound proteins. This interaction tethers the integration complex to specific genomic sites, resulting in a localized increase in integration (Bushman 2003). Examples of retrotransposons that recognize chromatin during integration include the *Schizosaccharomyces*

pombe Tf1 retrotransposon and the *S. cerevisiae* Ty3 retrotransposon, which integrate upstream of genes transcribed by RNA polymerase II and III (Pol II, Pol III), respectively (Chalker and Sandmeyer 1992; McLean, Wolfe et al. 1998). In both cases, transcription of target genes and proteins associated with transcription are required for target site choice (Yieh, Kassavetis et al. 2000; Yieh, Hatzis et al. 2002; Leem, Ripmaster et al. 2008; Majumdar, Chatterjee et al.). For the *S. cerevisiae* Ty5 retrotransposon, a six amino acid motif at the C-terminus of Ty5 IN binds the heterochromatin protein Sir4, resulting in integration into heterochromatin (Zhu, Dai et al. 2003; Tsankov, Thompson et al. 2010). Retroviruses also recognize chromatin during integration. HIV IN, for example, interacts with the transcription factor lens epithelium-derived growth factor (LEDGF), and this underlies HIV's preference to integrate into actively transcribed genes (Cherepanov, Maertens et al. 2003; Ciuffi, Llano et al. 2005).

Ty5 Background

The first retroelement for which a targeting mechanism was described in detail was the SaccharomycesretrotransposonTy5. Ty5 integrates preferentially into heterochromatin, which in yeast is found near the telomeres and silent mating loci (*HML* and *HMR*) (Zou, Ke et al. 1996; Zou, Kim et al. 1996; Zou and Voytas 1997). Ty5 IN selects integration sites using a six amino acid motif at the IN C-terminus (Gai and Voytas 1998; Xie, Gai et al. 2001). This IN targeting domain interacts with a protein component of heterochromatin, namely silent information regulator 4 (Sir4) (Xie, Gai et al. 2001; Zhu, Dai et al. 2003). The Ty5 IN/Sir4 interaction tethers the integration complex to target sites and results in Ty5's primary target site bias.

In this study, we applied high throughput DNA sequencing to characterize a large number of Ty5 insertions that we mapped to the *S. cerevisiae* genome. Whereas the majority of Ty5 elements integrated as predicted in heterochromatin, a secondary target site bias was revealed for both euchromatic and heterochromatic insertions. Logistic regression established that this secondary bias was influenced by chromosomal features characteristic of open chromatin,

including DNase hypersensitivity, lack of nucleosomes, the presence of transcription factors and epigenetic marks associated with gene transcription. We provide evidence suggesting that this secondary target site bias reflects sites that can be accessed by the Ty5 integration complex during integration.

Results

The Ty5 insertion dataset. To observe genome-wide patterns of Ty5 integration, we created an integrant library of approximately 400,000 independent transposition events. This library was derived from 16 separate Ty5 transposition assays – eight using the wild type YPH499 haploid strain and eight using the isogenic wild type diploid, YPH501. Ty5/host DNA junction fragments were recovered from each of the sixteen populations using linker-mediated PCR. Linkers were ligated to genomic DNA that had been digested with restriction enzymes. Four enzymes (each recognizing four bases) were used to maximize potential to recover sites and to minimize recovery bias. The genomic sequence at each insertion site was determined by pyrosequencing using the 454 GS FLX platform.

In total, approximately 337,000 sequencing reads were obtained (Table 1). Specific barcode sequences in the PCR primers made it possible to assign reads to one of the 16 transposition assays. Reads were excluded that 1) did not have a perfect match to a barcode and surrounding DNA or 2) had more than 4 mismatches to the primer. Further, insertions at a given position and orientation were only counted once in each pool. In total, approximately 160,000 reads passed our filters. Sequences sharing more than 98% sequence identity to a single site on the *S. cerevisiae* genome were designated as unambiguous insertions. Because Ty5 integrates preferentially into repetitive, subtelomeric regions, reads mapping to multiple sites in the genome (greater than 98% sequence identity) were also considered. These ambiguous insertions and were down-weighted by a factor equal to the number of sites to which the read mapped (i.e. each

ambiguous site was assigned a fraction of an integration event). Forty percent of the high quality reads were ambiguous.

Table 2-1: Ty5 Sequencing Pools

Strain name-Pool number	Ploidy	All Reads	Clean Reads	Basepairs Hosting Ambiguous Alignments	Basepairs Hosting Unambiguous Alignments
YPH499-1	Haploid	21960	10368	468	743
YPH499-2	Haploid	22050	11082	423	847
YPH499-3	Haploid	22559	10356	370	673
YPH499-4	Haploid	23351	10868	444	766
YPH499-5	Haploid	22102	10525	400	719
YPH499-6	Haploid	21367	9161	361	637
YPH499-7	Haploid	21361	9816	540	912
YPH499-8	Haploid	21779	10749	568	987
YPH501-1	Diploid	18605	9127	348	389
YPH501-2	Diploid	20365	9485	207	228
YPH501-3	Diploid	19292	8680	264	214
YPH501-4	Diploid	20889	9903	222	287
YPH501-5	Diploid	19572	9182	212	234
YPH501-6	Diploid	21967	10460	205	279
YPH501-7	Diploid	21014	10542	346	450
YPH501-8	Diploid	19237	8906	205	243
Not Assigned		6399	-	-	-

Ty5's primary target site bias. The majority of Ty5 insertions mapped to the ends of all 16 *S. cerevisiae* chromosomes (Figure 2-1, Figure 7-1, Figure 7-2). Thus the primary pattern of Ty5 integration matched what we predicted based on our previous work demonstrating the key role played by heterochromatin in target site choice (Zou, Ke et al. 1996; Zhu, Dai et al. 2003).

Because most Ty5 insertions were subtelomeric, for subsequent analyses the genome was split into two regions, designated euchromatin and heterochromatin. Heterochromatic regions began at the end of a chromosome and ended 10 kb centromere-proximal to the subtelomeric X repeat or one of the silent mating loci, *HML* or *HMR*. By this definition, heterochromatin constituted 4% of the genome and received 76% of the insertions. This insertion density is likely an underestimate, because reads mapping to the same position were excluded if they were derived from the same pool; such duplicate reads may represent independent insertions at the same site. Euchromatic regions comprised most of the chromosomes and were bounded by centromere-proximal points 40 kb distant from an X repeat, *HML* or *HMR*. This left a 30 kb buffer between heterochromatin and euchromatin to ensure signals were distinct. The euchromatin and buffer region constituted 88% and 7% of the genome, respectively. The rDNA and *MAT* were excluded from euchromatin, because the former is not accurately represented in the reference genome and the latter contained many ambiguous insertions due to duplicated sequences at the silent mating loci.

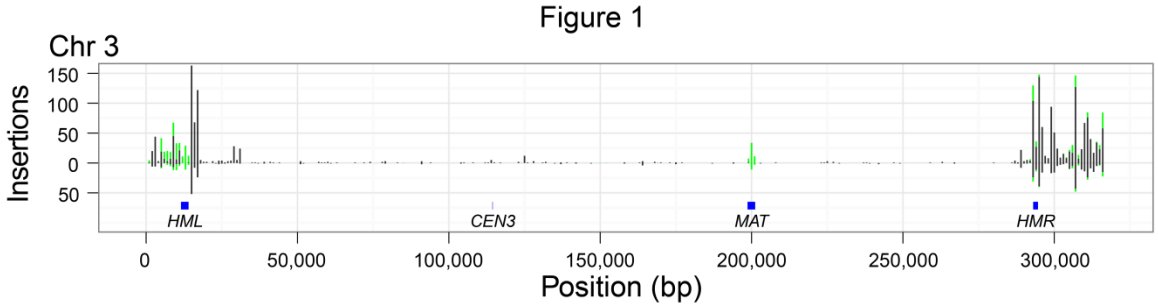

Figure 2-1 - Distribution of Ty5 insertions on Chromosome 3

The x-axis denotes position along the chromosome at 1000 bp resolution. Black bars indicate the number of unambiguous integrations at a particular site; stacked green (grey if greyscale) bars indicate additional ambiguous integrations. Bars above the x-axis indicate data from the haploid strain; bars below the x-axis denote data from the diploid.

Selection could influence the distribution of Ty5 insertions; for example, insertions may not be recovered if they occur in essential genes in haploid strains. To assess impacts of selection, Ty5 insertion sites were compared between the haploid and diploid populations. Both the haploid and diploid chromosomal distributions were nearly identical with a Pearson's correlation of 0.82 at 10 base pair resolution. Selection, therefore, does not play a significant role in global patterns of Ty5 integration.

Relationships between Ty5 insertions and chromosomal features. For *S. cerevisiae*, a large body of genome-wide data has accumulated describing, for example, distributions of various histone modifications, transcription factor binding sites or nucleosome occupancy (Table S1). To better understand factors that influence Ty5 target site choice, we used logistic regression to establish associations between insertions and these chromosomal features as well as DNA sequence landmarks such as open reading frames or specific gene classes (e.g. those transcribed by RNA pol III). Our implementation compared sites of observed integration (case) to a random subset of sites without integrations (control). The random distribution was corrected for possible recovery bias due to restriction site distribution. Additionally, the overall quality of the model was evaluated using ROC analysis, in particular the value of the area under the curve (AUC). Logistic regression was applied to the euchromatic and heterochromatic datasets separately (Figure 2-2). Both single- and multi-dimensional models were evaluated, and both gave the same overall conclusions. In the following paragraphs, we illustrate the major findings of one-dimensional logistic regression, using representative examples of euchromatic and heterochromatic Ty5 target sites. Details about the multi-dimensional models are provided in Figure 7-3.

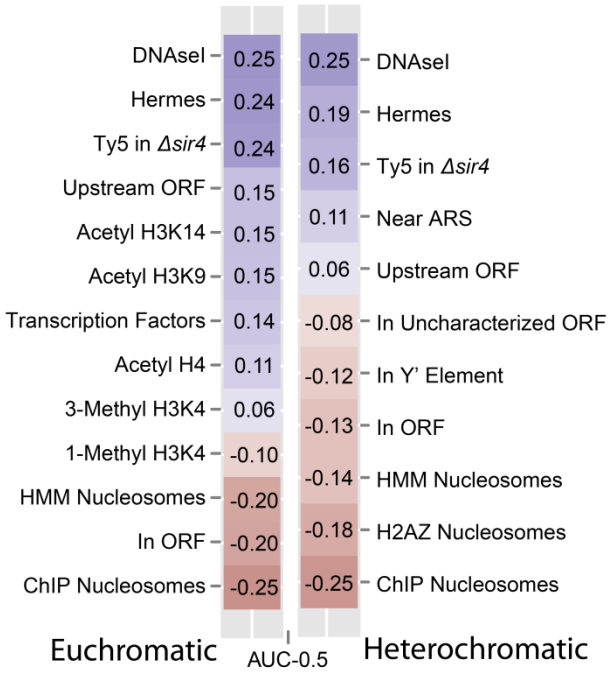

Figure 2-2 - AUCs from Ty5 Predictions

Associations between Ty5 insertions and chromosomal features. Heatmaps showing the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) Curve from logistic classifiers trained on single features. Actual values shown are AUC-0.5. As such, zero indicates a model of no predictive power whereas 0.5 and -0.5 indicate models of perfect predictive power. Positive AUCs signify features associated with case integrations; negative AUCs signify sites associated with control integrations. Heatmaps for insertions in euchromatin (left) and heterochromatin (right) were generated from separate models. Details of the datasets used for various chromosomal features can be found in Table S1.

Ty5 insertions in heterochromatin. Recently, a genome-wide map of Sir4 chromosomal occupancy was determined (Zill, Scannell et al.), and to our initial surprise, logistic regression did not reveal an association between Ty5 insertions and sites of Sir4. In Figure 2-3, we plot Ty5 insertions and Sir4 distribution at a few subtelomeric regions, and as can be seen, peaks of Sir4 and Ty5 insertions occur near the subtelomeric X repeats and the silencers flanking *HMR* (see also Figure 7-1 and Figure 7-2). As illustrated by these examples, Sir4 is highly localized, and sites of Sir4 occupancy are predictive of sites of Ty5 integration. However, because very little Sir4 is found elsewhere throughout the subtelomeric region (or the remainder of the genome), the

majority of insertions in heterochromatin (or euchromatin) have no clear link to Sir4 distribution. Our logistic regression model only considers chromosomal features at or near (e.g. within 1 kb) of a Ty5 insertion site, and so logistic regression did not reveal a strong Ty5/Sir4 association.

Sir4 aside, logistic regression identified several chromosomal features in heterochromatin that were positively or negatively associated with Ty5 insertions. Among these was a positive association (AUC-0.5 = 0.11) with 1 kb regions centered on known autonomously replicating sequences (ARSs), which often serve as sites of DNA replication (Figure 2-2) (Rehman and Yankulov 2009). The subtelomeric X repeats, which are bound in Sir4, also contain an ARS, and in our previous work, Ty5 insertions were considered targeted if they occurred within a 3 kb window centered on an X ARS (Zou, Ke et al. 1996). The high incidence of insertions near X repeats (Figure 2-3) likely explains the observed association with ARSs.

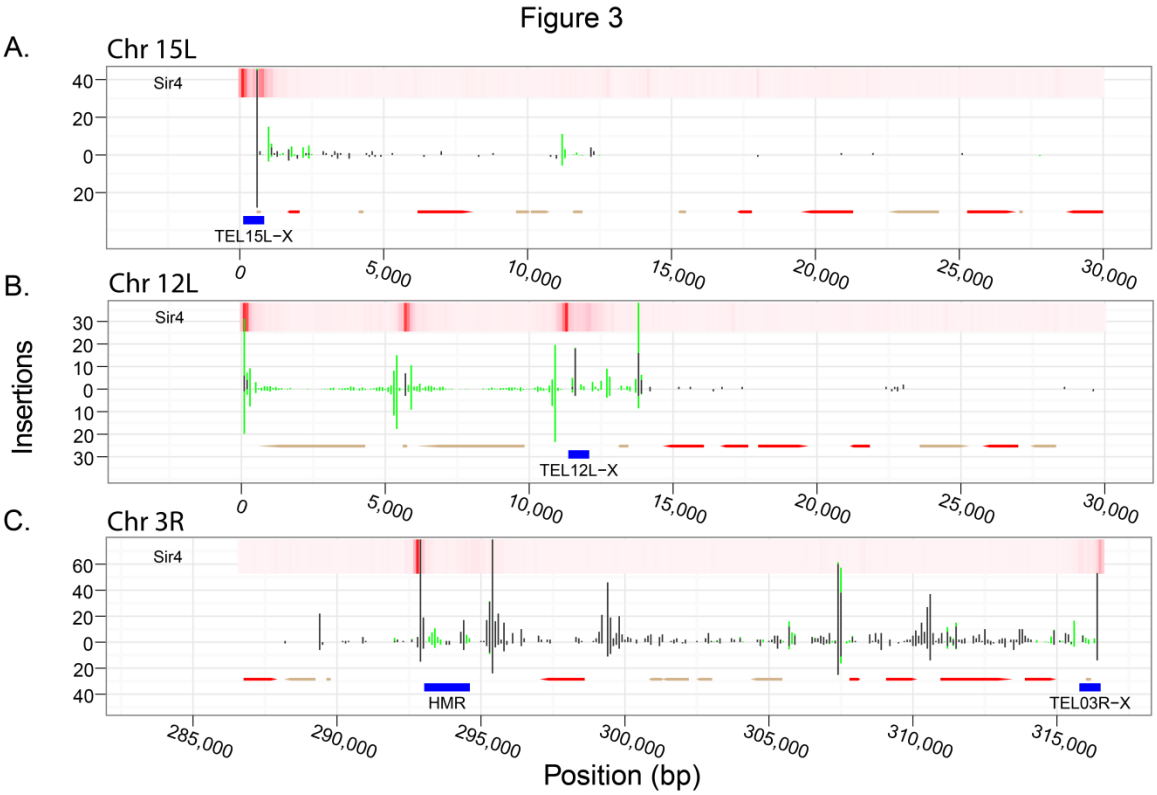

Figure 2-3 – Ty5 Integration Distribution at Select Telomeres

Representative heterochromatic domains and the location of verified (red) and uncharacterized (tan) ORFs are shown. Black and green (grey if greyscale) bars indicate the frequency of unambiguous and ambiguous integration events, respectively. Bars above the x-axis indicate integrations in the haploid strain; bars below the x-axis are integrations in the diploid. The heatmap at the top of the graph displays Sir4 occupancy; the color intensity was normalized to the chromosomal regions depicted.

A negative association (AUC-0.5 = -0.12) was identified between Ty5 insertions and Y' elements – repeats at the ends of some yeast chromosomes that are typically either 5.5 or 6.7 kb in length and encode a helicase (Louis and Haber 1992). The Y' coding region, in particular, was a coldspot for integration as illustrated for the two tandem Y' elements on chr 12L (Figure 2-3). Insertion hotspots, however, occurred on the centromere-proximal side of the Y' elements – the site of an X repeat – and at sites rich in Sir4 between the Y' elements and at the telomere itself. The coding sequences of Y' elements are bound by nucleosomes, and the Ty5 insertion hotspots flanking Y' elements lack nucleosomes (Lee, Tillo et al. 2007; Zhu and Gustafsson 2009). The pattern of Ty5 insertions is therefore consistent with the finding that nucleosome occupancy is a strong negative predictor of Ty5 insertion sites (Figure 2-2). Nucleosomes were represented in two different forms in the regression model: either as processed ChIP probe values (AUC-0.5 = -0.25) or as a ternary prediction from a hidden Markov model (HMM) trained on the ChIP data (AUC-0.5 = -0.14) (Lee, Tillo et al. 2007; Zhu and Gustafsson 2009). Nucleosomes were also avoided if they contained H2AZ (AUC-0.5 = -0.18), an H2 variant enriched in transcriptionally inactive genes (Li, Pattenden et al. 2005).

On chr 3, heterochromatic domains are found at the telomeres and silent mating loci, the latter of which are located up to 30 kb from the end of the chromosome. As illustrated for the right arm of chr 3 (Figure 2-3), in addition to peaks of Ty5 insertions near the silencers flanking *HMR* and at the X repeat, clusters of insertions occur throughout the region telomere-proximal to *HMR*, particularly in intergenic regions. Localized selection does not contribute to the distribution pattern, because none of the genes on the right arm of chr 3 are essential (project). Further, a

similar insertion distribution is observed in both haploid and diploid strains. Clustering of Ty5 insertions adjacent to coding sequences can also be seen in other subtelomeric regions (e.g. chr 12L, Figure 2-3). This pattern is consistent with the results of logistic regression indicating that heterochromatic insertions are slightly associated with upstream regions of genes (AUC-0.5 = 0.06) and strongly with DNAse hypersensitive sites (AUC-0.5 = 0.25), a feature characteristic of many promoters.

Ty5 insertions in euchromatin. Logistic regression performed on euchromatic insertions revealed a similarly pronounced association between Ty5 and regions flanking genes. As with heterochromatin, Ty5 insertions showed a strong positive association with DNase hypersensitive sites (AUC-0.5 = 0.25) and regions upstream of verified open reading frames (ORFs) (AUC-0.5 = 0.20). Other features characteristic of actively transcribed genes were also positively associated, such as H3 K14 and H3 K9 acetylation (AUC-0.5 = 0.15) (Pokholok, Harbison et al. 2005) and sites bound by transcription factors (AUC-0.5 = 0.14). Negative associations were similar to those in heterochromatin, namely Ty5 was less likely to be found in coding sequences (AUC-0.5 = -0.20) and sites bound by nucleosomes (AUC-0.5 = -0.20 HMM or -0.25 ChIP). Representative Ty5 hotspots in euchromatin are illustrated in Figure 7-4.

Ty5's secondary target site bias. Because Ty5 insertions in both euchromatin and heterochromatin were enriched in intergenic regions, we generated composite figures relating Ty5 insertions to ORFs in both of these chromatin environments (Figure 2-4). On average, insertions begin to occur near the start codon and peak approximately 100 bp upstream at a site corresponding to minimal nucleosome occupancy. Insertion frequency falls off to background levels approximately 1000 bp upstream of the translational start. A smaller peak of insertions is also observed in a nucleosome-poor region downstream of the ORFs. As indicated by the logistic regression analyses, Ty5 avoids integrating into the nucleosome-bound coding sequences. Subtle discrepancies distinguished euchromatin and heterochromatin integration patterns; for example,

there is a clear peak of Sir4 density downstream of ORFs in heterochromatin and an adjacent peak of Ty5 insertions. In the *Δsir4* strain, the Ty5 peak shifts to the site occupied by Sir4 in the wild type, suggesting that this site may now be more accessible to the integration complex.

One hypothesis to explain local Ty5 integration patterns is that there is a host protein like Sir4 that acts as a positive targeting determinant, drawing Ty5 insertions to promoter regions. To assess whether Sir4 itself contributes to local integration patterns, we evaluated a large dataset of Ty5 insertions recovered from a *Δsir4* strain. These insertions where generated to establish baseline patterns of Ty5 integration for 'calling card' experiments (Wang, Mayhew et al. 2011). A given transcription factor can be made a Ty5 calling card by fusing it to the domain of Sir4 that interacts with Ty5 IN (Wang, Johnston et al. 2007). Ty5 insertion sites in yeast strains expressing the calling cards identify chromosomal occupancy of the transcription factor. We treated Ty5 insertions in the *Δsir4* strain as a chromosomal feature and evaluated their association with insertions generated in wild type strains using logistic regression. The Ty5 insertions in *Δsir4* showed a significant positive association with insertions generated in wild type in both euchromatin (AUC-0.5 = 0.24) and heterochromatin (AUC-0.5 = 0.16) (Figure 2-2). Insertion sites in both strains were correlated (assuming 1 kb windows, Spearman's $\rho = 0.255$, $p < 2.2e\text{-}16$). This is evidenced in Figure 2-4, where insertions in *Δsir4* are mapped relative to ORFs. Secondary targeting patterns, therefore, are not due to Sir4, and if a different positive targeting determinant is responsible, it remains elusive.

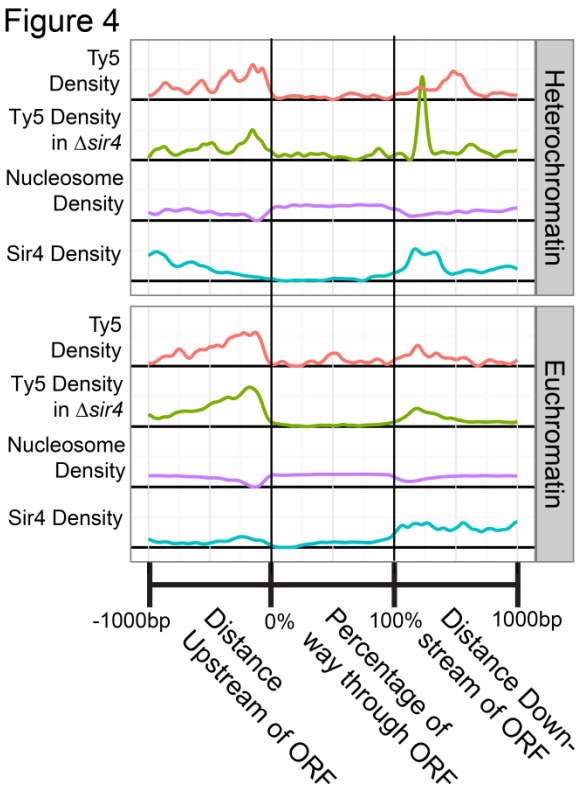

Figure 2-4 - Distribution of Ty5 Integration over ORFs

Ty5 insertions near verified ORFs. The x-dimension represents position in and around verified ORFs. To account for ORFs of different lengths, the region within the ORFs was scaled as a percentage of ORF length. Datasets were smoothed and scaled for easy comparison. As a result of scaling, all units are arbitrary and the integrals of all curves are equal.

An alternative hypothesis to explain secondary targeting patterns is that insertion hotspots simply reflect sites accessible to the Ty5 integration complex. This is consistent with DNase hypersensitivity being the strongest, positive predictor of Ty5 integration sites in both heterochromatin and euchromatin (Figure 2-2). Recently, a large number of insertion sites were recovered in yeast using the Hermes DNA transposon from housefly (Gangadharan, Mularoni et al.). Like Ty5, Hermes strongly prefers nucleosome-free regions. The Hermes dataset proved the second best predictor of Ty5 integration sites in both euchromatin (AUC-0.5 = 0.24), and heterochromatin (AUC-0.5 = 0.19) (Figure 2-2). Correspondence between Hermes insertions and Ty5 insertions in wild type and *Δsir4* strains can be visualized on a genome-wide level (Figure

2-1, Figure 7-1, Figure 7-2) and at select euchromatic sites (Figure 7-4). As with the Ty5 insertions in *Δsir4*, the distribution of Hermes insertions is highly correlated with the distribution of Ty5 insertions in wild type (assuming 1 kb windows, Spearman's $\rho = 0.257$, $p < 2.2e\text{-}16$). One explanation for the similarity in integration patterns of Hermes and Ty5 in wild type and *Δsir4* strains is that these preferred sites represent open chromatin where these mobile elements can gain access to DNA. This is further supported by the observation that Ty5 insertion sites are most positively associated with sites of DNase hypersensitivity and by our multi-dimensional model (Figure 7-3), which produces an AUC-0.5 of -0.30 using only features associated with open DNA. Access to DNA, therefore, is likely the basis for Ty5's secondary target site bias.

Discussion

The ability to recover large numbers of transposable element insertions using high throughput DNA sequencing technologies provides a powerful means to understand mechanisms underlying target site choice. Complementing the robust and quantitative measures of target specificity afforded by this approach is the wealth of genome-wide information that makes it possible to discern associations between mobile element insertions and specific chromosomal features. Pioneering work in this regard was performed with HIV, in which associations between insertion sites and various chromosomal features were assessed by computational approaches including logistic regression (Berry, Hannenhalli et al. 2006; Wang, Ciuffi et al. 2007). We adopted a similar approach with our dataset of over 14,000 Ty5 insertions and the extensive genome-wide datasets available for *S. cerevisiae*. One additional advantage of applying this approach in a model organism like yeast is that insertions can be readily recovered in various mutant backgrounds (e.g. *Δsir4*). The further use of genetic resources available for *S. cerevisiae* will undoubtedly lead to additional insights into mechanisms by which Ty5 and other yeast transposable elements select chromosomal integration sites.

Our genome-wide analysis reinforced what was previously known about Ty5's primary target site preference, namely that insertions predominantly occur in domains of heterochromatin. To our surprise, however, we did not observe a tight association between sites of Ty5 integration and Sir4 occupancy; rather insertions occurred throughout subtelomeric domains including regions largely devoid of Sir4. Our two-dimensional view of the genome and Sir4 occupancy, however, most certainly belies the actual architecture of subtelomereic regions. We believe that much of the subtelomeric DNA is actually within close proximity to sites enriched in Sir4 (Figure 2-5); therefore, once the Ty5 IN/Sir4 tether is established, integration can occur throughout the subtelomeric region. Alternatively, Ty5 IN could be loaded onto heterochromatin by Sir4 and then scan the subtelomeric regions for target sites.

Figure 5

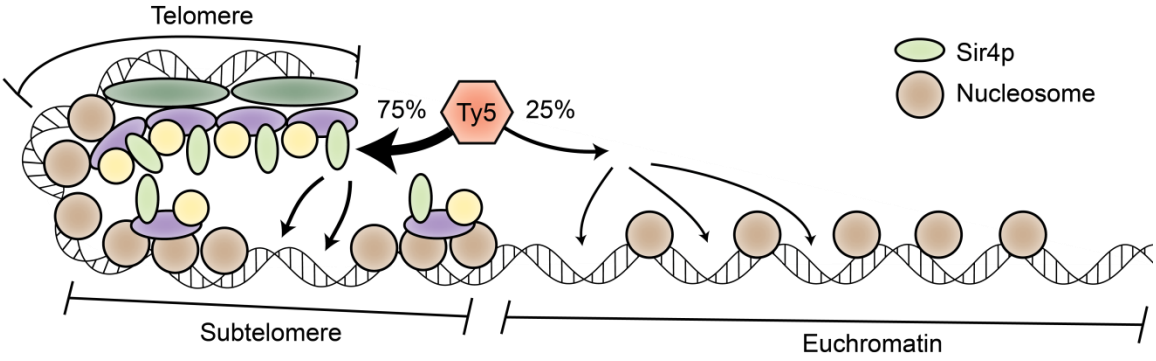

Figure 2-5 - Proposed Mechanism for Ty5 Integration Targeting

A model describing Ty5's primary and secondary target site biases. Ty5 IN interacts with Sir4, which localizes the integration complex to heterochromatin. This interaction results in Ty5's primary target site bias, namely the association of approximately 75% of Ty5 insertions with domains of heterochromatin. Ty5's secondary target site bias is determined by DNA access. Sites in heterochromatin are chosen that are nucleosome free and accessible to the integration complex. Access to DNA also dictates Ty5's preferred integration sites in euchromatin, resulting in integration primarily in nucleosome free regions flanking genes.

Ty5 integration patterns provide a readout for boundaries of heterochromatin on the yeast chromosomes. Probing chromatin is not a new role for Ty5, as changes in integration patterns have previously documented the chromatin dynamics that occur during aging, particularly the movement of Sir4 from the telomeres to the rDNA (Zhu, Zou et al. 1999). In addition, the recently developed 'calling card' approach is a clever implementation of Ty5's ability to mark chromosomal occupancy of proteins (Wang, Johnston et al. 2007). Ty5 calling cards are created by fusing the domain of Sir4 that interacts with Ty5 to a transcription factor, and Ty5 insertions mark chromosomal sites where the transcription factor is bound. Because many retroelements recognize specific chromatin features during integration, retroelements may increasingly prove valuable probes of chromatin dynamics.

Regardless of whether Ty5 integrates into euchromatin or heterochromatin, the chromosomal features influencing Ty5 target site choice were remarkably consistent. Ty5 insertions were associated with DNAse hypersensitive, nucleosome-free sites and other features

linked to transcription – a pattern we refer to as Ty5's secondary target bias. On average, Ty5 insertions peak in nucleosome-free windows approximately 100 bp upstream and downstream of coding sequences. A very similar pattern is observed for insertions generated in *Δsir4* strain indicating that this secondary target site bias is not due to Sir4. Hermes, a completely unrelated DNA transposon from the housefly, has an integration pattern correlated to Ty5's. Hermes is not adapted to life in its heterologous host and uses a very different enzyme to catalyze integration into the yeast genome. Hermes insertion sites, therefore, likely identify open chromatin, and this is consistent with their correlation with DNase hypersensitive sites (Spearman's $\rho = 0.715$, $p < 2.2e\text{-}16$). We believe that based on the data at hand, the most parsimonious explanation of Ty5's secondary target site bias is that it is dictated by accessibility of the Ty5 integration complex to DNA.

Secondary targeting patterns are not without consequence for genome structure and evolution. One consequence of integrating into nucleosome-free sites is that coding regions are often avoided, thereby limiting a negative consequence of transposition, namely insertional mutagenesis. It has been argued that heterochromatin, because it is gene-poor, provides a safe haven for Ty5 integration that minimizes deleterious consequences of transposition (Boeke and Devine 1998). It may be that integration into open chromatin provides an additional mechanism to avoid genes. That said, insertions in promoter regions, likely have consequences for the regulation of adjacent genes, which could have important evolutionary outcomes. Our proposed mechanism underlying Ty5's secondary target bias may underlie well-established associations between other mobile genetic elements and promoter regions (Guo and Levin ; Bellen, Levis et al. 2004; Liu, Yeh et al. 2009). Clearly, the discovery and initial characterization of Ty5's secondary target site bias as reported here reinforces the importance of chromatin in dictating retroelement target site choice.

Materials and Methods

Recovery of Ty5 insertions. Ty5 transposition assays were performed as previously described using the haploid and diploid strains YPH499 and YPH501, respectively (Zou, Ke et al. 1996). The donor Ty5 plasmid was pNK254, which contains a galactose-inducible Ty5 element with a marker gene to detect transposition. Each Ty5 transposition assay gave rise to a pool of approximately 25,000 Ty5 integrants. Genomic DNA was prepared from the pools and treated with two sets of restriction enzymes, AciI/TaqI and MspI/HinplI (Figure 7-5). Linker-mediated amplification of integration sites was performed using the protocol found in Ciuffi et al (2005) (Ciuffi, Mitchell et al. 2006). Digested DNA was ligated to a linker made up of two oligonucleotides, DVO4621 and DVO4622 (see Table 7-2 for linker sequences). To prevent amplification of the 5' LTR, DNA samples treated with AciI/TaqI were digested with AseI; samples treated with MspI/HinplI were digested with EcoRI. The first round of PCR amplification used the Ty5 LTR-specific primer DVO495 and the linker-specific primer DVO4632. The second round of PCR amplification used DVO4665 and one of several barcoded Ty5 LTR primers (DVO4666-DVO4681) (Table 7-2). PCR products were gel purified and fragments between 100 and 500 bp were sequenced using a 454 GLX sequencer.

Random control insertions. A total of 19,934 control insertions were produced *in silico* for euchromatin and 7,034 for heterochromatin. Each control insertion was the product of three random values: a restriction site value, a position value and an orientation value. These values select, respectively, a restriction site in the genome, a distance away from the restriction site and an orientation for the control insertion. The probability distribution function for a control insertion's position and orientation was calculated as the normalized frequency of recovered insertions relative to the restriction sites used in recovery. Control insertions were made to be

disjoint from known insertion sites. This process resulted in a set of control insertions whose restriction bias was similar to that of the recovered insertions.

Data annotation and analysis. Logistic regression was used to identify discriminative features for integration (Table 7-1). Regression models were trained using the *glm* log-linear regression function in the R statistical package (Team 2008; Friedman 2010). Our implementation compared the sites of observed integration (case) to a random subset of the sites without integrations (control). Logistic regression fits the following equation:

$$f(z) = \frac{1}{1 + e^{-z}}$$

where $f(z)$ is the class prediction and z is a linear function, $z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$, of the levels, x_i of the n chromosomal features.

Predictions from a logistic regression fall within the interval (0,1) with proximity to the endpoints indicating greater certainty of a class designation. This information was used to produce a ROC curve, a plot of the true positive rate vs. the false positive rate parameterized on a discrimination threshold. An area under a ROC curve (AUC-ROC or AUC) of 0.5 indicates a model with no predictive power while an AUC of 1.0 indicates perfect prediction. All AUC data presented herein is in the form of an AUC-0.5 where negative values indicate features showing a greater association with the control dataset.

3. Chapter Two: Ty1 Integration Site Selection

Ty1 Background

Although the yeast retrotransposon Ty1 is among most-studied mobile genetic element, the molecular mechanism underlying its target site choice remains elusive. Ty1 preferentially integrates upstream of genes transcribed by RNA Pol III (class III genes), including tRNA genes and 5S rRNA genes (Takamu and Oshima 1970; Ji, Moore et al. 1993). Targeting occurs within an ~750 bp window upstream of Pol III transcription start sites, and consistent with a chromatin tethering mechanism, targeting depends on the presence of the Pol III transcription complex.

Previous analyses of Ty1 target specificity monitored insertion patterns on a single chromosome (chr III) (Ji, Moore et al. 1993) or at a small number of known Ty1 targets (e.g. a subset of class III genes)(Bachman, Eby et al. 2004). A drawback to these studies is that analyses were restricted to a fraction of the genome, and the methods used to recover insertions made it difficult to obtain large numbers of independent insertions (32 on chr III; 836 at class III genes) (Ji, Moore et al. 1993; Bachman, Eby et al. 2004). To overcome these limitations, we applied linker-mediated PCR and high throughput sequencing to conduct a genome-wide survey of Ty1 integration patterns. We also took advantage of the wealth of genome-wide datasets for *S. cerevisiae*, and used machine learning (specifically logistic regression) to identify chromosomal features (e.g. histone modifications or specific transcription factors) associated with Ty1 insertion sites. Our analyses revealed that a specific surface of nucleosomes upstream of class III genes is a critical Ty1 targeting determinant, suggesting that histone modifications or proteins associated with nucleosomes upstream of class III genes are recognized by Ty1 IN and underlie this retrotransposon's target site bias.

Results

Generating, recovering and mapping Ty1 insertions. Ty1 integration events were generated using a modified version of the well-studied pGTy1*his3AI* element (called pGTy1*his3AI-*

SCUF)(Curcio and Garfinkel 1991). The 5' LTR of pGTy1*his3AI*-SCUF contains six nucleotide substitutions in the U5 region downstream of the initiation codon of the *GAG* ORF. Nucleotide changes were introduced so as not to alter the *GAG* amino acid sequence, and pGTy1*his3AI*-SCUF was found to transpose at frequencies comparable to the unmodified element (data not shown). The 6-nucleotide sequence tag is copied into the 3' LTR by reverse transcription, making it possible to distinguish the 3' LTR/genomic DNA junction of *de novo* Ty1 insertions from the 3' LTR/genomic DNA junction of pre-existing Ty1 elements in the genome. Reverse transcription of a spliced Ty1*his3AI* transcript produces a functional *HIS3* gene, which, when incorporated in the yeast genome, confers histidine prototrophy (Curcio and Garfinkel 1991). His^+ insertion events were recovered from three wild-type strains (YPH499, haploid **a** mating type; YPH501, diploid; BY4741, a derivative of YPH499 used for the genome-wide deletion project) and four mutant strains in the BY4741 background that affect Ty1 insertion frequency (*rrm3*Δ, *hos2*Δ, and *rtt109*Δ*)* or pattern (*rad6*Δ*)* (Supplementary Table 1). For each yeast strain tested, transposition was induced in 10 to 14 independent cultures, and ~10,000 His^+ colonies resulting from each induction were pooled. Genomic DNA was purified and digested with either AciI or TaqI. Linkers were annealed to the ends of the digested DNA, and 3' Ty1/genomic DNA junction fragments were amplified by PCR. PCR primers were specific to the linker and sequence modifications in the pGTy1*his3AI*-SCUF LTR. The primers had different DNA barcodes to distinguish between yeast strains and restriction enzyme digestions. All PCR products were pooled and sequenced by 454 pyrosequencing.

A single 454 run produced between 13,000 and 110,000 sequence reads per pool (Table 2). The data were processed using a pipeline to identify those sequences with a perfect match to the terminus of the Ty1 LTR and a 98% match to genomic DNA beginning within 3 bp of the end of the LTR. Approximately 19% of the sequence reads (154,408) passed these filters and could be mapped to the genome via BLAT (Table 2)(Kent 2002). Alignment revealed two distinct

sequence categories: approximately 85% of the insertions mapped unambiguously to unique sites in the genome; the remainder mapped to multiple positions. The majority of ambiguous hits were within endogenous Ty LTRs (see also below). It was possible to use the ambiguous hits in subsequent analyses by down-weighting each hit proportionally to the number of sites it mapped to in the genome. These normalized data were used principally to confirm and validate conclusions drawn from the unambiguous insertion dataset.

Table 3-1: Ty1 Sequencing Pools

		Insertion Events Recovered		
Strain Name	Restriction Enzyme	Reads	Ambiguous Alignments (at least 20bp long)	Unambiguous Alignments
BY4741	AciI	16891	4701	2508
	TaqI	18782	12754	3374
YPH499	AciI	111168	3501	5480
	TaqI	85851	9840	7173
hos2Δ	AciI	18974	3307	2858
	TaqI	17396	9763	3553
rad6Δ	AciI	19100	2642	2108
	TaqI	13560	7956	2407
rrm3Δ	AciI	15613	4743	1126
	TaqI	13352	14160	1351
rtt109Δ	AciI	21957	4490	2332
	TaqI	16395	13403	2761
YPH501	AciI	83793	3434	3691
	TaqI	73762	13835	5157

Genomic distribution of Ty1 insertions in wild type strains. Ty1 insertions mapped to all 16 chromosomes in a punctuate pattern, characterized by clusters of insertions upstream of class III genes (Figure 3-1, Figure 8-1, Figure 8-2). In addition, a small number of insertions were distributed throughout the genome. Pairwise comparisons between the diploid and two haploid strains failed to reveal a significant difference in the distribution of insertions (pairwise Pearson correlations for YPH499 vs. YPH501 = [.92, .93]). We conclude, therefore, that ploidy does not significantly influence targeting patterns. Variation in insertion patterns, however, was observed between YPH499 and BY4741. In particular, the *tE(UUC)C* and *tI(AAU)L1* loci received few to no insertions in BY4741, suggesting these genes are missing in this strain.

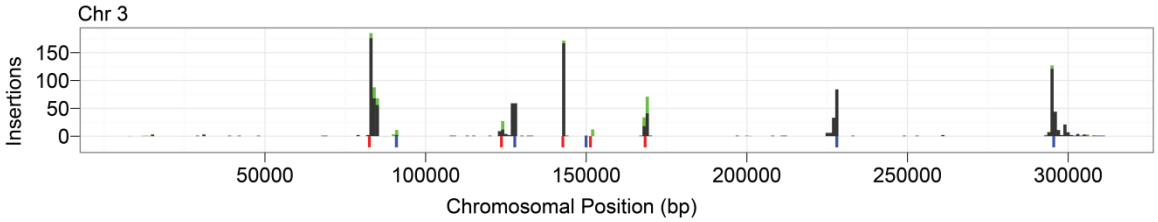

Figure 3-1 – Ty1 Integrations on Chromosome 3

The x axis denotes position along the chromosome at 1 kb resolution. Black bars indicate the number of unambiguous insertions at a particular site; stacked green bars indicate additional ambiguous insertions. Colored bars below the x axis indicate positions of class III genes; blue denotes genes transcribed from left to right; red denotes genes transcribed in the opposite direction.

Ty1 insertions were underrepresented in open reading frames (ORFs): only 4.86% of insertions occurred in verified ORFs in haploid cells, whereas random insertion would result in approximately 60% of insertions in verified ORFs ($p < 2.2e-16$). In the diploid strain, 5.02% of insertions occurred in ORFs, which does not differ significantly from the haploid ($p=0.59$). As such, we conclude that selection does not have a significant effect on the genomic distribution of Ty1 insertions.

We further analyzed the distribution of Ty1 insertions with respect to class III genes, which include 275 tRNA genes, *SNR6*, *RPR1*, *SCR1*, *SNR52*, *RNA170*, *ZOD1* and 100-200

tandem copies of *RDN5* (Harismendy, Gendrel et al. 2003; Roberts, Stewart et al. 2003)}. Whereas the 2000 bp upstream of all class III genes constitute less than 5% of the genome, those regions received 90% of the total Ty1 insertions. However, not all class III genes were equally targeted (Figure 3-2). A number of class III genes received zero insertions in all six independent experiments with wild-type strains, whereas other sites received as many as 561 insertions. Comparisons between the number of insertions at each class III gene and the appropriate random distribution (binomial: n=27382, p=1/288) indicates that Ty1 clearly prefers certain class III genes over others. This preference was consistent between yeast strains, with the YPH501 and YPH499 being more similar to each other than to BY4741 (pairwise Pearson correlations: YPH501/YPH499 = [.95, .97], BY4741/YPH499 = [0.84,.92], BY4741/YPH501 = [0.83,.92] at 99.9% confidence). The differences between YPH501/YPH499 and BY4741 appeared to be spread across all class III genes with the exception of the *tE(UUC)C* and *tI(AAU)L1* loci, as previously noted. We also correlated our BY4741 data with the results of BACHMAN *et al.* 2004 in terms of preference for specific class III gene targets (Spearman rho = 0.43, p=0.012). While the correlation was relatively poor, it was reasonable given differences in methodology.

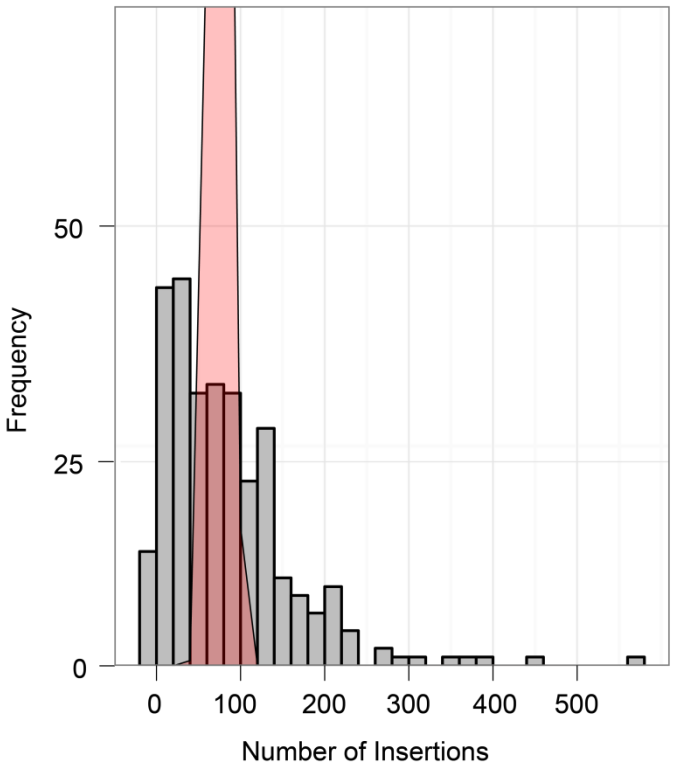

Figure 3-2 - Histogram of Ty1 insertion frequency per class III gene.

The x axis depicts the number of Ty1 insertions within a 2 kb window upstream of each class III gene in the *S. cerevisiae* genome. Values on the y axis indicate the number of class III genes with a given number of insertions. The curve denotes the pattern expected for random selection of class III gene targets.

Ty1 insertion at class III genes. Transcription of class III genes is required for targeted integration by Ty1 (Takamu and Oshima 1970), and this motivated investigation into the relationship between targeting patterns and Pol III occupancy at various class III genes. Two tRNA genes, *tT(UGU)H* and *tP(AAG)C*, have high levels of TFIIIB occupancy but low levels of Pol III (due to a premature termination signal and a sub-optimal initiation site sequence, respectively)(Soragni and Kassavetis 2008). These two sites received disparate levels of insertion: *tT(UGU)H* received at least 14 times more insertions than *tP(AAG)C. SNR6* has reduced levels of

TFIIIB and TFIIIC binding relative to tRNA genes, but a similar ratio of TFIIIC/B to Pol III as those seen at most tRNA genes (Soragni and Kassavetis 2008). Despite the modest reduction in Pol III transcription complexes, *SNR6* was a relatively hot target (156 insertions). On the other hand, the *ZOD1* locus has abnormally high levels of TFIIIC, modestly reduced levels of TFIIIB and little Pol III (Soragni and Kassavetis 2008). *ZOD1* was devoid of insertions, suggesting that a basal level of Pol III occupancy is important for targeting. The *S. cerevisiae* genome contains 8 loci called Extra TFIIIC (ETC) sites (*ETC1-8*) (Moqtaderi and Struhl 2004) that bind TFIIIC but not TFIIIB or Pol III (Simms, Dugas et al. 2008). All ETC sites received no insertions. We conclude, therefore, that while some subunits of Pol III transcription factor complexes discriminate targets from non-targets, none are significantly correlated with Ty1 insertion frequency. This suggests that while these particular subunits of Pol III are associated with target sites, they are not the primary targeting determinants. One conclusion, however, is clear: TFIIIC by itself, and probably TFIIIC and TFIIIB together, are not sufficient to direct Ty1 insertion.

Logistic regression to identify Ty1 targeting determinants. We were interested in further understanding the features important for targeting to class III genes as well as to sites elsewhere in the genome. Because numerous genomic features could affect Ty1 insertion patterns, we applied logistic regression to identify those features associated with Ty1's preferred target sites. The feature dataset was extensive and included genome-wide information on nucleosome position, histone modifications, and transcription factor occupancy (Supplementary Table 2). Our analysis treated each base pair in the genome as a potential insertion site and attempted to tell the difference between those with and without insertions. The quality of the models was evaluated using the area under the Receiver Operating characteristic Curve (AUC of the ROC curve)(Bradley 1996). We also trained our models on individual class III genes to identify features that distinguish hot and cold gene targets. These models, however, only

generated a subset of features with AUCs of low magnitude, implying the identified features are not essential for distinguishing class III gene targets (data not shown).

Logistic regression using the genome-wide datasets identified a small set of features associated with Ty1 insertions (Figure 3-3). As expected, these features included the region upstream of tRNA genes, which was almost perfectly predictive of a nucleotide that hosts insertions. Ty1 also preferred sites with H3K14 acetylation, the histone variant H2AZ, preexisting Ty LTRs, nucleosomes (predicted by hidden Markov modeling) and regulatory regions of genes transcribed by Pol II. Ty1 insertions avoided verified ORFs. The AUCs were stable regardless of whether one or two insertions were used as the minimum to define an insertion site. In the following sections we address in greater detail the genome-wide determinants of Ty1 targeting based on logistic regression.

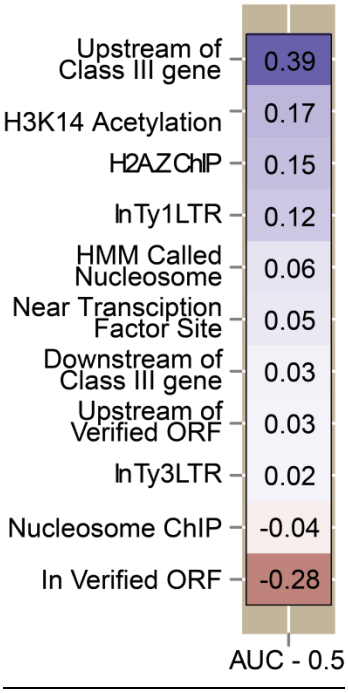

Figure 3-3 - Association of Ty1 insertions with chromosomal features.

Only a subset of features are shown for which significant positive (blue) or negative (red) AUC values were obtained by logistic regression. The color intensity denotes the strength of the association. Actual values shown are AUC-0.5. The analysis treated each base pair as a potential insertion site.

Ty1 insertions and nucleosomes. Logistic regression identified upstream regions of class III genes as most strongly predictive of insertion sites, and so Ty1 insertions upstream of class III genes were combined into a single distribution aligned on the start site of RNA Pol III transcription (Figure 3-4 a). This pattern, as previously noted (Bachman, Eby et al. 2004), is damped periodic with the amplitude attenuating with increasing distance from the start site. The amplitude reached background approximately 650 bp upstream of the transcription start. To better visualize the pattern, we applied spline smoothing to the combined data. Six distinct peaks were apparent, and the distances between peaks suggested three periods each with two peaks. The average period was 174 bp, similar to the 182 bp expected between nucleosomes. Because nucleosomes (as predicted by HMM modeling) were also predictive of Ty1 targets, we used a genome wide atlas of nucleosome positions to overlay nucleosome density onto the Ty1 insertion pattern (Lee, Tillo et al. 2007). This overlay revealed a tight association between the areas of lowest nucleosome density and the deepest troughs in insertion frequency. The more shallow insertion troughs were associated with the center of nucleosome-dense regions.

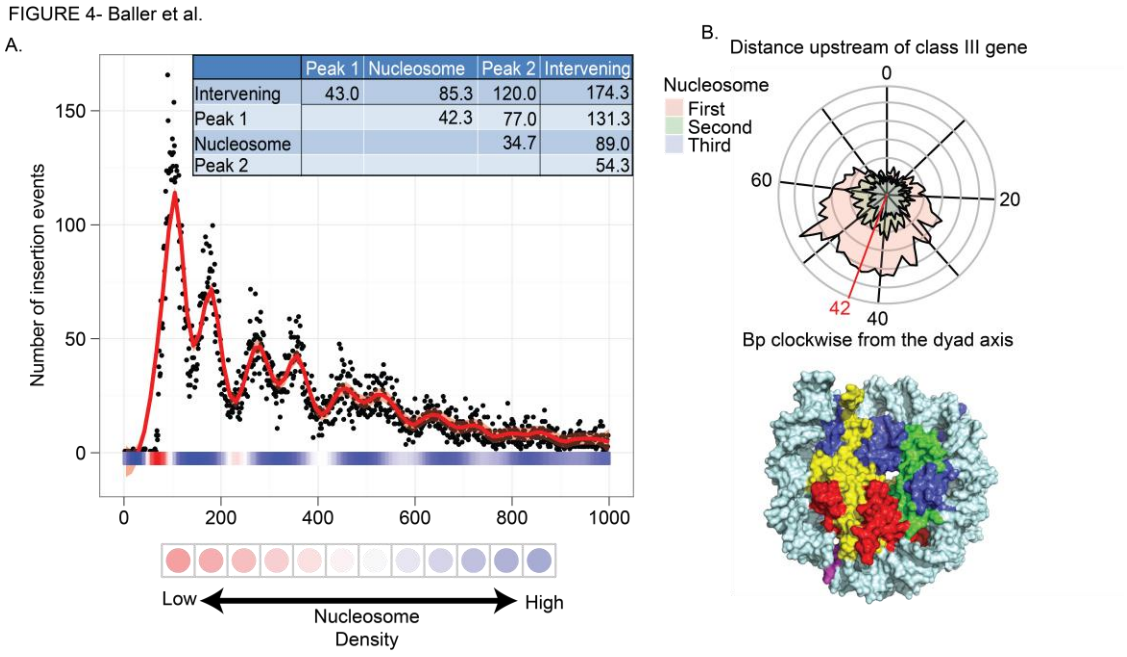

Figure 3-4 - Association of Ty1 insertions with nucleosomes.

A) Ty1 insertions upstream of class III genes were plotted in a single distribution relative to the start of transcription (position 0 on the x axis). Nucleosome density is depicted by the color of the x axis (Lee, Tillo et al. 2007). Blue denotes the presence of nucleosomes; red denotes the lack of nucleosomes. The intensity of the color indicates the strength of the signal. The y axis indicates the number of insertions per 10 bp. The red line in the graph depicts the spline-smoothed data. The spline identifies three periods, each with two peaks and two troughs. The deepest troughs (intervening) occur at approximately base positions 60, 220, 400 and 590. The other troughs occur within nucleosome-rich regions. Peak1 is the highest peak in each period; peak2 is the lowest. The inset provides calculated average distances (in bp) between features in the spline-smoothed data. B) A radial plot depicting the distribution of insertions relative to the wrapping of DNA in nucleosomes. Each rung of the radial plot denotes 20 insertions. The 0 point is the dyad axis of the nucleosome. The three colors indicate the three periods observed in panel A; 'first' denotes the plot of insertions that occurred within DNA bound by the first nucleosome upstream of the transcription start site; 'second' denotes insertions within the second nucleosome, etc. Note that the two peaks of insertions within a period are coincident on the nucleosome, and the red line indicates the coincident peaks of the spline-smoothed data. Below the radial plot is a space-filling model of a nucleosome. Yellow, H2A; red, H2B; blue, H3; green, H4. The position on the nucleosome-bound DNA of the coincident peaks of the spline-smoothed data is marked in pink.

A radial plot was used to represent the wrapping of DNA in nucleosomes (Figure 3-4 b). The two insertion peaks from each period mapped to the same region of the radial plot, indicating that they occurred in the same region of the nucleosome. We used a positional index to describe the position of DNA on the nucleosome. Position zero defines the nucleotide of the dyad axis on the face of the nucleosome with a single DNA helix. According to this index, the spline-smoothed peak of insertions was located 42 bp in the clockwise direction. This region is near the H2A:H2B interface. These results imply that the periodic insertion pattern is driven by an interaction of Ty1 IN with nucleosomes or nucleosome-associated factors. Modification of histone tails could be a contributing determinant; however, the location of tails in the crystal structures is not necessarily reliable.

Ty1 insertions and endogenous Ty elements. As described above, we did not exclude Ty1 insertions into repetitive DNA. This was particularly important in the analysis of insertions in endogenous Ty elements, particularly the Ty1 LTRs, which received numerous integration events. We mapped both ambiguous and unambiguous Ty1 insertions onto a canonical Ty1 LTR, identifying several peaks and troughs (Figure 3-5). Since most Ty1 LTRs are upstream of tRNA genes, we asked whether the observed pattern could be explained by positioned nucleosomes in

these regions. The distance between each LTR and nearby class III gene was determined and used to map onto the canonical LTR sequence the nucleosome occupancy peaks based on the periodic distributions of nucleosomes upstream of class III genes. The distribution of the nucleosomal peaks closely mirrors the distribution of Ty1 insertions into the LTRs. This suggests that the frequency and distribution of insertion into Ty1 LTRs is a consequence of positioned nucleosomes upstream of class III genes.

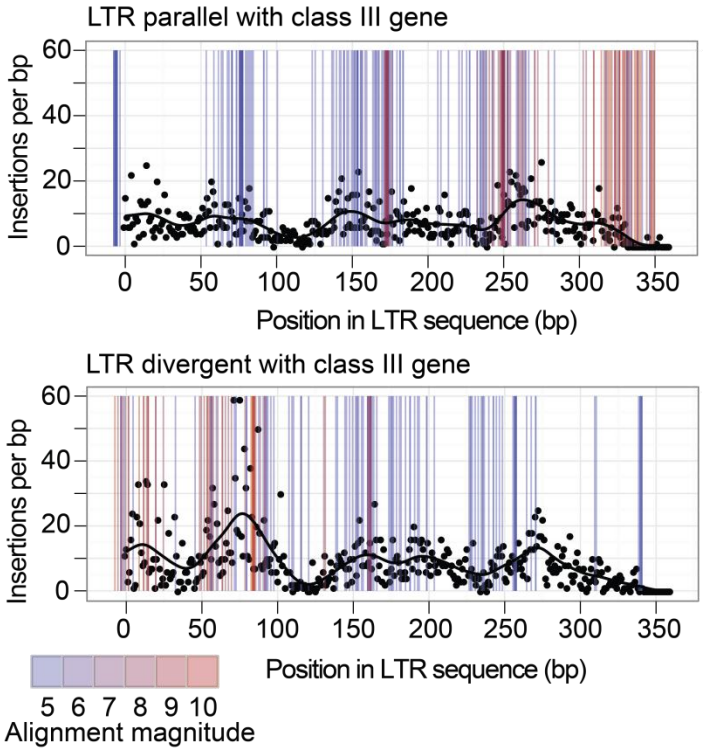

Figure 3-5 - Association of Ty1 insertions with endogenous Ty1 LTRs.

Ty1 insertions in a canonical Ty1 LTR were plotted. Separate plots were generated for insertions in LTRs (represented by black dots) in the same orientation (parallel, upper plot) or opposite orientation (divergent, lower plot) with respect to the direction of transcription of the adjacent class III gene. The distance was then calculated between the end of the LTR and the start of transcription of the adjacent class III gene. Using this distance, the expected position of integration peaks were plotted based on the data in Figure 4A. These expected peaks are shown in the plots as colored lines. The gradient of pink to blue color denotes the expected magnitude of the peaks observed at increasing distances from the start of transcription: pink represents the highest peak expected and blue the lowest. The alignment of the colored bars with the insertion peaks (black dots) suggests that the non-random distribution of insertions across the LTR is due to integration into phased nucleosomes upstream of class III gene targets.

Ty1 insertions and class II genes. Verified ORFs were the strongest negative predictor of Ty1 insertion sites, whereas positive predictors were factors associated with transcription and gene regulatory regions (Figure 3-3). For example, H3K14 acetylation was positively correlated with Ty1 insertions, and this epigenetic mark, which is mediated by *GCN5*, is associated with many highly transcribed genes (Pokholok, Harbison et al. 2005). We hypothesized that discrimination against ORFs may be due to targeting to upstream regions of genes transcribed by Pol II, similar to what was observed at class III genes. We therefore mapped insertions relative to verified ORFs, looking specifically at insertions that occurred within the ORF or either 1 kb upstream or downstream (Figure 3-6). The up- and downstream regions showed a symmetric pattern of insertions, with the first 400 bp on either side of the coding sequence receiving the fewest insertions. The increase in insertions adjacent to either end of the ORFs was coincident with the rise in nucleosome density. This pattern was consistent in both the haploid and diploid datasets, and underscores our previous arguments that avoidance of ORFs is not due to selection. Rather, the pattern of insertions up- and downstream of coding sequences is consistent with targeted insertion into nucleosome-rich regions flanking genes.

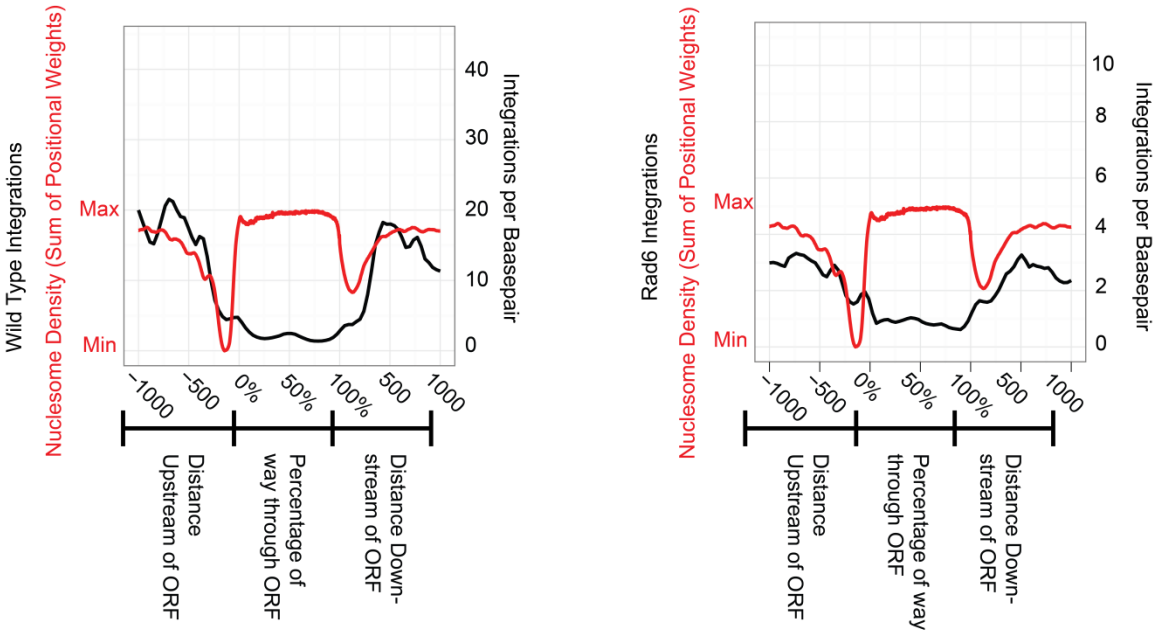

Figure 3-6 - Association of Ty1 insertions with class II genes.

The x axis describes the area within and around class II genes. Regions 1 kb up- and downstream of the coding region are shown. The falloff in insertions further from the ORF was due to intergenic regions shorter than 1000 bp. Coding regions are depicted as a normalized scale. The y axis describes nucleosome density (red) or the number of Ty1 insertions (black) in A) wild type of B) *rad6Δ* strains.

Ty1 insertion patterns in mutant backgrounds. The genome-wide pattern of Ty1 insertions was analyzed in four mutant backgrounds that have previously been shown to have altered levels or patterns of Ty1 transposition– *hos2Δ*, *rrm3Δ, rtt109Δ* and *rad6*Δ (Liebman and Newnam 1993; Mou, Kenny et al. 2006; Nyswaner, Checkley et al. 2008; Stamenova, Maxwell et al. 2009; Eaton, Galani et al. 2010). Previous work showed that the histone deacetylase, Hos2, acts at tRNA genes to promote Ty1 insertion (Mou, Kenny et al. 2006). Whereas Hos2 may increase the frequency of Ty1 insertion at class III targets, the genome-wide distribution of insertions in *hos2Δ* is not significantly different from wild type (pairwise Pearson correlations for BY4741 vs. *hos2Δ* = [.88, .88], p=.001). The frequency of insertion into verified ORFs was also equivalent to wild type (4.9%) (Figure 3-7 a) as was the distribution of insertions at different class III genes ([0.91,0.96] at 99.9% certainty) (Figure 3-7 b). Further, no discernable change in insertion pattern was observed upstream of class III genes: all six nucleosome-associated peaks

identified in the wild type were present in the *hos2Δ* with similar relative heights and spacing (Figure 3-7 c). This finding is consistent with the hypothesis that Hos2 influences integration efficiency and not integration specificity.

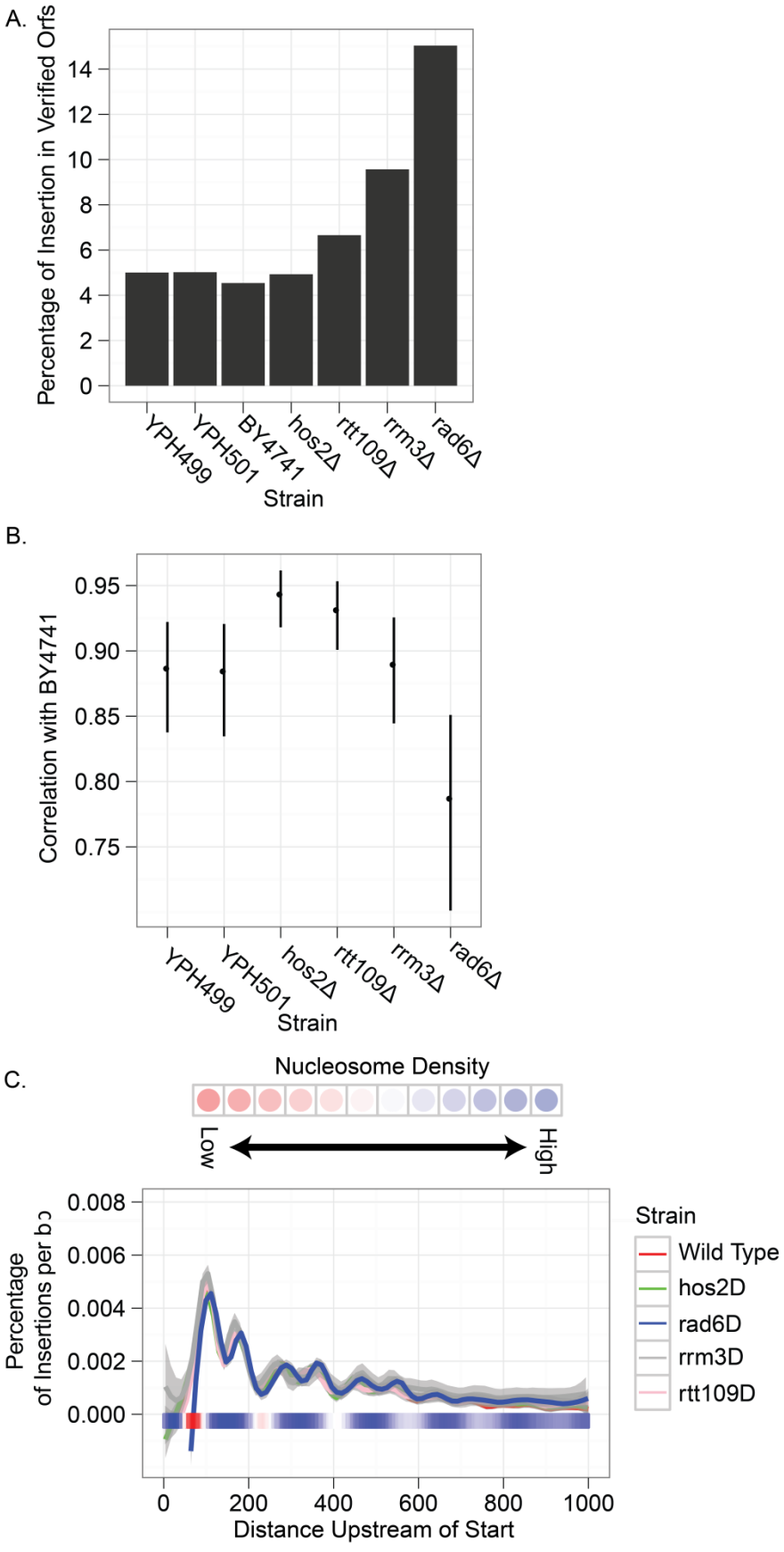
A.
Percentage of Insertion in Verified Orfs
0
2
4
6
8
10
12
14
YPH499
YPH501
BY4741
hos2Δ
rtt109Δ
rrm3Δ
rad6Δ
Strain
B.
Correlation with BY4741
0.75
0.80
0.85
0.90
0.95
YPH499
YPH501
hos2Δ
rtt109Δ
rrm3Δ
rad6Δ
Strain
C.
Nucleosome Density
Low
High
Percentage of Insertions per bp
0.000
0.002
0.004
0.006
0.008
0
200
400
600
800
1000
Distance Upstream of Start
Strain
Wild Type
hos2D
rad6D
rrm3D
rtt109D

Figure 3-7 - Distribution of Ty1 insertions in mutant strains.

A) Percentage of insertions that occurred in verified ORFs in different wild type and mutant strains. B) Pairwise Spearman correlations between strains based on the number of insertions in 2 kb windows upstream of class III genes. This provides a measure of the consistency in targeting between strains to particular class III genes. BY4741 serves as the reference strain. Error bars represent a p value of 0.001. C) Pattern of targeting upstream of class III genes in wild type and mutant backgrounds. The graph is the same as described in Figure 4A with the exception that only the spline-smoothed data is shown. Also, the y axis has been normalized with respect to the total number of insertions in the upstream region. Shading around each spline denotes error for the approximation at p = 0.05.

Rtt109 acetylates histone H3 on K56 and K9 residues, which is important for repression of Ty1 mobility, genome stability and cell survival of DNA damage (Scholes, Banerjee et al. 2001; Barry and Bell 2006; Mott and Berger 2007; Leinonen, Akhtar et al. 2012). The *rtt109Δ* background showed no significant changes from the wild type with respect to global distribution of insertions (pairwise Pearson correlations for BY4741 vs. *rtt109Δ* = [.92, .93], p=.001). The rate of insertion into verified ORFs was a moderate 6.6% (Figure 3-7 a) and the variance in class III gene target preference correlated strongly ([0.90, 0.95] at 99.9% confidence) with BY4741 (Figure 3-7 b). In addition, the pattern of insertion upstream of class III genes did not differ from wild type (Figure 3-7 c).

The Rrm3 "sweepase" is a DNA helicase that allows DNA replication forks to traverse non-nucleosomal protein:DNA complexes, such as the Pol III transcription complex on tRNA genes (Chen, Speck et al. 2008). The *rrm3*Δ mutation increases Ty1 mobility by promoting the insertion of multiple cDNA molecules, sewn together by recombination, into the genome (Stamenova, Maxwell et al. 2009). The global Ty1 integration patterns in *rrm3*Δ were less well correlated with wild type than the other two strains (pairwise Pearson correlations for BY4741 vs. *rrm3*= [.71, .72], p=.001), and a significantly higher percentage (9.6%, p= 4E-6) of insertions occurred in verified ORFs (Figure 3-7 a). However, patterns of insertion into upstream regions of class III genes correlated with BY4741 ([0.84, 0.92] at 99.9% certainty (Figure 3-7 b), and the

insertions that did go to class III genes reflected the wild type pattern with respect to nucleosome positioning (Figure 3-7 c).

The E2 conjugating enzyme Rad6 is involved in a number of aspects of DNA repair and genome stability (Game and Chernikova 2009). The *rad6Δ* background received a considerably higher frequency of Ty1 insertions into verified ORFs, with 15.0% of insertions going into verified ORFs (Figure 3-7 a). This is consistent with global targeting patterns, which showed the moderate correlation with wild type (pairwise Pearson correlations for BY4741 vs. *rad6Δ* = [.75, .75], p=.001). Higher frequency of integration into ORFs in *rad6Δ* is also consistent with previous studies that described higher levels of mutagenesis of *CAN1* and *URA3* by Ty1 (Liebman and Newnam 1993; Eaton, Galani et al. 2010). Despite this loosened target specificity, when Ty1 insertions were mapped with respect to the coding sequence of all class II genes, the pattern observed was similar to wild type, namely there was a preference for nucleosome-bound regions flanking genes (Figure 3-6 b). Ty1 showed a similar preference for class III gene targets as BY4741 in the *rad6Δ* background ([0.70,0.85] at 99.9% certainty) (Figure 3-7 b), and in the upstream regions of class III genes, the pattern of insertions grew, if anything, more pronounced (Figure 3-7 c). In addition to the pronounced six nucleosomal peaks, two more peaks 600 to 775 bp upstream of the transcription start site were evident. All peaks matched the magnitudes and spacing observed in wild type.

Discussion

The use of high throughput DNA sequencing to map large numbers of transposable element insertions is increasingly employed to understand how mobile elements interface with their host genome (Nawotka and Huberman 1988; Crawford, Chajara et al. 1995; Yang, Rhind et al. 2010)). In species such as *Saccharomyces cerevisiae*, the availability of genome-wide datasets for a large number of chromosomal features and functions (e.g. histone modifications or sites of DNA replication) makes it possible to relate insertion sites to diverse aspects of genome biology. Using

these resources, we undertook a rather straightforward approach: we used machine learning (specifically logistic regression) to assess relationships between various chromosomal features and Ty1 insertions to better understand how this mobile element selects integration sites. This use of this approach for ascertaining targeting determinants was pioneered for analysis of large datasets of retroviral insertions, and the analytic approaches we used were based on this previous work (Berry, Hannenhalli et al. 2006).

Class III genes are preferred Ty1 targets (Takamu and Oshima 1970), and fully 90% of the more than 150,000 mapped insertions occurred within a 2 kb window upstream of class III gene transcription start sites. Chromosomal localization of particular Pol III subunits did not explain targeting patterns. For example, Extra TFIIIC (ETC) sites (*ETC1-8*) (Moqtaderi and Struhl 2004) that bind TFIIIC but not TFIIIB or Pol III (Simms, Dugas et al. 2008) received no insertions. We conclude, therefore, that specific components of the Pol III complex are not targeted by Ty1, but rather Ty1 recognizes other feature(s) associated with sites of Pol III transcription. Our study also revealed wide variation in the number of insertions that occurred at different class III genes. Hot and cold targets were consistent between different wild type strains as well as with an earlier study that mapped a smaller number of Ty1 insertions at a subset of tRNA genes (Bachman, Eby et al. 2004) and with the large-scale analysis of Ty1 target site choice described in the companion study by Mularoni et al. (Mularoni, Zhou et al. 2011). Variation in insertion frequency at different class III genes, therefore, appears to be an inherent property of the targets. Insight into the underlying basis for Ty1's preference for different class III genes, however, was frustrated by our inability to identify a genomic feature(s) specifically associated with hot or cold targets.

The density of Ty1 insertions recovered by high throughput sequencing made it possible, for the first time, to comprehensively evaluate non-class III gene targeting. A strong negative association was observed between Ty1 insertions and verified ORFs. Selection was ruled out as

the basis for ORF-avoidance, because the frequency of insertion into ORFs was not significantly different between haploid and diploid strains. It would be expected that deleterious effects of an insertion would be mitigated, at least in part, by a second copy of the gene, leading to a higher frequency of ORF insertions in the diploid. Since selection did not significantly influence targeting patterns, this suggests that ORFs are not competent to receive Ty1 insertions either due to the absence of a targeting determinant or the presence of a repulsive factor. The ability of Ty1 to discriminate between coding and non-coding sequences likely has a selective advantage for Ty1, as it minimizes negative consequences of insertional mutagenesis and ensures host survival.

In addition to selection biases that might result from mutation of host genes, our experimental approach required expression of a *HIS3* reporter carried on Ty1 cDNA; biases in insertion site patterns may result if *HIS3* is not expressed in certain chromosomal environments. However, in previous work with the related yeast retrotransposon Ty5, for example, we found *HIS3* to be a very robust reporter for recovering insertions in heterochromatin – the preferred sites of Ty5 integration (Crawford, Chajara et al. 1995). Additionally, the experimental approach for recovering Ty1 insertions in the companion study by Mularoni et al. did not select cells harboring Ty1 integration events and yet produced a similar genome-wide pattern of insertions (Mularoni, Zhou et al. 2011).

One difference between our study and that of Mularoni et al. is that we did not recover Ty1 insertions in mitochondrial DNA. Because our insertion site dataset is smaller than Mularoni et al., mitochondrial insertions may be below our detection threshold. Based on the Mularoni et al data, 0.011% of sequenced reads matched mitochondrial DNA, suggesting that we should find only about 40 mitochondrial sites in our collection of more than 390,000 sequencing reads generated from wild type strains. In addition, Mularoni et al. suggest that some mitochondrial insertions may have occurred in DNA fragments released from shattered mitochondria, and a

subset of these events may not give rise to His+ cells, and therefore they would have not been recovered by our approach.

Ty1 and the nucleosome. Our analyses revealed the nucleosome as a new targeting determinant for Ty1. Logistic regression showed a significant positive association between Ty1 insertions and nucleosomes, especially for well-positioned nucleosomes (i.e. those predicted by hidden Markov modeling), such as those found upstream of tRNA genes. In contrast, a slightly negative association was observed between Ty1 insertions and nucleosomes using ChIP data. This is because the vast majority of nucleosomes genome-wide did not receive Ty1 insertions, but rather there was a distinct bias for specific nucleosomes that were targeted. In addition to the nucleosomes upstream of class III genes, nucleosomes flanking ORFs were much preferred over those located in coding sequences. The nucleosome preference also explains patterns of insertion observed in preexisting Ty1 elements, which are due in large part to their proximity to class III genes and associated, well-positioned nucleosomes.

Insertions into nucleosome-bound DNA did not distribute evenly, but instead were enriched at one end of the dyad axis. At this position, insertions struck both helices in both orientations, and the peak of insertions was the same as that observed by Mularoni et al. in their related study (Mularoni, Zhou et al. 2011). The pattern of insertions on the nucleosome is consistent with an interaction between Ty1 integrase and a specific histone modification or nucleosome-associated factor. A significant positive association was observed between Ty1 insertion sites and H3K14 acetylation; however, this modification is generally characteristic of transcriptionally active regions of the genome (Pokholok, Harbison et al. 2005), and so the association could be correlative. A strong positive association was also observed with the histone variant H2AZ, which is typically associated with promoter-proximal nucleosomes of both active and inactive genes in euchromatin (Raisner, Hartley et al. 2005; Liachko, Bhaskar et al. 2010; Papamichos-Chronakis, Watanabe et al. 2011). H2AZ replaces H2A in nucleosomes, and it is the

region of nucleosomal DNA near the H2A/H2B interface that is most highly targeted by Ty1. Further, *S. cerevisiae* strains lacking H2AZ show decreased levels of Ty1 transposition (Dakshinamurthy, Nyswaner et al. 2010), and decreases in levels of H2A and H2B alter patterns of integration at the *CAN1* locus (Rinckel and Garfinkel 1996). To evaluate more specifically a role for H2AZ in targeting, we performed logistic regression using models that test whether H2AZ is preferentially associated with hot or cold class III gene targets (data not shown). No significant association (positive or negative) was observed, and thus whether H2AZ has a specific role in targeting awaits further testing.

An alternative hypothesis to explain targeting to nucleosomes is that there exists an intermediary, bridging factor that links the Ty1 integration complex to nucleosome-bound DNA. Candidates include chromatin remodelers, some of which are known to affect Ty1 insertion patterns. For example, loss of ISW2 alters the periodic pattern of Ty1 insertion upstream of class III gene targets (Kaplan, Moore et al. 2009). However, this is likely due to changes in nucleosome positioning, as catalytically inactive ISW2 does not change overall targeting to tRNA genes (Brewer and Fangman 1987).

Nucleosomes are also preferred targets for retroviruses, due to distortion of nucleosome-bound DNA that allows access to retroviral integrase and promotes the integration reaction (Pryciak, Muller et al. 1992; Pryciak, Sil et al. 1992). Mapping of large numbers of genomic HIV and gammaretrovirus insertions revealed that they occur in a periodic fashion on the surface of the nucleosome, consistent with favored integration on the outward-facing DNA surface, a pattern not observed for Ty1 (Wang, Ciuffi et al. 2007; Roth, Malani et al. 2011). Like Ty1, however, insertions of HIV and gammaretroviruses were both associated with epigenetic modifications correlated with transcription. Ty1's preference for nucleosomal DNA stands in contrast to the related *S. cerevisiae* retrotransposon Ty5, which prefers nucleosome-free DNA for integration (Crawford, Chajara et al. 1995). Nucleosomes are also avoided by the DNA transposon Hermes

when it transposes in yeast (Nawotka and Huberman 1988). Clearly considerable variation exists with respect to how mobile elements interact with nucleosomes during transposition.

An association between Ty1 insertions and nucleosomes is also observed in the regions flanking class II genes. Nucleosomes are relatively abundant within the coding sequence; however, as mentioned above, coding sequences are particularly cold for Ty1 integration. In the first few hundred base pairs upstream and downstream of the coding sequence both nucleosomes and Ty1 insertions are largely absent, but further away from the coding sequence, the number of Ty1 insertions rise, coincident with the presence of nucleosomes. Our dataset of Ty1 insertions is too small to make more precise conclusions about the relationship between intergenic nucleosomes and Ty1; however, in light of the relationship between nucleosomes and tRNA genes, it is possible that a specific histone modification or chromatin factor present in the flanking regions of class II genes attracts Ty1 insertions. We propose that there is a common mechanism underlying targeting at both class II and class III genes and that the abundance of a particular factor – histone modification or bridging factor – determines degree of target competency. Said factor or modification, is particularly enriched in class III genes and is most abundant at the nucleosome closest to the start of transcription.

High throughput mapping of insertion sites in mutant strains. Another advantage of *S. cerevisiae* as an experimental system is the wealth of genetic resources that can be applied to better understand mechanisms of transposable element target specificity. As a first step in this direction, we mapped large numbers of Ty1 insertions in strains with mutations previously shown to impact frequency or specificity of Ty1 transposition, thereby allowing us to better describe the integration specificity phenotype. Neither loss of the histone deacetylase Hos2 nor the histone acetyltransferase Rtt109 had any impact on target site choice, although both are known to influence transposition frequency (Scholes, Banerjee et al. 2001; Mou, Kenny et al. 2006). Because the transposition defect in these strains occurs after cDNA synthesis, our data suggest

these proteins influence integration efficiency. The loss of the DNA helicase Rrm3 had a modest impact on target site choice, whereas loss of the E2 conjugating enzyme, Rad6, resulted in significantly higher numbers of insertions into ORFs (~5% for wild type vs 15% for *rad6Δ*). Increased mutagenesis in counter-selectable gene targets was previously observed in *rad6Δ* strains (Liebman and Newnam 1993; Eaton, Galani et al. 2010), and it appears that this loosened target specificity occurs genome wide. Because patterns of insertion near tRNA genes were largely unperturbed, the underlying determinants of nucleosomal targeting are intact in *rad6Δ* strains. This is consistent with Rad6 acting to strengthen the targeting signal, such that in its absence, some integrations go astray. Interestingly, one of the targets of Rad6 is H2A (Robzyk, Recht et al. 2000), and the loosening of target specificity may be due to altered modification of this protein.

Whereas our analysis of a handful of mutants did not allow us to make new conclusions about Ty1 targeting mechanisms, it nonetheless illustrates the potential for characterizing large numbers of insertions in mutant backgrounds to dissect Ty1 targeting mechanisms. Clearly one direction for future genetic studies will be to identify the factors that create the distinct nucleosomal surface upstream of genes transcribed by RNA Pol III that is such an attractive target for Ty1 integration.

Materials and Methods

Generating Ty1 insertions. Plasmid pGTy1*his3AI*-SCUF contains six nucleotide substitutions in the U5 region of the 5' LTR of Ty1-H3 downstream of the initiation codon of the *gag* ORF. The nucleotide substitutions are underlined in the following sequence, which comprises nucleotide 1-24 of *gag*: ATGGAATCCCAACAGCTTAGCCAA. Substitutions were introduced by overlap extension PCR using pGTy1*his3AId1* (Nyswaner, Checkley et al. 2008) as template DNA.

Plasmid pGTy1*his3AI*-SCUF DNA was transformed into strains YPH499, YPH501, BY4741 and *rrm3*Δ*::kanMX*, *hos2*Δ*::kanMX*, *rtt109*Δ*::kanMX* and *rad6*Δ*::kanMX* derivatives of BY4741. Independent Ura^+ transformants that supported a robust induction of Ty1*HIS3-SCUF* transposition were identified by growing patches of each isolate on SC-Ura 2% galactose 2% raffinose agar at 20˚C followed by replicating patches to 5-FOA-His plates. Selected pGTy1*his3AI*-SCUF transformants were grown overnight in SC-Ura 2% glucose broth at 30˚C. A 10µl aliquot of each culture was transferred to 1 ml SC-Ura 2% galactose, 2% raffinose 2% sucrose broth, and cultures were grown at 20˚C for 2 days. Cells were pelleted, resuspended in 0.2 ml ddH_20, transferred to YEPD agar and incubated at 30˚C for 16-18 hr. Cells were replicated to 5-FOA-HIS 2% glucose plates and incubated at 30˚C for two days. A 0.75 ml aliquot of ddH_20 was added to each plate, and cells were scraped from the agar into suspension using a sterile plastic scraper. The cell suspension was collected, and the agar surface was washed with an additional 0.75 ml aliquot of ddH_20. The cell suspensions were combined, and cells were pelleted; approximately 0.1 ml of cell pellet was obtained from each plate. Genomic DNA was prepared from individual pellets. Genomic DNA samples from 10 to 14 independently prepared cell pellets in each strain background were used for PCR. PCR amplification of the integration sites was based upon the linker mediated PCR protocol found in (Crawford, Chajara et al. 1995). Each sample was split, one fraction digested with AciI and the other with TaqI. Linkers were annealed and sequences with an adjacent Ty1 insertion were amplified by PCR. Barcoded primers were used in the PCR step to mark the source of the sequences (oligonucleotide sequences are available upon request).

DNA sequence processing. Raw 454 DNA sequence reads were sorted and cleaned using an in-house pipeline that employs the Smith-Waterman local sequence alignment algorithm to identify primer sequences (Smith and Waterman 1981). Reads were excluded that did not have a perfect match to a barcode and surrounding DNA or that had more than four mismatches to the

primer. Insertions at a given position and orientation were only counted once in each pool. Sequences were aligned to the genome using RazerS, a fast mapping algorithm capable of handling ambiguous insertions with no loss rate (Weese, Emde et al. 2009). For each read, only the highest quality maps with at least 98% similarity to genomic DNA were retained.

All data were housed in a relational database management system (RDBMS) with a many to many correspondence between reads and genome maps. Reads that mapped to a single genome location were labeled as unambiguous, whereas those that were related to more than one site were labeled as ambiguous. When multiple reads mapped to the same genomic location, reads from different pools or in different orientations were retained but reads from the same pool were collapsed with the least ambiguous read used as a representative.

Control sites were drawn randomly from the genome using a derived non-parametric distribution based on genomic sites for AciI and TaqI in the *S. cerevisiae* genome. The distribution was derived using the frequency and orientation of case sites in the vicinity of restriction sites. This process produced control sites with a similar bias to that of the case sites, thereby removing restriction enzyme recovery bias from the results of the logistic regression.

Data annotation and analysis. Nucleotide annotation of genomic features was based on information from the Saccharomyces Genome Database (Cherry, Adler et al. 1998), primary literature and in-house calculations (Supplementary Table 2). For features with a non-binary value, the values of overlapping features were summed. In cases of missing data, the nearest data points were averaged to interpolate the missing point. This process generated a vector of annotations for each site.

Analysis of insertion preference relied on logistic regression. Regression models were trained using the *glmnet* logistic regression function in the R statistical package (Team 2008; Friedman, Hastie et al. 2010). Models compared the set of experimentally derived integration sites (case) to a random subset of remaining sites (control), fitting equation (Eq. 1)

$$f(z) = \frac{1}{1+e^{-z}} \qquad [1]$$

where f(z) represents the class labels and z represents a linear function of form $z = \beta_0 + \beta_1 x_1$, with x being the level of the feature under investigation and β being the regression coefficents. Logistic models were evaluated using Receiver Operating Characteristic (ROC) analysis with 10 fold cross-validation. Evaluations are presented in the form of the Area Under the ROC Curve (AUC), or more precisely as AUC – 0.5. For AUC – 0.5, zero indicates a model with no predictive power and values at 0.5 or -0.5 indicate perfect prediction. The sign of the AUC indicates whether the feature is associated with case sites (positive) or control sites (negative). All visualization was handled in R using the ggplot2 graphics package (Wickham 2009).

4. Chapter Three: Prediction of Origins of Replication

Background

Origins of Replication

As the sites in the genome where genomic DNA begins replication, origins of replication are, by definition, found in all living organisms. The initiation of replication at an origin has been studied in archaea and eukaryotes and particularly well in prokaryotes *(Mott and Berger 2007).* Prokaryotes, including *Escherichia coli* and *Mycobacterium tuberculosis*, present the simplest replication process. Prokaryotes generally have a single origin of replication that fires every S phase. Initiation at prokaryotic origins is driven by a small number of highly conserved proteins. One such protein, DnaA, binds to a conserved sequence found, often in multiple copies, at the origin site. DnaA is responsible for local melting of the DNA strand and subsequent binding of the DnaB helicase. With the DnaB helicase in place, replication can progress. This basic theme holds true in both eukaryotic and archaeal organisms (Barry and Bell 2006; Chen, Speck et al. 2008). In fact there is considerable similarity between the DnaA prokaryotic initiator, Cdc6/Orc1 initiator in archaea and a number of the conserved ORC subunits in eukaryotes.

Eukaryotic origins of replication follow a process similar to that of prokaryotes. The origin recognition complex (ORC), 5 subunits of which are conserved in all eukaryotes (Chen, Speck et al. 2008; Duncker, Chesnokov et al. 2009), binds to a DNA motif in order to initiate replication. However, there are additional complexities generally not found in prokaryotes. Based on studies of the yeast *S. cerevisiae*, it is apparent that a degenerate 11bp sequence motif, referred to as the Autonomously Replicating Sequence (ARS) Consensus Sequence (ACS) motif, is necessary but not sufficient for ORC binding. A different, 50 bp motif sharing some characteristics has also been identified for *K. lactis* (Liachko, Bhaskar et al. 2010). Conservatively, there are about 6000 genomic sites that match the *S. cerevisiae* ACS motif; of those only about 253 have the ORC

complex bound (Eaton, Galani et al. 2010). Even then, only a subset of the ORC-bound sites initiates replication forks in any given S-phase. Whether the subset is static (Donaldson 2005), or stochastic (Czajkowsky, Liu et al. 2008) is still debated in the literature. However, recent work suggests that the older data supporting a static set of origins may also be compatible with the newer stochastic hypothesis (Yang, Rhind et al. 2010).

Biological Methodology

Debates about the mechanism of origin firing persist, in part, due to the laborious and relatively low-resolution technique used to validate origin sites *in vivo*. This technique, called two-dimensional non-denaturing gel electrophoresis, involves digestion of the genome by restriction enzymes, a primary separation by mass in a low-concentration agarose gel at low voltage and a secondary separation by molecular shape utilizing a high-concentration agarose gel at high voltage (Brewer and Fangman 1987). This is followed by detection of rare replication intermediates by hybridization to a radioactive DNA probe. The distinctive shapes of replication intermediates on the gel can distinguish the passage of a replication fork through a restriction fragment (Y-arcs) from the presence of an origin (bubble-arc) within the DNA fragment. This technique is relatively low-throughput and, though some enrichment can be done for forks/bubbles, is technically challenging because of the difficulty of detecting rare replication intermediates. Given that validation is troublesome, it would be useful to predict sites most likely to be origins, prioritize them for validation, and do so in a systematic way. Towards this end, I have applied a logistic regression model to the problem of origin prediction and done so in a way that allows for the scanning of whole genomes and that returns a result with nucleotide precision.

Unlike my retrotransposon analyses (Baller, Gao et al. 2011; Baller, Gao et al. 2012), this origin analysis utilizes primarily nucleosome occupancy data to predict potential sites. This approach was inspired by Eaton et al., who described arrayed nucleosomes and a distinctive nucleosome-

free region (NFR) at known origins of replication (Eaton, Galani et al. 2010) in *S. cerevisiae*. While Eaton et al. (Eaton, Galani et al. 2010) showed that a nucleosomal pattern was present at most known origins, whether the pattern was unique to origins was not addressed. Furthermore, there remained the question of whether the same pattern was conserved in other genomes. Towards this end, I evaluated the predictive power of nucleosome distributions within *S. cerevisiae* and *K. lactis*, another yeast species for which a large number of origins are known.

Results

Nucleosome occupancy data from a variety of yeast species was available from a single study, ensuring a consistent experimental methodology (Tsankov, Thompson et al. 2010). In this study nucleosomes were cross-linked to the DNA with formaldehyde, Micrococcal Nuclease was used to digest intervening DNA and resultant fragments were size selected for mononucleosomes by gel purification. The remaining fragments were sequenced using an Illumina 1G analyzer. The resultant datasets included a read count for each basepair of the respective genomes. The read count is assumed to be roughly proportional to average nucleosome occupancy at that position over a colony of cells. The analysis considered the 1024bp window of nucleosome data centered on each evaluated site. The plot of aggregate distribution at known origins (Figure 4-1) confirmed that nucleosome patterning was at least partly conserved between *S. cerevisiae* and *K.lactis*.

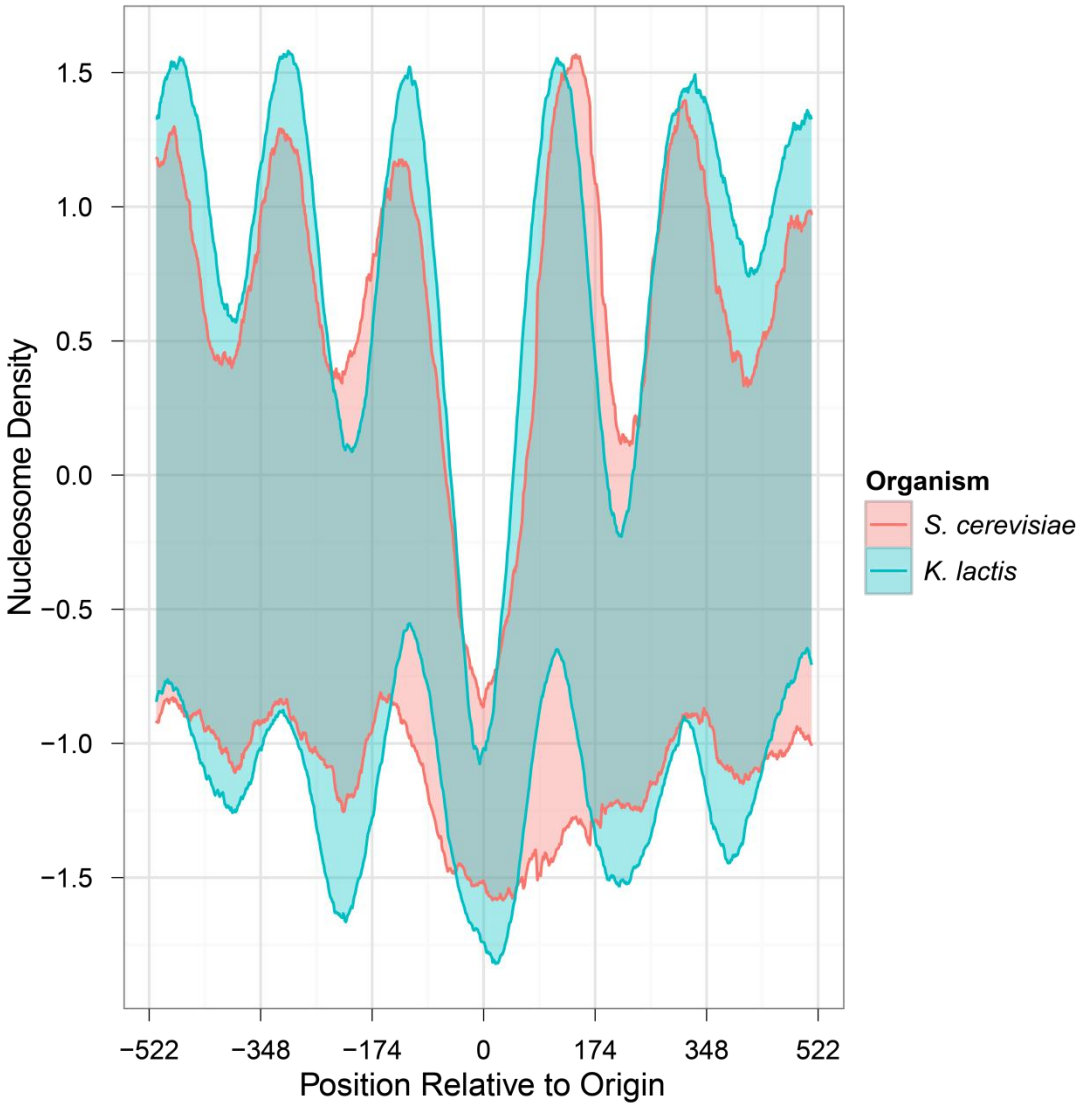

Figure 4-1 - Nucleosome Density at Origins of Replication

The x-axis indicates the distance from the center of the origin of replication identified by pink or blue for *S. cerevisiae* and *K. lactis*, respectively. The y-axis indicates the nucleosome density, determined by micrococcal nuclease digestion and high-throughput sequencing, at each position with the upper and lower bounds indicating the extent of the standard deviation around the mean.

Predictions were generated in a two-step process based on evidence that the model may have been over-fitting when a one step process was used.

First Pass

In the first step, a set of positive sites were generated by selecting the middle points of all known origins in either *S. cerevisiae* or *K. lactis*. Negative sites were selected at random from the remainder of the genome. In order to better capture patterns spanning multiple adjacent bps, the same data was transformed by fourier (FFT) and wavelet (decimated Least Asymmetry Daubechies type 4) (Sorensen, Jones et al. 1987; Daubechies 1990) and included in the modeling.

The number of features remaining in the *S. cerevisiae* trained, *S. cerevisiae* tested model (10 fold cross validation) were 10, 6 and 10 with AUCs of 0.905, 0.847 and 0.888 for the raw data, wavelet and Fourier transformed datasets respectively. The joint dataset containing all three feature sets only retained 2 features with an AUC of 0.894 (Figure 4-2). For the raw data model, 6 of the features represented points in the center of the trough and the remaining 4 points represented nucleosome peaks in the arrayed region (Table 9-1). For the wavelet model, a series of coarse features were retained (likely corresponding to the arrayed nucleosomes) (Table 9-2) and for the FFT model, the 6 lowest frequencies were retained (Table 9-3). The model using all available data used a raw data feature corresponding to the central trough and the FFT feature with a 170bp periodicity (Table 9-4).

The number of features remaining in the *K. lactis* trained, *K. lactis* tested model (10 fold cross validation) were 15, 7 and 5 with AUCs of 0.935, 0.899 and 0.934 for the raw data, wavelet and Fourier transformed datasets respectively. The joint dataset containing all three feature sets only retained 3 features with an AUC of 0.899 (Figure 4-3). For the raw data model, 5 of the features represented points in the center of the trough and the remaining 10 points represented nucleosome peaks in the arrayed region (Table 9-5). For the wavelet model a series of coarse features were retained (likely corresponding to the arrayed nucleosomes) (Table 9-6) and for the FFT model the 5 lowest frequencies, excluding the 341bp period were (Table 9-7). The model using all available

data used two raw data features corresponding to the central trough and the FFT feature with a 170bp periodicity (Table 9-8).

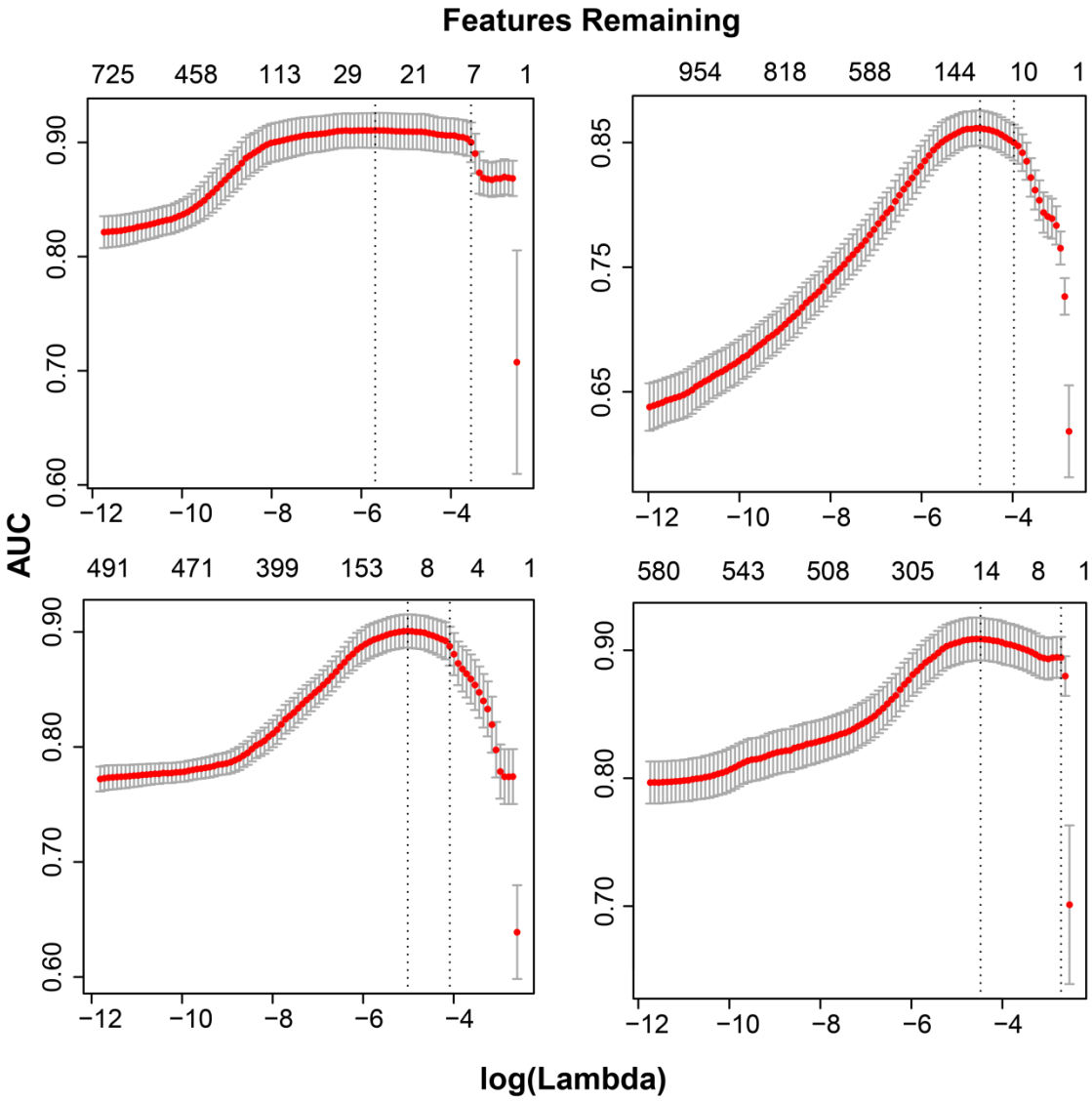

Figure 4-2 – *S. cerevisiae* LASSO regularization curves

These graphs show the LASSO regularization curves for different feature sets tested and trained in *S. cerevisiae.* Progression along the x-axis denotes increasing lambda resulting in a lower number of retained features. The y-axis denotes the AUC for a given lambda value with error bars from 10-fold cross-validation. The upper x-axis provides information about the number of features remaining in the model at a given lambda. Vertical dashed lines denote the highest AUC on the left and the

simplest model within 1 se of the maximum on the right. The panels are: (Top Left) raw data, (Top Right) wavelet data, (Bottom Left) Fourier data and (Bottom Right) all three feature sets.

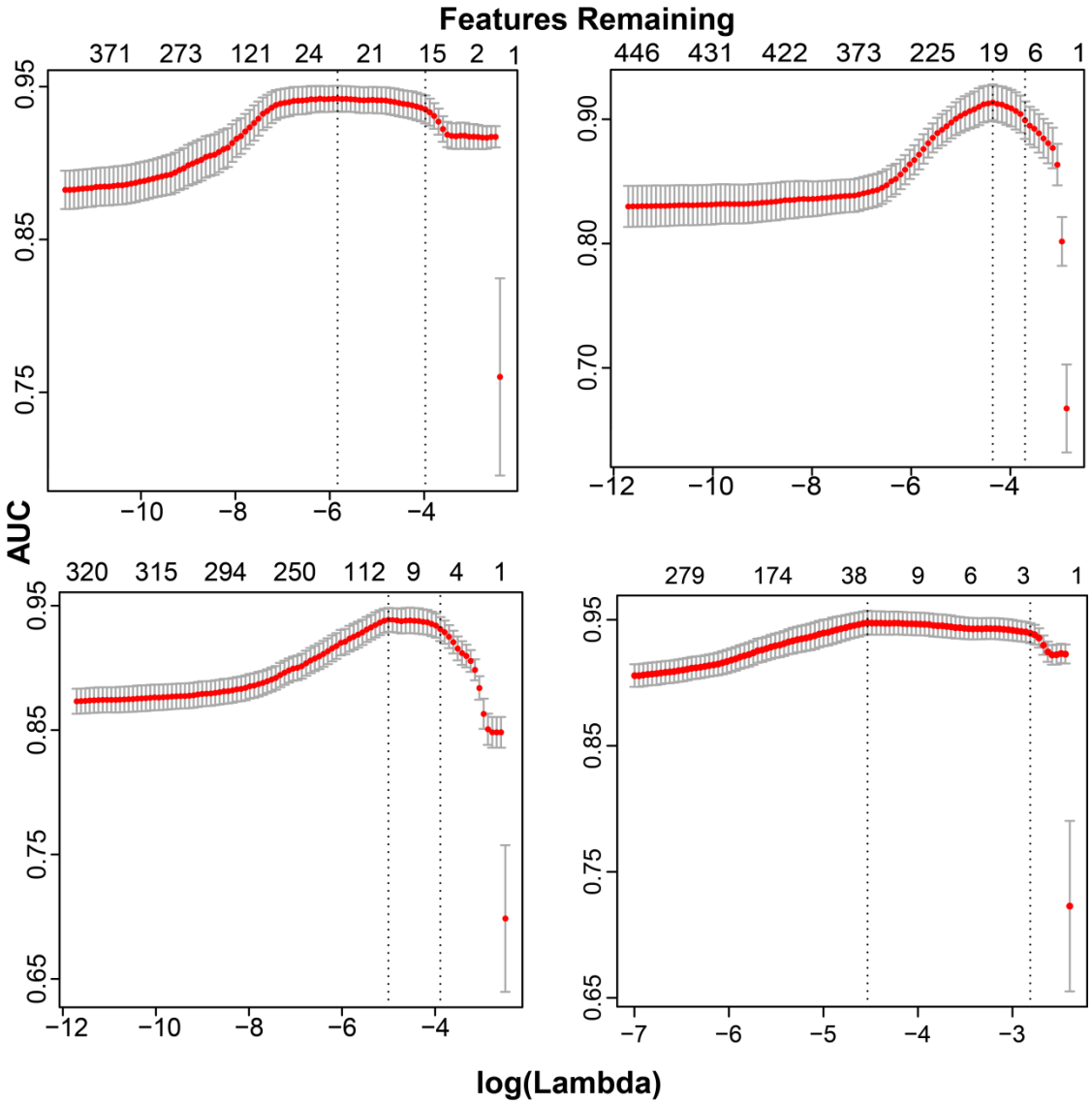

Figure 4-3 - *K. lactis* LASSO regularization curves

These graphs show the LASSO regularization curves for different feature sets tested and trained in *K. lactis.* Progression along the x-axis denotes increasing lambda resulting in a lower number of retained features. The y-axis denotes the AUC for a given lambda value with error bars from 10-fold cross-validation. The upper x-axis provides information about the number of features remaining in the model at a given lambda. Vertical dashed lines denote the highest AUC on the left and the simplest model within 1 se of the maximum on the right. The panels are: (Top Left) raw data, (Top Right) wavelet data, (Bottom Left) Fourier data and (Bottom Right) all three feature sets.

The resultant models, for both organisms, had high AUCs indicating a strong discriminative capability. However, the use or addition of transformed data failed to have a significant effect on the AUC and did not significantly simplify the models. As such, the use of the transforms was dropped from subsequent analyses.

However, the predictions were intended to enrich the list of potential origins in line for validation and to improve the resolution of predicted origins. For the model to be useful it was important to quantify the fraction of sites called as origins that actually are origins (Precision) with respect to the fraction of origins detected (Recall or True Positive Rate); this evaluation is encapsulated in precision-recall curves (Figure 4-4). The *S. cerevisiae* and *K. lactis* curves were nearly identical though the baseline proportions of origin sites to random sites were slightly different. These curves suggest that a 60% recall comes with a 50% precision. That is, 50% of the sites called as origins will be called erroneously. However, given that there are only 3000 negative cases used in these models and there are approximately 11.5 million bp in each of the genomes, the ratio of positive to negative cases used in model evaluation did not match the ratio of true origins to potential origin sites in the genome. This is relevant, as it is assumed that the negative sites selected for use in the model are representative of the genome as a whole. Thus if there are currently 100 sites that look sufficiently like origins to be falsely called as such, that number would be expect to increase proportional to the total number of negative sites. However, since adjacent basepairs show high dependency, it is safe to consider a smaller number of potential negative sites, perhaps 1.15 million or 115,000 sites. Even assuming an optimistic 100,000 sites,

there is a large resultant effect on the precision. With an assumption of 100,000 negative sites the actual precision is approximately 1% for 60% recall.

Inspection of the features surviving LASSO regularization indicated a possible cause of the low precision. A large number of the features, particularly in the raw data model, were based on the central trough of the nucleosome distribution. Regions of arrayed nucleosomes have troughs approximately every 174bp, thus a model that calls all troughs in arrayed regions as origins will show 174 fold enrichment relative to random selection. This indicates that a non-trivial component of the model's predictive power is derived from the central through but that central trough feature alone will result in low precision.

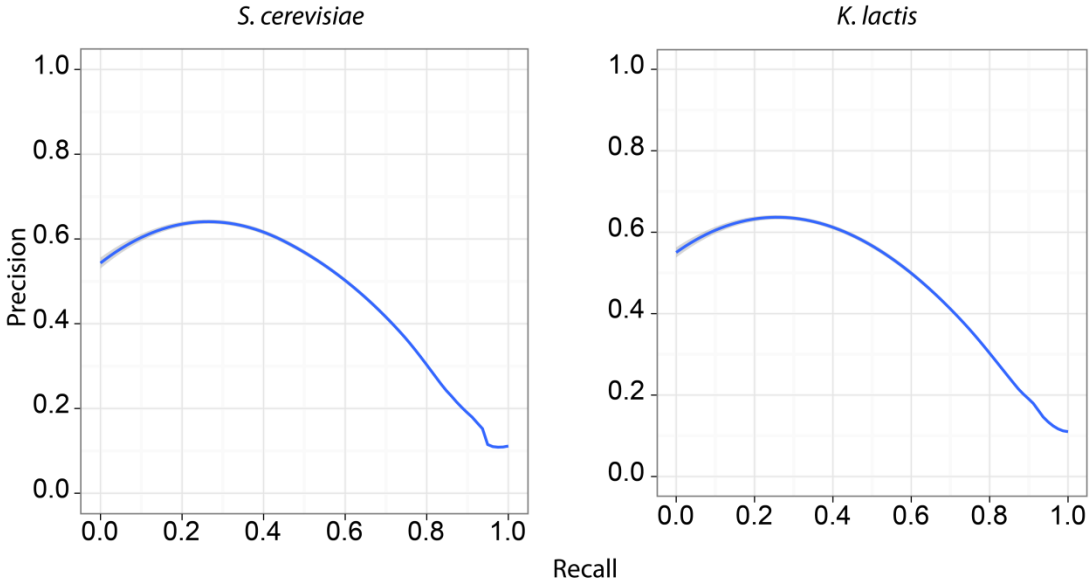

Figure 4-4 –Average Precision Recall Curve

A Loess curve generated from precision-recall curves under 10 fold cross-validation. The x-axis is the fraction of total positive cases correctly identified as positive. The y-axis is the fraction of cases identified as positive that actually are positive. The baseline for *S. cerevisiae* was 0.084 or 253 positives over 3000 negatives. The baseline precision for *K. lactis* was 0.049 or 148 positive over 3000 negatives.

Cross Species Prediction

We used the same models described above for cross species origin prediction, by applying them to the opposite dataset (Figure 4-5). When an *S. cerevisiae*-trained model was used to predict *K. lactis* origin positions, the AUC was 0.905. The precision-recall curve showed 40% precision at 60% recall, slightly worse than the same-species predictions. When a *K. lactis*-trained model was used to predict *S. cerevisiae* origin positions, the AUC was 0.923. The precision recall curve showed 75% precision at 60% recall, slightly better than the same species prediction. These results indicate that the K. lactis data produce better models. This is further supported by the *K. lactis* model outperforming the S. cerevisiae model in the same species analysis.

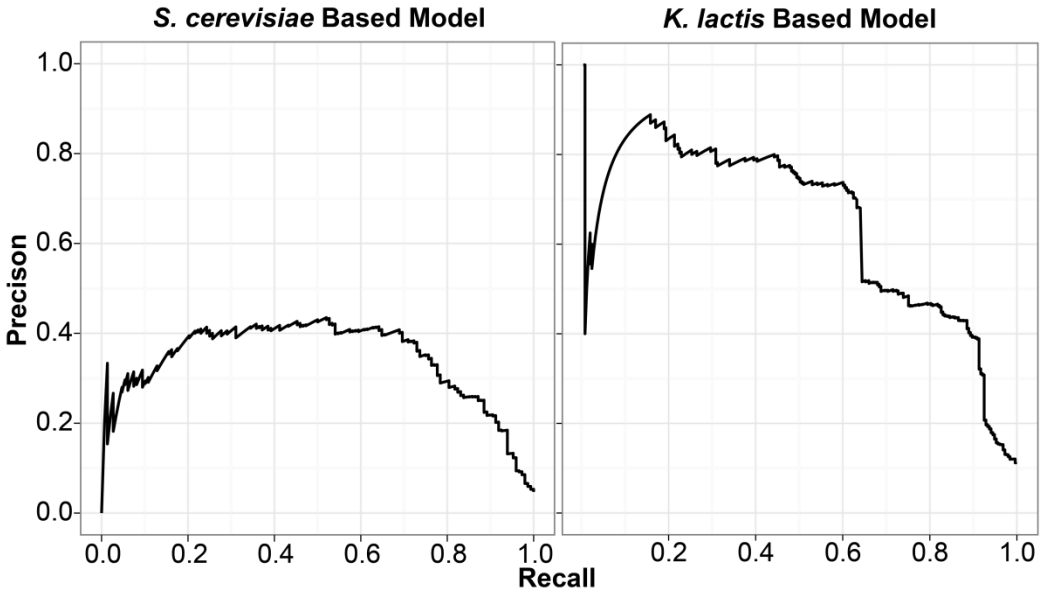

Figure 4-5 – Precision Recall Curves Based on Cross-Species Prediction

Precision recall curves generated from cross-species prediction data. The x-axis is the recall, the total fraction of positive cases identified as positive. The y-axis is the precision, the fraction of positive predictions that are true positives. The left frame is the curve from a *S. cerevisiae* model applied to *K. lactis* data. The right frame is the curve for a *K. lactis* model applied to *S. cerevisiae* data.

Second Pass

Based on the coefficients identified in the first pass (Table 9-1; Table 9-5) the first pass was highly focused on the central trough. As a result false positives in the first pass had deep central troughs, similar to those found at origin sites. To force consideration of other parts of the nucleosome pattern, a second logistic regression model was used to discriminate between sites falsely identified in the first pass and actual origin sites. It should be noted that false negatives from the first pass were excluded from analysis in the second pass in order to keep the precision recall curves accurate.

The second pass, tested and trained on *K. lactis* nucleosome data, produced only a minimal improvement to the original model with an AUC of 0.60 and did so by training on 27 different features of the raw nucleosomal data (Figure 4-6, Table 9-9). These features were scattered across the origin window and appear to refine the width of the peaks and troughs of the nucleosome pattern.

While the K. lactis chromosomes have not been annotated to the same extent as *S. cerevisiae*, there are a number of ChIP datasets available. One such dataset, an unpublished ORC occupancy study, was particularly applicable to this prediction problem. When ORC data was included in the feature set, the AUC increased to 0.93 based on a single feature, the maximal height of the ORC peak (Figure 4-7). The nucleosome data was ignored, suggesting that the ORC data was redundant with the nucleosome data at this point in the analysis. There was a precision of approximately 40% at 60% as compared to a 1.6% baseline for random selection, a considerable improvement over nucleosome data alone. The ORC data is of little benefit in the first pass due to the large numbers of ORC binding sites and the breadth, an average 3000 bp, of the sites. The model was trained using a subset of the total negative set and, as such, the full dataset baseline was actually 0.55%, meaning that the ultimate precision for this step was approximately 13%.

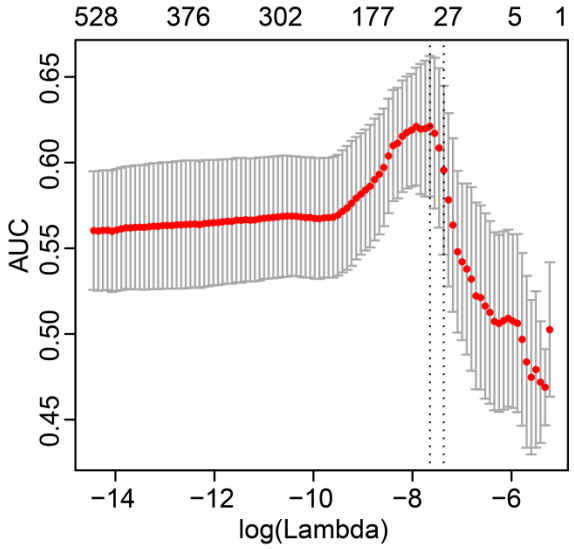

Figure 4-6 - Second Pass *K. lactis* Nucleosome Data

LASSO regularization curves for a second pass over the K. lactis nucleosome data. Progression along the x-axis denotes increasing lambda resulting in a lower number of retained features. The y-axis denotes the AUC for a given lambda value with error bars from 10-fold cross-validation. The upper x-axis provides information about the number of features remaining in the model at a given lambda. Vertical dashed lines denote the highest AUC on the left and the simplest model within 1 se of the maximum on the right.

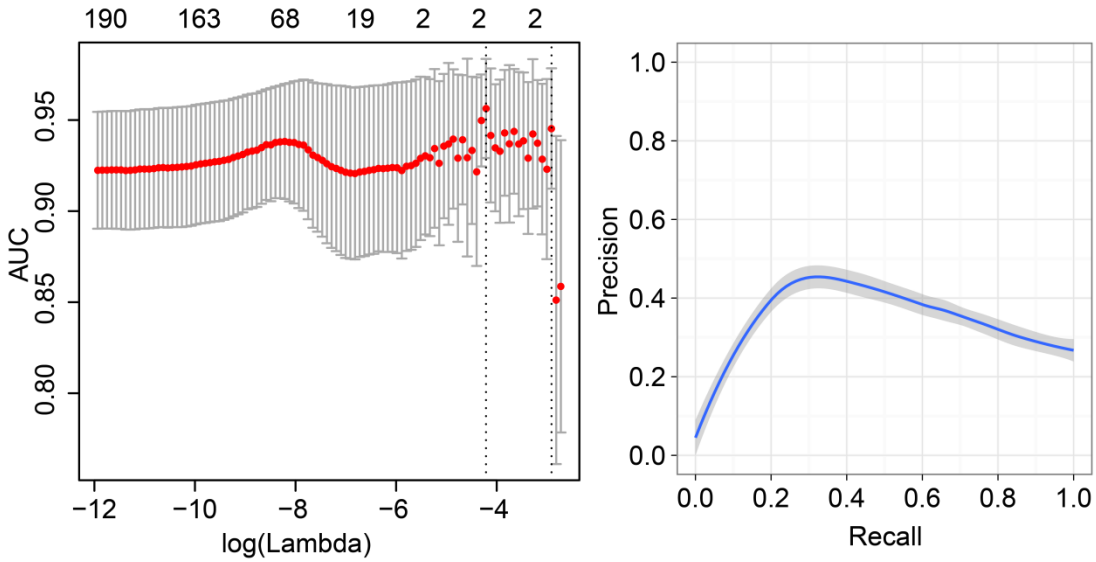

Figure 4-7 – Second Pass K. lactis Nucleosome and ORC data

A LASSO regularization curve (left) and precision recall curve (right) illustrate the predictive power of a second pass *K. lactis* model using nucleosome and ORC occupancy data. The baseline precision for random prediction in this curve is 0.016.

Discussion

All models described in the results showed predictive power well above random as evidenced by the AUCs greater than 0.9. Furthermore, the analysis suggests that a similar nucleosome pattern exists at origins in both *S. cerevisiae* and *K. lactis,* despite the very different size of the consensus sequence in the two organisms. Both models trained on similar features: a deep central trough with arrayed nucleosomal peaks on both sides of the trough. There were, however, some apparent differences. The primary difference was that the arrayed nucleosomes were less defined in the *S. cerevisiae* case. This was visually apparent in the aggregate nucleosome pattern (Figure 4-1). Fewer coefficients were assigned to arrayed nucleosome peaks and a greater proportion of the weight was assigned to the trough in the *S. cerevisiae* model as compared to the *K. lactis* model (Table 9-1, Table 9-5).

The use of Fourier and wavelet transforms had little to no effect on the total predictive power. However, when the same transforms were used on older, lower resolution nucleosome occupancy data they significantly improved predictive power (data not shown), that improvement made them roughly equivalent to the raw data models shown in this study. The improvement in prediction performance is attributable to the smoothing effect of both transforms, which removed artifacts from the nucleosome data. This result suggests that if lower quality data is being used, transforms may provide a significant improvement in predictive power.

Cross Species Prediction

Despite differences in retained features and predictive power between the *K. lactis* and *S. cerevisiae*-trained models, both models showed strong cross species prediction. Unexpectedly, the predictive power of the *K. lactis* model on *S.cerevisiae* data (AUC 0.923) was greater than the predictive power of the *S. cerevisiae* model on itself (AUC 0.905). In contrast, when the *S.*

cerevisiae model was used to predict *K. lactis* sites (AUC 0.905), it was more successful, following the same trend as when the *K. lactis* model was used to predict other *K. lactis* sites (AUC 0.935). These results suggest that either the *S. cerevisiae* nucleosome data or the list of known *S. cerevisiae* origins is inferior to those from *K. lactis*. Given that both nucleosome sets were generated in the same study, it is unlikely that the problem is in the nucleosome preparation; I propose that the difference lies in the origin lists for *S. cerevisiae* vs. *K. lactis*. A potential difference in the origin lists is the frequency with which origins fire: some weakly firing origins may have ORC bound in a smaller subset of the population leading to weaker nucleosome patterning. Extensive investigation of origin function in *S. cerevisiae* has identified not only sites that fire frequently but also sites that fire much less frequently. As such, the *S. cerevisiae* origin list may be suboptimal for training as it contains origins with activity below a useful level.

Final Predictive Strength

Because the *K. lactis* model showed prediction superior to that of the *S. cerevisiae* model, only the *K. lactis* model was used for the second pass. A second pass using the *K. lactis* nucleosome data and only the best origin candidates from the first pass had relatively little effect on origin prediction (AUC 0.60). This suggests that logistic regression has reached its limit for this application, though the use of purely discriminative models such as a Support Vector Machine (SVM) may be able to extract additional information.

The addition of non-nucleosome data, namely ORC complex occupancy, provided a considerable jump to predictive power bumping the AUC to 0.93. However, more importantly for full genome prediction, there was a commensurate increase in the precision to 40% for 60% recall from a baseline of 1.6% for random selection.

In the *K. lactis* genome there were 148 origins over a 10,729,447 bp genome. This amounts to 1 in 72,496 sites being origins when the whole genome is scanned, a highly imbalanced search. The

first pass improved this to 1 in 346 sites. The second pass, with ORC data, improved this to 1 in 10 with 36% total recall. The final set of predicted origins had a 7,249 fold enrichment over random selection.

The use of ORC binding sites to predict origin function for *K. lactis* is permissible as the ORC data was not used in the original identification of the origins. Additional work is required to identify other features capable of improving the predictive power of the model. Knowledge of the *S. cerevisiae* and prokaryotic origin sites suggests that DNA sequence plays an important role in origin activity. It is unclear whether DNA sequence motifs will be conserved between organisms or even amongst all origins in a given organism, but the work presented here prepares a foundation on which to ask that question.

5. Conclusion

The three chapters contained in this thesis describe applications of a general computational gramework for exploring chromatin and chromatin modification. The approach is based on the application of logistic regression to model differences between sites with and without a particular modification. While other types of models may provide greater predictive power or identify more nuanced relationships between chromatin modifications and features, the coefficients trained by the logistic regression model are easily interpretable and their interpretation aligns with our an abstract view of biological mechanisms. In this view the presence or absence of a protein or complex produces a binary or proportional response on regulated targets and that more complex behaviors are caused by interactions between multiple elements. LASSO, the regularization technique applied in all three chapters, helped to identify a succinct subset of features that best described the observed distribution of modification.

Both the Ty5 and Ty1 retrotransposon studies showed the elegance of this approach. Using a variety of *S. cerevisiae* chromosomal features including nucleosome, functional annotations and protein binding, I established a model for the observed distribution of each transposon. This was done primarily through evaluation of individual features independently, as multiple single dimensional models. Multidimensional models were used as well to examine whether features had redundant or synergistic effects.

For Ty5 I found that the canonical definition of a Ty5 target site explained only a fraction of the total distribution. Associations with chromosomal features revealed a strong preference for regions of the chromosome depleted of bound proteins within the heterochromatin of the telomere and subtelomere (Figure 2-2). This suggested a two-step biological mechanism in which the Sir4 telomeric protein attracted Ty5 to the subtelomeres and Ty5 then integrated into accessible sites (Figure 2-5). The two-step mechanism was further supported by continued

integration at sites of accessible DNA in the absence of Sir4 and at euchromatin sites shown to be devoid of Sir4 (Figure 2-4).

For Ty1, I found a novel pattern of integration at nucleosomes upstream of Pol III transcribed genes. Mapping the distribution of integration sites relative to the pattern of nucleosomes revealed two Ty1 integration peaks per nucleosome. Furthermore, mapping the Ty1 integration sites to the nucleosomes revealed that the two integration peaks aligned at a single face of the nucleosome (Figure 3-4). While the mechanism targeting Ty1 integration has yet to be ascertained, the pattern of integration provides some hints as to the mechanism by which integration sites are selected. While the nucleosomes themselves are rotationally symmetric about the dyad axis, the pattern of Ty1 integration on the nucleosome was rotationally asymmetric. Consistently, integration was skewed towards the Pol III binding site, suggesting that a recruiting factor was bound at the Pol III–transcribed gene, consistent with integration only occurring in the vicinity of Pol III–bound genes.

The logistic regression model was successful in identifying new targeting determinants for both Ty5 and Ty1 integration. However, the model was unable to determine why Ty1 prefers certain Pol III genes. Furthermore, the model only weakly identified the newly discovered nucleosome integration pattern. It is likely that the preference of Ty1 for particular Pol III genes is driven by some chromosomal feature not currently included in the model. For example, features such as chromatin remodeler activity or Pol III gene transcription rates may affect integration site preference. Both of these features have been difficult to evaluate. Some chromatin remodelers, such as Isw2 (Gelbart, Bachman et al. 2005), physically interact with a set of sites independent of their sites of remodeling, making ChIP analysis of the proteins uninformative. Similarly, due to sets of Pol III genes producing identical transcripts genome-wide expression studies fail to identify the precise source of some Pol III transcripts. As the *S. cerevisiae* feature

set is expanded and technical limitations are addressed, reanalysis of the Ty1 distribution may detect features guiding Pol III gene targeting.

Weak detection of Ty1's integration at nucleosomes was caused by integration on the sides of the nucleosome peak, effectively associating integration with non-extreme nucleosome occupancy values. Logistic regression models are unable to directly model such associations. This aspect of the model could be added using techniques developed for the origin of replication analysis: by including Fourier or wavelet transformed nucleosome data, the logistic regression model could be made to recognize local nucleosome patterning rather than simply the average nucleosome density in the vicinity of the integration site. Furthermore, the addition of transformed versions of existing features may improve the model. For instance, a feature representing the absolute difference of nucleosome occupancy from the mean occupancy would be expected to be highly predictive of Ty1 integration. Thus, there are some clear, feasible approaches that can be used to improve the predictive power of the model. The challenge is to devise a minimal set of transformations that have biological relevance and can be applied generally to a subset of features. A minimal set is important to avoid unnecessarily increasing the initial dimensionality of the logistic classifier. If the transforms must be customized to each feature they will have little use; the transforms will only capture what is known *a priori*.

In all, logistic regression was successful in modeling Ty1 and Ty5 targeting in *S. cerevisiae*. The method and features used are applicable to other retrotransposons and other chromosomal modifications occurring at single basepairs in the genome. Furthermore the methodology is generalizable to other model organisms such as *A. thaliana*, where a comparable variety of chromosomal feature sets are available. One such application in *A. thaliana* is the prediction of sites targetable using zinc-finger and TAL effector arrays. Both zinc-fingers and TAL effectors

target specific DNA sequences but their binding strength is thought to be affected by chromatin around the target site.

The origin of replication project applied the established model in a new way. Rather than ascribing potential mechanisms to a known distribution, a logistic classifier was used to predict the pattern of modification. However, even in a prediction based approach, the model was able to identify particular features indicative of a modification site. The application of the model showed that nucleosome patterning at origins is conserved between the yeasts *K. lactis* and *S. cerevisiae* and that nucleosome data alone can significantly enrich for sites with origins of replication. When predicting origin sites, the model focused on two components of the nucleosome pattern, a large central trough in nucleosome density and nucleosomes arrayed with a 170bp periodicity. The origin predictions exhibit lower precision than would be preferred but the future addition of more features to the model, particularly sequence-based features, may improve precision further. Regardless, the origin project has shown that the logistic regression methodology, as originally applied to Ty retrotransposons, can be extended to a prediction context. Additionally, the origin project differs from the Ty projects in that the origin project focuses on events spanning larger segments of the genome than a single basepair.

The origin problem required, and will continue to require, the development of techniques for identifying and characterizing localized patterns in chromosomal features. While the results presented in chapter three focused largely on the raw data, the joint dataset containing the Fourier, wavelet and raw data trained on only two features, highlighting the role of the trough and arrayed nucleosomes in origin prediction. The effectiveness of the wavelet transform is possibly hampered by the use of a fixed 1024bp window. The 1024bp window size was chosen for fast computation both at the transformation step and the subsequent model fitting. A larger window size would take advantage of the multiscale capabilities of the wavelet transform but lead to

greater computation time. Some minor changes to the processing pipeline could alleviate some of this computational burden. The use of a non-decimated wavelet transform, due to its translational invariance, would reduce redundant calculations incurred when scanning the chromosome, thereby reducing transformation time. Additionally, screening of the transformed feature set would allow for the removal of features with minimal variance, reducing the number of features provided to the logistic classifier. The techniques developed for the origin problem should also be applicable to other projects where a multi-basepair locus is being identified. For instance prediction of neo-centromere locations (Ketel, Wang et al. 2009), new centromeres formed when canonical centromeres are disrupted or fragile sites (Durkin and Glover 2007), regions of the genome particularly prone to breakage or translocation.

Together, these applications support the concept of a single, generalized model capable of predicting testable targeting mechanisms from the vast quantities of chromosomal data available. Analysis of new distributions in *S. cerevisiae* is simply a matter of loading the distribution of interest and applying the existing feature set. Since important basic properties of the chromosomes, such as DNA remaining linear or circular, are conserved between the domains of life, the computational framework described herein should be applicable to Archaea and Bacteria in addition to Eukaryotes. Application of the methodology in other organisms is only hampered by the time necessary to collect the available chromosomal feature sets into a single, accessible repository. However, due to major differences in chromatin structure between eukaryotes and prokaryotes this thesis primarily focused on eukaryotic applications.

As our understanding of the landscape of chromatin in model organisms such as *S. cerevisiae* and *A. thaliana* becomes more complete, new analyses will be required. Rather than simply investigating a single chromosomal feature with respect to the others, we will be able to take a more holistic approach; looking for clusters of similar patterns within the chromatin and

sequence. From a computational viewpoint, this manifests as a switch from supervised methods like logistic regression to unsupervised methods such as clustering and associative rule mining. Unsupervised approaches are currently being applied to particular subsets of the data such as histone modifications, gene expression and nucleosome occupancy. However, little work has been done to integrate all of the available data into a single model.

The unsupervised learning methods provide a different readout than the supervised methods described previously. Rather than trying to find a way of describing a specified distribution in terms of other available information, an unsupervised method would group genomic locations with similar patterns. Some of the patterns will be recognizable. For example, there exist distinct nucleosome, acetylation and methylation patterns known to be associated with highly transcribed genes. The key to this approach is that it would provide a systemic look at the patterns in the cell rather than focusing on known pattern. Once the obvious patterns are annotated, it will remain to identify smaller subpatterns. The subpatterns could identify points of interest in the non-coding DNA or could elucidate mechanistic differences between genes with different expression levels.

From a biological perspective a complete picture of the chromatin landscape for particular model organisms is startlingly near. There are still particular difficulties to work out, like how to find the sites of action of proteins with only tenuous interactions with DNA and nucleosomes. Counter to these difficulties, sequencing technology continues to drop in price, allowing for more experimentation and more data available for analysis. The real challenge remains in the computational arena: gathering the data, combining the data, and designing a system capable of considering each piece of data in its proper context.

6. Bibliography

Avery, O. T., C. M. Macleod, et al. (1944). "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii." The Journal of experimental medicine **79**(2): 137-158.

Bachman, N., Y. Eby, et al. (2004). "Local definition of Ty1 target preference by long terminal repeats and clustered tRNA genes." Genome Res **14**(7): 1232-1247.

Bailey, T. L., M. Boden, et al. (2009). "MEME SUITE: tools for motif discovery and searching." Nucleic acids research **37**(Web Server issue): W202-208.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology **2**: 28-36.

Bailey, T. L. and M. Gribskov (1998). "Combining evidence using p-values: application to sequence homology searches." Bioinformatics **14**(1): 48-54.

Baller, J. A., J. Gao, et al. (2012). "A nucleosomal surface defines an integration hotspot for the Saccharomyces cerevisiae Ty1 retrotransposon." Genome research.

Baller, J. A., J. Gao, et al. (2011). "Access to DNA establishes a secondary target site bias for the yeast retrotransposon Ty5." Proceedings of the National Academy of Sciences of the United States of America **108**(51): 20351-20356.

Barrett, T., D. B. Troup, et al. (2009). "NCBI GEO: archive for high-throughput functional genomic data." Nucleic acids research **37**(Database issue): D885-890.

Barry, E. R. and S. D. Bell (2006). "DNA replication in the archaea." Microbiology and molecular biology reviews : MMBR **70**(4): 876-887.

Beauregard, A., M. J. Curcio, et al. (2008). "The take and give between retrotransposable elements and their hosts." Annual review of genetics **42**: 587-617.

Bellen, H. J., R. W. Levis, et al. (2004). "The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes." Genetics **167**(2): 761-781.

Bensic, M., N. Sarlija, et al. (2005). "Modelling small-business credit scoring by using logistic regression, neural networks and decision trees." Intelligent Systems in Accounting, Finance and Management **13**(3): 133-150.

Bentley, D. R., S. Balasubramanian, et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-59.

Berry, C., S. Hannenhalli, et al. (2006). "Selection of target sites for mobile DNA integration in the human genome." PLoS computational biology **2**(11): e157.

Berry, C., S. Hannenhalli, et al. (2006). "Selection of target sites for mobile DNA integration in the human genome." PLoS Comput Biol **2**(11): e157.

Blackburn, E. H. and J. G. Gall (1978). "A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena." Journal of molecular biology **120**(1): 33-53.

Blackburn, E. H. and J. W. Szostak (1984). "The molecular structure of centromeres and telomeres." Annual review of biochemistry **53**: 163-194.

Bloom, K. and A. Joglekar (2010). "Towards building a chromosome segregation machine." Nature **463**(7280): 446-456.

Boeke, J. D. and S. E. Devine (1998). "Yeast retrotransposons: finding a nice quiet neighborhood." Cell **93**(7): 1087-1089.
Bouchard, G. and B. Triggs (2004). THE TRADE-OFF BETWEEN GENERATIVE AND DISCRIMINATIVE CLASSIFIERS. COMPSTAT. Prague.
Bradley, A. P. (1996). "The Use of the Area Under The ROC Curve in the Evaluation of Machine Learning Algorithms." Pattern Recognition **30**(7): 1145-1159.
Brewer, B. J. and W. L. Fangman (1987). "The localization of replication origins on ARS plasmids in S. cerevisiae." Cell **51**(3): 463-471.
Bushman, F. D. (2003). "Targeting survival: integration site selection by retroviruses and LTR-retrotransposons." Cell **115**(2): 135-138.
Campos, E. I. and D. Reinberg (2009). "Histones: annotating chromatin." Annual review of genetics **43**: 559-599.
Cedar, H. and Y. Bergman (2009). "Linking DNA methylation and histone modification: patterns and paradigms." Nature reviews. Genetics **10**(5): 295-304.
Chalker, D. L. and S. B. Sandmeyer (1992). "Ty3 integrates within the region of RNA polymerase III transcription initiation." Genes Dev **6**(1): 117-128.
Chang, B., Y. Chen, et al. (2007). "JMJD6 is a histone arginine demethylase." Science **318**(5849): 444-447.
Chen, Z., C. Speck, et al. (2008). "The architecture of the DNA replication origin recognition complex in Saccharomyces cerevisiae." Proceedings of the National Academy of Sciences of the United States of America **105**(30): 10326-10331.
Cherepanov, P., G. Maertens, et al. (2003). "HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells." J Biol Chem **278**(1): 372-381.
Cherry, J. M., C. Adler, et al. (1998). "SGD: Saccharomyces Genome Database." Nucleic Acids Res **26**(1): 73-79.
Ciuffi, A. and F. D. Bushman (2006). "Retroviral DNA integration: HIV and the role of LEDGF/p75." Trends Genet **22**(7): 388-395.
Ciuffi, A., M. Llano, et al. (2005). "A role for LEDGF/p75 in targeting HIV DNA integration." Nat Med **11**(12): 1287-1289.
Ciuffi, A., R. S. Mitchell, et al. (2006). "Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts." Mol Ther **13**(2): 366-373.
Clapier, C. R. and B. R. Cairns (2009). "The biology of chromatin remodeling complexes." Annual review of biochemistry **78**: 273-304.
Cohn, M., M. J. McEachern, et al. (1998). "Telomeric sequence diversity within the genus Saccharomyces." Current genetics **33**(2): 83-91.
Cortes, C. and V. Vapnik (1995). "Support-Vector Networks." Machine Learning, **20**(3): 273-297.
Crawford, N., A. Chajara, et al. (1995). "Targeting platelets containing electro-encapsulated iloprost to balloon injured aorta in rats." Thrombosis and haemostasis **73**(3): 535-542.
Crick, F. (1970). "Central dogma of molecular biology." Nature **227**(5258): 561-563.
Curcio, M. J. and D. J. Garfinkel (1991). "Single-step selection for Ty1 element retrotransposition." Proc Natl Acad Sci U S A **88**(3): 936-940.
Czajkowsky, D. M., J. Liu, et al. (2008). "DNA combing reveals intrinsic temporal disorder in the replication of yeast chromosome VI." Journal of molecular biology **375**(1): 12-19.
Dakshinamurthy, A., K. M. Nyswaner, et al. (2010). "BUD22 affects Ty1 retrotransposition and ribosome biogenesis in Saccharomyces cerevisiae." Genetics **185**(4): 1193-1205.

Daubechies, I. (1990). "The Wavelet Transform, Time-Frequency Localization and Signal Analysis." IEEE Transactions on Information Theory **36**(5).
Dhalluin, C., J. E. Carlson, et al. (1999). "Structure and ligand of a histone acetyltransferase bromodomain." Nature **399**(6735): 491-496.
Dietterich, T. G. (2000). "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization." Machine Learning, **40**(2): 139-157.
Donaldson, A. D. (2005). "Shaping time: chromatin structure and the DNA replication programme." Trends in genetics : TIG **21**(8): 444-449.
Dorn, J. F. and P. S. Maddox (2011). "Kinetochore dynamics: how protein dynamics affect chromosome segregation." Current opinion in cell biology.
Dreiseitl, S. and L. Ohno-Machado (2002). "Logistic regression and artificial neural network classification models: a methodology review." J Biomed Inform **35**(5-6): 352-359.
Du, L. L., T. M. Nakamura, et al. (2006). "Histone modification-dependent and -independent pathways for recruitment of checkpoint protein Crb2 to double-strand breaks." Genes & development **20**(12): 1583-1596.
Duan, Z., M. Andronescu, et al. (2010). "A three-dimensional model of the yeast genome." Nature **465**(7296): 363-367.
Duncker, B. P., I. N. Chesnokov, et al. (2009). "The origin recognition complex protein family." Genome biology **10**(3): 214.
Durkin, S. G. and T. W. Glover (2007). "Chromosome fragile sites." Annual review of genetics **41**: 169-192.
Eaton, M. L., K. Galani, et al. (2010). "Conserved nucleosome positioning defines replication origins." Genes & development **24**(8): 748-753.
Friedel, M., S. Nikolajewa, et al. (2009). "DiProDB: a database for dinucleotide properties." Nucleic acids research **37**(Database issue): D37-40.
Friedman, J. (2010). "Regularization paths for generalized linear models via coordinate descent." Journal of Statistical Software **33**: 1-22.
Friedman, J., T. Hastie, et al. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of statistical software **33**(1): 1-22.
Frith, M. C., N. F. Saunders, et al. (2008). "Discovering sequence motifs with arbitrary insertions and deletions." PLoS computational biology **4**(4): e1000071.
Fuda, N. J., M. B. Ardehali, et al. (2009). "Defining mechanisms that regulate RNA polymerase II transcription in vivo." Nature **461**(7261): 186-192.
Gai, X. and D. F. Voytas (1998). "A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin." Mol Cell **1**(7): 1051-1055.
Game, J. C. and S. B. Chernikova (2009). "The role of RAD6 in recombinational repair, checkpoints and meiosis via histone modification." DNA repair **8**(4): 470-482.
Gangadharan, S., L. Mularoni, et al. "Inaugural Article: DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo." Proc Natl Acad Sci U S A **107**(51): 21966-21972.
Gardner, K. E., C. D. Allis, et al. (2011). "Operating on chromatin, a colorful language where context matters." Journal of molecular biology **409**(1): 36-46.
Gelbart, M. E., N. Bachman, et al. (2005). "Genome-wide identification of Isw2 chromatin-remodeling targets by localization of a catalytically inactive mutant." Genes Dev **19**(8): 942-954.

Gerbi, S. A. and A. K. Bielinsky (1997). "Replication initiation point mapping." Methods **13**(3): 271-280.
Gilbert, D. M. (2002). "Replication timing and transcriptional control: beyond cause and effect." Current opinion in cell biology **14**(3): 377-383.
Goodier, J. L. and H. H. Kazazian, Jr. (2008). "Retrotransposons revisited: the restraint and rehabilitation of parasites." Cell **135**(1): 23-35.
Guo, Y. and H. L. Levin "High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in Schizosaccharomyces pombe." Genome Res **20**(2): 239-248.
Harismendy, O., C. G. Gendrel, et al. (2003). "Genome-wide location of yeast RNA polymerase III transcription machinery." The EMBO journal **22**(18): 4738-4747.
Hecht, A., T. Laroche, et al. (1995). "Histone H3 and H4 N-termini interact with SIR3 and SIR4 proteins: a molecular model for the formation of heterochromatin in yeast." Cell **80**(4): 583-592.
Hesselberth, J. R., X. Chen, et al. (2009). "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting." Nat Methods **6**(4): 283-289.
Hoerl, A. E. and R. W. Kennard (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics **12**(1).
Huet, J., R. Schnabel, et al. (1983). "Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type." The EMBO journal **2**(8): 1291-1294.
Ioshikhes, I., S. Hosid, et al. (2011). "Variety of genomic DNA patterns for nucleosome positioning." Genome research **21**(11): 1863-1871.
Ji, H., D. P. Moore, et al. (1993). "Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences." Cell **73**(5): 1007-1018.
Johnson, S. C. (1967). "Hierarchical clustering schemes." Psychometrika **32**(3): 241-254.
Jolliffe, I. T. (2002). Principal component analysis. New York, Springer.
Jung, D. and F. W. Alt (2004). "Unraveling V(D)J recombination; insights into gene regulation." Cell **116**(2): 299-311.
Kadonaga, J. T. (2004). "Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors." Cell **116**(2): 247-257.
Kalliomaa-Sanford, A. K., F. A. Rodriguez-Castaneda, et al. (2012). "Chromosome segregation in Archaea mediated by a hybrid DNA partition machine." Proceedings of the National Academy of Sciences of the United States of America.
Kaplan, N., I. K. Moore, et al. (2009). "The DNA-encoded nucleosome organization of a eukaryotic genome." Nature **458**(7236): 362-366.
Kastenmayer, J. P., L. Ni, et al. (2006). "Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae." Genome research **16**(3): 365-373.
Kates, M., D. Kushner, et al. (1993). The Biochemistry of archaea (archaebacteria). Amsterdam ; New York, Elsevier.
Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-664.
Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome research **12**(6): 996-1006.
Ketel, C., H. S. Wang, et al. (2009). "Neocentromeres form efficiently at multiple possible loci in Candida albicans." PLoS genetics **5**(3): e1000400.
Koster, D. A., A. Crut, et al. (2010). "Cellular strategies for regulating DNA supercoiling: a single-molecule perspective." Cell **142**(4): 519-530.

Kouzarides, T. (2007). "Chromatin modifications and their function." Cell **128**(4): 693-705.
Kurdistani, S. K., S. Tavazoie, et al. (2004). "Mapping global histone acetylation patterns to gene expression." Cell **117**(6): 721-733.
Kyrpides, N. C. and C. A. Ouzounis (1999). "Transcription in archaea." Proceedings of the National Academy of Sciences of the United States of America **96**(15): 8545-8550.
Lander, E. S. (2011). "Initial impact of the sequencing of the human genome." Nature **470**(7333): 187-197.
Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.
Le Cessie, S. and J. C. Van Houwelingen (1992). "Ridge Estimators in Logisitic Regression." Applied Statistics **41**(1): 191-201.
Lee, W., D. Tillo, et al. (2007). "A high-resolution atlas of nucleosome occupancy in yeast." Nat Genet **39**(10): 1235-1244.
Lee, W., D. Tillo, et al. (2007). "A high-resolution atlas of nucleosome occupancy in yeast." Nature genetics **39**(10): 1235-1244.
Leem, Y. E., T. L. Ripmaster, et al. (2008). "Retrotransposon Tf1 is targeted to Pol II promoters by transcription activators." Mol Cell **30**(1): 98-107.
Leinonen, R., R. Akhtar, et al. (2012). "European Nucleotide Archive." NAR Molecular Biology **40**(D1).
Lengronne, A., P. Pasero, et al. (2001). "Monitoring S phase progression globally and locally using BrdU incorporation in TK(+) yeast strains." Nucleic acids research **29**(7): 1433-1442.
Lesage, P. and A. L. Todeschini (2005). "Happy together: the life and times of Ty retrotransposons and their hosts." Cytogenet Genome Res **110**(1-4): 70-90.
Li, B., S. G. Pattenden, et al. (2005). "Preferential occupancy of histone variant H2AZ at inactive promoters influences local histone modifications and chromatin remodeling." Proc Natl Acad Sci U S A **102**(51): 18385-18390.
Liachko, I., A. Bhaskar, et al. (2010). "A comprehensive genome-wide map of autonomously replicating sequences in a naive genome." PLoS genetics **6**(5): e1000946.
Lieberman-Aiden, E., N. L. van Berkum, et al. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." Science **326**(5950): 289-293.
Liebman, S. W. and G. Newnam (1993). "A ubiquitin-conjugating enzyme, RAD6, affects the distribution of Ty1 retrotransposon integration positions." Genetics **133**(3): 499-508.
Lipford, J. R. and S. P. Bell (2001). "Nucleosomes positioned by ORC facilitate the initiation of DNA replication." Molecular Cell **7**(1): 21-30.
Liu, C. L., T. Kaplan, et al. (2005). "Single-nucleosome mapping of histone modifications in S. cerevisiae." PLoS biology **3**(10): e328.
Liu, L. F. and J. C. Wang (1987). "Supercoiling of the DNA template during transcription." Proceedings of the National Academy of Sciences of the United States of America **84**(20): 7024-7027.
Liu, S., C. T. Yeh, et al. (2009). "Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome." PLoS Genet **5**(11): e1000733.
Livny, J., Y. Yamaichi, et al. (2007). "Distribution of centromere-like parS sites in bacteria: insights from comparative genomics." Journal of bacteriology **189**(23): 8693-8703.

Loh, W.-Y. (2011). "Classification and Regression Trees." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1**(1): 14-23.
Long, W. J., J. L. Griffith, et al. (1993). "A comparison of logistic regression to decision-tree induction in a medical domain." Computers and biomedical research, an international journal **26**(1): 74-97.
Louis, E. J. (2002). "Are Drosophila telomeres an exception or the rule?" Genome biology **3**(10): REVIEWS0007.
Louis, E. J. and J. E. Haber (1992). "The structure and evolution of subtelomeric Y' repeats in Saccharomyces cerevisiae." Genetics **131**(3): 559-574.
Majumdar, A., A. G. Chatterjee, et al. (2010). "Determinants that specify the integration pattern of retrotransposon Tf1 in the fbp1 promoter of Schizosaccharomyces pombe." J Virol **85**(1): 519-529.
Mardis, E. R. (2008). "Next-generation DNA sequencing methods." Annual review of genomics and human genetics **9**: 387-402.
Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.
Marsolier-Kergoat, M. C. and A. Goldar (2012). "DNA replication induces compositional biases in yeast." Molecular biology and evolution **29**(3): 893-904.
Masai, H., S. Matsumoto, et al. (2010). "Eukaryotic chromosome DNA replication: where, when, and how?" Annual review of biochemistry **79**: 89-130.
Mavrich, T. N., I. P. Ioshikhes, et al. (2008). "A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome." Genome research **18**(7): 1073-1083.
Maxam, A. M. and W. Gilbert (1992). "A new method for sequencing DNA. 1977." Biotechnology **24**: 99-103.
Maxwell, P. H., C. Coombes, et al. (2004). "Ty1 mobilizes subtelomeric Y' elements in telomerase-negative Saccharomyces cerevisiae survivors." Molecular and Cellular Biology **24**(22): 9887-9898.
McLean, M. J., K. H. Wolfe, et al. (1998). "Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes." Journal of molecular evolution **47**(6): 691-696.
Metzger, E. and R. Schule (2007). "The expanding world of histone lysine demethylases." Nature structural & molecular biology **14**(4): 252-254.
Meyne, J., R. L. Ratliff, et al. (1989). "Conservation of the human telomere sequence (TTAGGG)n among vertebrates." Proceedings of the National Academy of Sciences of the United States of America **86**(18): 7049-7053.
Misteli, T. (2007). "Beyond the sequence: cellular organization of genome function." Cell **128**(4): 787-800.
Mitchell, R. S., B. F. Beitzel, et al. (2004). "Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences." PLoS Biol **2**(8): E234.
Monnig, C. A. and R. T. Kennedy (1994). "Capillary electrophoresis." Analytical chemistry **66**(12): 280R-314R.
Moqtaderi, Z. and K. Struhl (2004). "Genome-wide occupancy profile of the RNA polymerase III machinery in Saccharomyces cerevisiae reveals loci with incomplete transcription complexes." Mol Cell Biol **24**(10): 4118-4127.

Mott, M. L. and J. M. Berger (2007). "DNA replication initiation: mechanisms and regulation in bacteria." Nature reviews. Microbiology **5**(5): 343-354.
Mou, Z., A. E. Kenny, et al. (2006). "Hos2 and Set3 promote integration of Ty1 retrotransposons at tRNA genes in Saccharomyces cerevisiae." Genetics **172**(4): 2157-2167.
Mularoni, L., Y. Zhou, et al. (2011). "Ty1 integration targets specific nucleosomal DNA." Genome Res **In press**.
Nawotka, K. A. and J. A. Huberman (1988). "Two-dimensional gel electrophoretic method for mapping DNA replicons." Molecular and Cellular Biology **8**(4): 1408-1413.
Nyswaner, K. M., M. A. Checkley, et al. (2008). "Chromatin-associated genes protect the yeast genome from Ty1 insertional mutagenesis." Genetics **178**(1): 197-214.
Omberg, L., J. R. Meyerson, et al. (2009). "Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression." Molecular systems biology **5**: 312.
Orlando, V. (2000). "Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation." Trends in biochemical sciences **25**(3): 99-104.
Papamichos-Chronakis, M., S. Watanabe, et al. (2011). "Global regulation of H2A.Z localization by the INO80 chromatin-remodeling enzyme is essential for genome integrity." Cell **144**(2): 200-213.
Pereira, S. L., R. A. Grayling, et al. (1997). "Archaeal nucleosomes." Proceedings of the National Academy of Sciences of the United States of America **94**(23): 12633-12637.
Pinkel, D., R. Segraves, et al. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays." Nature genetics **20**(2): 207-211.
Pokholok, D. K., C. T. Harbison, et al. (2005). "Genome-wide map of nucleosome acetylation and methylation in yeast." Cell **122**(4): 517-527.
project, S. "Saccharomyces Genome Database." from http://www.yeastgenome.org/.
Pruss, D., F. D. Bushman, et al. (1994). "Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core." Proc Natl Acad Sci U S A **91**(13): 5913-5917.
Pryciak, P. M., H. P. Muller, et al. (1992). "Simian virus 40 minichromosomes as targets for retroviral integration in vivo." Proc Natl Acad Sci U S A **89**(19): 9237-9241.
Pryciak, P. M., A. Sil, et al. (1992). "Retroviral integration into minichromosomes in vitro." Embo J **11**(1): 291-303.
Raisner, R. M., P. D. Hartley, et al. (2005). "Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin." Cell **123**(2): 233-248.
Rehman, M. A. and K. Yankulov (2009). "The dual role of autonomously replicating sequences as origins of replication and as silencers." Curr Genet **55**(4): 357-363.
Richmond, T. J. and C. A. Davey (2003). "The structure of DNA in the nucleosome core." Nature **423**(6936): 145-150.
Rinckel, L. A. and D. J. Garfinkel (1996). "Influences of histone stoichiometry on the target site preference of retrotransposons Ty1 and Ty2 in Saccharomyces cerevisiae." Genetics **142**(3): 761-776.
Roberts, D. N., A. J. Stewart, et al. (2003). "The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships." Proc Natl Acad Sci U S A **100**(25): 14695-14700.
Robzyk, K., J. Recht, et al. (2000). "Rad6-dependent ubiquitination of histone H2B in yeast." Science **287**(5452): 501-504.

Roth, S. L., N. Malani, et al. (2011). "Gammaretroviral integration into nucleosomal target DNA in vivo." Journal of virology **85**(14): 7393-7401.
Rudkin, G. T. and B. D. Stollar (1977). "High resolution detection of DNA-RNA hybrids in situ by indirect immunofluorescence." Nature **265**(5593): 472-473.
Sadeh, R. and C. D. Allis (2011). "Genome-wide "re"-modeling of nucleosome positions." Cell **147**(2): 263-266.
Salih, F., B. Salih, et al. (2007). "Sequence-directed mapping of nucleosome positions." Journal of biomolecular structure & dynamics **24**(5): 489-493.
Sandman, K., S. L. Pereira, et al. (1998). "Diversity of prokaryotic chromosomal proteins and the origin of the nucleosome." Cellular and molecular life sciences : CMLS **54**(12): 1350-1364.
Sandmeyer, S. (2003). "Integration by design." Proc Natl Acad Sci U S A **100**(10): 5586-5588.
Sanger, F., S. Nicklen, et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proceedings of the National Academy of Sciences of the United States of America **74**(12): 5463-5467.
SantaLucia, J., Jr. and D. Hicks (2004). "The thermodynamics of DNA structural motifs." Annual review of biophysics and biomolecular structure **33**: 415-440.
Satchwell, S. C., H. R. Drew, et al. (1986). "Sequence periodicities in chicken nucleosome core DNA." Journal of molecular biology **191**(4): 659-675.
Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.
Scholes, D. T., M. Banerjee, et al. (2001). "Multiple regulators of Ty1 transposition in Saccharomyces cerevisiae have conserved roles in genome maintenance." Genetics **159**(4): 1449-1465.
Shalon, D., S. J. Smith, et al. (1996). "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization." Genome research **6**(7): 639-645.
Shampay, J., J. W. Szostak, et al. (1984). "DNA sequences of telomeres maintained in yeast." Nature **310**(5973): 154-157.
Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nature biotechnology **26**(10): 1135-1145.
Shun, M. C., Y. Botbol, et al. (2008). "Identification and characterization of PWWP domain residues critical for LEDGF/p75 chromatin binding and human immunodeficiency virus type 1 infectivity." Journal of virology **82**(23): 11555-11567.
Simms, T. A., S. L. Dugas, et al. (2008). "TFIIIC binding sites function as both heterochromatin barriers and chromatin insulators in Saccharomyces cerevisiae." Eukaryot Cell **7**(12): 2078-2086.
Simpson, R. T. (1990). "Nucleosome positioning can affect the function of a cis-acting DNA element in vivo." Nature **343**(6256): 387-389.
Slotkin, R. K. and R. Martienssen (2007). "Transposable elements and the epigenetic regulation of the genome." Nature reviews. Genetics **8**(4): 272-285.
Smale, S. T. and J. T. Kadonaga (2003). "The RNA polymerase II core promoter." Annual review of biochemistry **72**: 449-479.
Smith, L. M., J. Z. Sanders, et al. (1986). "Fluorescence detection in automated DNA sequence analysis." Nature **321**(6071): 674-679.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." Journal of molecular biology **147**(1): 195-197.
Soragni, E. and G. A. Kassavetis (2008). "Absolute gene occupancies by RNA polymerase III, TFIIIB, and TFIIIC in Saccharomyces cerevisiae." J Biol Chem **283**(39): 26568-26576.
Sorensen, H., D. Jones, et al. (1987). "Real-valued Fast Fourier Transform Algorithms." IEEE Transactions on Acoustics, Speech and Signal Processing **35**(6): 849-863.
Stamenova, R., P. H. Maxwell, et al. (2009). "Rrm3 protects the Saccharomyces cerevisiae genome from instability at nascent sites of retrotransposition." Genetics **182**(3): 711-723.
Strahl, B. D. and C. D. Allis (2000). "The language of covalent histone modifications." Nature **403**(6765): 41-45.
Sweetser, D., M. Nonet, et al. (1987). "Prokaryotic and eukaryotic RNA polymerases have homologous core subunits." Proceedings of the National Academy of Sciences of the United States of America **84**(5): 1192-1196.
Syvanen, A. C. (2001). "Accessing genetic variation: genotyping single nucleotide polymorphisms." Nature reviews. Genetics **2**(12): 930-942.
t Hoen, P. A., Y. Ariyurek, et al. (2008). "Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms." Nucleic acids research **36**(21): e141.
Takamu, I. and Y. Oshima (1970). "ALLELISM TESTS AMONG VARIOUS HOMOTHALLISM CONTROLLING GENES AND GENE SYSTEMS IN SACCHAROMYCES." Genetics **64**(2): 229.
Talbert, P. B. and S. Henikoff (2006). "Spreading of silent chromatin: inaction at a distance." Nature reviews. Genetics **7**(10): 793-803.
Team, R. D. C. (2008). R: a language and environment for statistical computing. Vienna, Austria.
Teixeira, M. T. and E. Gilson (2005). "Telomere maintenance, function and evolution: the yeast paradigm." Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology **13**(5): 535-548.
Thorne, A. W., F. A. Myers, et al. (2004). "Native chromatin immunoprecipitation." Methods in molecular biology **287**: 21-44.
Trifonov, E. N. (2010). "Nucleosome positioning by sequence, state of the art and apparent finale." Journal of biomolecular structure & dynamics **27**(6): 741-746.
Tsankov, A. M., D. A. Thompson, et al. (2010). "The role of nucleosome positioning in the evolution of gene regulation." PLoS biology **8**(7): e1000414.
Valouev, A., J. Ichikawa, et al. (2008). "A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning." Genome research **18**(7): 1051-1063.
Wang, G. P., A. Ciuffi, et al. (2007). "HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications." Genome research **17**(8): 1186-1194.
Wang, G. P., A. Ciuffi, et al. (2007). "HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications." Genome Res **17**(8): 1186-1194.
Wang, H., M. Johnston, et al. (2007). "Calling cards for DNA-binding proteins." Genome Res **17**(8): 1202-1209.
Wang, H., D. Mayhew, et al. (2011). "Multiplexed identification of the genomic targets of DNA-binding proteins." Genome Research **In press**.

Wang, J., X. Dai, et al. (2010). "Identifying the combinatorial effects of histone modifications by association rule mining in yeast." Evolutionary bioinformatics online **6**: 113-131.
Wang, X., G. O. Bryant, et al. (2011). "An effect of DNA sequence on nucleosome occupancy and removal." Nature structural & molecular biology **18**(4): 507-509.
Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature reviews. Genetics **10**(1): 57-63.
Watson, J. D. and F. H. Crick (1953). "The structure of DNA." Cold Spring Harbor symposia on quantitative biology **18**: 123-131.
Weese, D., A. K. Emde, et al. (2009). "RazerS--fast read mapping with sensitivity control." Genome Res **19**(9): 1646-1654.
Wetterstrand, K. A. (2011, 07/2011). "DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program." Retrieved January 6th, 2012, from www.genome.gov/sequencingcosts.
Wickham, H. (2009). Ggplot2 : elegant graphics for data analysis. New York, Springer.
Woodcock, C. L. and R. P. Ghosh (2010). "Chromatin higher-order structure and dynamics." Cold Spring Harbor perspectives in biology **2**(5): a000596.
Xie, W., X. Gai, et al. (2001). "Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p." Mol Cell Biol **21**(19): 6606-6614.
Xu, W., J. G. Aparicio, et al. (2006). "Genome-wide mapping of ORC and Mcm2p binding sites on tiling arrays and identification of essential ARS consensus sequences in S. cerevisiae." BMC genomics **7**: 276.
Yang, S. C., N. Rhind, et al. (2010). "Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing." Molecular systems biology **6**: 404.
Yieh, L., H. Hatzis, et al. (2002). "Mutational analysis of the transcription factor IIIB-DNA target of Ty3 retroelement integration." J Biol Chem **277**(29): 25920-25928.
Yieh, L., G. Kassavetis, et al. (2000). "The Brf and TATA-binding protein subunits of the RNA polymerase III transcription factor IIIB mediate position-specific integration of the gypsy-like element, Ty3." J Biol Chem **275**(38): 29800-29807.
Yu, G. L., J. D. Bradley, et al. (1990). "In vivo alteration of telomere sequences and senescence caused by mutated Tetrahymena telomerase RNAs." Nature **344**(6262): 126-132.
Zemach, A., I. E. McDaniel, et al. (2010). "Genome-wide evolutionary analysis of eukaryotic DNA methylation." Science **328**(5980): 916-919.
Zhu, C., K. J. Byers, et al. (2009). "High-resolution DNA-binding specificity analysis of yeast transcription factors." Genome research **19**(4): 556-566.
Zhu, X. and C. M. Gustafsson (2009). "Distinct differences in chromatin structure at subtelomeric X and Y' elements in budding yeast." PLoS One **4**(7): e6363.
Zhu, Y., J. Dai, et al. (2003). "Controlling integration specificity of a yeast retrotransposon." Proc Natl Acad Sci U S A **100**(10): 5891-5895.
Zhu, Y., S. Zou, et al. (1999). "Tagging chromatin with retrotransposons: target specificity of the Saccharomyces Ty5 retrotransposon changes with the chromosomal localization of Sir3p and Sir4p." Genes Dev **13**(20): 2738-2749.
Zill, O. A., D. Scannell, et al. "Co-evolution of transcriptional silencing proteins and the DNA elements specifying their assembly." PLoS Biol **8**(11): e1000550.
Zillig, W., P. Palm, et al. (1988). "Comparative evaluation of gene expression in archaebacteria." European journal of biochemistry / FEBS **173**(3): 473-482.

Zou, S., N. Ke, et al. (1996). "The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci." Genes Dev **10**(5): 634-645.
Zou, S., J. M. Kim, et al. (1996). "The Saccharomyces retrotransposon Ty5 influences the organization of chromosome ends." Nucleic Acids Res **24**(23): 4825-4831.
Zou, S. and D. F. Voytas (1997). "Silent chromatin determines target preference of the Saccharomyces retrotransposon Ty5." Proc Natl Acad Sci U S A **94**(14): 7412-7416.

7. Appendix I: Supplemental Ty5 Information

Table 7-1: Chromosomal features evaluated in Ty5 study.

Feature	Source	Range[3]
Within 1000bp Upstream of tRNA Gene	SGD[1]	Binary
Within 1000bp Downstream of tRNA Gene	SGD	Binary
Within 100bp Upstream of tRNA Gene	SGD	Binary
Within 100bp Downstream of tRNA Gene	SGD	Binary
Within tRNA Gene	SGD	Binary
Nucleosome Density (Chip Levels)	Lee et al. *Nature Genetics* **39**, 1235 - 1244 (2007)	(-∞,∞)
Nucleosome Density (HMM Calls)	Lee et al. *Nature Genetics* **39**, 1235 - 1244 (2007)	Ternary
Within 1000 bp Downstream of Verified ORF	SGD	Binary
Within 100 bp Downstream of Verified ORF	SGD	Binary
Within 1000 bp Upstream of Verified ORF	SGD	Binary
Within 100 bp Upstream of Verified ORF	SGD	Binary
Within Verified ORF	SGD	Binary
Within Ty1 LTR	SGD	Binary
Within Ty2 LTR	SGD	Binary
Within Ty3 LTR	SGD	Binary
Within Ty4 LTR	SGD	Binary

Within Ty5 LTR	SGD	Binary
Within 500bp of any Transciption Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of YAP6 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MSN2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MSN4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of PHO2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of FHL1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of ABF1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of DIG1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SWI4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SWI5 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SWI6 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of FKH1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of FKH2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of ACE2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary

Within 500bp of AFT2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of DIG1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of BAS1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MOT3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of NRG1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of DAL82 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of CBF1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SUT1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of INO2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of UME6 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RTG3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of GCN4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of PHD1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RAP1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SUM1 Transcription	MacIsaac et al. *BMC*	Binary

Factor Binding Site	*Bioinformatics*, **7:**113 (2006)	
Within 500bp of MCM1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of CIN5 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of YAP5 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MOT3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of NRG1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of CBF1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RTG3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of STP4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of GLN3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of GCR1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of GCR2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of ROX1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of TYE7 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of STE12 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary

Within 500bp of SNT2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SPT2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of REB1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of TEC1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of DAL80 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of ADR1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MAC1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of ARR1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of YAP7 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MET31 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MET32 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MET4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RME1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of HSF1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SFP1 Transcription	MacIsaac et al. *BMC*	Binary

Factor Binding Site	*Bioinformatics*, **7:**113 (2006)	
Within 500bp of STP1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RCS1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MBP1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of YOX1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of YHP1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of INO4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of HAP5 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of HAP2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of HAP4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SKN7 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SPT23 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of ARG80 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of ARG81 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of PUT3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary

Within 500bp of HAP1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SOK2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of XBP1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of YAP1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of LEU3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SKO1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RPN4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of CST6 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of CAD1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of PDR3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RGT1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of IME1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RDS1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of GAL4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of GAL80 Transcription	MacIsaac et al. *BMC*	Binary

Factor Binding Site	*Bioinformatics*, **7:**113 (2006)	
Within 500bp of GZF3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of HAP3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of STB4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of GTS1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of ASH1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of NDD1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of THI2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of PHO4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of STB2 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of YDR520C Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RLM1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of GAT1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of AZF1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MATA1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary

Within 500bp of CHA4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of OPI1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of STB1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of STB5 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RFX1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of HAC1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of GAT3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RPH1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SNF1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of PDR1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of IXR1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of MIG1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of YRR1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SIP4 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of SMP1 Transcription	MacIsaac et al. *BMC*	Binary

Factor Binding Site	*Bioinformatics*, **7:**113 (2006)	
Within 500bp of DAL81 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of UGA3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of ARO80 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RLR1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of YML081W Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of ZAP1 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of YAP3 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Within 500bp of RIM101 Transcription Factor Binding Site	MacIsaac et al. *BMC Bioinformatics*, **7:**113 (2006)	Binary
Level of H3K14Acetylation	Pokholok et al. *Cell*, 122(4): 517-527 (2005)	(-∞,∞)
Level of H4 Acetylation	Pokholok et al. *Cell*, 122(4): 517-527 (2005)	(-∞,∞)
Level of H3K9 Acetylation	Pokholok et al. *Cell*, 122(4): 517-527 (2005)	(-∞,∞)
Level of H3K4 1 Methylation	Pokholok et al. *Cell*, 122(4): 517-527 (2005)	(-∞,∞)
Level of H3K4 2 Methylation	Pokholok et al. *Cell*, 122(4): 517-527 (2005)	(-∞,∞)
Level of H3K4 3 Methylation	Pokholok et al. *Cell*, 122(4): 517-527 (2005)	(-∞,∞)

Sir4 Level	Pokholok et al. *Cell*, 122(4): 517-527 (2005)	(-∞,∞)
Log Sir4 Level	Pokholok et al. *Cell*, 122(4): 517-527 (2005)	[0,∞)
Within 1000 bp of ARS	OriDB[2]	Binary
Within ARS	OriDB	Binary
Within 1000 bp downstream of Uncharacterized ORF	SGD	Binary
Within 1000 bp upstream of Uncharacterized ORF	SGD	Binary
Within Uncharacterized ORF	SGD	Binary
Within 1000 bp downstream of Dubious ORF	SGD	Binary
Within 1000 bp upstream of Dubious ORF	SGD	Binary
Within Dubious ORF	SGD	Binary
Within Y prime	SGD	Binary
Within 1000 bp upstream of Y prime	SGD	Binary
Within 1000 bp downstream of Y prime	SGD	Binary
Ty5 Integration Frequency in *Δsir4* background 1000bp window	Wang et al. *in press* (2011)	[0,∞)
Hermes Integration Frequency 1000bp window	Gangadharan et al. *PNAS* 107(51): 21966–21972 (2010)	[0,∞)
DNAseI Sensitivity 1000bp window	Hesselberth et al. *Nature Methods* **6**,283 -289 (2009)	[0, ∞)

SGD is the Saccharomyces Genome Database found at www.yeastgenome.org.
[2] OriDB is the DNA Replication Origin Database found at www.oridb.org.
[3] Binary indicates a two state feature (either in feature or not in feature), ternary a three state feature; (a, b) represent a range of continuous values, excluding endpoints, between 'a' and 'b'; [a,b) represent a range of continuous values, square brackets indicate that an endpoint is included.

Table 7-2: Oligonucleotide primers used in Ty5 study.

Oligonucleotide Name	*Sequence*
DVO4621	5'-/Phos/CGGTCCCTTAAGCGGAG/3AmM/
DVO4622	5'-GTAATACGACTCACTATAGGGCTCCGCTTAAGGGAC
DVO495	5'-CCATAGTTTCTGTGTACAAGAGT
DVO4632	5'-GTAATACGACTCACTATAGGGC
DVO4665	5'-GCCTTGCCAGCCCGCTCAG AGGGCTCCGCTTAAGGGAC
DVO4666	5'-GCCTCCCTCGCGCCATCAG actgactg TCCCAACAGCTTAGCCAAC
DVO4667	5'-GCCTCCCTCGCGCCATCAG actgacgt TCCCAACAGCTTAGCCAAC
DVO4668	5'-GCCTCCCTCGCGCCATCAG actgatcg TCCCAACAGCTTAGCCAAC
DVO4669	5'-GCCTCCCTCGCGCCATCAG actgagct TCCCAACAGCTTAGCCAAC
DVO4670	5'-GCCTCCCTCGCGCCATCAG actgctag TCCCAACAGCTTAGCCAAC
DVO4671	5'-GCCTCCCTCGCGCCATCAG actgcatg TCCCAACAGCTTAGCCAAC
DVO4672	5'-GCCTCCCTCGCGCCATCAG actgcagt TCCCAACAGCTTAGCCAAC
DVO4673	5'-GCCTCCCTCGCGCCATCAG actgtacg TCCCAACAGCTTAGCCAAC
DVO4674	5'-GCCTCCCTCGCGCCATCAG ctgagact TCCCAACAGCTTAGCCAAC
DVO4675	5'-GCCTCCCTCGCGCCATCAG ctgagcat TCCCAACAGCTTAGCCAAC
DVO4676	5'-GCCTCCCTCGCGCCATCAG ctgactag TCCCAACAGCTTAGCCAAC
DVO4677	5'-GCCTCCCTCGCGCCATCAG ctgacatg TCCCAACAGCTTAGCCAAC
DVO4678	5'-GCCTCCCTCGCGCCATCAG ctgacagt TCCCAACAGCTTAGCCAAC
DVO4679	5'-GCCTCCCTCGCGCCATCAG ctgacgat TCCCAACAGCTTAGCCAAC
DVO4680	5'-GCCTCCCTCGCGCCATCAG ctgatacg TCCCAACAGCTTAGCCAAC

DVO4681	5’-GCCTCCCTCGCGCCATCAG ctgatcag TCCCAACAGCTTAGCCAAC

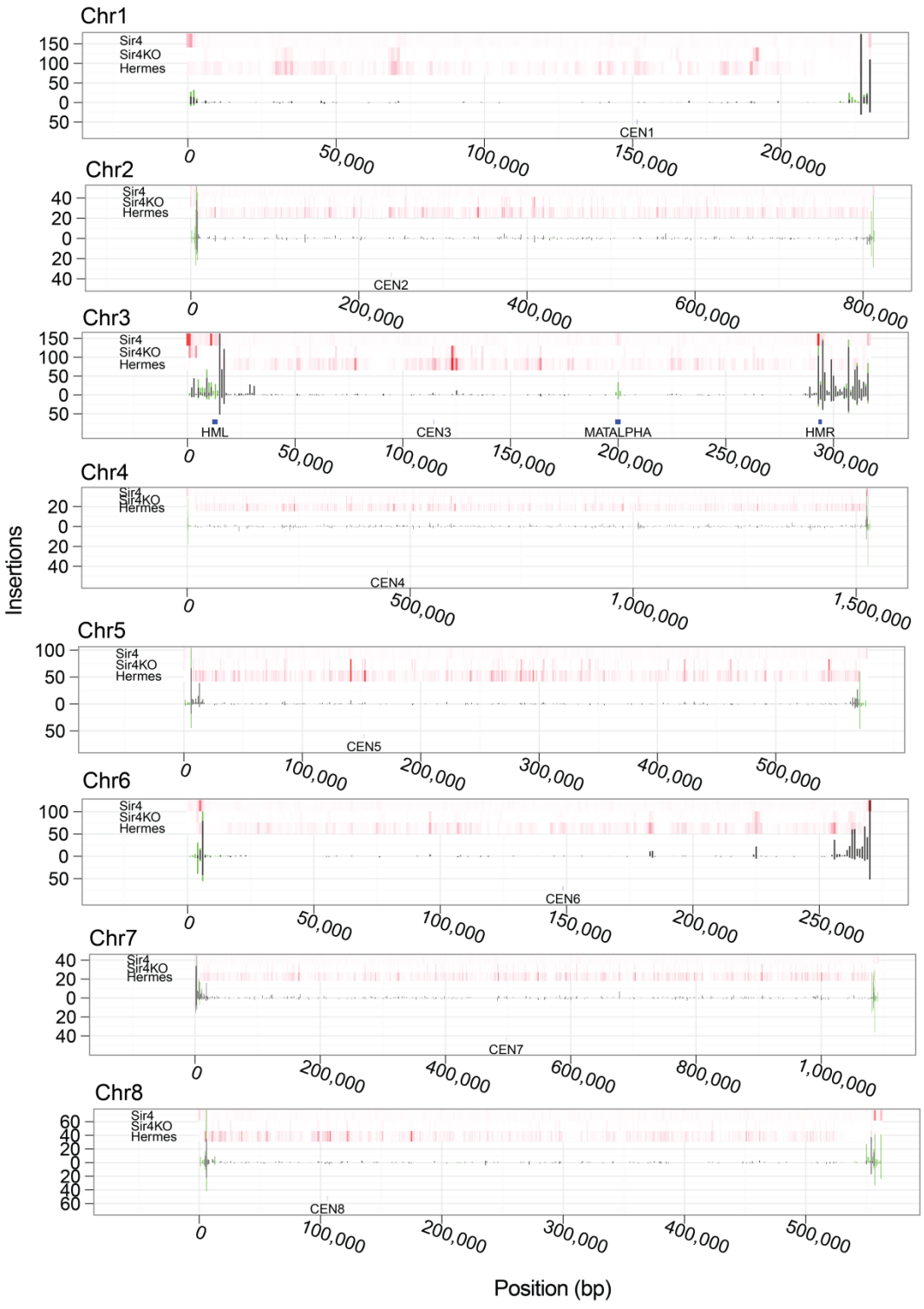

Figure 7-1 – Distribution of Ty5 at Chromosomes 1 through 8

The x-axis denotes position along the chromosome at 1000 bp resolution. Black bars indicate the number of unambiguous integrations at a particular site; stacked green bars indicate additional ambiguous integrations.

Bars above the x-axis indicate data from the haploid strain; bars below the x-axis denote data from the diploid. Red markings above the bar plot indicate prevalence of Sir4, Ty5 insertions in a *Δsir4* strain and Hermes insertions. The intensity of the red color is scaled on a genome-wide basis, such that the brightest red color represents the hottest insertion sites in the genome.

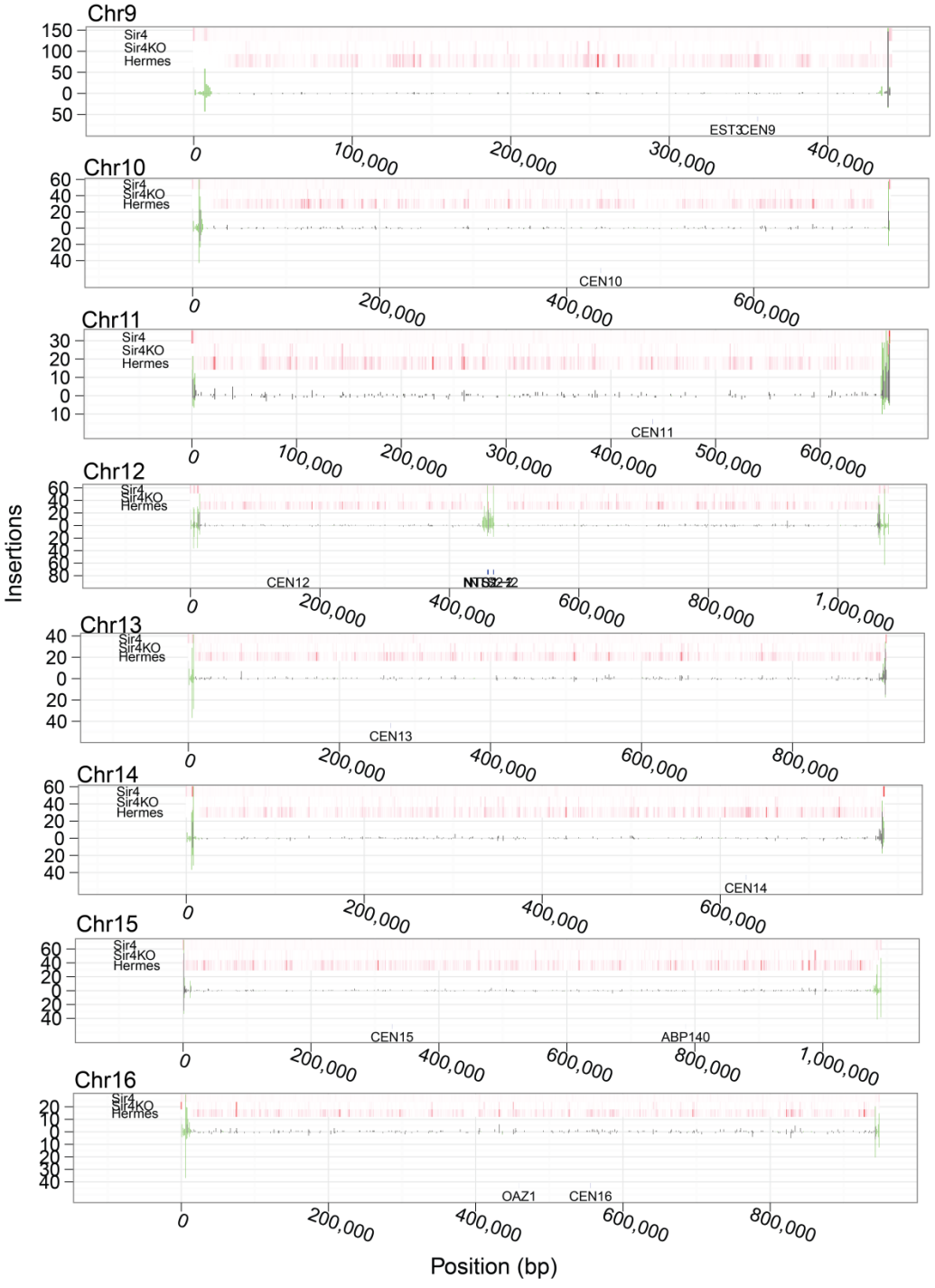

Figure 7-2 – Distribution of Ty5 on Chromosomes 9 through 16

The x-axis denotes position along the chromosome at 1000 bp resolution. Black bars indicate the number of unambiguous integrations at a particular site; stacked green bars indicate additional ambiguous integrations. Bars above the x-axis indicate data from the haploid strain; bars below the x-axis denote data from the diploid. Red markings above the bar plot indicate prevalence of Sir4, Ty5 insertions in a *Δsir4* strain and Hermes insertions. The intensity of the red color is scaled on a genome-wide basis, such that the brightest red color represents the hottest insertion sites in the genome.

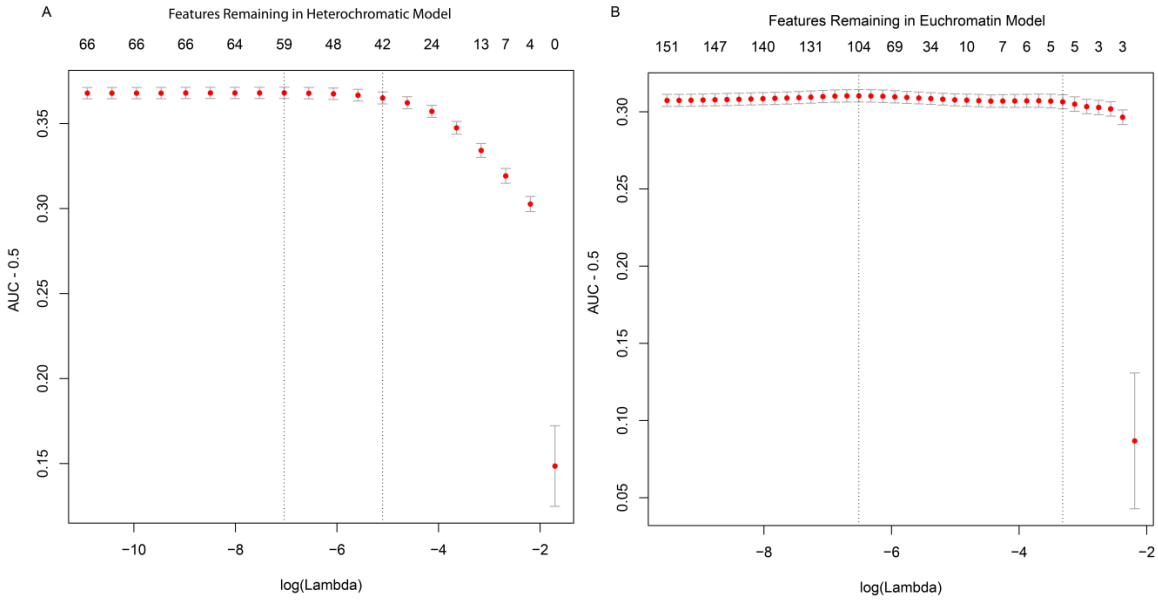

Figure 7-3 – Number of Features in Euchromatin and Heterochromatin models under LASSO regularization

Plot of AUC versus regularization parameter (lambda) for the multidimensional logistic regression on euchromatic data (panel A) and heterochromatic data (panel B). The upper scale for the x-axis shows the number of features remaining when a given regularization coefficient (lambda) is used. The y-axis shows the AUC – 0.5 for each model. Multi-dimensional models were performed to take into account the cooperative or redundant effects between features. These models used all features concurrently. To increase interpretability and prevent over fitting of the model, the multi-dimensional model was regularized on the L1-norm using the Least Absolute Shrinkage and Selection Operator (LASSO). Both the euchromatic and heterochromatic multi-dimensional models showed improvement over their single dimensional counterparts (Supp. LASSO curves). The euchromatic multi-dimensional model achieved an AUC – 0.5 of 0.303 using five features, a significant improvement over the best single dimensional result of 0.25 (Fig. 2). Similarly the heterochromatic multi-dimensional model achieved an AUC – 0.5 of 0.367 though it uses 42 features. A more restrictive lambda value reduced the model to four features and an AUC – 0.5 of 0.303.

The features retained in the euchromatic multi-dimensional model showed that the Ty5 pattern of integration could be described as abundant in sites of Hermes integration, DNAse sensitivity and sites where Ty5 integrates in a sir4Δ background. Ty5 is scarce in sites occupied by nucleosomes and inside verified ORFs. LASSO regularization tends to remove the less predictive of a set of redundant features. As such, this result suggests that the DNAse sensitivity, *Δsir4* and Hermes signals are complementary rather than redundant. All of these features support the theory that Ty5

prefers to integrate into open DNA, and that it is open DNA, rather than interaction with any of these features, that is responsible for site selection.

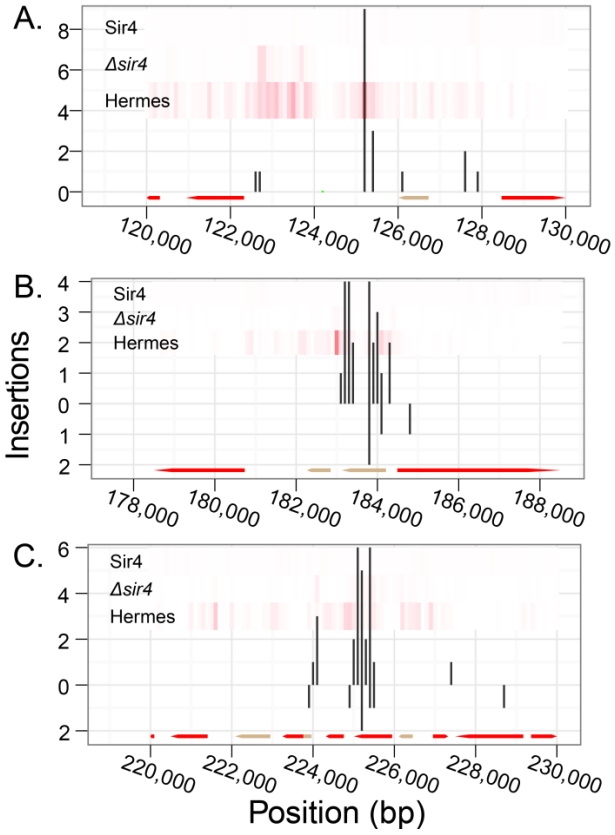

Figure 7-4 – Integration hotspots at Select Euchromatin Loci

A selection of euchromatic Ty5 integration hotspots and their proximity to verified (red) and uncharacterized (tan) ORFs. Black bars indicate frequency of integration events. Bars above the line indicate integrations in the haploid strains, whereas bars below the line are from the diploid strains. Red markings above the bar plot indicate prevalence of Sir4, Ty5 insertions in a *Δsir4* strain and Hermes insertions. The intensity of the red color is scaled on a genome-wide basis, such that the brightest red color represents the hottest insertion sites in the genome.

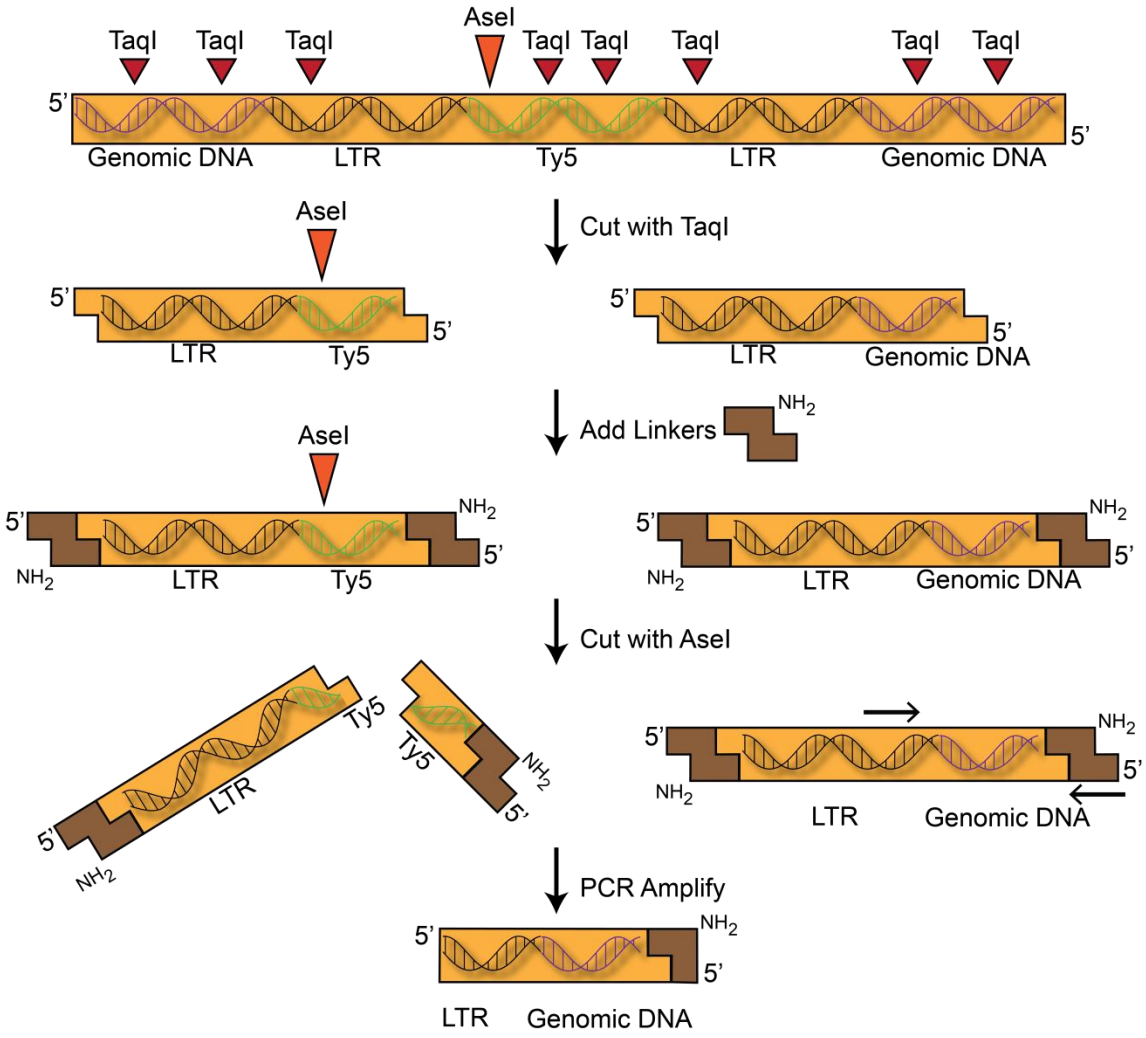

Figure 7-5 – Read Amplification Strategy

Schematic of linker mediated PCR used to amplify genomic sequences adjacent to sites of Ty5 integration. TaqI, a restriction enzyme with a 4bp recognition sequence, was used for illustrative purposes. Genomic DNA containing Ty5 insertions is initially digested with TaqI. Linkers are added to the digested DNA fragments. An additional restriction digestion is performed with AseI to destroy 5' LTR junction fragments. The remaining 3' LTR junction fragments are amplified by PCR and subjected to DNA sequencing.

8. Appendix II: Supplemental Ty1 Information

Table 8-1: *S. cerevisiae* strains used in Ty1 study

Strain Name	Genotype	Reference
BY4741	*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*	GIAEVER *et al.* 2002
BY4741, *hos2Δ*	*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 hos2ΔTkanMX*	GIAEVER *et al.* 2002
BY4741, *rad6Δ*	*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 rad6ΔTkanMX*	GIAEVER *et al.* 2002
BY4741, *rrm4Δ*	*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 rrm4ΔTkanMX*	GIAEVER *et al.* 2002
BY4741, *rtt109Δ*	*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 rtt109ΔTkanMX*	GIAEVER *et al.* 2002
YPH499	*MATa ura3-52 lys2-801_amber ade2-101_ochre trp1-Δ63 his3-Δ200 leu2-Δ1*	SIKORSKI and HIETER 1989
YPH501	*MATa*/*MATα ura3-52/ura3-52 lys2-801_amber/lys2-801_amber ade2-101_ochre/ade2-101_ochre trp1-Δ63/trp1-Δ63 his3-Δ200/his3-Δ200 leu2-Δ1/leu2-Δ1*	SIKORSKI and HIETER 1989

Table 8-2: Ty1 Primers

Primer Name	Primer Sequence[1]	Strains and Restriction Enzyme Treatment
Dvo4864	GCCTCCCTCGCGCCATCAG agctactg TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT ACTG) for Ty1 LTR from 499 AciI
Dvo4865	GCCTCCCTCGCGCCATCAG agctacgt TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT ACGT) for Ty1 LTR from 499 TaqI
Dvo4866	GCCTCCCTCGCGCCATCAG agctatcg TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT ATCG) for Ty1 LTR from 501 AciI
Dvo4867	GCCTCCCTCGCGCCATCAG agctatgc TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT ATGC) for Ty1 LTR

		from 501 TaqI
Dvo4868	GCCTCCCTCGCGCCATCAG agctctag TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT CTAG) for Ty1 LTR from 4741 AciI
Dvo4869	GCCTCCCTCGCGCCATCAG agctctga TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT CTGA) for Ty1 LTR from 4741 TaqI
Dvo4870	GCCTCCCTCGCGCCATCAG agctcgat TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT CGAT) for Ty1 LTR from hos2 delta AciI
Dvo4871	GCCTCCCTCGCGCCATCAG agctcgta TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT CGTA) for Ty1 LTR from hos2 delta TaqI
Dvo4872	GCCTCCCTCGCGCCATCAG agctcatg TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT CATG) for Ty1 LTR from rtt109 delta AciI
Dvo4873	GCCTCCCTCGCGCCATCAG agctcagt TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT CAGT) for Ty1 LTR from rtt109 delta TaqI
Dvo4874	GCCTCCCTCGCGCCATCAG agctgact TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT GACT) for Ty1 LTR from rrm3 delta AciI
Dvo4875	GCCTCCCTCGCGCCATCAG agctgatc TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT GATC) for Ty1 LTR from rrm3 delta TaqI
Dvo4876	GCCTCCCTCGCGCCATCAG agctgtac TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT GTAC) for Ty1 LTR from rad6 delta AciI
Dvo4877	GCCTCCCTCGCGCCATCAG agctgtca TCCCAACAGCTTAGCCAAC	forward primer with barcode (AGCT GTCA) for Ty1 LTR from rad6 delta TaqI
Dvo4665	GCCTTGCCAGCCCGCTCAGAGGGCTCCGCTTA AGGGAC	reverse primer for linker
Dvo4621	/Phos/CGGTCCCTTAAGCGGAG/3AmM/	Part I of linker
Dvo4622	GTAATACGACTCACTATAGGGCTCCGCTTAAGG	Part II of linker

GAC

[1]Lowercase sequence denotes barcode

Table 8-3: List of Features Used in Ty1 Study

Feature	Source	Range[3]
In Subtelomeric Region	SGD[1]	Binary
In Subtelomeric X repeats	SGD	Binary
Near Abf1p Binding Site	SGD	Binary
In Dubious ORFs	SGD	Binary
In Verified ORF	SGD	Binary
Within 1000bp Upstream of Verified ORFs	SGD	Binary
Within 1000bp Downstream of Verified ORFs	SGD	Binary
In Uncharactrized ORFs	SGD	Binary
In a Polymerase I Transcribed Gene Exon	SGD	Binary
In a Polymerase I Transcribed Gene Intron	SGD	Binary
In a Polymerase III Transcribed Gene	SGD	Binary
In a tRNA Gene	SGD	Binary
In an endogenous Ty LTR Sequence	SGD	Binary
In an endogenous Ty Element	SGD	Binary
In a Predicted Autonomously Replicating Sequence (ARS)	SGD	Binary
Within 1000bp Upstream of Predicted ARS	SGD	Binary
Within 1000bp Downstream of Predicted ARS	SGD	Binary
In Subtelomeric Y' Element	SGD	Binary
Within 1000bp Upstream of Polymerase III Transcribed Gene	SGD	Binary
Within 1000bp Downstream of Polymerase III Transcribed Gene	SGD	Binary
Within 1000bp Upstream of tRNA Gene	SGD	Binary
Within 1000bp Downstream of tRNA Gene	SGD	Binary
Near YAP6 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MSN2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MSN4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near PHO2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near FHL1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near ABF1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near DIG1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SWI4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SWI5 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SWI6 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near FKH1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary

Near FKH2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near ACE2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near AFT2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near DIG1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near BAS1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MOT3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near NRG1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near DAL82 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near CBF1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SUT1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near INO2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near UME6 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RTG3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near GCN4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near PHD1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RAP1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SUM1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MCM1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near CIN5 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near YAP5 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MOT3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near NRG1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near CBF1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RTG3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near STP4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near GLN3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near GCR1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near GCR2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near ROX1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near TYE7 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near STE12 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SNT2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SPT2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near REB1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near TEC1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near DAL80 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near ADR1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MAC1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near ARR1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near YAP7 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MET31 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary

Near MET32 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MET4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RME1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near HSF1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SFP1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near STP1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RCS1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MBP1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near YOX1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near YHP1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near INO4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near HAP5 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near HAP2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near HAP4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SKN7 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SPT23 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near ARG80 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near ARG81 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near PUT3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near HAP1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SOK2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near XBP1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near YAP1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near LEU3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SKO1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RPN4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near CST6 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near CAD1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near PDR3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RGT1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near IME1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RDS1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near GAL4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near GAL80 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near GZF3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near HAP3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near STB4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near GTS1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near ASH1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near NDD1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near THI2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary

Near PHO4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near STB2 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near YDR520C Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RLM1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near GAT1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near AZF1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MATA1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near CHA4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near OPI1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near STB1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near STB5 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RFX1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near HAC1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near GAT3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RPH1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SNF1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near PDR1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near IXR1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near MIG1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near YRR1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SIP4 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near SMP1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near DAL81 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near UGA3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near ARO80 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RLR1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near YML081W Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near ZAP1 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near YAP3 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near RIM101 Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Near Any Transcription Factor Binding Site	MacIsaac et al. (2006)	Binary
Nucleosome ChIP Weights	Lee et al. (2007)	(-∞,∞)
Nucleosome Hidden Markov Model Calls	Lee et al. (2007)	Ternary
DNAse Sensitivity	Hesselberth et al. (2009)	(-∞,∞)
In a Predicted Autonomously Replicating Sequence (ARS)	OriDB[2]	Binary
Within 1000bp Upstream of Confirmed ARS	OriDB	Binary
Within 1000bp Downstream of Confirmed ARS	OriDB	Binary
H4 Acetylation Weight	Pokholok et al. (2005)	(-∞,∞)
H3K9 Acetylation Weight	Pokholok et al. (2005)	(-∞,∞)
H3K14 Acetylation Weight	Pokholok et al. (2005)	(-∞,∞)

H3K36me3 Methylation Weight	Pokholok et al. (2005)	(-∞,∞)
H3K4me1 Methylation Weight	Pokholok et al. (2005)	(-∞,∞)
H3K4me2 Methylation Weight	Pokholok et al. (2005)	(-∞,∞)
H3K4me3 Methylation Weight	Pokholok et al. (2005)	(-∞,∞)
H3K79me3 Methylation Weight	Pokholok et al. (2005)	(-∞,∞)

SGD is the Saccharomyces Genome Database found at www.yeastgenome.org.

[2] OriDB is the DNA Replication Origin Database found at www.oridb.org.

[3] Binary indicates a two-state feature (either in feature or not in feature), ternary indicates a 3 state feature; (a, b) represent a range of continuous values, excluding endpoints, between 'a' and 'b'; [a,b) represents a range of continuous values, square brackets indicate that an endpoint is

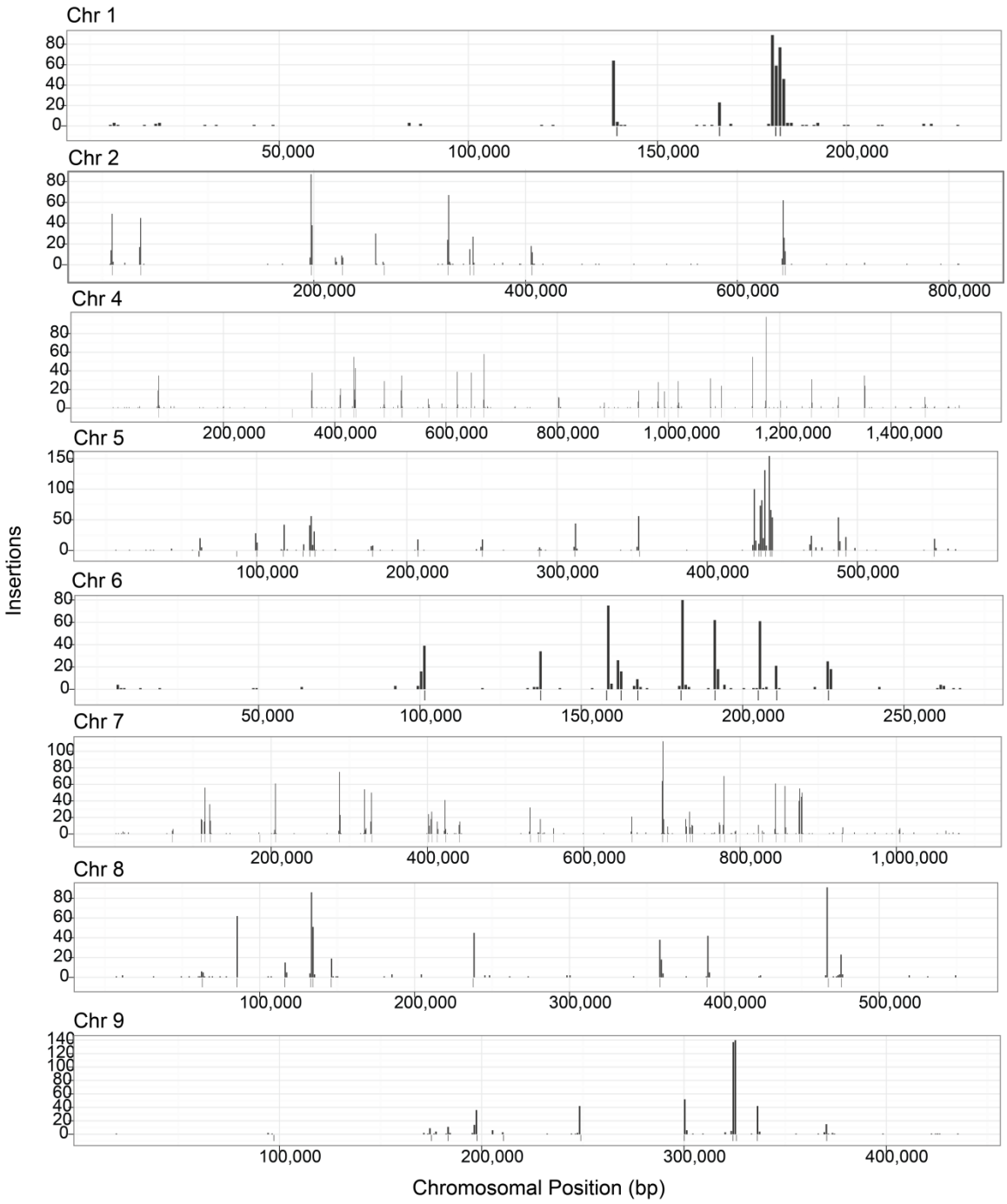

Figure 8-1 – Distribution of Ty1 on Chromosomes 1 through 9

The distribution of insertions on chr III is shown in Fig. 1. The x axes indicate base positions on the chromosome, and they are scaled differently for each chromosome. The y axes indicate the number of insertions within 1 kb windows. Bars below the x axis indicate positions of class III genes.

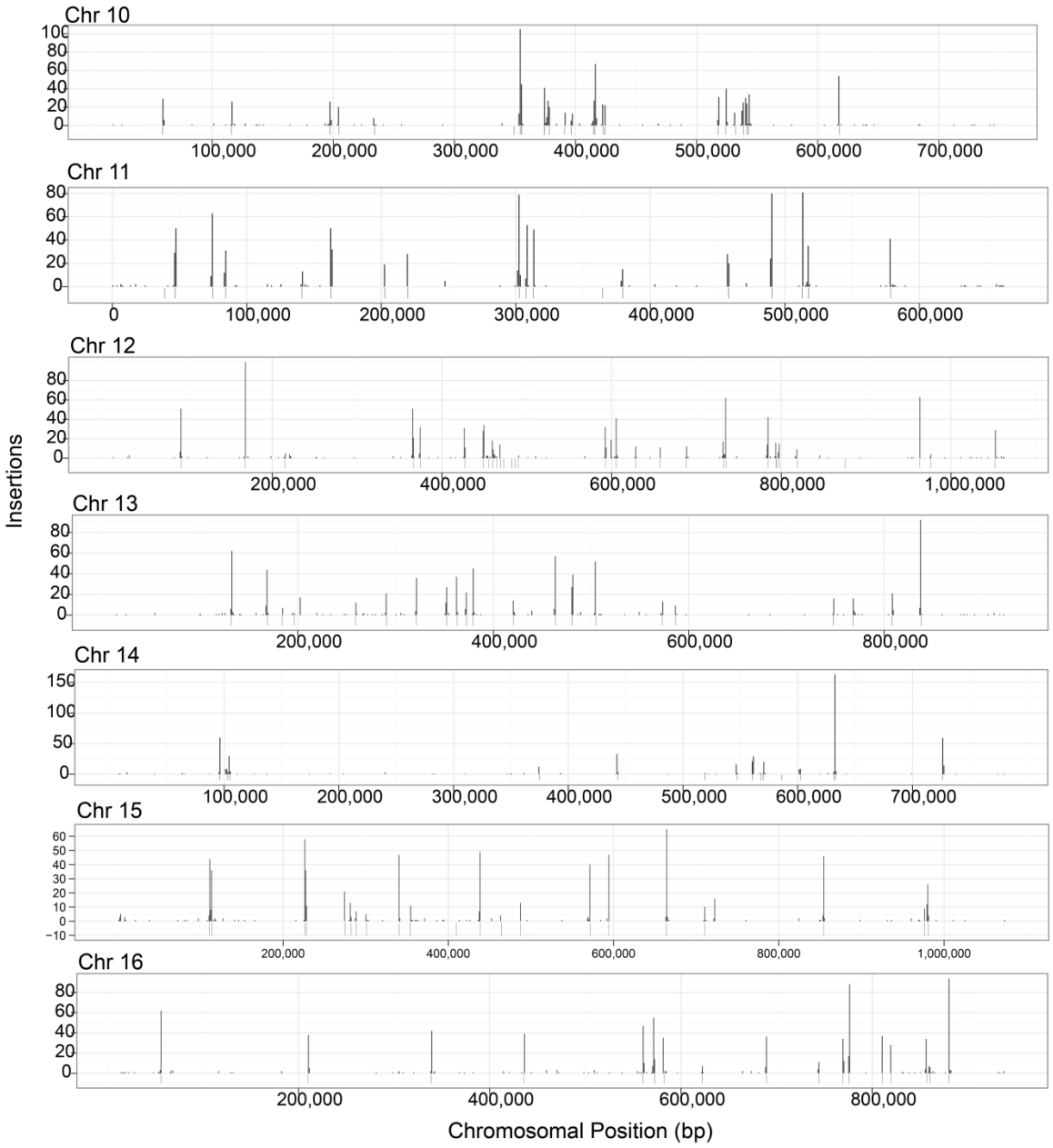

Figure 8-2 – Distribution of Ty1 on Chromosomes 10 through 16

The x axes indicate base positions on the chromosome, and they are scaled differently for each chromosome. The y axes indicate the number of insertions within 1 kb windows. Bars below the x axis indicate positions of class III genes.

9. Appendix III: Supplemental Origin Information

Table 9-1: Coefficients for *S. cerevisiae* Raw Data Model

Feature Position	Feature Weight
-499	0.01881885
-318	0.05944005
-150	0.18317072
2	-0.12330220
8	-0.39516071
9	-0.64796717
10	-0.03275735
15	-0.16279991
16	-0.02829790
117	0.10943531

Table 9-2: Coefficients for *S. cerevisiae* Wavelet Data Model

Coefficient Number	Feature Weight
1030	-0.08213653
1037	0.02194773
1045	0.01567097
1046	-0.02966764
1050	0.01714496
1059	0.08918273
1060	0.03699924
1068	0.01917605

1073	-0.02883062
1076	-0.04628602

Table 9-3: Coefficients for *S. cerevisiae* FFT Data Model

Period	Feature Weight
1024	-0.0002986458
512	0.0025718440
341.3	-0.0010787793
256	0.0030543237
204.8	-0.0017432159
170.7	0.0052592150

Table 9-4: Coefficients for *S. cerevisiae* All Data Model

Coefficient	Feature Weight
Fourier Period 170.6	0.0005303897
Raw Position 9	-0.2000452470

Table 9-5: Coefficients for *K. lactis* Raw Data Model

Feature Position	Feature Weight
-486	0.023830709
-319	0.032423006
-115	0.089902366

-114	0.022262444
-16	-0.103226050
-15	-0.050603120
-14	-0.555465177
-5	-0.442318641
15	-0.566878789
116	0.005972398
213	-0.038880131
224	-0.053904061
226	-0.044876149
228	-0.055137417
506	0.158501489

Table 9-6: Coefficients for *K. lactis* Wavelet Data Model

Coefficient Number	Feature Weight
1033	-0.023119334
1046	-0.048558482
1047	0.019004992
1059	0.046856546
1060	0.047107202
1068	0.007722621
1076	-0.036862934

Table 9-7: Coefficients for *K. lactis* FFT Data Model

Period	Feature Weight
1024	-0.0005344339
512	0.0026637716
256	0.0005353598
204.8	-0.0025702169
170.7	0.0044460580

Table 9-8: Coefficients for *K. lactis* All Data Model

Coefficient	Feature Weight
Position 14 Raw	-0.0427902400
Position -6 Raw	-0.4374355529
FFT Period 170.7	0.0006081609

Table 9-9: Coefficient for *K. lactis* Second Pass - No ORC (By Raw Position)

-495	0.025449
-450	-0.10305
-435	-0.0828
-420	0.074888
-418	0.032797
-407	0.002426
-387	-0.20978
-243	0.219562
-216	-0.16087
-141	-0.13547
-18	0.238764
6	-2.96225

15	-0.31868
21	1.457949
37	-0.7833
56	0.363508
160	-0.13156
202	0.517584
226	-0.87217
292	0.114423
393	-0.06256
471	0.692398
485	-0.69341
486	-0.1244