January 2009

Calling Cards For DNA-Binding Proteins

Haoyi Wang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Molecular Cell Biology

Dissertation Examination Committee:
Mark Johnston, Chair
Robi Mitra
John Majors
Gary Stormo
Sheila Stewart
Douglas Chalker

"CALLING CARDS" FOR DNA-BINDING PROTEINS

by

Haoyi Wang

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirement for the degree
of Doctor of Philosophy

August 2009

Saint Louis, Missouri

Abstract

Organisms respond to their environment by altering patterns of gene expression. This process is orchestrated by transcription factors, which bind to specific DNA sequences near genes. In order to understand the regulatory networks that control transcription, the genomic targets of all transcription factors under various conditions and in different cell types must be identified. This remains a distant goal, mainly due to the lack of a high-throughput, *in vivo* method to study protein-DNA interactions. To fill this gap, I developed transposon "Calling Cards" for DNA-binding proteins. I endowed DNA binding proteins with the ability to direct the insertion of a transposon into the genome near to where they bind. The transposon becomes a "Calling Card" that marks the visit of a DNA-binding protein to the genome. I demonstrated that the Calling Card method is accurate and robust. I combined Calling Cards with "next generation" DNA sequencing technology to increase the sensitivity, specificity, and resolution of the method. This improved method ("Calling Card-Seq") allows for multiple transcription factors to be analyzed in a single experiment, greatly increasing sample throughput. I used Calling Card-Seq to study transcription factors of the yeast *S. cerevisiae* that have not been well-characterized, and I successfully identified DNA sequence recognition motifs and target genes for many of them. Calling Card-Seq will enable a systematic exploration of transcription factor binding under many different environments and growth conditions in a way that has heretofore not been possible. This dissertation describes my work developing this method, as well as several interesting results obtained using this method to study the gene regulatory networks of the yeast *S. cerevisiae*.

Acknowledgement

Many people helped me during my graduate study. First, I would like to thank my mentor Mark Johnston for welcoming me to his lab and getting me started on this very exciting project. Mark taught me a lot about yeast genetics and showed me how he addresses problems as a geneticist. He was always encouraging and supportive. He gave me total freedom to explore many possibilities with intellectual and financial support. I also want to thank my co-mentor Robi Mitra. As a bioengineer, Rob has perspectives and expertise different from Mark. He taught me a lot about thinking about questions in a quantitative way. It was so much fun to work with Mark and Rob. I thoroughly enjoyed every weekly meeting we had. They also gave me a lot of advice and help in my career planning. Working with Mark and Rob made me feel the great fun of doing science, and helped me make up my mind to pursue an academic career.

I also need to thank the other members of my thesis committee, John Majors, Gary Stormo, Sheila Stewart, and Douglas Chalker. They have given me important input and guidance towards finishing this dissertation. Douglas Chalker provided me valuable advice when I first started working with Ty5. Sheila Stewart gave me suggestions on the project of developing *piggyBac* Calling Cards in mammalian cells. Gary Stormo and John Majors shared with me their insights on transcriptional regulation. It has always been a pleasure to have our meeting and discuss science with them.

Mark Johnston lab is like a big family. Jim Dover taught me almost every experimental technique working with yeast. He is also a great friend to discuss political and philosophical issues, which often ended up as a fierce debate. Xuhua Chen has been working with me on "Calling Cards" projects for one and a half year. She is very smart, responsible, and hard working. Thanks to her help, I was able to explore more questions. Jeff Sabina, Chris Hittinger, Victoria Brown-Kennerly, and other lab members gave me advice during my graduate study. It was great to brainstorm new ideas with them. I had my first kid last year, and the whole lab gave me great supports. All the female members of the lab, Yanfang Jiang, Victoria Brown-Kennerly, Linda Riles, Jessica Jackson, Su-Wen Ho, and Xuhua Chen gave my wife and me much advice about having a kid, and prepared a wonderful baby shower for us.

I would like to thank the Robi Mitra lab, where I spent the last half year after Mark left for Denver. David Mayhew started work on "Calling Cards" projects one and a half years ago. He developed an analysis pipeline for the large amount of sequencing data and a very nice software GNASHY for visualizing the data. He also contributed significantly on data analysis. Without his help, I wouldn't have been able to finish Chapters 3, 4, and 5. All the other Rob lab members are good friends. I would also like to thank Jessica Hoisington-Lopez, Laura Langton, Brian Koebbe, and Jim Dover for their expert assistance with the use of Illumina Genome Analyzer. It has been a great pleasure to work with all the other members of the Center for Genome Sciences, who have given me too much help to mention in this limited space.

There are many other friends that have made my graduate school experience enjoyable. Shuo Luo, a neuroscience student, and DaMa, a developmental biology student, entered the DBBS program the same year as I did. We have been good friends since then. They provided me both scientific input and great friendship during graduate school. There are many friends in the Chinese Students and Scholars Association

(CSSA). We have worked together to contribute to the Chinese community at school, and I am very proud of it.

My big family in China has been very supportive. I am sorry that I didn’t go back to China more often to visit my grandparents and other relatives. My parents provided me with all the freedom and support to pursue my career. Both of them played great role models for me. My wife Yilin Yu not only takes care of me and our son, but also helped me with some replica plating work. As a pianist, she knows how to replicate many plates without fatigue. Last but not least, I would like to thank my son, Xiaoyu Wang, for being so cute.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Transcriptional Regulation

The genetic information of organisms is encoded in the DNA sequence of their genomes. Although every cell of a multi-cellular organism has the same genome, expression of each gene is regulated differently in different types of cells, causing them to develop different morphologies and functions, and form different tissues and organs. Similarly, single celled organisms respond to environmental changes by regulating gene expression dynamically.

Gene expression is regulated primarily at the transcriptional level (Latchman 1997). The transcriptional regulatory program of each gene is written in the DNA sequence of its promoter. One of the best studied examples is the *GAL1-10* promoter in the yeast *Saccharomyces cerevisiae.* This divergent promoter lies upstream of the *GAL1* and *GAL10* genes and regulates transcription of both genes in response to environmental cues. In the presence of glucose, both genes are shut down almost completely; galactose induces expression of both genes about 1,000 fold (Johnston and Davis 1984). This promoter drives expression of other genes in response to galactose (Guarente et al. 1982; Johnston and Davis 1984; Sopko et al. 2006; West et al. 1984), demonstrating that this regulation is encoded in the *GAL1-10* promoter sequences.

Three common features are apparent in the promoters of most protein-coding genes: the transcription start site, the “TATA box”, and specific sequences recognized by DNA-binding proteins (transcription factors) (Lee and Young 2000). The TATA box is an AT-rich sequence located 25 to 120bp upstream of transcription start site (Struhl 1995). The TATA-box binding protein (TBP) contacts the minor groove of the TATA sequence and bends the DNA (Kim et al. 1993a; Kim et al. 1993b) to stimulate transcription initiation. Different sequence-specific transcription factors (TFs) bind to specific sequences in promoters to activate or repress transcription.

Transcription factors are DNA-binding proteins that recognize and bind to specific short DNA sequences. They contain a DNA-binding domain and one or more activation or repression domains (Kadonaga 2004). The modular nature of transcription factors allows each domain to be functional when expressed as part of a chimeric protein. This modularity has been exploited in many applications, such as the yeast two hybrid screen (Fields and Song 1989; Uetz et al. 2000), and artificial Zinc-finger transcription factors (Bae et al. 2003) and nucleases (Santiago et al. 2008).

Transcription factors function as transcriptional activators and/or repressors. Activators stimulate gene transcription by recruiting to promoters the general transcription apparatus (Ptashne and Gann 1997) as well as chromatin-modifying complexes (Agalioti et al. 2000); repressors inhibit transcription by binding to activators (Leuther and Johnston 1992) or by competing for activator binding sites and/or by recruiting histone deacetylases (Ayer 1999). The level of expression of each gene *in vivo* depends on the balance of activators and repressors bound to its promoter.

Because of their essential role in regulating gene expression, overexpressing or deleting transcription factors often cause dramatic phenotypic changes. Several groups have demonstrated that mammalian cells can be induced to acquire different cell fates by over-expressing or deleting one or more transcription factors. For example, ectopic expression of the myogenic transcription factor Myod converted fibroblasts into muscle cells (Weintraub et al. 1989). Deletion of Pax5 dedifferentiated mouse B cells into progenitor cells (Nutt et al. 1999). Perhaps the best example is the ground-breaking work of Yamanaka, who was able to reprogram somatic cells and convert them into pluripotent stem cells by over-expressing four transcription factors (Takahashi and Yamanaka 2006). These studies highlight the crucial role transcriptional regulation plays in important biological processes such as development.

Gene regulatory networks can be quite complex. A single transcription factor often regulates a large set of genes to coordinate their expression. Furthermore, some promoters recruit multiple transcription factors that act in a combinatorial fashion to control gene expression. The situation is further complicated by the presence of feedback loops in which the gene targets of a transcription factor also regulate its expression. To understand transcriptional regulation, many methods have been developed to study protein-DNA interactions. Here I review the major experimental approaches.

Methods to Study Protein-DNA interactions

In the 1980s, researchers began to comprehensively analyze gene promoters in eukaryotes. The dissection of the promoters of the *HIS3* gene in yeast (Struhl 1981), the herpes simplex virus thymidine kinase gene (McKnight and Kingsbury 1982), and others led to discovery of several *cis*-acting elements such as the heat shock element (Pelham 1982) and the glucocorticoid response element (Chandler et al. 1983). Identifying the transcription factors that bound these elements became a major goal.

The most widely applied method for identifying and characterizing DNA-binding proteins is the **Electrophoretic Mobility Shift Assay** (EMSA) (Fried and Crothers 1981;

Garner and Revzin 1981). This method is based on the fact that DNA-protein complexes move through polyacrylamide gels more slowly than unbound DNA fragments.

DNA fragments containing the binding sites of one or multiple transcription factors are radioactively or fluorescently labeled. A purified DNA-binding protein or crude cell extract is combined with labeled DNA fragments in a buffer with appropriate salt concentration and pH to allow proteins binding to DNA. This mixture is loaded onto a polyacrylamide gel and subjected to electrophoresis. The gel is then imaged to visualize the positions of the protein-DNA complexes and the unbound DNA fragments. Bound DNA fragments migrate more slowly than unbound DNA, so if a mobility shift is observed (relative to a no-protein control), it signals a protein-DNA interaction. The specificity of the protein-DNA interaction is confirmed by its competition with unlabeled probe (Carthew et al. 1985). Here, unlabeled DNA fragments are added to the binding reaction as a specific competitor for protein binding. The amount of the mobility-shifted band should diminish as more unlabeled DNA is added. However, adding unlabeled DNA fragments with unrelated sequences or mutations in the transcription factor binding sites (nonspecific competitors) will not diminish the amount of the shifted band. To confirm the identity of the proteins present in the protein-DNA complex, specific antibodies can be used to perform a super shift assay (Kristie and Roizman 1986). The antibody is added to the binding reaction, and if the antibody recognizes the protein, an antibody-protein-DNA complex will be formed and cause a further shift (super shift) relative to the protein-DNA complex.

EMSA is relatively simple and fast. Its high sensitivity allows for the detection of femtomole quantities of transcription factors. It can also be used to study the kinetic and thermodynamic properties of protein-DNA interactions (Chodosh et al. 1986; Fried and Crothers 1981). However, as an *in vitro* assay, EMSA does not always recapitulate the protein-DNA interactions that occur *in vivo*. Another problem with the EMSA technique is that there are many variables in the binding reaction and in the gel electrophoresis, so experiments studying specific protein-DNA interactions usually need to be optimized individually, making it difficult to use this method in a high throughput manner.

Another widely used method for studying protein-DNA interactions is **DNase I Footprinting Analysis** (Galas and Schmitz 1978). The basic idea of this method is as follows: when a protein binds to a DNA sequence, it protects the DNA phosphodiester backbone from being hydrolyzed by DNase I. When the end-labeled DNA fragments that were cleaved by DNase I are separated by gel electrophoresis, a “footprint” of the binding can be observed.

In practice, the following protocol is used: first, a DNA fragment containing the protein-binding sites is radioactively labeled at one end. A series of dilutions of the DNA-binding protein of interest is prepared to cover a wide range of concentrations. In separate tubes, an increasing amount of the protein is added to the DNA fragments. Next, an amount of DNase I sufficient to partially cleave the DNA is added. The DNA is then ethanol-precipitated, resuspended in gel loading buffer, and applied to a polyacrylamide gel. After electrophoresis, the gel is dried and exposed to film to visualize the DNA bands. The intensity of each band is then determined from the audioradiogram. Since protein binding protects specific sites in the DNA fragment from enzymatic digestion, the fractional protection (*f*) of the DNA by the protein can be inferred from the band intensities. By plotting *f* versus protein concentration, a binding curve (and equilibrium constant) can be obtained for different protein-binding sites. DNA fragments with multiple binding sites can still be analyzed by this methodology -- by analyzing all of the binding curves, equilibrium constants for intrinsic and cooperative binding can be determined (Brenowitz et al. 1986).

This assay is highly quantitative, but it requires using purified DNA-binding protein, since the exact concentrations of both protein and DNA need to be known to derive equilibrium constants. With minor modifications, this method can also be used to identify crude extracts or partial fractions with protein binding activity during the purification of the DNA binding protein.

Since transcription factors are present at low abundances in most eukaryotic cells, it is difficult to purify them using conventional chromatographic methods. Sequence-specific **DNA-Affinity Chromatography** (Kadonaga and Tjian 1986) was developed and widely applied to purify transcription factors that bind to defined DNA elements. In this method, a protein extract is run through an affinity matrix coupled to a specific DNA sequence. The particular DNA-binding protein binding to this sequence will be retained by the matrix. The DNA binding protein is then eluted from the matrix by applying a solution that disrupts non-covalent binding (e.g. a high salt buffer).

To perform DNA-Affinity Chromatography, the DNA sequence recognized by the DNA-binding protein must be known, and the optimal protein-DNA binding conditions (i.e. pH, ionic strength) need to be determined using methods such as the EMSA or DNase I footprinting assays previously described. To avoid contamination with nonspecific DNA binding proteins, non-specific competitor DNA must be added. Various non-specific competitor DNA sequences, such as poly dI-dC and calf thymus DNA, should be tested using DNase I footprinting or EMSA to determine a concentration that does not interfere with the specific protein-DNA interaction. It is also helpful to use two or more different DNA-affinity columns, each containing protein-binding sites with different flanking sequences.

Many sequence-specific DNA-binding proteins have been identified by DNA-affinity chromatography, including Sp1 (Briggs et al. 1986), AP-1 (Lee et al. 1987), and HSF (Wu et al. 1987). This method can purify proteins to more than 95 percent homogeneity. However, for each protein-DNA interaction, a careful optimization is required, and usually, the protein needs to be partially purified using conventional chromatography beforehand. These requirements make the universal application of this method difficult.

Since purifying transcription factors is difficult, several methods were developed to identify the genes that encode these factors without the need for protein purification. One method is the **Southwestern Screen**, which was developed to identify clones encoding DNA-binding proteins from cDNA libraries (Singh et al. 1988).

In this method, mRNAs are first reverse transcribed into cDNA and cloned into the λgt11 vector to make a phage library, which is then plated on a lawn of E. coli cells. Once plaques become visible, a nitrocellulose filter saturated with IPTG is overlaid on each plate to induce expression of recombinant protein from each plaque. The proteins are then immobilized on the filter, which is then probed with end-labeled DNA fragments containing a protein-binding site of interest. After the filter is washed to remove nonspecifically bound probes, the filter is processed for autoradiography. The clones that express proteins that specifically bind to the DNA probes are identified by correlating the radioactivity on the filter to the position of plaques on the plates. Usually it takes more than one round of screening and plaque purification to get *bona fide* positive clones.

The Southwestern Screen was the first method that could identify genes encoding specific DNA-binding proteins without protein purification. This enabled the identification of numerous transcription factors, including MBP-1 (Singh et al. 1988) and CREB (Hoeffler et al. 1988). However, this method also has some limitations. This method will only identify cDNA clones that are highly expressed and whose protein products are folded in a functional form in *E. coli* cells. The protein must also bind the DNA probes strongly enough to withstand multiple washes. Finally, proteins that bind DNA cooperatively with a cofactor will not be identified by this method.

Another well-established method for identifying transcription factors that interact with specific DNA sequences is the **Yeast One-Hybrid Screen** (Y1H) (Meijer et al. 1998; Wang and Reed 1993). This method screens a cDNA library for transcription factors that bind to specific DNA sequences to drive a reporter gene.

The first step is to make a reporter construct. The DNA sequence of interest, the "bait", is cloned upstream of reporter gene *HIS3*. The bait sequence can be a defined transcription factor binding site, a complete promoter, or any DNA sequence thought to be bound by a protein. This construct is then integrated into the yeast genome to make the reporter strain. Since reporter genes are often "leaky" – they display some background expression even in the absence of a cDNA clone due to transcription factors native to yeast -- it is important to inhibit the background expression of *HIS3* reporter gene using 3-amino-triazole (3-AT), a competitive inhibitor of His3 enzyme. For each bait strain, one must determine the optimal concentration of 3-AT to inhibit background expression of the reporter gene.

A library of cDNAs are cloned in frame to sequences that encode a strong transcriptional activation domain, such as that of Gal4 or VP16 of Herpes virus, and driven by a constitutively active promoter such as the *ADH1* promoter. The plasmid library is then transformed into the yeast reporter strain on a scale large enough to saturate the library. Proteins that bind to the bait sequence will activate the *HIS3* reporter gene, enabling the yeast cells to survive on a synthetic media plate lacking histidine (and containing the appropriate amount of 3-AT). A typical Yeast One-Hybrid Screen generates tens to hundreds of positives, many of which are false positives due to nonspecific DNA-binding proteins. Therefore, it is often necessary to screen these positives in control reporter strains that contain different bait sequences to discriminate true positives from false positives. Further verification by an independent method such as EMSA or DNase I footprinting is necessary to confirm the identification of the cDNA clones encoding transcription factors.

The Yeast One-Hybrid method has been used to identify genes encoding transcription factors from yeast (Li and Herskowitz 1993), plants (Kim et al. 1997), and animals (Wang and Reed 1993). It has also been successfully used to reveal the architecture of transcriptional regulatory networks in *C. elegans* (Deplancke et al. 2004; Deplancke et al. 2006; Vermeirssen et al. 2007). Unlike the biochemical methods introduced earlier, this *in vivo* assay does not require specific optimization for each protein-DNA interaction. Yeast cells serve as a better test tube to ensure the proper folding and modification of eukaryotes transcription factors, so the Yeast One-Hybrid often has a higher rate of success than the *E. coli*-based Southwestern screen. Its simplicity allows for high-throughput analysis of many DNA sequences and transcription factors. One shortcoming of the Yeast One-Hybrid Screen is that transcription factor-DNA binding is not queried at the native genomic locus. In addition, factors that bind cooperatively are unlikely to be identified using this method.

The recent sequencing of the complete genomes of many model organisms has resulted in the identification and annotation of most of their protein-coding genes. By comparing the amino acid sequences of uncharacterized genes to the sequences of known transcription factors, it is possible to identify new transcription factors (for example, by looking for DNA-binding domains in uncharacterized open reading frames (ORFs)) (Luscombe et al. 2000; Vaquerizas et al. 2009). This type of **Computational Prediction** has been performed across all completely sequenced genomes to predict sequence-specific transcription factors, and the results were deposited in the DBD database (Wilson et al. 2008). There are also numerous studies predicting transcription factors for individual organisms, such as *Saccharomyces cerevisiae* (Cherry et al. 1998) (SGD), *Caenorhabditis elegans* (Reece-Hoyes et al. 2005), and *Drosophila melanogaster* (Adryan and Teichmann 2006). These genome-wide computational analyses are powerful, although these predictions may contain a fair number of false positives and false negatives, because these studies make their predictions based on DNA-binding domain homology between the query proteins and well-characterized transcription factors. Hence, any protein with a domain similar to a known DNA-binding domain, for example, a protein with an RNA-binding domain, could be falsely predicted as DNA-binding proteins. This could result in false positive predictions. Conversely, novel classes of transcription factors not yet recognized could be omitted from this analysis, due to the lack of significant homology to known transcription factors. In practice, it is important to confirm computational predictions by experimental methods.

Due, in large part, to the methods discussed above, a reasonably comprehensive catalogue of DNA binding proteins has been compiled for most model organisms. The next challenge is to understand where these transcription factors bind *in vivo*. Although biochemical binding analyses can quantitatively describe the *in vitro* binding of transcription factors, such assays do not provide enough information to predict *in vivo* binding, mainly because DNA is packaged into chromatin. Chromatin is DNA wrapped in a chain of basic units of proteins called the nucleosome, which consists of dimmers of four different histone proteins. Within each nucleosome, 146 bp DNA is wrapped around one histone octamer. In general, DNA packaged into chromatin is inaccessible to transcription factors. In response to environmental signals, histones and DNA are modified covalently and/or non-covalently and the chromatin structure changes dynamically to regulate different protein-DNA interactions.

Chromatin Immunoprecipitation (ChIP) is the most widely used method to map the *in vivo* binding of individual transcription factors. In this method, the protein of interest is immunoprecipitated along with the DNA bound to it (Gilmour and Lis 1984; Solomon et al. 1988; Solomon and Varshavsky 1985).

Cells are first treated with formaldehyde to crosslink proteins to DNA as well as to other proteins. Next, a whole cell extract is made and subjected to sonication, which shears the DNA into short fragments (usually about 500 bp). The transcription factor of interest, along with the DNA bound to it, is then immunoprecipitated with a specific antibody. After reverse the protein-DNA crosslinks, the immunoprecipitated DNA is released and identified by quantitative PCR, by hybridization to a microarray, or by high-throughput DNA sequencing.

To identify the genomic loci bound by a transcription factor using quantitative PCR (**ChIP-qPCR**), primers are designed to amplify a specific genomic region of interest. PCR is then performed using immunoprecipitated DNA as template and input DNA (i.e. not immunoprecipitated) as a control. By quantifying the amount of DNA that is “pulled down” from the locus of interest relative to a control locus in both the ChIP and control samples, protein-DNA interactions can be detected. However, ChIP-qPCR can only analyze a small number of predefined genomic loci.

DNA fragments enriched by ChIP can also be hybridized to microarray (**ChIP-chip**). This allows for the identification of all of the genomic regions that are bound by a given transcription factor (Reid et al. 2000; Ren et al. 2000). To implement this protocol, ChIP samples are amplified by PCR, labeled with fluorescent dye, and hybridized to an oligonucleotide microarray. To ensure that all bound fragments will be amplified by the PCR reaction, universal adapters are ligated to the immunoprecipiated DNA fragments. The immunoprecipitated sample and the input sample are then labeled with different dyes and hybridized together to a genome tilling microarray. Any genomic regions that were bound by a given protein will display a high fluorescence signal in the ChIP channel relative to the control channel on the microarray. This method has been used to identify targets of DNA-binding proteins in yeast (Harbison et al. 2004; Lee et al. 2002; Ren et al. 2000) and higher eukaryotes (Boyer et al. 2005; Zeitlinger et al. 2007). The ChIP-chip method is powerful, but limited by the availability and design of microarray. For example, a microarray containing probes for all of the intergenic regions in a genome can not be used to detect binding events in introns and exons.

Recently, second-generation sequencing technologies have become widely available, which has greatly reduced the cost of DNA sequencing. Several groups have used these platforms to readout ChIP samples (**ChIP-seq**) (Johnson et al. 2007; Robertson et al. 2007). Samples from standard ChIP experiments are ligated with linkers and amplified by the PCR. The PCR products are then loaded on an Illumina flowcell, and millions of single DNA molecules are amplified clonally. By taking pictures of millions of spots on the flowcell at the end of each cycle of incorporation of nucleotides labeled with different fluorescent dyes, millions of DNA molecules are sequenced simultaneously. ChIP-seq generally performs better than ChIP-chip because this method is not limited by the availability and design of microarrays, and avoids many technical issues associated with microarrays, such as the cross hybridization of the DNA sample to different probes.

The ChIP-based methods are powerful because they are flexible. They can be used to analyze a wide variety of DNA-binding proteins in many different organisms and cell types. However, ChIP methods provides only a snapshot of DNA-binding at the time the protein is precipitated, which means that each experiment measures the binding of each transcription factor under on culture condition or in one cell type at one point in time. This makes it difficult and expensive to exhaustively catalogue the binding patterns of a large number of transcription factors under a variety of conditions.

Another method for mapping protein-DNA interactions *in vivo* is **Dam Identification** (DamID) (van Steensel et al. 2001; van Steensel and Henikoff 2000). The idea is to attach a DNA-binding protein to *E coli* DNA adenine methyltransferase (Dam), which will methylate the genomic regions bound by the protein. Dam methylates the adenine in the sequence GATC, a base not normally methylated in eukaryotes.

To perform DamID, a transcription factor-Dam fusion is expressed in a eukaryotic cell, and genomic DNA is extracted and digested with Dpn I. The digested DNA fragments are then size selected by sucrose gradient centrifugation. As a control, samples are also prepared from cells expressing only Dam. The experimental and control samples are labeled with different fluorescence dyes and hybridized to microarray. Since Dpn I only cuts methylated GATC sites, and since the transcription factor-Dam fusion methylates only GATC sites that are close to a binding site, any genomic targets of the transcription factor will produce a higher signal on microarray.

This method employs a very interesting concept, but has serious limitations. This method does not directly detect a transcription factor binding site, but instead detects the methylation of GATC sites close to the binding sites, resulting in a resolution of roughly 2-5 kb (van Steensel et al. 2001). This relatively poor resolution often makes it difficult to identify the recognition sequence of DNA-binding proteins. In addition, nonspecific methylation often occurs, and this leads to false positive events.

All of the methods introduced above can be used to identify protein-DNA interactions, but they do not paint an accurate picture of chromatin structure at a given locus. Chromatin can form complex structures that often bring regions of DNA that are separated by tens of thousands of base pairs into close proximity of one another. For example, a transcription factor binding to an enhancer element can activate genes hundreds of kbs away or even on a different chromosome (Spilianakis et al. 2005); this is thought to be achieved through DNA looping. How can we identify and analyze long range interactions that depend on chromatin structure? The most successfully applied method for analyzing chromosome conformation is **Chromosome Conformation Capture** (3C) (Dekker et al. 2002). This method captures the interactions between chromosomal regions by crosslinking DNA fragments that are in close proximity, ligating them together, amplifying them, and then identifying them.

Cells are first treated with formaldehyde to induce covalent crosslinks between interacting chromosome regions. Cross-linked chromatin is then digested with a restriction enzyme, and the resulting DNA fragments are ligated under reaction conditions that favor intramolecular ligation (i.e. at low concentrations of crosslinked DNA). To analyze a genomic region of interest, sets of primers are designed to prime unidirectionally at the 5' end of the restriction sites. PCR is performed with different primer sets to detect the interactions frequencies between different chromatin segments. Including control samples taken from the cells that were not crosslinked is crucial to the success of this method.

The 3C method has been successfully used to study chromatin-looping events in yeast and higher eukaryotes (Dekker et al. 2002; Spilianakis and Flavell 2004; Spilianakis et al. 2005; Vakoc et al. 2005).

Applications of the methods described above, and of others not presented, have provided insight into complex transcriptional regulations and specific protein-DNA interactions. But to fully dissect transcriptional regulation networks, new high-throughput method for identification of protein-DNA interactions under various culture conditions or in multiple cells types *in vivo* is needed.

Focus of the Dissertation

The overall goal of my work is to develop a high-throughput *in vivo* method to identify target genes and recognition sequences of transcription factors, which allows testing under various culture conditions and in multiple cell types.

Chapter 2 describes the establishment of the "Calling Card" method based on microarray readout. Chapter 3 describes the development of an early version of "Calling Card-seq" method and applying it to study poorly characterized yeast TFs. Chapter 4 summarizes several major improvements of "Calling Card-seq" method, including using pair-end next generation sequencing technology as a readout and sample multiplexing. Chapter 5 describes the implementation of "Calling Cards" in mammalian cells. Chapter 6 summarizes the advantages and weakness of Calling Card methods and presents the future potential of the method.

Chapter 2: Calling Cards for DNA-binding Proteins

Haoyi Wang, Mark Johnston, and Robi David Mitra

Department of Genetics, Washington University, School of Medicine, 4444 Forest Park Parkway, St. Louis, MO 63108

Comparison of the genome sequences of related species reveals conserved sequence motifs that are assumed to be functional regulatory sequences, but experimental verification of their function has been hindered by the lack of in vivo method detecting transcription factors binding to the genome. I devised a method for identifying the genomic targets of DNA-binding proteins that exploits the ability of the Sir4 protein to recruit the Ty5 integrase. This chapter presents the proof of the principle of the “Calling Card” method.

This chapter is a reproduction of a manuscript published in *Genome Research* in 2007. This work was done in collaboration with Mark Johnston and Robi Mitra. Robi Mitra, Mark Johnston and I designed all the experiments and wrote the paper. I did all the experiments. Robi Mitra performed the microarray data analysis.

ABSTRACT

Identifying genomic targets of transcription factors is fundamental for understanding transcriptional regulatory networks. Current technology enables identification of all targets of a single transcription factor, but there is no realistic way to achieve the converse: identification of all proteins that bind to a promoter of interest. We have developed a method that promises to fill this void. It employs the yeast retrotransposon Ty5, whose integrase interacts with Sir4 protein. A DNA-binding protein fused to Sir4 directs insertion of Ty5 into the genome near where it binds; the Ty5 becomes a "Calling Card" the DNA-binding protein leaves behind in the genome. We constructed customized Calling Cards for seven transcription factors of yeast by including in each Ty5 a unique DNA sequence that serves as "molecular bar code". Ty5 transposition was induced in a population of yeast cells, each expressing a different transcription factor-Sir4 fusion and its matched, bar-coded Ty5, and the Calling Cards deposited into selected regions of the genome were identified, revealing the transcription factors that visited that region of the genome. In each region we analyzed we found calling cards for only the proteins known to bind there: in the *GAL1-10* promoter we found only calling cards for Gal4; in the *HIS4* promoter we found only Gcn4 calling cards; in the *PHO5* promoter we found only Pho4 and Pho2 calling cards. We discuss how Ty5 Calling Cards might be implemented for mapping all targets of all transcription factors in a single experiment.

INTRODUCTION

Transcription factors program gene expression by binding to specific sites in the genome and regulating chromatin-modifying enzymes and the transcriptional apparatus. Knowledge of the sites in the genome bound by each transcription factor is necessary for a full understanding of transcriptional regulation. Chromatin immunoprecipitation can be used to identify the sites in the genome to which any DNA-binding protein binds by using the DNA that co-precipitates with it to probe a microarray of DNA fragments that tile the genome (called the "ChIP-chip" method; (Horak and Snyder 2002; Ren et al. 2000). However, there is currently no realistic way to do the converse: identify all the proteins that bind to a particular region of the genome. To fill this gap in technology, we developed a new method for identifying protein-DNA interactions.

Our method exploits the retrovirus-like transposon Ty5 of bakers' yeast. After Ty5 mRNA is reverse transcribed and converted into a double-stranded cDNA, the Ty5 integrase carries it to the nucleus and catalyzes its insertion into the genome. Copies of Ty5 are found in the *S. cerevisiae* and *S. paradoxus* genomes near telomeres and the silent copies of the mating-type genes (Zou et al. 1996; Zou et al. 1995) because the Ty5 integrase interacts with Sir4, an integral component of the chromatin in these regions of the genome (Xie et al. 2001; Zhu et al. 2003). Fusion of Sir4 to a DNA-binding protein causes Ty5 to integrate into DNA near the binding sites for that protein (Zhu et al. 2003) (Fig. 2.1A). We have exploited this property of Ty5 to develop a method for identifying the proteins that bind to any selected region of the yeast genome. This method also provides a convenient alternative to the ChIP-chip technique for identifying the targets of any selected DNA-binding protein.

RESULTS

Principle of the method. When a DNA-binding protein fused to Sir4 binds to a site in the genome, it recruits the Ty5 integrase and thereby directs insertion of Ty5 into the genome. If the Ty5 carries a unique sequence "bar code", it becomes a "Calling Card" that uniquely identifies the TF that directed its insertion. If we provide each DNA-binding protein with a bar-coded Ty5 calling card and induce transposition in a mixture of such strains, each carrying a different TF-Sir4 fusion and its matched Ty5 calling card, we should be able to identify all the proteins that bind to a particular region of the genome by recovering the Ty5 elements that were deposited there and reading the bar code sequences they carry (Fig. 2.3).

Identification of targets of individual transcription factors. Before attempting to implement this method, we had to confirm that DNA-binding proteins reliably direct the insertion of Ty5 near their binding sites in the genome. We did this for Gal4, a DNA-binding protein with well-characterized targets in the genome. We fused the Gal4 DNA-binding domain (Gal4DBD) to a fragment of Sir4 (amino acid 951 to 1200) that contains its Ty5 integrase-interacting domain (Xie et al. 2001; Zhu et al. 2003). This Gal4DBD-Sir4 fusion protein was expressed in a yeast strain lacking *SIR4* and carrying a Ty5 element under the control of the *GAL1* promoter. Growth of this strain on galactose results in transcription of the Ty5 element, which is reverse transcribed into DNA that is competent to integrate into the genome (Zou et al. 1996). The Ty5 also carries a *HIS3* gene with an artificial intron that interrupts its coding sequence and which therefore becomes functional only after this artificial intron is spliced out of the mRNA, thereby providing a selection for cells in which the Ty5 has integrated into the genome (Curcio and Garfinkel 1991; Zou et al. 1996).

To identify the regions of the genome into which Gal4DBD-Sir4 directed Ty5 insertion, we recovered the DNA immediately flanking the Ty5 and determined its nucleotide sequence. Genomic DNA from each His^{+} FOA^{r} colony was cleaved with restriction enzyme Hinp1I, which cuts near the end of Ty5, and the resulting fragments were ligated in dilute solution to favor their recircularization. The sequence of the junction of the Ty5 and genomic DNA was determined after its amplification by inverse PCR (Ochman et al. 1988) (Fig. 2.1B). Among 96 independent transposition events in cells expressing Gal4DBD-Sir4, 76 occurred in promoters of known targets of Gal4: 39 upsteam of *GAL1-10*, 35 upstream of *GAL7*, one upstream of *GCY1* and one upstream of *FUR4*. Almost all of these insertions are within 35 bp of a Gal4 binding site ($\mathrm{CGGN_{11}CCG}$). The 15 genes not known to be bound by Gal4 into whose promoters we found Ty5 to transpose are not likely to be *bona fide* Gal4 targets because only one contains a Gal4 binding site, and their known or predicted functions do not make them good candidates for targets of Gal4. Five Ty5 transposition events occurred in the telomeres and into other Ty elements in the genome, as previously observed (Zhu et al. 1999). The strong enrichment of Ty5 integration events near known Gal4 binding sites validated the use of Ty5 to mark TF binding sites.

The relatively small number of transposition events analyzed in this initial experiment makes it difficult to determine if the transpositions within promoters not known to be regulated by Gal4 represent previously unrecognized Gal4 targets or are the background false positives of this method. To enable analysis of many more Ty5 insertions, we employed a more efficient method to identify their sites of insertion. Yeast cells representing about 5,000 Ty5 transposition events directed by Gal4-Sir4 were pooled and their genomic DNA was extracted and digested with 3 different restriction endonucleases with 4 base-pair recognition sequences that are present 300 to 1000 base-pairs from the end of Ty5. The resulting fragments (containing Ty5 sequence on one end and the adjacent genomic sequence on the other end) were ligated in dilute solution to favor their circularization and amplified by inverse PCR using primers complementary to the end of Ty5. The PCR products (of variable size) were labeled with Cy5 and used to probe a microarray of oligonucleotides that tile the yeast genome to identify regions of the genome flanking the Ty5 insertions (see Methods for details).

Seven regions known to be bound by Gal4 (*GAL1-GAL10*, *GAL7*, *GAL2*, *GAL3*, *FUR4*, *GCY1*, *PCL10*) (SCPD, http://rulai.cshl.edu/SCPD; TRANSFAC, http://www.gene-regulation.com/pub/databases.html#transfac; (Ren et al. 2000) are among the top 20 hybridization signals (see Methods for details of the analysis of the hybridization signals); two other known Gal4-regulated genes, *MTH1* and *GAL80* rank in the top 60 hybridization signals. (The data for all the genes that pass our significance criteria is provided in Supplemental Table 2.1).

Eight of the 13 promoters among the top 20 hybridization signals on the array that are not known to be Gal4 targets contain at least one Gal4-binding site (CGGN$_{11}$CCG) (Table 2.1). In an attempt to validate binding of Gal4 to these 13 promoters that are not known to be Gal4 targets, we immunoprecipitated Gal4 (*via* the myc epitope it carries) and tested for co-precipitation of those regions of the genome. Three of the 13 promoters (*SFL1-RUP1, YPL067C-YPL066W, YOR084W*) were clearly enriched in the sample immunoprecipitated from cells with the myc-tagged Gal4 compared to cells with an untagged Gal4 (Fig. 2.2A). Indeed, Gal4 regulates expression of these genes (Fig. 2.2B). Expression of the divergently transcribed genes flanking two of these promoters (*SFL1* -- *RUP1,* and *YPL067C* --*YPL066W)* is induced by galactose *via* Gal4 (Fig. 2.2B, compare lane 3 to lane 1 and lane 4 to lane 2); interestingly, Gal4 regulates expression of *YOR084W* in an unexpected way: it seems to repress its expression (compare lane 4 to lane 3, and lane 6 to lane 5). Although 10 of the 13 potential Gal4 targets were not confirmed by the chromatin immunoprecipitation experiments, five of them have Gal4-binding sites and therefore could be Gal4 targets.

To estimate the sensitivity and specificity of the method we turned to Gcn4, because it has a well-characterized DNA binding specificity (Oliphant et al. 1989), many known targets in the genome (Natarajan et al. 2001; Pokholok et al. 2005), and many genes are known that are unlikely to be its target (Pokholok et al. 2005). In addition, Gcn4 was used to determine the sensitivity and specificity of the ChIP-chip method (Pokholok et al. 2005), enabling a direct comparison of the two methods. Using the same approach as for Gal4, we determined where in the genome Gcn4-Sir4 deposits Ty5. About 300 regions of the genome displayed significant hybridization to the array (see Methods for the criteria for significance). Twelve known Gcn4 targets are among the top 20 signals; the remaining eight all have perfect or recognizable Gcn4-binding sites (several of these genes are especially propitious Gcn4 targets because they encode enzymes involved in amino acid biosynthesis) (Table 2.1).

To estimate the specificity and sensitivity of this assay, we determined how many known Gcn4 target genes (defined by Pokholok et al. 2005) were not identified by our method ("false negatives") and how many regions of the genome that are unlikely to be Gcn4 targets (also as defined by Pokholok et al. 2005) turned up in our assay ("false positives"). Fifty-one percent of the known or likely Gcn4 target genes hybridized strongly enough to probes on the DNA microarray to pass our criteria for a positive signal. This false negative frequency of 49% comes at a false positive frequency of 2.5%. This is somewhat higher than the 25% false negative frequency of the ChIP-chip method (at a false positive frequency of 1%), which is perhaps not surprising since the reference sets of Gcn4 target genes chosen by Pokholok et al. (2005) are partially based on results from ChIP-chip experiments. It should be noted, however, that this false positive rate means that a substantial proportion of our 300 potential Gcn4 targets are likely to be false positives (2.5% of 6000 = ~150 false positives). Some of these are derived from recombination of the Ty5 calling card with Ty5 elements and LTRs, and can be easily recognized by their location (usually near the telomeres) in the genome. (The data for all the genes that pass our significance criteria is provided in Supplemental Table 2.2).

Identification of the proteins that bind to any selected region of the genome. With the confidence these results provided that DNA-binding proteins carrying Sir4 direct insertion of Ty5 into the genome near where they bind, we proceeded to test if the calling cards can be used to reveal which proteins bind to a particular region of the genome. We manufactured Ty5 calling cards containing 20 base-pair oligonucleotides that serve as "molecular bar codes" for seven transcriptional regulators fused to Sir4: Gal4, Gal80, Ste12, Bas1, Pho2, Gcn4, and Pho4. Yeast cells were co-transformed with a plasmid encoding a TF-Sir4 fusion and a plasmid carrying its matched Ty5 calling card (Fig. 2.3). These seven strains were pooled and Ty5 transposition was induced by growing them on galactose-containing medium. We recovered the calling cards deposited in three different promoters by performing PCR with oligonucleotide primers complementary to Ty5 and to the regions flanking the promoters of interest (Fig. 2.3, see Methods for details). The identity of the "bar codes" in these PCR products was determined by using them to probe a mini-array of the bar code sequences (Fig. 2.3, see Methods for details). In each of the three promoters we analyzed we found calling cards for only those proteins known to bind to them (Fig. 2.4): in the *GAL1-10* promoter we only found Ty5 elements carrying the Gal4 bar code (Fig. 2.4A); in the *HIS4* promoter, we found only Gcn4 bar codes (Fig. 2.4B) (Tice-Baldwin et al. 1989). In the *PHO5* promoter we found only bar codes corresponding to Pho4 and Pho2, and only when transposition was induced in cells starved for phosphate (Fig. 2.4C,D), as expected because Pho4 and Pho2 bind to *DNA* only when phosphate is scarce (Barbaric et al. 1996; Oshima et al. 1996). This pilot experiment suggests that Ty5 can be used to identify proteins that bind to any region of the genome.

DISCUSSION

We have exploited the properties of the Ty5 transposon to provide DNA-binding proteins with "calling cards" that reveal the places in the genome they visit. We validated this method with 7 different DNA-binding proteins, and found that we could successfully identify the proteins that bind to different promoters. The method proved to be robust: it identified the proteins known to bind to the *GAL1-10*, *HIS4*, and *PHO5* promoters. Based on these results we are confident we can implement calling cards for all ~200 DNA-binding proteins of yeast, which would enable identification of all the proteins that bind to any particular region of the genome under a variety of growth conditions by a simple PCR followed by hybridization to a microarray of oligonucleotide bar codes. We are confident calling cards can also be implemented for non-DNA-binding, chromatin associated proteins, because we used calling cards to identify a known target of Mth1, which is recruited to promoters of *HXT* genes by the Rgt1 transcriptional repressor (data not shown).

This method fills a gap in technology for characterizing DNA-binding proteins. Currently we can identify the targets of any particular DNA-binding protein with the ChIP-chip technique, but to do the converse—identify the proteins that bind to a particular region of the genome—one would have to perform a ChIP-chip experiment on all DNA-binding proteins of an organism. Our calling card method promises to make this feasible.

Our method also provides an alternative to the ChIP-chip method for the genome-wide identification of targets of transcription factors, and can serve as an independent verification of the results obtained with the ChIP-chip method. Indeed, we were able to discover previously unidentified targets of Gal4, probably the best characterized transcription factor of yeast, perhaps because our method is very different from those that employ chromatin immunoprecipitation.

The calling card technology could be improved in several ways. Probably most important is to increase the number of transposition events sampled. For practical reasons we have been harvesting 3,000 to 5,000 independent transposition events in each experiment, but it should not be difficult to scale up the experiment and obtain more. This may be necessary because we did not find in the *HIS4* promoter bar codes for Pho2 and Bas1, which are known to bind there (Tice-Baldwin et al. 1989). We identified two Pho2 bar codes among 18 that we analyzed by direct DNA sequencing in a preliminary experiment, suggesting that binding of these proteins would have been detected by hybridization to the microarray with a larger number of Ty5 transposition events. The number of transposition events could also be increased by improving the Ty5 transposition efficiency, which is relatively low compared to other Ty elements. This could also allow a shorter time of induction of transposition. Second, expression of the Ty5 calling card from the *GAL1* promoter limits the conditions that can be tested. It would be better to use a different promoter, such as one that is activated by a gratuitous inducer like tetracycline (Belli et al. 1998; Berens and Hillen 2003). Third, it has been speculated that the region of Sir4 that interacts with the Ty5 integrase also interacts with other proteins, which might interfere with the method in some cases. A clever solution to this potential problem—use of a heterologous pair of protein interaction domains on the DNA-binding protein and the integrase—was implemented by Zhu et al. (2003). That would also allow the method to be applied with a *SIR4* strain, which would avoid the possibility of disruption of chromatin structure in certain regions of the genome. Fourth, fusing Sir4 to a DNA-binding protein could interfere with its ability to bind to DNA. This problem can be minimized by fusing Sir4 to each end of the protein (in different constructs). Finally, insertion of a calling card into a promoter could, in some cases, disrupt expression of the gene, which might prevent recovery of those cells. This problem can easily be solved by using a diploid strain.

We would like to reduce the false positive and the false negative rates of our method. We empirically determined the significance cutoff using lists of genes that are likely or unlikely to be Gcn4 targets, as was done for the ChIP-chip method (Pokholok et al. 2005). This cutoff was applied to all experiments. We arbitrarily chose a significance cutoff that yielded 2.5% false positives, which results in a 49% false negative rate. Similar performance (4% false positives and ~24% false negatives) was sufficient for application of the ChIP-chip method to genome-wide analysis of transcription factor targets in yeast (Harbison et al. 2004). Of course, the false positive rate can be reduced by increasing the cutoff, but that comes at the expense of a higher false negative rate. Advances in the experimental approach are likely to be necessary for significant improvement in the specificity and sensitivity of our method (Gabriel et al. 2006; Wheelan et al. 2006). One reason for this high false positive rate might be the large number of cycles of the inverse PCR required to provide enough probe for hybridization to the DNA microarrays, which may result in over amplification of some of the non-specific insertions. Stochastic amplification of non-specific insertions in the inverse PCR ("jackpotting") could also contribute to the problem. Both problems should be ameliorated by performing the inverse PCR on individual molecules in a water/oil emulsion (Griffiths and Tawfik 2006). In addition, the low transposition efficiency of Ty5 in our experiments may exacerbate the "jackpotting" problem, so the false positive rate will likely be improved if we can sample more transposition events.

By coupling the calling card method to next-generation (massively parallel) sequencing technologies, it may be possible to identify genome-wide the binding locations of all yeast transcription factors in a single experiment. Induction of transposition of the calling cards in a library of strains representing all ~200 DNA-binding protein-Sir4 fusions with their corresponding calling cards, followed by recovery of each calling card along with the adjacent genomic DNA would enable determination of the sequences of *both* the bar-code identifiers of the DNA-binding proteins *and* the adjacent genomic sequence, thereby revealing both *where* in the genome proteins bind and *which* proteins bind there. This would be equivalent to performing a ChIP-chip experiment for each of the 200 DNA-binding proteins. Several novel DNA sequencing methods have recently been developed that offer the throughput needed for this implementation of the calling card method (Margulies et al. 2005; Shendure et al. 2005). This would enable us to examine the regulatory network of yeast under a large number of different conditions. Finally, we note that transposons are present throughout the tree of life, so it may be possible to implement calling cards for DNA-binding proteins in species other than yeast.

METHODS

Strains and growth media

The *sir4* deletion mutant yDV561 (*MATa*, *ura3*-52, *trp1*-63, *his3*-200, *leu2*-1, *lys2*-801, *ade2*-101, *sir4::KanMX*) obtained from Dan Voytas (Zhu et al. 2003) was the host strain for Ty5 transpostion. Chromatin immunoprecipitation was done from extracts of strain Z1319 (*MATa*, *ade2*-1, *trp1*-1, *can1*-100, *leu2*-3,112, *his3*-11,15, *ura3*, *GAL*$^+$, *psi*$^+$, *GAL4*::18-Myc), (Ren et al. 2000).Yeast strain BY4743 (*MATa*/*MATα* *his3*Δ1/*his3*Δ1 *leu2*Δ0/*leu2*Δ0 *ura3*Δ0/*ura3*Δ0 *met15*Δ0/*MET15* *LYS*2/*lys2*Δ0) and homozygous *gal4* deletion strain (Saccharomyces Genome Deletion Project, #31044)

(*MAT***a***/MAT***α** *his3*Δ1/*his3*Δ1 *leu2*Δ0/*leu2*Δ0 *ura3*Δ0/*ura3*Δ0 *met15*Δ0/*MET15* *LYS*2/*lys2*Δ0 *gal4*Δ0 /*gal4*Δ0) (Brachmann et al. 1998; Giaever et al. 2002) were used for reverse transcription PCR to measure gene expression. Yeast cells were grown in complete synthetic media with the addition of 2% glucose or galactose, unless described otherwise.

Construction of plasmids

To construct pBM5037 (Gal4DBD-Sir4-Myc), the region of *SIR4* encoding amino acids 951 to 1200 was amplified in a PCR and fused to the Gal4 DNA binding domain (amino acid 1 to147 plus amino acid 877 to 881) in pOBD2 by "gap repair" (Ma et al. 1987; Wach et al. 1994). Three copies of the Myc epitope were amplified using PCR and fused to the C-terminus of Gal4DBD by gap repair. The entire ORF of each transcription factor was amplified in a PCR and used to replaced Gal4DBD by homologous recombination. Gal4DBD-Sir4-Myc was linearized by cutting with XhoI (cuts once C-terminal to Gal4DBD coding sequence) to serve as the recipient plasmid for gap repair to construct all the other TF-Sir4 fusions.

The plasmid pSZ293 with Ty5 expressed from the *GAL1* promoter was obtained from Dan Voytas (Zhu et al. 2003). The XhoI-NotI fragment that includes *GAL1*::Ty5 was inserted between the XhoI and NotI sites of pRS316 (Sikorski and Hieter 1989) to generate pBM4735. AcaI and FseI sites were engineered adjacent to the 3'long terminal repeat (LTR) to allow insertion of the 20bp "bar codes". The bar codes that identify each transcription factor were those developed for each gene in the Yeast Gene Knockout

(YKO) collection (Yuan et al. 2005). Double-stranded oligonucleotides with the bar code sequences were inserted between the engineered AcaI and FseI sites of the Ty5.

Induction of Ty5 transposition and inverse PCR

Since Ty5 is driven by the *GAL1* promoter, transposition was induced by culturing cells in galactose medium for two to three days at room temperature. After induction, cells were plated on Glu –His media to select for cells with transposition events. His^+ cells were replica plated on –His, FOA-containing media to eliminate His^+ colonies due to recombination of reverse-transcribed Ty5 with the transposon donor plasmid.

To map sites of Ty5 integration directed by Gal4-Sir4, 96 His^+ FOA^r colonies were grown in YPD and their genomic DNA extracted, digested by Hinp1I (1μg in a 20μl reaction). 5μl of digested DNA was then ligated overnight at 15°C in 100μl to encourage self-circularization. 5μl of the ligated DNA was used as template for inverse PCR with primers that anneal to Ty5 sequences (OM6313 and OM6188 were used to amplify the genomic region on the right side of Ty5 integration; OM6458 and OM4960 were used to amplify the genomic region on the left side).

For hybridization of the inverse PCR products to the yeast genome tiling array, we pooled 3,000 to 5,000 His^+ FOA^r colonies for each sample, extracted the total DNA, digested it with three different enzymes (Hinp1I, HpaII, and Taq1), and ligated them in dilute solution. Using two pairs of primers, the genomic region on the left side (primers OM6609 and OM6458) and right side (primers OM6610 and OM6456) of Ty5 was amplified from each enzyme digested sample. The PCR products were purified (using the

Qiagen PCR purification kit), and the same amount of product from digestion with each restriction endonuclease were pooled. 1.6 μg of PCR products were labeled with Cy5 using Invitrogen's BioPrime Array CGH Genomic Labeling Module (cat# 18095-011), and the genomic DNA, sonicated into 0.5 to 1 kb fragments, was labeled with Cy3. The Cy3 and Cy5 labeled samples were combined and hybridized to an Agilent yeast whole genome tilling array.

For the experiments employing "bar-coded" Ty5 elements, we cultured in glucose medium lacking uracil and tryptophan seven strains, each carrying a different TF-Sir4 fusion and its matched bar-coded Ty5 element. Once the OD_{600} of each culture reach approximately one, 100μl of cells of each strain were pooled and Ty5 transposition was induced and selected for as described above. Genomic DNA was extracted from about 3,000 His^+ FOA^r colonies and used as the template in a PCR with promoter-specific primers. To amplify all the calling cards deposited within a particular promoter, we used a primer that anneals to sequences flanking the promoter and a primer (OM6606) that anneals to sequences within Ty5. 600ng PCR products for each promoter were purified, labeled with Cy5, and hybridized to a mini-array of bar code oligonucleotides, using Genisphere's Array 900DNA Cy3 and Cy5 labeling kits (cat # RDNA130 and RDNA140). Probes on the mini-array are 60bp oligonucleotides consisting of three copies of the 20bp bar code sequence. Each probe was printed in quadruplicate on the array. In addition oligonucleotides of the LTR sequence were printed to serve as a positive control; three unrelated bar code oligonucleotides served as negative control.

Primers for PCR

OM6313: TAAGCTCGGAATTCGAGCTC

OM6188: ACAAGGAAAACATAGAGCAGC

OM6458: AGGTTATGAGCCCTGAGAG

OM4960: CGTAGTGAATTACGATCTAGC

OM6609: CTTTTGGGTTATCACATTCAAC

OM6610: ATCGTAATTCACTACGTCAAC

OM6456: CCCATAACTGAATACGCATG

OM6606: AAGATCGAGTGCTCTATCGC

DNA sequencing

The ABI Prism BigDye Terminator Cycle Sequencing Ready Reaction kit was used for DNA sequencing. 100ng of PCR product or 1μg of plasmid DNA was used as the template. The products of the reaction were separated and detected on an ABI 310 genetic analyzer.

Microarray analysis

We used two methods to identify the regions of the genome where calling cards were deposited due to the binding of the TF-Sir4 fusion protein. Each method requires a different type of hybridization control.

The Rosetta Error Model

We used the Rosetta error model to analyze the transcription factors Gal4 and Gcn4. In these experiments, our control was a *sir4D* strain containing a plasmid expressing Ty5 (pBM4735), but with no plasmid expressing a TF–Sir4 fusion. We induced transposition and performed inverse PCR as described above. We labeled the control reaction with the Cy3 (green) dye, the experimental reaction with the Cy5 (red) dye, pooled the reactions, hybridized them to the microarray, and imaged the slide. For each probe, we subtracted the intensity value observed in the control channel from the intensity value observed in the experimental channel. We then assigned each probe a p-value that gives the probability of the observed intensity difference, assuming no calling card was deposited at that location. As did Pokholok et al. (2005), we used the Rosetta error model, to calculate this p-value. In this model, the difference in intensities between two technical replicates is assumed to be normally distributed, and the variance of this distribution increases with average probe intensity.

We chose our significance cutoff empirically by using the published test sets of positive and negative targets for Gcn4 (Pokholok et al. 2005). We selected a p-value threshold which minimized the rate of false negatives at a false positive rate of 2.5%. This cutoff resulted in a false negative rate of 55%. If a gene is within 250bp of a significant probe then it is considered a target of the transcription factor that is being analyzed.

The Maximum Likelihood Estimate of DNA Concentration (MLEDC) Method

The Rosetta error model works well when the distribution of intensities in the control channel is similar to the distribution of background intensities in the experimental channel. However, we observed a significant increase in integration “hot spots” when no TF-Sir4 fusion protein is present, rendering the Rosetta error model inadequate. We therefore developed a second way to analyze the calling card data. Using labeled genomic DNA as a control, we estimated the concentration of DNA present at each locus after recovery of calling cards and flanking genomic DNA. The maximum likelihood value of DNA concentration is proportional to the average ratio of experimental to control intensities. We ranked the probes based on their average ratio and empirically selected a cutoff as described above. We selected a threshold which minimized the rate of false negatives at a false positive rate of 2.5%. This cutoff resulted in a false negative rate of 49%. Since this is slightly better than the Rosetta error model, the data were analyzed using the MLEDC method.

To understand better the nature of our false negatives, we manually examined the intensities of these genes in the MLEDC analysis – the majority of these features displayed little to no fluorescence in the red channel, suggesting that these features were categorized as negatives because no transposition event had occurred in these samples, and not due to inaccurate assumptions in our error model. Data from probes covering telomere regions were ignored (because Ty5 can insert into these regions of the genome due to homologous recombination with Ty5 elements that reside there. *HIS3* probes were also excluded because *HIS3* sequences from the Ty5 calling cards are present in the inverse PCR product.

For the bar-code array experiments, the raw intensity of each probe on the array was normalized by dividing it by the raw intensity of a probe containing LTR sequence. To eliminate the random hopping background, we applied a stringent criteria: if the probe gets a ratio over 0.1 only in one experiment out of three biological replicates, we count it as a random event and exclude it from the data.

Chromatin IP

Chromatin immunoprecipitation was performed as previously described (Orlando 2000). Cultures were grown in minimal medium with galactose. Bound proteins were crosslinked to DNA in vivo by addition of formaldehyde, followed by cell lysis and sonication to shear DNA. Individual transcription factors were immunoprecipitated with antibody to their Myc epitope tag, followed by reversal of the crosslinks. DNA immunoprecipitated from a Myc-tagged strain and from a control strain with no Myc tag were used as template to amplify the promoter of interest.

Reverse transcription PCR

Wild type and Gal4 deletion strains were cultured in 50ml YP medium with 2% glucose, 2% galactose plus 5% glycerol, or 5% glycerol as carbon source. When the cultures reached an OD_{600} of 1.5, the cells were harvested and their RNA extracted. The same amount of RNA from each sample was treated with DNAse and then reverse transcribed into cDNA using SuperScript™ II Reverse Transcriptase from Invitrogen. The cDNA served as the template in a PCR employing primers that amplify 200-300bp of coding sequence of the genes of interest. 25 cycles were used for each PCR. Primers amplifying *ACT1* was used as the loading control for each sample.

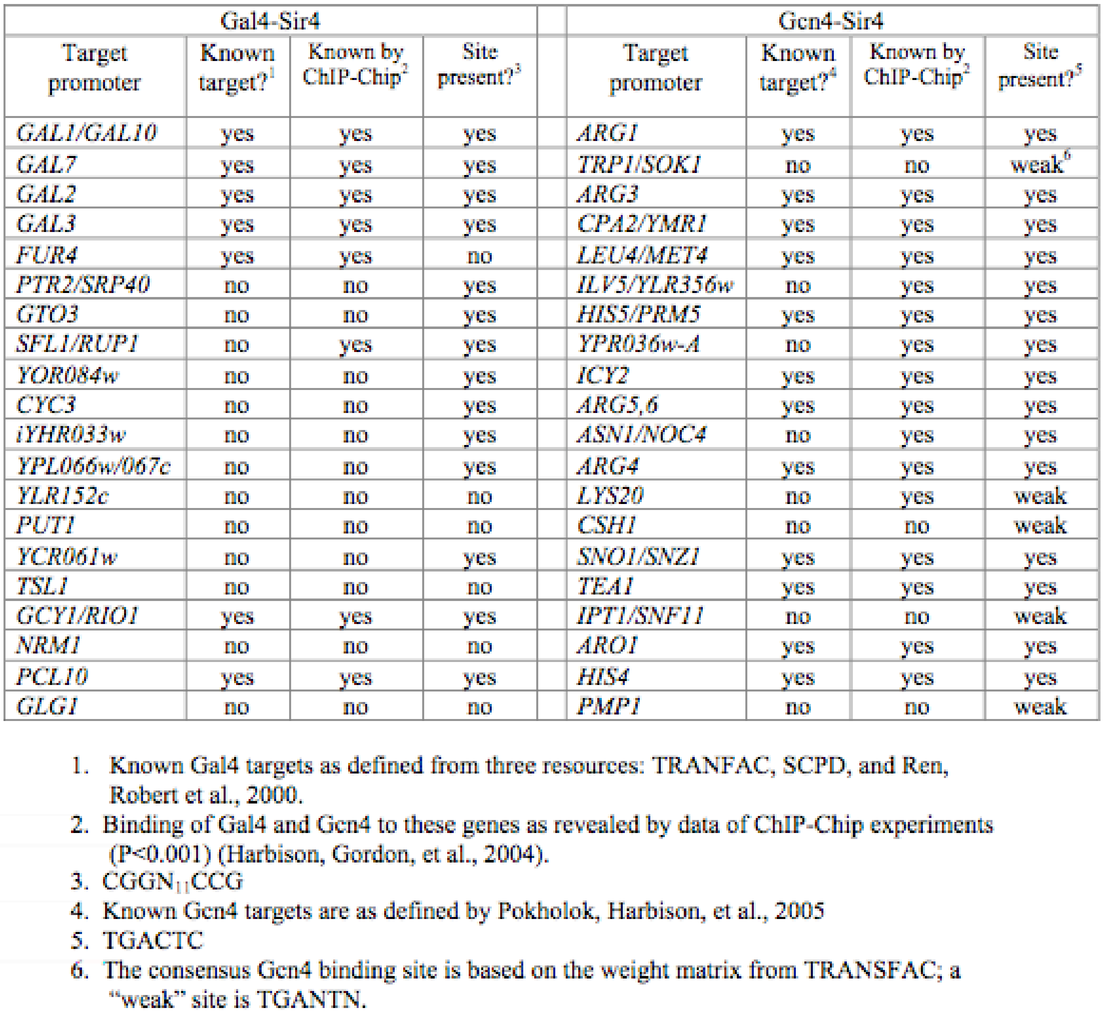
ACKNOWLEDGEMENTS

We are especially grateful to Dan Voytas (Iowa State University) for generously providing reagents and advice, and to Seth Crosby and Michael Heinz (Washington

University) for their expert assistance with the use of DNA microarrays. We also thank Doug Chalker (Washington University) for helpful suggestions, and Rick Young and Nancy Hannett (Whitehead Institute, MIT) for providing strains expressing myc-tagged Gal4 and Gcn4. This work was supported by funds provided by the James S. McDonnell Foundation.

TABLES

Table 2.1. Top 20 targets of Gal4 and Gcn4



Gal4-Sir4					Gcn4-Sir4			
Target promoter	Known target?[1]	Known by ChIP-Chip[2]	Site present?[3]		Target promoter	Known target?[4]	Known by ChIP-Chip[2]	Site present?[5]
GAL1/GAL10	yes	yes	yes		*ARG1*	yes	yes	yes
GAL7	yes	yes	yes		*TRP1/SOK1*	no	no	weak[6]
GAL2	yes	yes	yes		*ARG3*	yes	yes	yes
GAL3	yes	yes	yes		*CPA2/YMR1*	yes	yes	yes
FUR4	yes	yes	no		*LEU4/MET4*	yes	yes	yes
PTR2/SRP40	no	no	yes		*ILV5/YLR356w*	no	yes	yes
GTO3	no	no	yes		*HIS5/PRM5*	yes	yes	yes
SFL1/RUP1	no	yes	yes		*YPR036w-A*	no	yes	yes
YOR084w	no	no	yes		*ICY2*	yes	yes	yes
CYC3	no	no	yes		*ARG5,6*	yes	yes	yes
iYHR033w	no	no	yes		*ASN1/NOC4*	no	yes	yes
YPL066w/067c	no	no	yes		*ARG4*	yes	yes	yes
YLR152c	no	no	no		*LYS20*	no	yes	weak
PUT1	no	no	no		*CSH1*	no	no	weak
YCR061w	no	no	yes		*SNO1/SNZ1*	yes	yes	yes
TSL1	no	no	no		*TEA1*	yes	yes	yes
GCY1/RIO1	yes	yes	yes		*IPT1/SNF11*	no	no	weak
NRM1	no	no	no		*ARO1*	yes	yes	yes
PCL10	yes	yes	yes		*HIS4*	yes	yes	yes
GLG1	no	no	no		*PMP1*	no	no	weak

1. Known Gal4 targets as defined from three resources: TRANFAC, SCPD, and Ren, Robert et al., 2000.
2. Binding of Gal4 and Gcn4 to these genes as revealed by data of ChIP-Chip experiments ($P<0.001$) (Harbison, Gordon, et al., 2004).
3. CGGN_{11}CCG
4. Known Gcn4 targets are as defined by Pokholok, Harbison, et al., 2005
5. TGACTC
6. The consensus Gcn4 binding site is based on the weight matrix from TRANSFAC; a "weak" site is TGANTN.

FIGURES

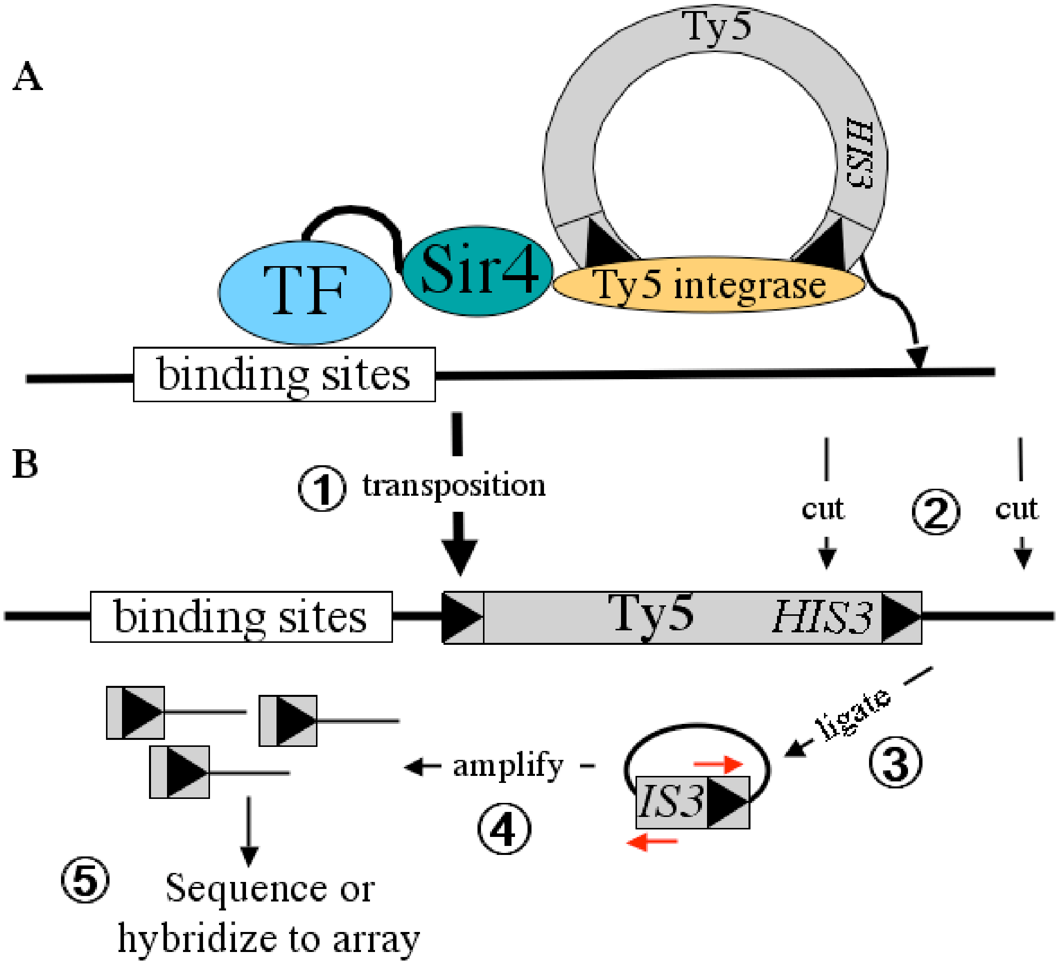


Figure 2.1. Identification of genomic targets of DNA-binding proteins using Ty5. (A) Sir4 fused to a DNA-binding protein causes Ty5 to integrate into the genome near the binding sites for that transcription factor (TF). (B) After ① Ty5 transposition, genomic DNA is ② cleaved with a restriction enzyme that cuts near the end of Ty5 and ③ ligated in dilute solution to favor recircularization of the fragments. This is ④ followed by amplification of the circular DNA that contains the end of the transposon and flanking genomic DNA by an "inverse PCR" (PCR primers labeled in red) and ⑤ the identity of the flanking genomic DNA is determined by DNA sequencing or hybridization to a DNA microarray.

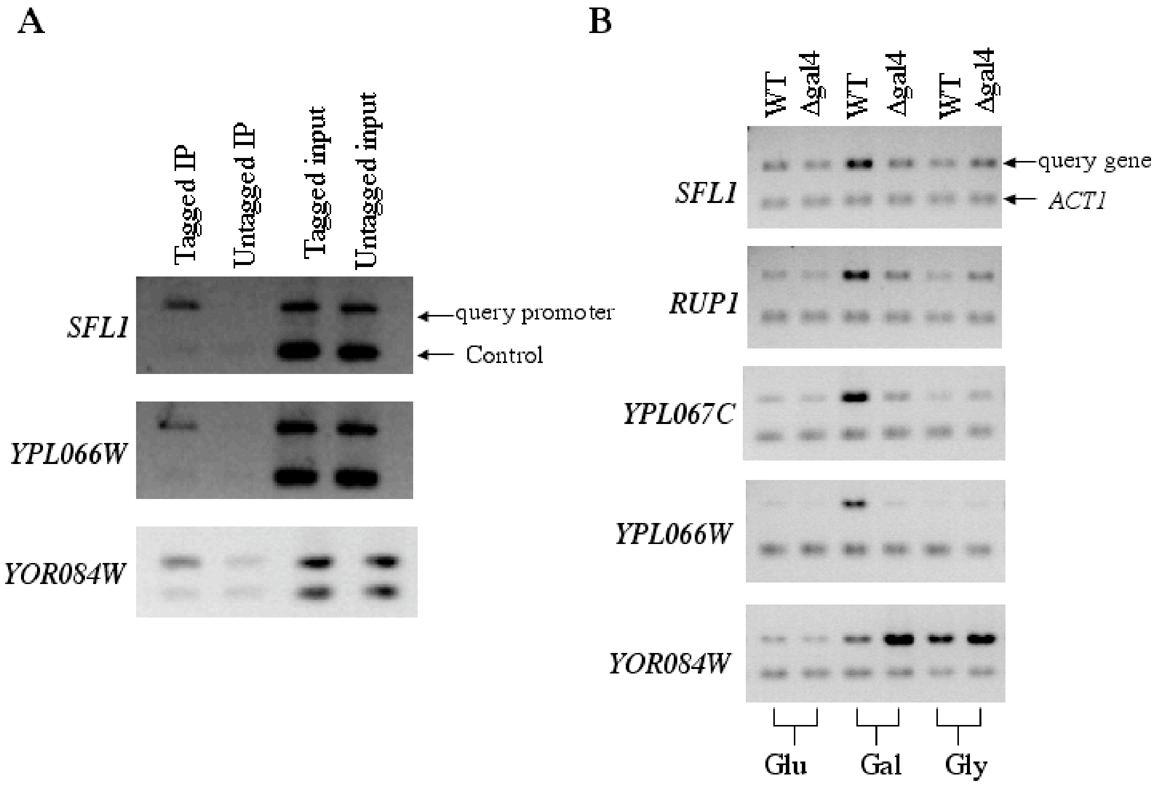


Figure 2.2. Verification of novel Gal4 targets. (A) ChIP assay for Gal4 binding. Chromatin was crosslinked to protein by treatment with formaldehyde, and Gal4 tagged with the 18-myc epitope, which was precipitated with anti-myc antibody. The precipitated DNA was released from protein and detected by PCR as described in Methods, using primers specific for sequences upstream of the indicated putative Gal4 targets (query promoter) and primers specific for the *GAL4* promoter (control promoter) that amplify a 150-bp fragment. (B) RT-PCR analysis compared the expression of novel Gal4 target genes in wild-type FM393 cells versus *gal4Δ* cells grown on different carbon sources. Cells were grown to saturation in YPD and then diluted 100 times in fresh 2% glucose, 2% galactose plus 5% glycerol, or 5% glycerol. Cells were harvest once they reach mid-log phase (OD_{600} = 1.5 to 2.0), total RNA was prepared, and RT-PCR was performed on the indicated targets. Control reactions lacking reverse transcriptase produce no PCR products (data not shown).

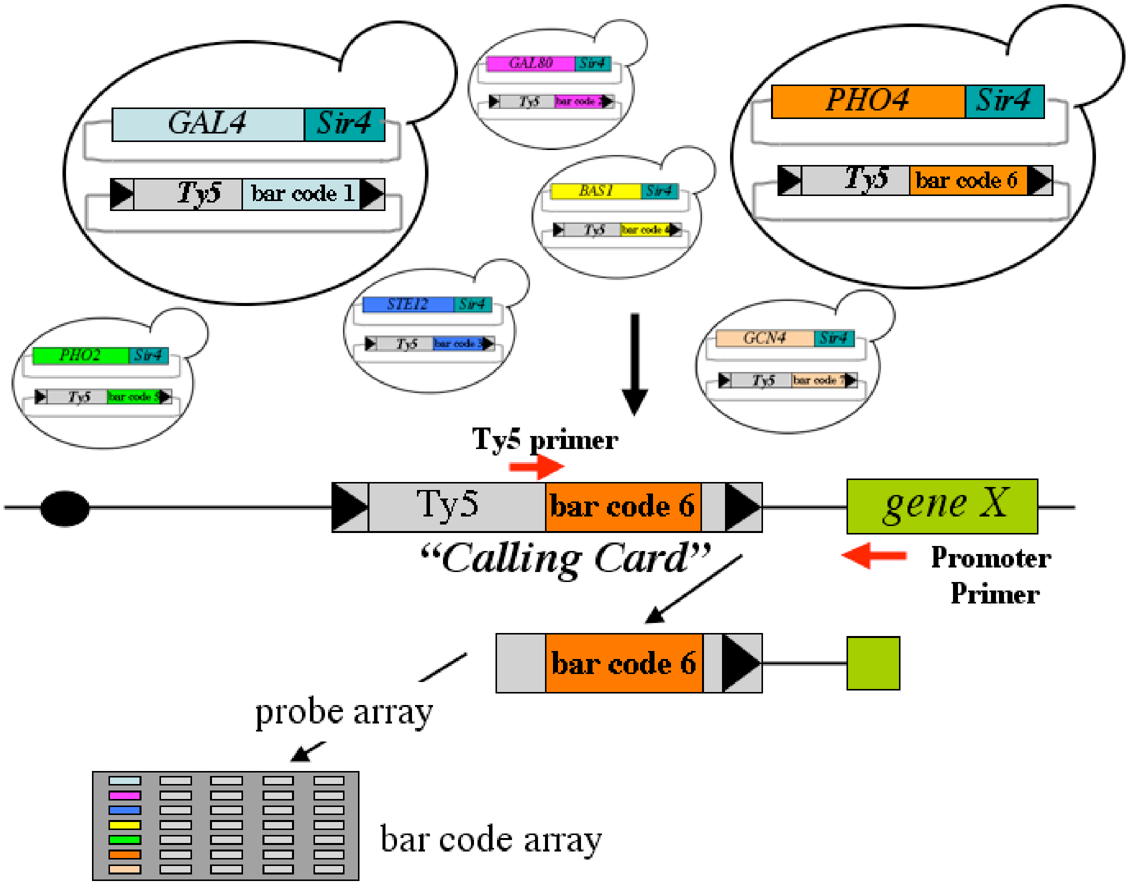


Figure 2.3. "Calling cards" for DNA-binding proteins.

For each of seven transcription factors fused to Sir4 (Gal4, Gal80, Ste12, Bas1, Pho2, Gcn4, and Pho4), a unique 20 base-pair oligonucleotide was inserted into Ty5 to serve as a "molecular bar code", thereby transforming Ty5 into a "calling card" that the TF leaves behind when it visits a site in the genome. Each strain was co-transformed with a plasmid encoding a TF-Sir4 fusion and a plasmid carrying its matched Ty5 calling card. After transposition, the calling cards deposited in the promoters of interest were recovered by a PCR with Ty5 and promoter specific primers.

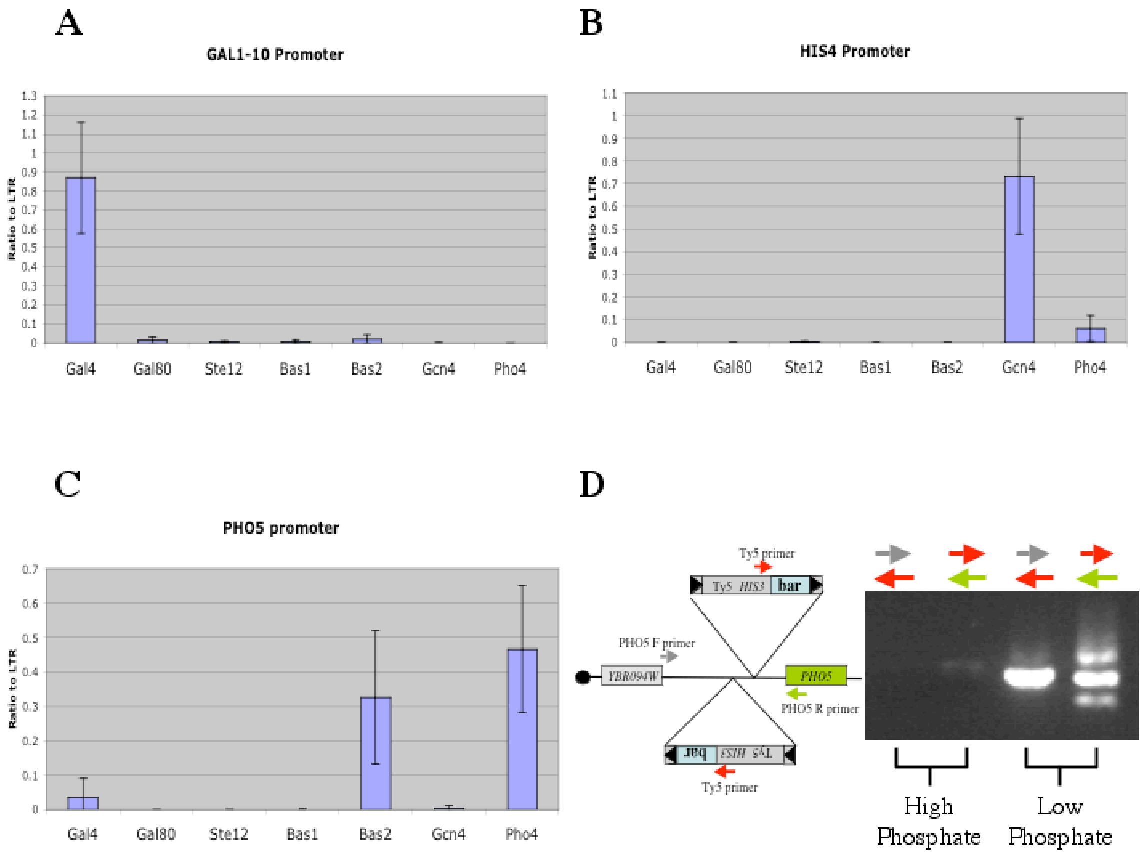


Figure 2.4. "Calling Cards" deposited in three promoters. The PCR products from three promoters were hybridized to the bar code array. Shown is the ratio of the intensity of hybridization of each bar code to the intensity of hybridization to an LTR probe on the array. (A) In the *GAL1-10* promoter, only the Gal4 bar code is enriched. (B) In the *HIS4* promoter, only the Gcn4 bar code is enriched. (C) In the *PHO5* promoter, only Pho2 and Pho4 bar codes are enriched. (D) When transposition was induced in media rich in phosphate (YPD), the *PHO5* specific primers produced no PCR product, but when transposition was induced in cells grown in low phosphate media the *PHO5* specific primers produced abundant PCR products, which contain only Pho4 and Pho2 bar codes, as revealed by hybridization to the bar code array.

SUPPLEMENTAL FIGURES AND TABLES

Supplemental Table 2.1. Positive Targets of Gal4. Columns 1 and 2: All Gal4 targets above our significance cutoff are listed. Genes flanking a divergent promoter are listed in the same row. Columns 3: The ratio of hybridization intensity of the Ty5 inverse PCR product to the hybridization intensity of the genomic control (Red vs Green). Genes flanking a divergent promoter are listed in the same row. Columns 4: The Log10 ratio of expression of each gene in gal4D vs. GAL4 in cells grown on galactose from the data in (Ideker et al. 2001). Among all 115 promoters we identified, 47 drive genes that show expression change over two fold.

1. Systematic Name	2. Standard Name	3. Ratio (Red/Green)		4. EXPRESSION RATIOS: gal4D+gal vs. reference (wt+gal)	
YBR019C/YBR020W	GAL10/GAL1	1659.87629	2815.39682	-1.917	-1.875
YBR018C	GAL7	1218.68558		-1.97	
YLR081W	GAL2	198.916056		-0.59	
YDR009W	GAL3	55.694058		-1.01	
YBR021W	FUR4	49.430548		-0.704	
YKR092C/YKR093W	SRP40/PTR2	35.585561	35.585561	0.009	0.06
YMR251W	GTO3	29.079802		-0.093	
YOR140W/YOR138C	SFL1/RUP1	27.626516	11.157659	-0.058	-0.284
YOR084W		26.433427		0.374	
YAL039C	CYC3	26.325149		0.311	
YPL067C/YPL066W	YPL067C/YPL066W	21.0688	11.442911	-0.689	0.06
YLR152C		20.899318		0.328	
YLR142W	PUT1	20.526141		0.225	
YCR061W		19.676886		0.193	
YML100W	TSL1	18.9693		-0.327	
YOR119C/YOR120W	RIO1/GCY1	15.42399	15.42399	-0.227	-0.402
YNR009W	NRM1	14.379649		-0.258	
YGL134W	PCL10	13.418209		-0.586	
YKR058W	GLG1	13.310884		0.002	
YEL017C-A/YEL017W	PMP2/GTT3	12.701274	12.701274	-0.156	0.239
YMR037C	MSN2	11.81432		-0.08	
YMR083W	ADH3	11.79541		0.244	
YER035W	EDC2	11.648407		0.138	
YER153C/YER154W	PET122/OXA1	11.584539	11.584539	0.035	-0.131

YBR043C	QDR3	11.515219		0.031	
YOL110W/YOL111C	SHR5/MDY2	11.504362	11.504362	0.213	0.037
YPL262W/YPL263C	FUM1/KEL3	11.487273	11.487273	0.168	0.29
YER130C/YER131W	YER130C/RPS26B	11.40851	11.40851	-0.955	-0.36
YDR270W	CCC2	11.018255		0.267	
YBR112C/YBR114W	CYC8/RAD16	10.988151	10.988151	-0.029	-0.626
YKL085W	MDH1	10.38967		-0.099	
YMR318C	ADH6	10.371853		-0.745	
YBR017C	KAP104	10.343763		-0.191	
YIL056W/YIL057C	VHR1/YIL057C	10.263454	7.583475	0.279	0.689
YNL160W	YGP1	10.194146		0.198	
YPL265W	DIP5	10.127431		0.44	
YAL038W	CDC19	9.962189		-0.001	
YAL039C	CYC3	9.962189		0.311	
YGR250C/YGR251W	YGR250C/YGR251W	9.952245	9.952245	-0.105	-0.482
YDR345C	HXT3	9.902566		0.966	
YJL047C-A		9.712614		#N/A	
YJL048C	UBX6	9.712614		0.153	
YDR284C/YDR285W	DPP1/ZIP1	9.649109	9.649109	0.158	0
YNL073W/YNL074C	MSK1/MLF3	9.579458	9.579458	-0.092	-0.272
YOR348C/YOR349W	PUT4/CIN1	9.335449	9.335449	0.867	0.551
YMR251W-A	HOR7	9.237532		#N/A	
YBR015C/YBR016W	MNN2/YBR016W	9.233806	9.233806	0.015	0.092
YPR160W	YPR159C-A/GPH1	9.132703	9.132703	0.259	0.259
YMR135C/YMR136W	GID8/GAT2	9.035585	9.035585	0.09	-0.076
YPR194C	OPT2	8.952033		-0.956	
YPR196W		8.952033		0.668	
YBR083W	TEC1	8.864551		-0.363	
YFR034C	PHO4	8.86368		-0.264	
YKL086W/YKL087C	SRX1/CYT2	8.863341	8.863341	#N/A	-0.055
YMR280C/YMR281W	CAT8/GPI12	8.829097	8.829097	0.971	0.072
YPR148C/YPR149W	YPR148C/NCE102	8.797662	8.797662	0.126	0.477
YAL060W	BDH1	8.777637		0.364	
YMR043W	MCM1	8.747654		-0.638	
YML051W	GAL80	8.656285		-0.624	
YDR277C	MTH1	8.632937		0.288	
YGR202C/YGR203W	PCT1/YGR203W	8.62398	7.776986	-0.137	-0.145
YGR253C/YGR254W	PUP2/ENO1	8.619022	8.619022	-0.126	0.045
YLR327C/YLR328W	TMA10/NMA1	8.613617	8.613617	1.335	-0.169
YOL086C	ADH1	8.403336		0.377	
YDL181W	INH1	8.344742		0.074	
YBR066C	NRG2	8.256566		-0.31	
YJR127C	RSF2	8.07203		0.091	
YLR257W		7.974075		-0.214	
YER001W	MNN1	7.850364		-0.961	

YLR355C/YLR356W	ILV5/YLR356W	7.817386	7.817386	-0.389	0.224
YDL047W/YDL048C	SIT4/STP4	7.676721	7.676721	0.002	#N/A
YER152C		7.6017		0.065	
YNR036C		7.463903		0.303	
YDR524C-B		7.432196		#N/A	
YDR525W-A	SNA2	7.432196		#N/A	
YJR001W	AVT1	7.332991		0.239	
YPR036W-A		7.097926		#N/A	
YBR008C	FLR1	7.018194		0.002	
YGR086C	PIL1	7.014496		0.338	
YDR247W	VHS1	6.886465		0.342	
YNL239W/YNL240C	LAP3/NAR1	6.756187	6.756187	-0.107	0.217
YGL190C	CDC55	6.616684		0.031	
YMR008C/YMR009W	PLB1/ADI1	6.536338	6.536338	0.061	-0.347
YBR009C/YBR010W	HHF1/HHT1	6.492409	6.492409	0.208	0.132
YGR143W	SKN1	6.463317		0.263	
YER073W	ALD5	6.415061		-0.297	
YDR077W	SED1	6.36566		0.47	
YGR191W	HIP1	6.329531		-0.099	
YPL134C	ODC1	6.200612		0.378	
YML075C	HMG1	6.177683		0	
YOL059W/YOL060C	GPD2/MAM3	6.171103	6.171103	-0.323	0
YDR368W	YPR1	6.15217		-0.018	
YDR216W	ADR1	6.094167		0.381	
YDR275W	BSC2	5.98319		-0.031	
YDR406W	PDR15	5.971804		0.009	
YDR072C/YDR073W	IPT1/SNF11	5.961185	5.961185	0.214	0.006
YGL009C	LEU1	5.935324		-0.114	
YBL032W/YBL033C	HEK2/RIB1	5.903688	5.903688	0.004	-0.004
YEL044W	IES6	5.89115		-0.181	
YHR082C/YHR083W	KSP1/SAM35	5.833539	5.833539	-0.071	-0.06
YMR253C		5.732263		0.057	
YGL178W/YGL179C	MPT5/TOS3	5.665327	5.665327	-0.234	-0.819
YNL055C	POR1	5.621353		0.084	
YNL015W	PBI2	5.619322		0.21	
YNR002C	ATO2	5.617467		0.638	
YGL006W-A		5.599102		#N/A	
YBR067C	TIP1	5.595879		#N/A	
YER088C	DOT6	5.547613		0.042	
YOL084W	PHM7	5.518819		1.109	
YPR144C/YPR145W	NOC4/ASN1	5.496217	5.496217	0.041	-0.648
YNL277W	MET2	5.488248		0.009	
YNL277W-A		5.488248		#N/A	
YPR074C	TKL1	5.472925		-0.431	
YKL065W-A		5.41555		#N/A	

Supplemental Table 2.2. Positive Targets of Gcn4. Columns 1 and 2: All Gcn4 targets above our significance cutoff are listed. Genes flanking a divergent promoter are listed in the same row. Columns 3: The ratio of hybridization intensity of the Ty5 inverse PCR product to the hybridization intensity of the genomic control (Red vs Green). Genes flanking a divergent promoter are listed in the same row. Columns 4: The Log10 ratio of expression of each gene in GCN4/ gcn4D in 100mM 3AT from the data in (Natarajan et al. 2001). Among all 287 promoters we identified, 131 drive genes that show expression change over two fold.

1. Systematic Name	2. Standard Name	3. Ratio (Red/Green)		4. GCN4/ gcn4D in 100mM 3AT Log10(ratio)	
YOL058W	ARG1	864.617282		1.968	
YDR006C/YDR007W	SOK1/TRP1	283.286773	113.986814	-0.13	0.017
YJL088W	ARG3	176.719532		1.498	
YJR109C/YJR110W	CPA2/YMR1	117.771916	117.771916	1.403	0.364
YNL103W/YNL104C	MET4/LEU4	58.944321	58.944321	0.239	0.913
YLR355C/YLR356W	ILV5/YLR356W	51.538875	51.538875	0.441	0.478
YIL116W/YIL117C	HIS5/PRM5	46.888159	46.888159	1.009	-0.029
YPR036W-A		44.691971		#N/A	
YPL250C	ICY2	42.85757		0.855	
YER069W	ARG5,6	38.817951		1.257	
YPR145W/YPR144C	ASN1/NOC4	35.28644	5.70876	1.328	-0.551
YHR018C	ARG4	34.320831		1.191	
YDL182W/YDL183C	LYS20/YDL183C	33.827906	33.827906	1.161	0.488
YBR161W	CSH1	28.73738		-0.279	
YMR095C/YMR096W	SNO1/SNZ1	24.999277	24.999277	1.915	1.916
YOR337W	TEA1	22.877059		0.576	
YDR072C/YDR073W	IPT1/SNF11	22.332516	22.332516	-0.092	-0.12
YDR127W	ARO1	21.556988		0.7	
YCL030C	HIS4	20.31851		1.27	
YCR024C-A	PMP1	19.542954		-0.095	
YLR081W	GAL2	19.540851		-0.731	
YPL252C	YAH1	18.988625		0.595	
YOR316C-A/YOR317W	YOR316C-A/FAA1	18.8817	18.8817	#N/A	-0.093
YGL184C	STR3	18.789571		1.256	
YAL040C	CLN3	18.545158		-0.171	
YEL036C	ANP1	18.36383		-0.126	

YOR302W		18.000993		NaN	
YDR009W	GAL3	16.501795		0.371	
YPR194C	OPT2	15.807286		-0.17	
YPR196W		15.807286		0.051	
YDR158W	HOM2	15.658263		1.064	
YLR436C	ECM30	15.433134		-0.108	
YHR179W	OYE2	15.212846		-0.075	
YBR248C	HIS7	15.168653		0.776	
YPL111W/YPL112C	CAR1/PEX25	15.063686	5.867922	0.574	0.06
YGR161C/YGR161W-C	RTS3/YGR161W-C	15.003036	15.003036	0.33	#N/A
YBR218C	PYC2	14.840874		0.569	
YNL220W/YNL221C	ADE12/POP1	14.793264	14.793264	0.624	0.48
YER052C	HOM3	14.783925		0.806	
YMR121C	RPL15B	14.413732		-0.385	
YHR161C/YHR162W	YAP1801/YHR162W	14.189431	14.189431	0.223	0.233
YBL043W	ECM13	14.107664		0.885	
YBL045C/YBL044W	COR1/YBL044W	14.107664	9.361832	0.007	0.137
YKL163W/YKL164C	PIR3/PIR1	14.068411	14.068411	0.038	0.234
YER070W	RNR1	14.044459		-0.969	
YJR126C	VPS70	13.901097		0.016	
YGL009C	LEU1	13.827912		1.154	
YMR251W-A	HOR7	13.730826		0.34	
YNR056C	BIO5	13.561945		0.888	
YML119W/YML120C	YML119W/NDI1	13.353615	13.353615	0.033	0.047
YER114C	BOI2	13.196136		-0.131	
YER073W	ALD5	13.163852		1.204	
YDR379C-A/YDR380W	YDR379C-A/ARO10	13.160895	13.160895	#N/A	0.631
YBR068C	BAP2	13.044104		1.095	
YDR034C	LYS14	13.005359		0.376	
YBR083W	TEC1	12.774836		-0.134	
YBR066C	NRG2	12.772751		-0.462	
YLR120C	YPS1	12.755639		0.026	
YGR033C/YGR034W	TIM21/RPL26B	12.534669	12.534669	-0.167	-0.44
YIL056W/YIL057C	VHR1/YIL057C	12.527303	9.439361	0.921	-1.148
YFR034C	PHO4	12.362853		0.373	
YDR354W	TRP4	12.237617		1.06	
YPR138C	MEP3	11.999414		0.406	
YGL180W	ATG1	11.984838		1.058	
YEL072W	RMD6	11.973074		0.03	
YEL073C		11.973074		0.405	
YER001W	MNN1	11.887629		0.043	
YJL210W/YJL212C	PEX2/OPT1	11.633803	11.633803	0.105	-0.174
YDR085C	AFR1	11.546603		0.211	
YDR449C/YDR450W	UTP6/RPS18A	11.491882	11.491882	-0.776	-0.354
YGL125W	MET13	11.451879		0.837	

YNL042W	BOP3	11.376159		0.216	
YHR143W	DSE2	11.349735		-0.218	
YOR188W	MSB1	11.133476		0.071	
YLR254C	NDL1	11.055889		-0.071	
YDR115W		10.995425		-0.097	
YHR087W		10.985946		0.539	
YOR388C/YOR389W	FDH1/YOR389W	10.914296	10.914296	-0.037	-0.117
YOR376W-A		10.844308		#N/A	
YLR152C		10.736324		1.134	
YER033C/YER034W	ZRG8/YER034W	10.68286	10.68286	0.379	0.017
YLR300W	EXG1	10.643695		-0.248	
YOL125W/YOL126C	TRM13/MDH2	10.438132	10.438132	-0.158	0.33
YCL018W	LEU2	10.375358		1.062	
YGR286C	BIO2	10.330991		-0.372	
YDL181W	INH1	10.300291		-0.167	
YDR247W	VHS1	10.297816		-0.043	
YOR119C/YOR120W	RIO1/GCY1	10.289717	10.289717	-0.314	0.271
YMR195W	ICY1	10.286544		0.402	
YNL106C	INP52	10.185926		0.005	
YBR069C	TAT1	10.07541		-0.439	
YLR327C/YLR328W	TMA10/NMA1	10.036819	6.396684	0.726	-0.097
YBR112C/YBR114W	CYC8/RAD16	10.024695	10.024695	0.083	0.121
YDR077W	SED1	9.989611		0.013	
YJL100W/YJL101C	LSB6/GSH1	9.967383	9.967383	0.226	-0.028
YOR230W	WTM1	9.958198		0.363	
YKL217W/YKL218C	JEN1/SRY1	9.91338	9.91338	0.617	1.316
YKR092C/YKR093W	SRP40/PTR2	9.909185	9.909185	-0.471	-0.294
YER124C/YER125W	DSE1/RSP5	9.864819	9.864819	-0.222	-0.068
YOR032C/YOR032W-A	HMS1/YOR032W-A	9.779805	9.779805	0.003	#N/A
YBR043C	QDR3	9.761184		0.587	
YHR207C/YHR208W	SET5/BAT1	9.688371	9.688371	0.087	0.95
YBR055C/YBR056W	PRP6/YBR056W	9.577594	9.577594	-0.035	0.245
YMR043W	MCM1	9.576239		-0.166	
YGR067C		9.507606		0.036	
YMR216C/YMR217W	SKY1/GUA1	9.462513	9.462513	-0.201	-0.478
YHR082C/YHR083W	KSP1/SAM35	9.456434	9.456434	-0.189	0.094
YLR304C	ACO1	9.353934		-0.304	
YGL178W/YGL179C	MPT5/TOS3	9.228035	9.228035	0.039	-0.282
YLR108C/YLR109W	YLR108C/AHP1	9.219747	9.219747	-0.43	0.083
YJL115W/YJL116C	ASF1/NCA3	9.212585	9.212585	-0.169	-1.106
YDL066W/YDL067C	IDP1/COX9	9.124502	9.124502	1.015	0.218
YDR113C	PDS1	9.066534		-0.192	
YDR508C/YDR510W	GNP1/SMT3	9.028619	9.028619	-0.117	-0.15
YKL185W	ASH1	8.98123		-0.218	
YDR298C/YDR299W	ATP5/BFR2	8.889266	8.889266	-0.182	-0.616

YNL160W	YGP1	8.879042		0.008	
YGL256W/YGL257C	ADH4/MNT2	8.861143	8.861143	-0.248	0.026
YPL232W	SSO1	8.819513		-0.189	
YIL164C	NIT1	8.811423		1.103	
YBR198C/YBR199W	TAF5/KTR4	8.796924	8.796924	-0.024	0.02
YCR052W	RSC6	8.791436		0.178	
YJR095W	SFC1	8.784889		-0.026	
YGR154C/YGR155W	GTO1/CYS4	8.694651	8.694651	0.626	0.087
YHR001W	OSH7	8.616857		0.03	
YJR025C	BNA1	8.511246		1.249	
YDR026C		8.420734		-0.073	
YBR222C	PCS60	8.391525		0.142	
YOR226C/YOR227W	ISU2/YOR227W	8.314935	8.314935	0.184	0.286
YKL109W/YKL110C	HAP4/KTI12	8.305314	8.305314	-0.53	0.158
YNL015W	PBI2	8.282111		0.182	
YGL263W	COS12	8.238373		0.001	
YML100W-A		8.220135		-2	
YDR525W-A	SNA2	8.214472		#N/A	
YLR297W		8.21102		0.006	
YDL124W		8.187917		0.213	
YOR108W	LEU9	8.129759		0.588	
YDR043C/YDR044W	NRG1/HEM13	8.116672	8.116672	-0.572	-0.402
YDR384C/YDR385W	ATO3/EFT2	8.102436	8.102436	0.396	-0.195
YOL059W/YOL060C	GPD2/MAM3	8.048883	8.048883	0.231	0.322
YDL025C		8.005017		0.831	
YHR098C/YHR099W	SFB3/TRA1	7.983706	7.983706	-0.042	-0.044
YNL178W	RPS3	7.974924		-0.388	
YJL159W/YJL160C	HSP150/YJL160C	7.957417	7.957417	0.063	1.881
YNL067W-B		7.956627		#N/A	
YNL068C	FKH2	7.956627		-0.052	
YNR069C/YNR070W	BSC5/YNR070W	7.948566	7.948566	0.71	0.444
YAL062W/YAL063C	GDH3/FLO9	7.916592	7.916592	0.14	0.22
YDR264C/YDR265W	AKR1/PEX10	7.8218	7.8218	0.076	0.05
YPL265W	DIP5	7.739394		-0.381	
YGR124W	ASN2	7.688705		0.663	
YOR316C	COT1	7.679683		0.064	
YLR453C	RIF2	7.636573		0.083	
YLR454W	FMP27	7.636573		0.284	
YMR019W	STB4	7.575595		0.454	
YML005W/YML006C	TRM12/GIS4	7.555755	7.555755	-0.113	-0.191
YMR135C/YMR136W	GID8/GAT2	7.511422	7.511422	0.545	0.375
YBR018C	GAL7	7.447956		0.137	
YIL135C	VHS2	7.416529		0.045	
YGR250C/YGR251W	YGR250C/YGR251W	7.374293	7.374293	-0.123	-0.287
YGL234W/YGL236C	ADE5,7/MTO1	7.373234	7.373234	0.243	0.279

YDR216W	ADR1	7.34785		-0.491	
YHL007C	STE20	7.321875		-0.176	
YLR257W		7.294998		-0.02	
YBL029W/YBL029C-A	YBL029W/YBL029C-A	7.28716	7.28716	-0.123	#N/A
YDR096W	GIS1	7.186331		-0.077	
YOL011W/YOL012C	PLB3/HTZ1	7.170094	7.170094	-0.261	-0.066
YGR146C-A		7.145991		#N/A	
YMR318C	ADH6	7.140242		-0.012	
YLR335W	NUP2	7.138019		0.032	
YBR147W		7.110749		2	
YOR084W		7.092835		-0.087	
YJR016C	ILV3	7.088666		0.808	
YOR273C/YOR274W	TPO4/MOD5	7.087329	7.087329	0.272	-0.235
YGL006W-A		7.080465		#N/A	
YDR345C	HXT3	7.05448		-0.179	
YOR246C/YOR247W	YOR246C/SRL1	7.034496	7.034496	-0.039	0.019
YOR267C	HRK1	7.032372		0.228	
YGL055W/YGL056C	OLE1/SDS23	7.030654	7.030654	-0.034	0.1
YKL096W	CWP1	7.023233		-0.378	
YLR353W	BUD8	7.00301		0.284	
YBR145W	ADH5	6.978796		1.6	
YMR083W	ADH3	6.973759		-0.328	
YBR296C	PHO89	6.956433		0.329	
YBR296C-A/YBR297W	YBR296C-A/MAL33	6.956433	6.956433	#N/A	0.138
YER145C/YER146W	FTR1/LSM5	6.942147	6.942147	-0.595	-0.129
YDR146C/YDR147W	SWI5/EKI1	6.934058	6.934058	0.041	-0.141
YLR295C	ATP14	6.933284		-0.123	
YOR086C	TCB1	6.915386		0.083	
YOL119C	MCH4	6.910191		0.714	
YBR201C-A/YBR202W	YBR201C-A/CDC47	6.859449	6.859449	#N/A	0.072
YNL144C		6.821764		0.282	
YNL098C	RAS2	6.807516		0.078	
YLR347C	KAP95	6.781618		-0.286	
YPL274W	SAM3	6.771976		0.111	
YGR043C	NQM1	6.715228		0.545	
YOL084W	PHM7	6.695717		0.078	
YOR298C-A/YOR299W	MBF1/BUD7	6.661735	6.661735	#N/A	0.022
YMR062C/YMR063W	ECM40/RIM9	6.64839	6.64839	1.308	0.254
YDL110C	TMA17	6.645949		0.269	
YDL173W/YDL174C	YDL173W/DLD1	6.636262	6.636262	0.068	0.016
YGR144W	THI4	6.620413		0.172	
YJR047C/YJR048W	ANB1/CYC1	6.611498	6.611498	-0.323	-0.261
YLR314C	CDC3	6.590342		-0.078	
YOR130C	ORT1	6.587842		1.255	
YDR259C	YAP6	6.576594		-0.249	

YJL082W	IML2	6.545709		0.14	
YGR097W	ASK10	6.535877		0.257	
YGR143W	SKN1	6.525726		0.063	
YMR041C/YMR042W	ARA2/ARG80	6.496035	6.496035	0.045	0.338
YGR253C/YGR254W	PUP2/ENO1	6.488969	6.488969	0.077	0.043
YAL060W	BDH1	6.456611		0.206	
YJL153C	INO1	6.453988		-0.195	
YGR197C/YGR198W	SNG1/YPP1	6.440248	7.154007	0.388	0.217
YBR067C	TIP1	6.411985		-0.388	
YGR233C/YGR234W	PHO81/YHB1	6.378556	6.378556	-0.174	-0.44
YOR152C/YOR153W	YOR152C/PDR5	6.343844	6.343844	0.071	-0.148
YDR341C		6.33818		0.559	
YJL133W	MRS3	6.304023		-0.178	
YHR022C/YHR023W	YHR022C/MYO1	6.246243	6.246243	-0.232	0.191
YOR357C/YOR358W	SNX3/HAP5	6.214597	6.214597	0.133	0.325
YGL121C	GPG1	6.211917		0.63	
YPL132W/YPL133C	COX11/RDS2	6.211859	6.211859	-0.267	-0.193
YBR143C	SUP45	6.197579		-0.252	
YJL186W/YJL187C	MNN5/SWE1	6.196528	6.196528	-0.215	-0.532
YLR154C	RNH203	6.184603		0.096	
YEL007W		6.153175		-0.067	
YMR296C/YMR297W	LCB1/PRC1	6.151077	6.151077	-0.051	0.118
YLR110C	CCW12	6.138536		-0.079	
YER091C/YER092W	MET6/IES5	6.135969	6.135969	0.328	-0.001
YBR084W	MIS1	6.118916		-0.56	
YJL184W/YJL185C	GON7/YJL185C	6.115718	6.115718	0.352	0.396
YPL088W/YPL089C	YPL088W/RLM1	6.112178	6.112178	-0.187	-0.138
YMR106C	YKU80	6.087006		0.324	
YML075C	HMG1	6.083259		-0.402	
YGR132C/YGR133W	PHB1/PEX4	6.059961	6.059961	-0.016	0.151
YDL049C	KNH1	6.05406		-0.052	
YPL092W	SSU1	5.994355		1.502	
YLR256W	HAP1	5.952326		-0.06	
YBR182C	SMP1	5.951808		-0.562	
YBR182C-A/YBR183W	YBR182C-A/YPC1	5.951808	5.951808	#N/A	-0.008
YBR249C/YBR250W	ARO4/SPO23	5.95083	5.95083	0.785	0.273
YHR075C/YHR076W	PPE1/PTC7	5.94914	5.94914	0.297	0.446
YFR055W	IRC7	5.941356		0.375	
YDL085C-A/YDL085W	YDL085C-A/NDE2	5.930921	5.930921	#N/A	1.697
YMR315W		5.902302		0.361	
YJR001W	AVT1	5.878744		0.032	
YOL086C	ADH1	5.876039		0.227	
YGR023W	MTL1	5.864753		0.169	
YLR179C/YLR180W	YLR179C/SAM1	5.8368	5.8368	-0.01	-0.487
YCL024W/YCL025C	KCC4/AGP1	5.835818	5.835818	-0.312	0.61

YML088W	UFO1	5.82415		-0.158	
YBR085C-A		5.815821		#N/A	
YHR094C	HXT1	5.811337		0.161	
YER053C-A		5.810103		#N/A	
YPR006C	ICL2	5.804223		0.158	
YJR108W	ABM1	5.803131		-1.079	
YLL028W	TPO1	5.798704		-0.014	
YNL124W/YNL125C	NAF1/ESBP6	5.793228	5.793228	0.362	0.427
YML028W	TSA1	5.772001		-0.063	
YKL120W	OAC1	5.765064		0.815	
YPL135W	ISU1	5.73117		0.277	
YPL137C	GIP3	5.73117		0.056	
YBR053C/YBR054W	YBR053C/YRO2	5.679188	5.679188	0.108	-0.386
YPL262W/YPL263C	FUM1/KEL3	5.675374	5.675374	0.321	-0.545
YDL022W	GPD1	5.666252		0.269	
YIL130W/YIL131C	ASG1/FKH1	5.664883	5.664883	-0.135	-0.15
YBR162C	TOS1	5.65858		0.137	
YBR162W-A	YSY6	5.65858		-0.035	
YGR121C/YGR122W	MEP1/YGR122W	5.652507	5.652507	0.077	0.091
YJL200C	ACO2	5.646987		0.631	
YDR490C/YDR492W	PKH1/IZH1	5.629497	5.629497	0.332	0.202
YKR091W	SRL3	5.623176		0	
YMR253C		5.621853		-0.065	
YJL112W	MDV1	5.620316		-0.195	
YGR282C	BGL2	5.62003		-0.122	
YBR046C	ZTA1	5.560574		1.023	
YDL047W/YDL048C	SIT4/STP4	5.539314	5.539314	-0.139	-0.077
YLR130C	ZRT2	5.521985		-0.259	
YBR126C	TPS1	5.517144		0.235	
YJR154W		5.510697		0.884	
YNL097C-B		5.506022		#N/A	
YHR048W		5.476629		0.302	
YGL157W		5.448736		0.054	
YLR267W	BOP2	5.427243		0.862	
YHR019C/YHR020W	DED81/YHR020W	5.417231	5.417231	0.442	0.277

Chapter 3: Applying the "Calling Card-seq" method to study poorly characterized yeast transcription factors

Haoyi Wang, David Mayhew, Xuhua Chen, Mark Johnston, and Robi David Mitra

Department of Genetics, Washington University, School of Medicine, 4444 Forest Park Parkway, St. Louis, MO 63108

I next sought to improve the Calling Card method by coupling it with next-generation sequencing technology. I developed "Calling Card-seq", which uses massively parallel DNA sequencing to map the locations of calling cards that have been integrated into the genome. This method has several advantages over the microarray-based method described in Chapter 2. First, Calling Card-seq maps calling card insertion sites to a single base pair resolution, something that cannot be achieved using a microarray. Second, Calling Card-seq more accurately identifies the gene targets of a transcription factor. Finally, Calling card-seq can analyze multiple transcription factors simultaneously. This is accomplished by tagging Ty5 transposons with a DNA barcode and co-transforming the barcoded transposons with different TF-sir4 fusion constructs. By harvesting calling cards and then sequencing the barcodes and the flanking genomic DNA, it is possible to determine the location of each calling card as well as the identity of the transcription factor that deposited it into the genome.

I developed two protocols for performing Calling-Card-seq. The first, which I describe in this chapter, requires the use of an engineered Ty5 transposon with modified LTRs. This construct did not transpose as efficiently as wild-type Ty5, so I later developed a second, more efficient, Calling Card-seq protocol that uses the wild-type transposon (This is described in Chapter 4).

In this chapter, I also describe the application of the (first) Calling Card-seq protocol to study poorly characterized yeast transcription factors. I constructed 89 TF-Sir4 fusions, 62 of which have unknown sequence recognition motifs.

I designed these experiments in collaboration with Mark Johnston and Rob Mitra. Xuhua Chen and I performed all the experiments. David Mayhew was responsible for all of the computational analyses.

ABSTRACT

Sequence-specific transcription factors (TFs) regulate gene expression in response to signals from the environment. Despite concentrated efforts to identify position specific weight matrices (PSWM) and target genes for all yeast TFs, the DNA-binding specificities for one-third of these are still not known. To fill this gap in our knowledge of protein-DNA interactions, we analyzed the Calling Cards deposited by 62 yeast TFs that have no identified PSWM. To allow for sample multiplexing, we "bar-coded" the Calling Cards and developed a method to determine, in a single Illumina sequencing read, the DNA sequence of the bar code and the genomic region flanking a Calling Card. We used this method to analyze Gal4p and Leu3p and successfully mapped Calling Cards deposited in promoters known to be targets of these transcription factors. However, we identified several technical limitations that prohibited us from recovering more than a few hundred Calling Cards deposited by each TF, which was not enough to enable the

accurate prediction of binding motifs and target genes. This led us to develop the more efficient Calling-Card-seq method described in Chapter 4.

INTRODUCTION

Organisms respond to environmental changes and developmental cues by changing their gene expression. In many cases this is accomplished by altering the function of transcription factors (TFs) that bind to specific DNA sequences near particular genes and activate or repress gene expression. Identification of the target genes and the DNA sequence recognized by each TF is essential for understanding how transcription is regulated.

As an important eukaryotic model organism, the bakers' and brewers' yeast *Saccharomyces cerevisiae* has been studied intensively to dissect its transcriptional regulatory networks (Harbison et al. 2004; Ideker et al. 2001; MacIsaac et al. 2006). *In vivo* Chromatin immunoprecipitation-DNA microarray ("ChIP-chip") experiments and *in vitro* protein binding to DNA mircroarrays (PBMs) experiments have been employed in an attempt to identify the DNA sequence recognition motifs of all two hundred yeast TFs (Badis et al. 2008; Zhu et al. 2009). However, almost one third of yeast TFs did not yield their binding sites in this way, and many of the motifs predicted from the different methods disagree. The target genes of still more TFs remain ill-defined. The current incomplete catalog of DNA-protein interactions prevents prediction of the expression pattern of a gene from its promoter sequence, and hinders engineering of gene expression.

We reported the development of the “Calling Card” method for accurate and robust identification of protein-DNA interactions (Wang et al. 2007). We have combined the Calling Card method with massively parallel DNA sequencing technology to study yeast TFs with unknown DNA sequence recognition motifs. We applied this method to identify the targets of more than 40 TFs. In the course of this work, we identified technical limitations that prevented us from predicting DNA sequence recognition motifs and target genes of these TFs. Our solution to these problems is the use of paired-end sequencing, described in Chapter 4.

RESULTS

Making TF-Sir4 library

We selected 62 TFs without known DNA sequence recognition motifs (Table 3.1) (Harbison et al. 2004; Ho et al. 2006). As a control set, we chose 27 TFs with known recognition sequences. *SIR4* was fused to all 89 of these genes (see Methods for details). For all TF-Sir4 constructs, correct sequences from both junctions have been obtained, which are enough to cover the whole ORF for 43 TFs in our collection. Since three copies of myc were fused after Sir4 in our constructs, we were able to detect TF-Sir4-3xMyc expression in yeast by Western blot hybridization using anti-myc antibodies. We applied Western blot hybridization to 62 TFs, and confirmed expression of TF-Sir4 fusion proteins with correct sizes for 40 of them (Table 3.1). For those that we couldn’t detect strong protein expression, the junction sequences confirmed correct cloning. We think this is largely due to cellular regulation that keeps these TFs in a concentration below our Western detection. However, this relatively low protein concentration is not expected to prevent them from depositing Calling Cards. Indeed, we could not detect Gcn4-Sir4 using Western hybridization, but it was able to deposit Calling Cards in the genome that allowed mapping of Gcn4 target genes.

Engineering of Ty5 for Illumina sequencing

To determine the DNA sequence of the Calling Cards with the Illumina sequencer, we placed the 33bp sequencing priming site (Seq 1) immediately adjacent to the DNA fragment to be sequenced, since the read length is only 36bp long. Because the Ty5 Calling Card encodes multiple proteins essential for transposition (Zou et al. 1995), we were limited in where we could place the priming site (Seq 1). Seq 1 points towards a restriction enzyme cleavage site that will be ligated to the end of the flanking genomic sequence by self-ligation (Fig 3.1). After enzyme digestion, self-ligation, and an inverse PCR, we would read the sequence of the restriction enzyme cleavage site closest to the Caling Card integration site. Thus, the resolution of the mapping of Calling Cards in this way is similar to that obtained by hybridization to a DNA microaray, which does not reveal the exact point of insertion.

To overcome this problem, we need to generate a shorter genomic DNA fragment that is self-ligated to Seq 1, allowing us to sequence through to the end of the Calling Card (Fig 3.1). The end of LTR sequence (GTCAACA) can easily be converted to an MmeI recognition sequence (TCCRAC) by inserting a C or by changing GT to TC at the end of the LTR. MmeI cuts the DNA 18bp away from its recognition sequence, so after self-ligation and inverse PCR the 36bp sequence reads will include 17 bp of the genomic DNA flanking the Calling Card, which is enough to uniquely map it to the yeast genome, and extend into the LTR sequence to reveal the exact point of insertion of the Calling Card (Fig 3.1). In addition, we also inserted into the Calling Card short DNA sequence “bar codes” (4 bp) to enable multiplexing of samples. Thus, the 36 bp sequence from obtained from the Illumina squencer provided both the information of insertion site of the Calling Card and the bar code identifier of the TF that put the Calling Card there.

Ty5-MmeI was able to transpose, but with efficiency five- to ten-fold lower than the wild type Ty5 (Fig 3.2).

Multiplexed mapping of Calling Cards

Since different samples were bar coded uniquely, we pooled samples of four TFs and sequence on one lane of Illumina flowcell. A few hundred independent Calling Card insertions were identified for each TF. For Gal4-Sir4 and Leu3-Sir4, we identified tight clusters of Ty5 insertions within the promoters of their known target genes (Fig 3.3 A, B). Within these promoters, Calling Cards are highly enriched near the binding sites for the TFs (Fig 3.3 A, B). To determine the sensitivity and specificity, we plot receiver operating characteristic (ROC) curves (Lusted 1971) for each of the data sets (Fig 4.3 C, D). The method appears to be quite specific, but a few hundred of insertions do not provide enough data to achieve good sensitivity. We applied this MmeI-based protocol on more than 40 TFs of yeast, and for many of them we observed unique insertion patterns. Even though the relatively small number of Calling Cards recovered is insufficient to identify target genes and sequence recognition motifs with confidence, these data are a good reference for the results of experiments using the improved protocol described in Chapter 4. These data are summarized in table 3.2.

DISCUSSION

To map Calling Cards insertions using massively parallel DNA sequencing technology, we engineered the LTR of the Ty5 Calling Card and developed a working protocol. The strength of this protocol is that we can obtain both the genomic sequence flanking Calling Cards and their bar codes in a single direction sequencing read on the Illumina sequencer. In addition, this protocol is not likely to produce bias through restriction enzyme digestion and amplification in the PCR, since the templates for inverse PCR are all the same size (Fig 3.1). The uniform product size of 130bp is optimal for Illumina sequencing. Using this method, we mapped hundreds of Calling Card insertions for each 40 TFs and identify target genes of a few poorly characterized TFs. However, the alteration of the LTR sequence, which is necessary for this protocol, dramatically reduced Ty5 transposition efficiency (Fig 3.2). This low transposition efficiency made it difficult to recover enough independent Calling Card insertions to be able to confidently identify sequence recognition motifs and target genes, making multiplexing experiments inefficient. Furthermore, this protocol involves MmeI digestion and blunt-end self-ligation, neither of which is very efficient. Consequently, the sensitivity of these data is limited by only a few hundred independent insertions mapped for each sample.

As described in Chapter 4, we developed an improved protocol that requires neither changing Ty5 LTR sequence, MmeI digestion, nor blunt end ligation. In addition, changing marker gene from His3AI to *HIS3*, Ty5 transposition efficiency is increased five-fold over wild type Ty5 (Fig 3.2). Coupling this improvement with pair-end sequencing, we were able to map more than 5,000 insertions for each TF and achieve good sensitivity (Fig 3.3 C, D). We are in the process of applying this pair-end protocol to study all TF-Sir4 constructs in our library.

METHODS

Strains and media

All the experiments were done on the diploid *sir4* deletion mutant, YM7635 (*MATa /MATalpha his3Δ1/ his3Δ1 leu2Δ0/ leu2Δ0 ura3Δ0/ ura3Δ0 met15Δ0/MET15 lys2Δ0/LYS2 sir4::Kan/ sir4::Kan trp1::Hyg/ trp1::Hyg*). It was grown in complete synthetic media containing 2% glucose or galactose.

TF-Sir4 library construction

I made each TF-Sir4 chimera by the "gap repair" method (Ma et al. 1987; Wach et al. 1994). All TF-Sir4 fusions were derived from pBM5037 (Gal4DBD-Sir4-Myc) (Wang et al. 2007). I designed one pair of universal primers that hybridize to attB sites (OM6883 & OM6884)for amplification of all the ORFs in the yeast ORF collection, which are flanked by attB sites (Gelperin et al. 2005). I amplified each TF ORF (specific primers were designed to amplify from genomic DNA each of the 13 TFs not included in the ORF collection or which failed to amplify). The resulting PCR products along with the recipient vector pBM5037 linearized by XhoI digestion were then used to transform yeast cells Trp^{+}. DNA was extracted from Trp^{+} positive clones and introduced into E. coli to get purified plasmids. Plasmid DNA was confirmed by Sanger sequencing using primers reading both junctions of the cloned ORFs. We also applied Western blot hybridization of randomly selected constructs to confirm the expression of TF-Sir4 chimeric proteins.

Ty5 LTR mutagenesis and engineering

To use MmeI in the protocol, I needed to eliminate the existing MmeI site in the middle of the Ty5 LTR, and introduce a new MmeI cleavage site at the end of the LTR. I used the QuickChange Kit (Stratagene) to make the necessary nucleotide changes. I tried two strategies to engineer the MmeI recognition sequence into the LTR: first, I inserted a "C" to convert the sequence GTCAACA to GTCCAACA (MmeI site is marked in red). Because the "ATG" start codon of the gene encoding the Ty5 Gag protein is inside the 5' LTR (Ke et al. 1999), the insertion of "C" will shift the reading frame at the 5' LTR and abolish translation of all Ty5 proteins. As expected, adding this "C" to form an MmeI site prevented transposition of the Ty5 element, so Ty5 proteins had to be provided in *trans* on a helper plasmid. This Ty5 helper cannot transpose because it lacks a 3' LTR, but it produces integrase to mobilize Ty5-MmeI elements. I also changed GTCAACA into TCCAACA by converting "GT" to "TC". This change does not shift the translational reading frame, so Ty5-MmeI is able to transpose autonomously. In both cases, Ty5-MmeI transposed with similarly low efficiencies, about five- to ten-fold lower than the wild type Ty5 (Fig 3.2).

Oligonucleotide OM8006 was used for knocking out the existing MmeI site within the 5' and 3' LTR sequences of Ty5. Oligonucleotides OM8005 and OM8004 were used for inserting one "C" in the 5' and 3' LTRs respectively. Oligonucleotides OM8114 and OM8113 were used to convert "GT" to "TC" in the 5' and 3' LTRs. Ty5 donor plasmid pBM5218 (Wang et al. 2008a) was used as template. The QuickChange site-directed mutagenesis kit (Stratagene) was used following the manufacturer's protocol. The Ty5 element containing a single MmeI site at the end of both 5' and 3' LTR was named pBM5196.

To insert the Illumina sequencing primer site and DNA sequence bar codes into pBM5196, a 66 bp sequence that includes the sequences of the Illumina adaptor P5, the Illumina sequencing primer 1, a 4bp bar code, and TaqI sites were cloned between FseI-AscI sites that lie between the 3' LTR and the *His3AI* gene in the Calling Card (Fig 3.1). Ten different 4bp bar codes used to make ten bar-coded Ty5 Calling Cards.

Induction of Ty5 transposition and inverse PCR

Since expression of Ty5 is driven by the *GAL1* promoter, transposition was induced by culturing cells in medium containing galactose for two to three days at room temperature, after which cells were plated on Glu -His media to select for cells with transposition events. Cells were then serially replica plated onto –His, FOA-containing media twice to select for cells with Calling Cards in their genome.

Approximately one thousand colonies were harvested for each TF and their genomic DNA was extracted. Each DNA sample was digested with TaqI and MmeI. The DNA overhang was then made blunt using End-It DNA end-repair kit (Epicentre) following manufacturer's protocol. Blunt-ended fragments were ligated overnight at 15°C in dilute solution to encourage self-circularization. After ethanol precipitation, self-ligated DNA was resuspended in ddH_2O and used as template for an inverse PCR. Primers that anneal to Ty5 and Illumina P5 adaptor sequences (OM8162 and OM8163) were used to amplify the genomic regions flanking Ty5 integrations and the bar codes within Ty5. PCR products were separated by electrophoresis through a 3% agrose gel. DNA fragments of approximately 130 bp were cut out of gel and purified using QIAquick Gel Extraction Kit (Qiagen), and diluted to 10nM. Samples of four TFs were pooled and submitted for Illumina sequencing.

Primers

OM6883:

ATA CAA TCA ACT CCA AGC TTG AAG CAA GCC TCC TGA AAG GGCGCGCC AAC AAG TTT GTA CAA AAA AGC AG

OM6884:

TTT GGG TTT GCT AGA ATT AGT ATC ACT ATG CGA CAC TCT ATC AAC CAC TTT GTA CAA GAA AG

OM8004: CGTAATTCACTACGTCCAACAGGATCCACTAGTTC

OM8005: CGTAATTCACTACGTCCAACAGGTTATGAGCCCTG

OM8006: CAAACCTCCGATCCGAGAGTACTTAAGAAACCATAG

OM8113: AGATCGTAATTCACTACTCCAACAGGATCCACTAGTTCTAG

OM8114: AGATCGTAATTCACTACTCCAACAGGTTATGAGCCCTGAG

OM8162: AATGATACGGCGACCACCGAGATCT ACACTCTTTCCCTACACGAC

OM8163: CAAGCAGAAGACGGCATACGA ATCGTAATTCACTACGTCCAAC

Sequence map back

DNA sequence reads were filtered by requiring a correct LTR sequence at the end of each read and an appropriate barcode at the beginning of each read. The 16 bp genomic sequences were mapped using a hash table of all possible 16 bp sequences from the yeast genome. Reads that uniquely locate the site of insertion were passed as correct. Independent insertions were required to have at least 10 reads to be considered real.

ACKNOWLEDGEMENTS

We are especially grateful to Jessica Hoisington-Lopez, Laura Langton, Brian Koebbe, and Jim Dover (Washington University) for their expert assistance with the use of Illumina Genome Analyzer. This work was supported by NIH grants R21RR023960 and 5P50HG003170-03, and funds provided by the James S. McDonnell Foundation.

TABLES

Table 3.1. TF-Sir4 library

Gene Name	System	NamMotif	ORF length	Protein MW (Da)	Western	Sequence Confirmed?
HMRA1	YCR097W	known	487	14,803	N	yes (whole orf)
Rpi1	YIL119C		1224	46,622	y (weak)	yes (whole orf)
MATALPHA2	YCR039C	known	633	24,282	N	yes (whole orf)
Rtg1	YOL067C	GGTCAC	534	19,016	y	yes (whole orf)
Stb1	YNL309W	RRACGCSAA	1263	45,683	y (weak)	yes (both junctions)
KAR4	YCL055W		1008	38,672	y	yes (whole orf)
LYS14	YDR034C		2373	89,396	y (very weak)	yes (both junctions)
SFG1	YOR315W		1041	39,021	N	yes (whole orf)
Rds1	YCR106W	CGGCCG	2499	95,689	y (very weak)	yes (both junctions)
Yap6	YDR259C	TTACTAA	1152	43,597	y (weak)	yes (whole orf)
Hms1	YOR032C		1305	48,871	y	yes (whole orf)
HMRA2	YCR096C	known	360	13,882	y	yes (whole orf)
Hir2	YOR038C		2628	98,444	y (very weak)	yes (both junctions)
Met32	YDR253C	AAACTGTGG	576	21,518	y	yes (whole orf)
Ndt80	YHR124W		1884	71,479	N	yes (both junctions)
Mth1	YDR277C		1302	49,060	y	yes (both junctions)
Rgm1	YMR182C		636	23,855	y	yes (whole orf)
Yap3	YHL009C	TTACTAA	993	37,955	y	yes (whole orf)
Sfl1	YOR140W		2301	83,317	y	yes (both junctions)
YJL103C	YJL103C		1857	70,381	y (very weak)	yes (both junctions)
Arg80	YMR042W	wGACkC	534	19,487	y	yes (whole orf)
CTH1	YDR151C		978	36,772	y	yes (whole orf)
Dal82	YNL314W	GAAAATTGCGTT	768	29,079	y	yes (whole orf)
Met28	YIR017C	TCACGTG	564	21,590	N	yes (whole orf)
Met31	YPL038W	AAACTGTGG	534	19,557	y (weak)	yes (whole orf)
Pdr1	YGL013C	CCGCGG	3207	121,793	N	yes (both ends)
NRG2	YBR066C		663	25,009	y (very weak)	yes (whole orf)
NHP10	YDL002C		612	23,857	y	yes (whole orf)
Hap5	YOR358W	CCAAT	729	27,675	y	yes (whole orf)
Yap1	YML007W	TTASTMA	1953	72,532	N	yes (both junctions)
YDR026C	YDR026C	TTACCCGGM	1713	66,352	y (very weak)	yes (both junctions)
Swi5	YDR146C	KGCTGR	2130	79,774	N	yes (both junctions)
Wtm2	YOR229W		1404	51,951	y	yes (both junctions)
Uga3	YDL170W	CCGNNNNCGG	1587	61,223	N	yes (both junctions)

Opi1	YHL020C	TCGAAYC	1215	46,065	y (very weak)	yes (whole orf)
Hap3	YBL021C	CCAAT	435	16,154	y	yes (whole orf)
Reh1	YLR387C		1299	49,689	N	yes (whole orf)
Cin5	YOR028C	TTACTAA	888	32,975	N	yes (whole orf)
Aft2	YPL202C	...AAAGTG CACCC	1251	47,104	y (weak)	yes (whole orf)
Mac1	YMR021C	GAGCAAA	1254	46,516	y	yes (whole orf)
LEE1	YPL054W		906	33,596	y (very weak)	yes (whole orf)
TUP1	YCR084C		2142	78,307	n	yes (both junctions)
PZF1	YPR186C		1290	50,027	y	yes (whole orf)
Ste12	YHR084W	ATGAAAC	2067	77,866	y (weak)	yes (both junctions)
Tye7	YOR344C	CANNTG	876	32,689	N	yes (whole orf)
TOS4	YLR183C		1470	55,467	N	yes (whole orf)
Tos8	YGL096W		831	31,257	N	yes (whole orf)
Yrr1	YOR162C	TttTGTTAC SCr	2433	92,467		yes (both junctions)
SET5	YHR207C		1581	60,547	y	yes (both junctions)
YRM1	YOR172W		2361	91,083	y (weak)	yes (both junctions)
Hir1	YBL008W		2523	93,889	y (weak)	yes (both junctions)
Ime4 methyltransferase	YGL192W		1803	69,395	y (weak)	yes (both juntions)
Haa1	YPR008W		2085	76,670	y (weak)	yes (both junctions)
Msn1	YOL116W		1149	43,060	y (size large)	yes (whole orf)
YDR049W	YDR049W		1899	72,733	n	yes (both junctions)
Dat1	YML113W		747	27,067	y (size large)	yes (whole orf)
Mal13	YGR288W		1422	54,325	y	yes (whole orf)
Cup9	YPL177C		921	34,653		yes (whole orf)
FLO8	YER109C		2400	86,648	n	yes (both junctions)
HCM1	YCR065W		1695	63,647	n	yes (both junctions)
Stp2	YHR006W		1626	60,792	y	yes (both junctions)
Mig3 YER028C			1185	43,119	y (weak)	yes (whole orf)
CUP2	YGL166W		678	24,425		yes (whole orf)
YPR013C	YPR013C		954	35,358		yes (whole orf)
GAT2	YMR136W	known?	1683	63,138		yes (both junctions)
Rts2	YOR077W		699	27,054		yes (whole orf)
Rdr1	YOR380W		1641	61,287		yes (both junctions)
GAT4	YIR013C		366	13,245		yes (whole orf)
Hms2	YJR147W		1077	41,192		yes (whole orf)
STP3	YLR375W		1032	37,718		yes (whole orf)
UME1	YPL139C		1383	51,021		yes (whole orf)
Mig2	YGL209W		1149	42,048		yes (whole orf)
SEF1	YBL066C		3447	127,991		yes (both junctions)
Aca1	YER045C		1470	54,592		yes (both junctions)
Mal33	YBR297W		1407	54,193		yes (both junctions)
Hir3	YJR140C		4947	191,678		yes (both junctions)
Sut2	YPR009W		807	30,257		yes (whole orf)

Fzf1	YGL254W		900	33,994		yes (whole orf)
Mbf1	YOR298C		456	16,404		yes (whole orf)
Hal9	YOL089C		3093	117,925		yes (both junctions)
RDS2	YPL133C		1341	50,081		yes (both junctions)
GIS1	YDR096W		2685	99,480		yes (both junctions)
TEA1	YOR337W		2280	86,832		yes (both junctions)
Sip3	YNL257C		3690	142,818		yes (both junctions)
YKL222C	YKL222C		2118	82,247		yes (both junctions)
Upc2	YDR213W	SRE(CTCGTATAAGC)	2742	100339		yes (both junctions)
YIL130W	YIL130W		2895	108,780		yes (both junctions)
IMP2'	YIL154C		1041	39,070		yes (whole orf)
YLR278C	YLR278C		4026	151,277		yes (both junctions)

Table 3.2. Review of Calling Card-seq data

TF	Run	Lane	Insertions[a]	Our Status[b]	Bulyk[c]	Hughes[d]
LEE1	54	5	314	putative	n	n
LEU3	54	5	384	known	y	y
SFG1	54	5	472	targets	n	n
YPR013C	54	5	140	diffuse	y	y
HMS1	63	2	469	putative	n	n
RPI1	63	2	384	putative	n	n
RTS2	63	2	290	diffuse	n	n
YRM1	63	2	972	putative	y	y
KAR4	67	8	424	few targets	n	n
NRG2	67	8	749	few targets	n	n
REH1	67	8	322	diffuse	n	n
RGM1	67	8	947	putative	n	y
NHP10	73	8	463	putative	n	y
TOS4	73	8	164	diffuse	n	n
TOS8	73	8	361	putative	n	y
MSN1	73	8	318	few targets	n	n
MAL13	73	3	342	diffuse	n	n
PAT1	73	3	758	targets	n	n
MIG3	73	3	459	diffuse	y	y
YLR278C	73	3	366	targets	n	y
CUP2	75	1	327	diffuse	n	n
MTH1	75	1	240	diffuse	n	n
GAT4	75	1	730	targets	y	y
HMS2	75	1	253	diffuse	n	n
YPR013C	75	2	185	few targets	y	y
STP3	75	2	398	few targets	n	y
UME1	75	2	276	diffuse	n	n
ACA1	75	2	333	targets	n	n
CUP9	79	7	314	few targets	y	y
SEF1	79	7	612	putative	n	n
YRR1	79	7	780	putative	y	y
MIG2	79	8	164	diffuse	y	y
WTM2	79	8	65	bad	n	n
LYS14	79	8	26	bad	y	y
HIR2	79	8	39	bad	n	n
HIR3	84	1	297	targets	n	n
UPC2	84	1	300	few targets	n	n
GIS1	84	1	439	diffuse	n	y
HAL9	84	1	361	few targets	y	y
SUT2	84	2	272	few targets	y	n

FZF1	84	2	214	diffuse	n	y
MBF1	84	2	229	diffuse	n	n
RDS2	84	2	294	targets	y	y

[a] Number of independent Ty5 insertions identified for each TF.

[b] Status of data analysis: “Putative” means we predicted target genes and motif. “Diffuse” indicates failure of predicting target genes. “Targets” means we were able to predict target genes but no motif was found.

[c] Whether there is motif predicted in (Zhu et al. 2009).

[d] Whether there is motif predicted in (Badis et al. 2008).

FIGURES

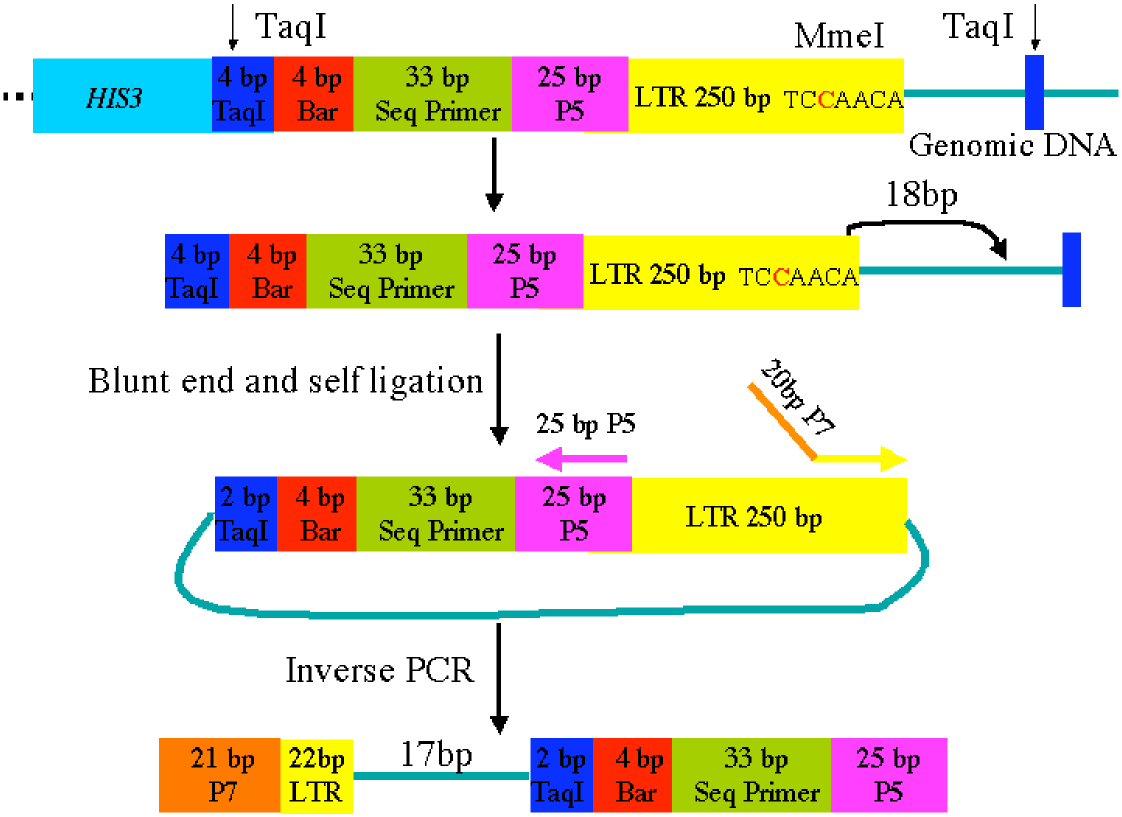


Figure 3.1. Sequence genomic DNA and bar code in single direction. Genomic DNA is first digested with restriction enzyme TaqI, and then MmeI that cuts 18bp downstream of the end of Ty5. DNA fragments are blunt-ended and ligated in dilute solution to favor recircularization, followed by amplification of the flanking genomic DNA and bar code by an "inverse PCR" (PCR primers containing Illumina sequencing primers and adaptor sequences). The identity of inverse-PCR products is then determined by Illumina sequencing.

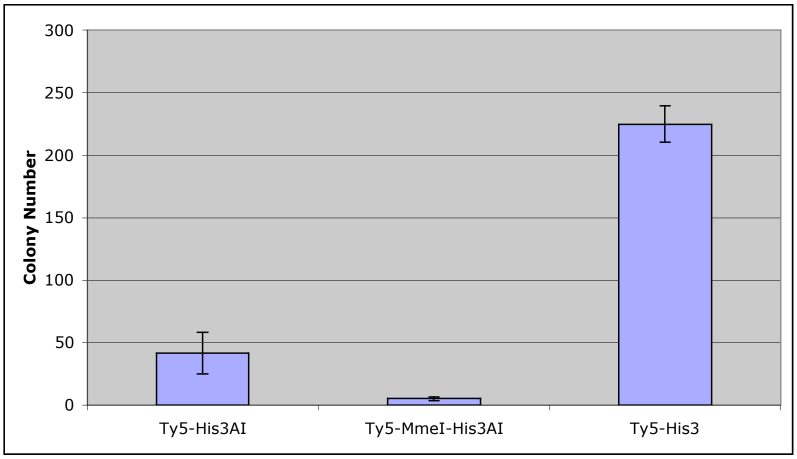


Figure 3.2. Transposition efficiency of different Ty5 constructs. Yeast strain YM7635 was transformed with Gcn4-Sir4 and different Ty5 constructs. Single colony from each transformation was cultured in 1 ml Glu –Trp –Ura media until reaching OD600 at 1. Cells were cultured in 2 ml of Gal –Trp –Ura media at room temperature for 20 hrs. After galatose induction, all cells were plated on YPD plates and then replica to Glu –His +5FOA plates. Colony numbers are counted and plotted. For each construct, mean of three experiments were shown.

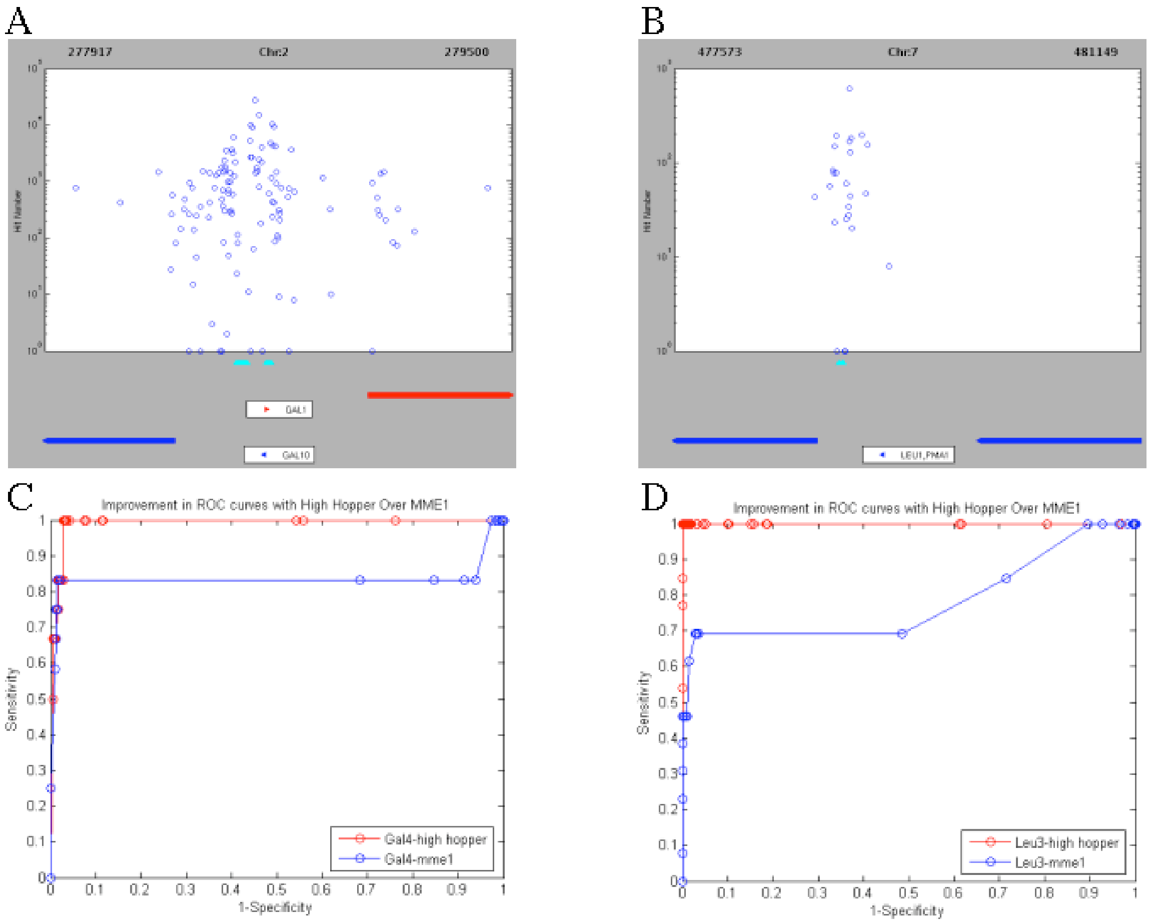


Figure 3.3. Using Calling Cards to map Gal4, Leu3 target genes. (A) In Gal4-Sir4 experiment, Ty5 insertions (indicated by the blue circles) were clustered above known Gal4 binding sites (indicated by the teal triangles). (B) In Leu3-Sir4 experiment, Ty5 insertions were clustered above known Leu3 binding sites (indicated by the teal triangles). (C) ROC curve of Gal4 data produced using MmeI protocol (indicated by the blue line) and Pair-End protocol (indicated by the red line) (described in Chapter 4). (D) ROC curve of Leu3 data produced using MmeI protocol (indicated by the blue line) and Pair-End protocol (indicated by the red line).

Chapter 4: Identification of target genes of multiple yeast transcription factors in a single experiment using the Calling Card method and next generation sequencing

Haoyi Wang, David Mayhew, Xuhua Chen, Mark Johnston, and Robi David Mitra

Department of Genetics, Washington University, School of Medicine, 4444 Forest Park Parkway, St. Louis, MO 63108

In chapter 3, I described experiments demonstrating the feasibility of Calling Card-seq. However, the low hopping efficiency of the engineered Ty5 transposon limited the utility of the method. So, I developed an improved Calling Card-seq protocol compatible with the wild-type Ty5 transposon, which allowed us to recover an order of magnitude more "calling cards" in each experiment. Using this improved protocol, I demonstrated the multiplexed analysis of 8 transcription factors, and show that these results are accurate and reproducible. Minor additional improvements should enable the simultaneous analysis of hundreds of transcription factors.

This work was done in collaboration with David Mayhew, Xuhua Chen, Mark Johnston, and Robi David Mitra. Mark Johnston, Robi Mitra, and I designed the experiments. Xuhua Chen and I performed all the experiments. David Mayhew did all the computational analyses.

ABSTRACT

We describe a method to determine the genomic targets of many transcription factors in a single experiment. We endow DNA-binding proteins with the ability to

direct the insertion of a transposon into the genome near to where they bind. The transposon becomes a “Calling Card” that marks the visit of a DNA-binding protein to the genome. By ”barcoding” transposons with sequence identifiers matched to each DNA-binding protein, every Calling Card is marked with a signature that indicates which protein deposited it in the genome. This enables the simultaneous analysis of multiple DNA-binding proteins. We recover the Calling Cards and determine the sequence of their bar codes and of the flanking genomic DNA by massively-parallel DNA sequencing. To demonstrate the feasibility of this method, we determined the targets of eight transcription factors in a single experiment. This method promises to enable determination of the genomic targets of many transcription factors under many different environmental conditions.

INTRODUCTION

Transcription factors (TFs) regulate gene expression in response to environmental changes and developmental signals. Identification of the target genes for each TF under different conditions is essential for understanding transcriptional regulation. Genome wide chromatin immunoprecipitation on chip (ChIP-chip) experiments have been done to study two hundred and three transcription factors in *Saccharomyces cerevisiae* (Harbison et al. 2004). With extensive filtering and statistical testing and the help of phylogenetic alignment, recognition sequences (motifs) were identified for 98 of these transcription factors (Harbison et al. 2004; MacIsaac et al. 2006). These sequence motifs have not been verified *in vivo* due to the lack of an alternative high-throughput *in vivo* method. More important, the experiments were performed under only a few conditions (Harbison et al.

2004), which could account for the failure to identify target genes and motifs for a significant portion of TFs.

We previously reported a method that employs the retrotransposon Ty5 as a “Calling Card” to mark the visits of TFs to their targets in the genome (Wang et al. 2007). We used Calling Cards to analyze the genome-wide binding of the transcription factors Gal4 and Gcn4, benchmarking our results to those obtained with the ChIP-chip method. Those experiments demonstrated that the method is accurate and specific. Importantly, we were able to identify several targets of Gal4 not revealed by the chromatin-IP method, so Calling Cards provide an orthogonal way to identify the targets of transcription factors (Wang et al. 2007).

The Calling Card method involves fusing to the TF a piece of the Sir4 protein that interacts with the Ty5 transposase. This chimeric protein recruits the Ty5 transposase, which directs integration of the Ty5 transposon into the genome near its binding sites (Wang et al. 2007; Zhu et al. 2003) (Fig 4.1). The transposon Calling Cards are harvested from genomic DNA by digestion with three different restriction endonucleases. The digested DNA is circularized by self-ligation with DNA ligase, then amplified in an inverse PCR, using primers complementary to the transposon sequence. The DNA sequences of the genomic regions immediately flanking the Calling Card are then identified. In our first implementation of the method, we mapped the genome sequences flanking the Calling Cards by hybridization to an oligonucleotide microarray (“Calling Card-chip”). But the great potential of this method for multiplexing incited us to apply “next generation” DNA sequencing to map Calling Card insertion sites and to read out their barcodes (“Calling Card-Seq”). Here we describe the application of this method to map the targets of eight TFs in a single experiment.

RESULTS

Calling Card-seq accurately maps transcription factor binding in vivo

We first applied Calling Card-Seq to three well studied TFs: Gal4, Gcn4, and Leu3. For each TF, we mapped over five thousand independent Ty5 insertions. The genome-wide pattern of Ty5 insertions is dramatically different between yeast strains with and without a TF-Sir4 chimeric protein (Fig 4.2A). Using control data obtained from a strain with no TF-Sir4, we built a null model for the tendency of Ty5 to integrate into each promoter region, which we assumed to be a function of nucleosome occupancy, promoter size, and Ty5 hotspots. We modeled the insertion number at each promoter as a poisson distribution with the null model setting the expectation. The probability of observing a certain number of Ty5 insertions or greater within each promoter, expressed as a P-value, can be calculated for each TF (see Methods for details). We adjusted the stringency with which we identified target genes by using different P-value cutoffs. To determine the sensitivity and specificity of the method, we plotted receiver-operator curves (ROC) (Lusted 1971) for each set of data (see Methods for the definition of positive and negative data sets for each TF) (Fig 4.2B). ROC curves are plots of the sensitivity of the method (how many known targets are identified) versus the false positive rate (or 1 - specificity) for different statistical cutoffs. The area under a receiver operator curve (AUC) provides a measure of the accuracy of a method: an area of 1 indicates the method is perfectly accurate; an area of 0.5 indicates that the method is performing as expected by chance. Each data point in the plot represents the sensitivity and specificity of the data using a different P-value cutoff, from most to least stringent. As shown, in figure 4.2B, the AUC for the Calling Card-Seq method is 0.99, 0.84, and 0.99 for Gal4, Gcn4, and Leu3 respectively, suggesting that the method is highly accurate.

In our initial Calling Card-Seq experiments, we observed that Calling Card insertions were highly enriched around TF binding sites in promoters (Fig 4.2C). Since we can determine the locations of Ty5 transposons with single nucleotide resolution, we characterized the distribution of Ty5 insertions around known protein binding sites. A plot of the frequency of Calling Cards deposited by Gcn4 as a function of distance from known Gcn4-binding sites (Fig 4.2D) reveals that most Gcn4-directed insertions (>60%) occurred within 100 base-pairs of a known Gcn4 binding site. The same pattern is observed for Gal4 and Leu3 (supplemental Fig 4.1). There were strikingly few insertions directly into the binding site (note the sharp dip in the histogram from -5 to +5 bases in Fig 4.2D). Presumably, this is because the transcription factor sterically blocks integration at those nucleotides. The tight distribution of insertion events around binding sites means that the Calling Cards method provides a high resolution map of transcription factor binding. Since a large number of Calling Cards are inserted in the promoters of *bona fide* gene targets, and these insertions are centered around TF binding sites, it is possible to accurately estimate the location of a binding site based on the distribution of Calling Cards (Kharchenko et al. 2008). This makes it relatively straightforward to infer the position specific weight matrix (PSWM) of a transcription factor using Calling Cards data. We searched for a PSWM for each TF by analyzing the DNA sequence flanking

Calling Card insertion sites with the AlignACE algorithm (see Methods for details). Previously known motifs for all three TFs were successfully identified (supplemental Fig 4.2). We conclude that the Calling Card method can be used to determine the recognition sequences of transcription factors at single nucleotide resolution, in addition to identifying *in vivo* gene targets.

We noticed that in many cases Calling Cards were deposited into the promoters of genes adjacent to those to which the TF was bound. For example, in cells expression Gal4-Sir4, Calling Cards are found in the promoter of *FUR4*, the gene immediately downstream of *GAL1* (Fig 4.3). Known Gal4p sites are marked as teal triangles near the x-axis. The red line plots the transcription factor binding potential based on the known Gal4p recognition sequence weight matrix on a $\log_{10}$ scale. This probability was computed using GOMER, an algorithm that uses an explicit equilibrium model to output a binding probability based on DNA sequence and a weight matrix (Granek and Clarke 2005). In Fig 4.3, we see that a large number of Calling Cards are inserted at the *GAL1-10* and *GAL7* promoters, which is expected, since they contain several strong Gal4p sites. However, Calling Cards were also deposited at the promoter of *FUR4*, which contains no known Gal4p sites and is not predicted to bind Gal4p (as indicated by its low GOMER potential). This observation is not an artifact of the Calling Cards method, since ChIP-chip experiments on Gal4 also detected DNA from the *FUR4* promoter (Ren et al. 2000). We have seen this pattern of Calling Cards deposition throughout the genome with several different transcription factors (Supplemental table 4.1). We imagine that Gal4 bound to the *GAL1* promoter makes contact with the transcription apparatus at the *FUR4* promoter, with looping of the intervening DNA. The fact that *FUR4* expression is significantly decreased in a *gal4* mutant is consistent with the possibility that *FUR4* expression is indeed stimulated by Gal4p (Hughes et al. 2000; Ren et al. 2000). This may explain why many adjacent genes in yeast appear to be co-regulated (Cohen et al. 2000).

Calling Card-seq allows for sample multiplexing

The Calling Cards method offers the possibility of multiplexing if unique sequence identifiers ("bar codes") are included in the Ty5. We explored the degree to which the method can be multiplexed with seven TFs whose consensus recognition sequence motifs were not identified in ChIP-chip experiments (Harbison et al. 2004; MacIsaac et al. 2006). Gal4 was included as a positive control. Eight yeast strains, each carrying a different TF-Sir4 fusion paired with a Ty5 transposon carrying a unique 5bp "bar code" were pooled and the Calling Card protocol was performed (Figure 3.4) (see Methods for details). Two "paired-end" sequencing reads were obtained for each recovered Calling Card. The first sequencing read reveals the genome sequence immediately flanking the Calling Card; the second "paired-end" read is of the unique sequence "barcode" that identifies the TF that deposited the Calling Card at that location (Fig 4.1 and Fig 4.4). In this way, we determined where in the genome the Calling Card landed, and which TF put it there. Since a 5bp sequence can encode 1024 different barcodes, there is the potential to analyze tens or even hundreds of TFs in a single experiment (Fig 4.4). From a single lane on an Illumina GAII flowcell we obtained over six million reads, 95% of which contain intact bar codes and which map uniquely in the yeast genome. For seven of the eight TFs we were able to map more than 4,500 independent insertions. (We mapped 1,600 independent Calling Cards deposited by

Rgm1). For the Gal4 control we identified all previously known target genes and deduced the sequence of its binding site. The multiplexed calling card method is highly reproducible, as the correlation of two replicate experiments is extremely high (Fig 4.5A and supplemental Fig 4.3).

We were able to predict recognition sequence motifs for three of the seven poorly characterized TFs (Yrm1, Rgm1, and Sef1). We compared our data with two recent sets of data obtained with protein binding mircroarrays (PBMs) (Badis et al. 2008) (Zhu et al. 2009). The PSWM we predicted for Yrm1 is almost identical to that predicted by these two studies that employed PBM (Fig 4.5B) (Badis et al. 2008; Zhu et al. 2009). For Rgm1, we predicted AGGGGNGGGG (Fig 4.5B), compared to CAGGGG predicted by (Badis et al. 2008) (Zhu et al. (2009) were unable to identify a recognition motif for Rgm1). We believe that CAGGGG is only a half binding-site for this protein. Although Rgm1 binds to this motif *in vitro*, it most likely prefers AGGGGNGGGG *in vivo*. Similar discrepancies between recognition motifs identified *in vitro* and recognition motifs determined *in vivo* are also observed for Hsf1, Sip4, Skn7, Stp4, and Uga3 (Badis et al. 2008; Harbison et al. 2004; MacIsaac et al. 2006). We were able to predict a high information content recognition motif for the zinc-cluster transcription factor Sef1 that is characteristic for such proteins (Fig 4.5B), which was not discovered in the other studies (Badis et al. 2008; Zhu et al. 2009). We were unable to predict a binding sequence with reasonable information content for the remaining four TFs (Kar4, Rpi1, Sfg1, and Lee1), like in previous studies (Badis et al. 2008; Harbison et al. 2004; MacIsaac et al. 2006; Zhu et al. 2009). However, we were able to predict target genes of six out of the seven TFs (Supplemental Table 4.2). We were not able to identify sequence recognition motifs or target genes for Lee1, a putative TF containing a CCCH zinc-finger. This type of zinc-finger, apparent in many proteins, is thought to be an RNA-binding domain (Carballo et al. 1998; Lai et al. 1999), but it has also been reported to be a DNA-binding domain (He et al. 2009). The pattern of deposition of Calling Cards by Lee1 is very similar to that in the no-TF control strain and the correlation of two replicate experiments is poor (Supplemental Fig 4.3), suggesting that Lee1 does not bind to DNA.

We are confident of the target genes we predicted for these six TFs because of the highly clustered insertions of Calling Cards within the promoters of these target genes. The Calling Cards insertions for each TF are dramatically different from each other and from the no-TF control, in a reproducible way (Supplemental Fig 4.3). To verify our target gene predictions, we compared the targets we predicted for Yrm1 to the targets predicted from expression profiling and *in vivo* chromatin IP experiments (Lucau-Danila et al. 2003). 19 of the 23 targets predicted by Danila et al. are found in our target gene list (using p<0.01 cutoff) (supplemental table 4.2). Quite a few of the remaining target genes we predicted are involved in response to drug and chemical stimulus (*YPR036W-A*, *YHK8*, *OYE3*, *AAD3*, *RSB1*, *ZWF1*, *YPP1*), and are therefore likely to be targets of Yrm1. We also looked for enrichment of GO terms for the target genes of Kar4, which is known to be required for gene regulation in response to pheromones (Kurihara et al. 1996; Lahav et al. 2007)). Predicted Kar4 targets genes are highly enriched in conjugation, as expected for a protein involved in karyogamy (11/32, P=1.01E-11), and in reproduction (14/32, P=1.82E-10). Sef1 target genes are highly enriched for acetyl-CoA metabolic processes (1.31E-9). Rpi1 target genes are enriched for cyclin-dependent protein kinase activity (P=2.35E-5) and response to heat shock (P=4.86E-4). Sfg1 target genes are enriched for regulation of cyclin-dependent protein kinase activity (P=1.33E-7). Rgm1 target genes are enriched for carbohydrate metabolic process (P=2.11E-5). All these significantly enriched GO terms agree with the limited information on each TF in SGD, which suggests to us that the target genes we predicted for these TFs are biologically relevant.

These results demonstrate that the Calling Card method is a robust and efficient method for high-throughput analysis of transcription factor binding to DNA.

Calling Card-seq is functional when TF-Sir4 is expressed from native genomic locus

Since many environmental signals directly regulate the transcription of different TFs, overexpressing the fusion protein from a heterologous promoter makes it difficult to measure native TF binding in different conditions. Therefore, I created several yeast strains that expressed TF-Sir4 fusion proteins from the TFs' native genomic loci. We observed a pattern of Calling Card deposition similar to that obtained when the TF-Sir4 was expressed on a plasmid (Fig 4.6), suggesting that native levels of expression is sufficient for the Calling Card method for yeast TFs.

DISCUSSION

We have described a method for mapping targets of DNA-binding proteins that is robust, reliable, accurate, and sensitive to environmental changes. The ability to multiplex tens, possibly hundreds of transcription factors will enable a systematic exploration of transcription factor binding under many different environments and growth conditions in a way that has heretofore not been possible. This proof of principle study identified recognition motifs for three TFs based on *in vivo* data. We were unable to identify any PSWM with reasonable information content for three TFs, but we observed tight clustering of Calling Cards in the promoters of a subset of genes, allowing us to predict the targets of these DNA-binding proteins. These target genes provide clues to the functions of these poorly characterized TFs. How do these TFs recognize specific promoters without revealing an obvious DNA sequence with high information content? Perhaps their recognition sequences are degenerate, or perhaps they rely on specific chromatin structure or some other epigenetic information. Another possibility is that they have partner proteins that determine their specificity, though we would expect to capture the binding motifs of the cofactors.

Our observation of presumed DNA looping events is interesting. We found clusters of Calling Cards in the promoters of genes neighboring those where a TF is bound. One possible explanation of this phenomenon is that TF-Sir4 bound to a promoter recruits a large amount of Ty5 integrase to this region, which causes insertions of Calling Cards in neighboring promoters. However, we observed that Calling Cards were deposited in a small subset of genes neighboring the binding site of the TF, and mostly only to one side of the binding site. For example, Gal4 seems to direct insertion of Calling Cards only to one side of where it binds (in the *FUR4* promoter), and not within the *KAP104* promoter on the other side of the *GAL1-10,7* gene cluster. These observations suggest an alternative model: two neighboring promoters are brought to close proximity through DNA looping.

One possible limitation of the Calling Cards method is that fusing a potion of Sir4 (250 amino acids) to a TF could disrupt its folding, modifications, or DNA-binding ability. We have so far applied this method to more than 20 TFs with several different types of DNA-binding domains, and the method seems to be working well with all of them, suggesting that attachment of Sir4 to a DNA-binding protein seldom incapacitates it. Another potential limitation of this method is that Calling Cards may not be able to be recovered from some genes because they inactivate the gene's promoter. We believe that is unlikely, partly because mutations that abolish promoter function are relatively rare, but mostly because we use diploid cells for our experiments. Indeed, we found that, in no TF-Sir4 experiment, Calling Cards were deposited into promoters of essential genes at the same frequency as they were deposited into the promoters of non-essential genes (average number of insertions per essential gene promoter = .83 +/- 1.8; average number of insertions per non-essential gene promoter = 0.91 +/- 2.0; total number of insertions = 6671; p-value = 0.18). We conclude that there is no restriction in the types of promoters from which Calling Cards can be recovered.

The Calling Card-seq method is relatively easily implemented. The protocol employs general techniques of molecular biology, such as DNA cloning, restriction digest, DNA ligation, and PCR. Its simplicity should allow it to be implemented in most molecular biology laboratories. In our experience, different people obtain highly reproducible data using the method. The method is also flexible. It can be multiplexed, for example to map in a single experiment the genome-wide DNA-binding patterns of one TF in different mutant strains by "bar coding" each strain. In addition, Calling Card-seq is cost-effective. Ten to twenty TFs can be analyzed on a single lane of an Illumina GAII flowcell.

As a high throughput *in vivo* method, Calling Card-seq has several advantages over *in vitro* methods for mapping gene targets of DNA-binding proteins. Calling Card-seq allows genome-wide mapping of protein-DNA interactions in a throughput that makes testing multiple conditions feasible. The *in vivo* DNA-binding data can be used to identify efficiently PSWM and target genes. This is preferable to *in vitro* determination of PSWMs because that do not accurately predict *in vivo* binding sites due to the dynamic nature of chromatin. More important, many TFs bind to different targets under different conditions, and all these protein-DNA interactions must be studied in the relevant environmental context to be able to make biologically relevant conclusions.

For poorly characterized TFs, it is necessary to overexpress the TF-Sir4 fusion from a plasmid, to enable identification of its targets when the activating condition is unknown. We found that expression of TF-Sir4 fusion proteins from their native loci produced high quality DNA-binding data. Our goal is to multiplex all ~200 TFs of yeasts, each as a TF-Sir4 expressed from its native genomic locus, and test many different growth conditions. To be able to achieve this, we will need to make modest improvements in Ty5 transposition efficiency to provide a higher throughput and shorter induction time. Application of our method promises to bring us closer to having a complete list of target genes and sequence recognition motifs of all yeast TFs under many different conditions.

METHODS

Strains and media

All the experiments on Gal4, Gcn4, Leu3, and eight TFs multiplexing were done in diploid yeast strain with *sir4* deletion, YM7635 (*MATa /MATalpha his3Δ1/ his3Δ1 leu2Δ0/ leu2Δ0 ura3Δ0/ ura3Δ0 met15Δ0/MET15 lys2Δ0/LYS2 sir4::Kan/ sir4::Kan* *trp1::Hyg/ trp1::Hyg*) . Haploid strain YM7691 (*MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 sir4::Kan trp1::Hyg GCN4::sir4*) was used in the native expression experiment. Yeasts were grown in complete synthetic media with the addition of 2% glucose or galactose. For amino acid starvation, 10mM 3-AT was added to the media.

Construction of plasmids

All TF-Sir4 fusion constructs were derived from pBM5037 (Gal4DBD-Sir4-Myc) (Wang et al. 2007). The entire ORF of each TF was amplified in a PCR and used to replace Gal4DBD by homologous recombination by cotransformation of yeast cells with Gal4DBD-Sir4-Myc linearized by cleavage with XhoI (cuts once in Gal4DBD coding sequence) for “gap repair” (Ma et al. 1987; Wach et al. 1994).

Ty5 donor plasmid pBM5249 is derived from plasmid pBM5218 (Wang et al. 2008a) (encodes the Ty5 transposon with *URA3* as the selectable marker). The *HIS3*AI marker within Ty5 is exchanged into *HIS3*. A 34 bp sequence containing partial Illumina sequencing primer 2, 5bp bar code 1, and Hinp1I, HpaII, and TaqI recognition sequences were cloned between the FseI and PacI sites located between the 3’ LTR and the *HIS3* marker. All other bar coded Calling Cards were derived from pBM5249.

Induction of Ty5 transposition and inverse PCR

For multiplexing experiments, each strain transformed with one TF-Sir4 construct and a uniquely bar coded Calling Card were grown to saturation individually in 5 ml Glu –Trp –His media. Cultures of all eight strains were pooled and plated on 50 Gal –Trp –His plates and incubated for 3 days at room temperature to induce Ty5 transposition.

After induction, cells were replica-plated to YPD media and grown for one day to allow cells to lose the Ty5 donor plasmid. Cells were then serially replica plated onto –His, FOA-containing media twice to select for cells containing Calling Cards in their genome. To map the locations of the Calling Cards in the genome, all His^{+} FOA^{r} colonies were harvested and their genomic DNA extracted. Each DNA sample was divided into three aliquots, each digested by Hinp1I or HpaII or TaqI individually. Digested DNA was then ligated overnight at 15°C in dilute solution to encourage self-circularization. After ethanol precipitation, self-ligated DNA was resuspended in ddH_2O and used as template in an inverse PCR. Primers that anneal to Ty5 sequences (OM6313 and OM6188) were used to amplify the genomic regions flanking Ty5 integrations and the bar codes within Ty5, as well as adding adapter sequences that allow the PCR products to be sequenced on the Illumina GA analyzer. The PCR products were purified using the QIAquick PCR Purification Kit (Qiagen) and diluted to 10nM. For each sample, the same amount of PCR product from digestion with each restriction endonuclease was pooled and submitted for Illumina sequencing.

Mapping Paired-end sequences to the genome

Sequence reads were filtered by requiring a correct LTR sequence in the first paired read and an appropriate barcode and digestion site in the second paired read. The genomic fragments of both reads were mapped using a hash table of all possible 16 bp sequences from the yeast genome. Reads in which a combination of both paired end reads could uniquely locate the site of insertion were passed as correct. Independent insertions were required to have at least 10 reads to be considered real.

Target gene calling and motif finding

Promoter size was defined as the 1000 bp 5’ of the transcription start site of a gene, or until the next gene on either strand, but with a minimum size of 200 bp. The average number of insertions from triplicate experiments with no transcription factor, plus pseudocounts, was used to create a null model for the tendency of Ty5 to insert in a specific promoter. Insertions at each promoter were modeled with a Poisson distribution, with the values from the null model (scaled by the ratio of total insertions between experiments) used as the expected value for the individual Poisson distribution. P-values were assigned by calculating the cumulative distribution function for the number of insertions at each promoter for the transcription factor directed Ty5 insertions.

Clusters of insertions were determined by integrating windows of 50 bp in both directions from every insertion and creating a cutoff for the minimum number of overlaps required. An initial cutoff for number of overlapping insertion frames was determined by requiring that every promoter with a p-value less than 0.001 have at least one cluster. The sequences corresponding to these clusters of insertions were then searched for overrepresented sequences using AlignACE. The cutoff would subsequently be adjusted both up and down to see if higher cutoffs would converge to a single motif or lower cutoffs would reveal a new motif. Binding potentials for the five highest information content motifs from AlignACE were generated using GOMER across the genome. The motif that most accurately predicted binding as determined by area under receiver operator curve was selected. If no statistically significant motif was identified by AlignACE or no motif from AlignACE could accurately predict transcription factor binding, then no motif was called for that transcription factor.

Calculation of sensitivity and specificity of Calling Cards

To calculate sensitivity, positive lists for known targets were created from all genes that had at least one known binding site for that transcription factor listed from SGD in their promoter. To calculate specificity, negative lists were generated for Gal4, Gcn4 and Leu3. 900 genes were randomly selected from a list that contained genes whose promoters (1) had a p-value greater than 0.05 in Harbison's data set (Harbison et al. 2004), (2) were not within 2 genes in either direction of a strong target (p-value < 0.0001 in Harbison's data set), and (3) did not contain strong or weak binding sites for the known PWM of that transcription factor as defined by being in the lower half of all promoters as ranked by GOMER with the default Gaussian parameters.

ACKNOWLEDGEMENTS

We are especially grateful to Jessica Hoisington-Lopez, Laura Langton, Brian Koebbe, and Jim Dover (Washington University) for their expert assistance with the use of Illumina Genome Analyzer. This work was supported by NIH grants R21RR023960 and 5P50HG003170-03, and funds provided by the James S. McDonnell Foundation.

FIGURES

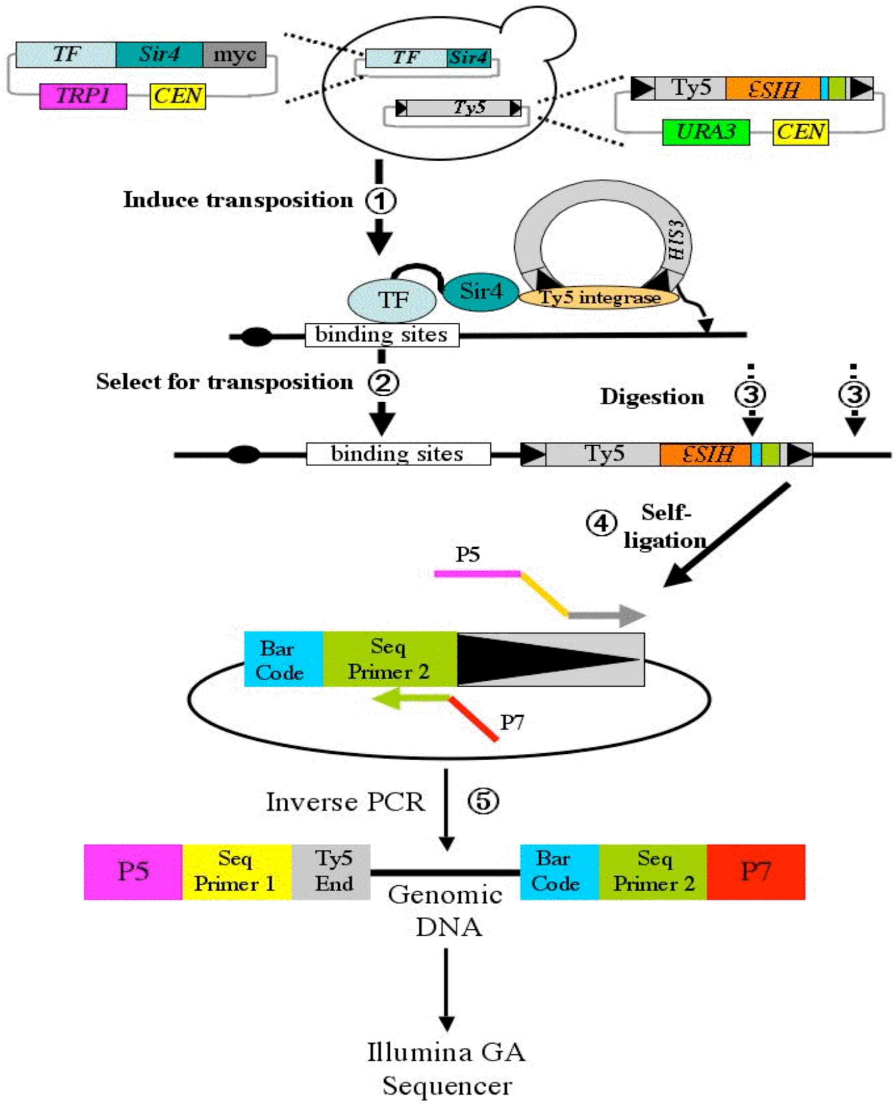


Figure 4.1. Calling Card-seq method. Sir4 fused to a DNA-binding protein causes Ty5 to integrate into the genome near the binding sites for that transcription factor (TF). After ① Ty5 transposition, cells that have undergone Ty5 transposition are selected ②. Genomic DNA is cleaved with restriction enzymes that cut near the end of Ty5 ③ and ligated in dilute solution to favor recircularization of the fragments ④. This is followed by amplification of the circular DNA that contains the end of the transposon and flanking genomic DNA by an "inverse PCR" ⑤ (PCR primers containing Illumina sequencing primers and adaptors). The identity of inverse-PCR products is then determined by Illumina sequencing.

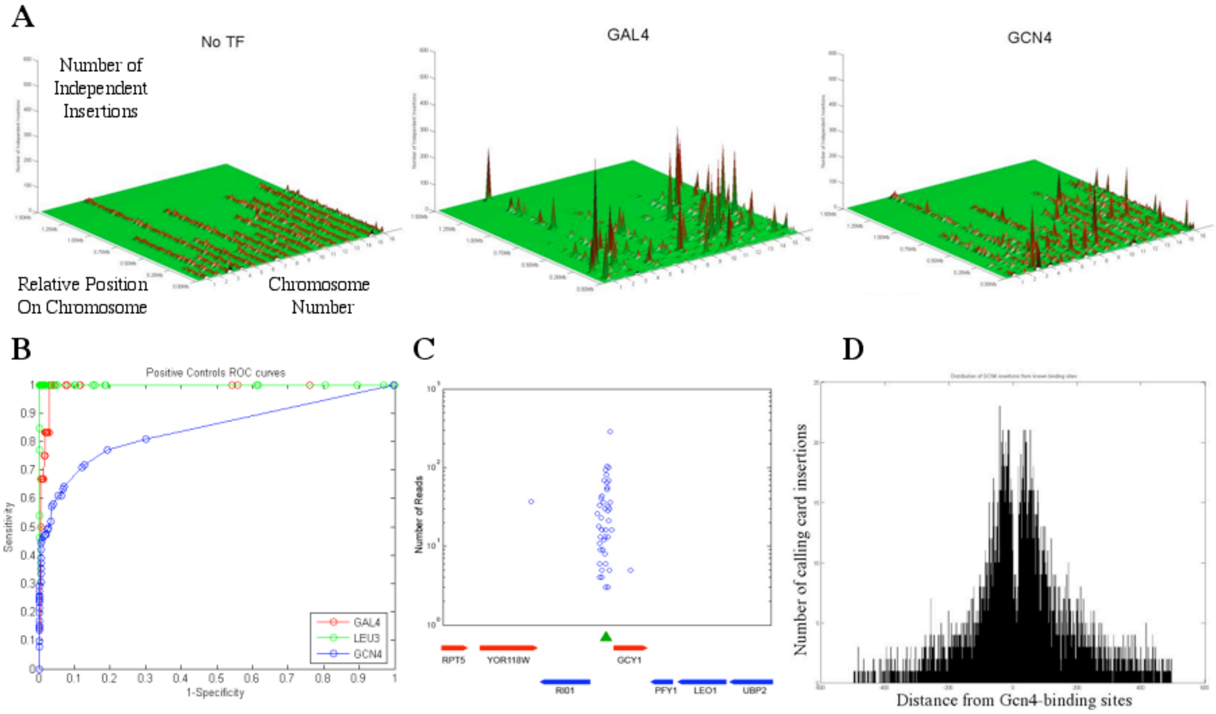


Figure 4.2. Calling Card-seq is accurate in predicting target genes and binding motifs. (A) Genome-wide Ty5 insertion patterns of no-TF control, Gal4, and Gcn4. (B) ROC curve of Gal4 (red), Leu3 (green), and Gcn4 (blue) data. (C) Ty5 integrations are enriched around Gal4 binding sites in *GCY1* promoter (indicated by the green triangle). (D) The distribution of Gcn4-directed Ty5 insertions around known Gcn4p binding sites is plotted. The x-axis specifies the distance from a known binding site, and the y-axis gives the number of insertion events.

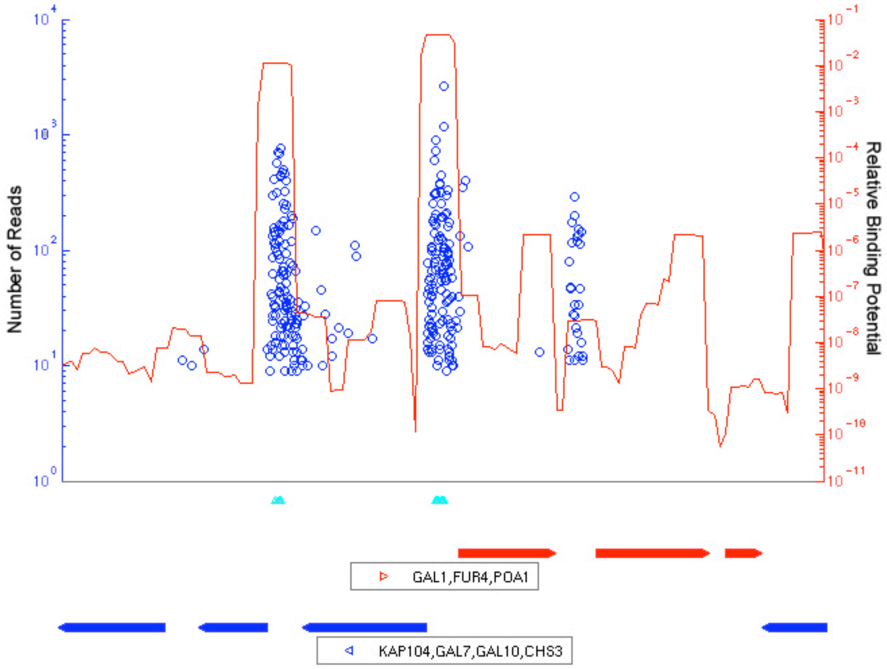


Figure 4.3. Putative Looping between the *GAL1* promoter and the *FUR4* promoter. Gal4p binds in the *GAL1* promoter (Gal4 binding sites are indicated by teal triangles) and also at the *FUR4* promoter, despite the fact that the thermodynamic binding potential (Granek and Clarke 2005) of the *FUR4* promoter is not high (red curve). The observed binding may be due to a looping event that brings the *FUR4* promoter in close proximity to the *GAL1* promoter.

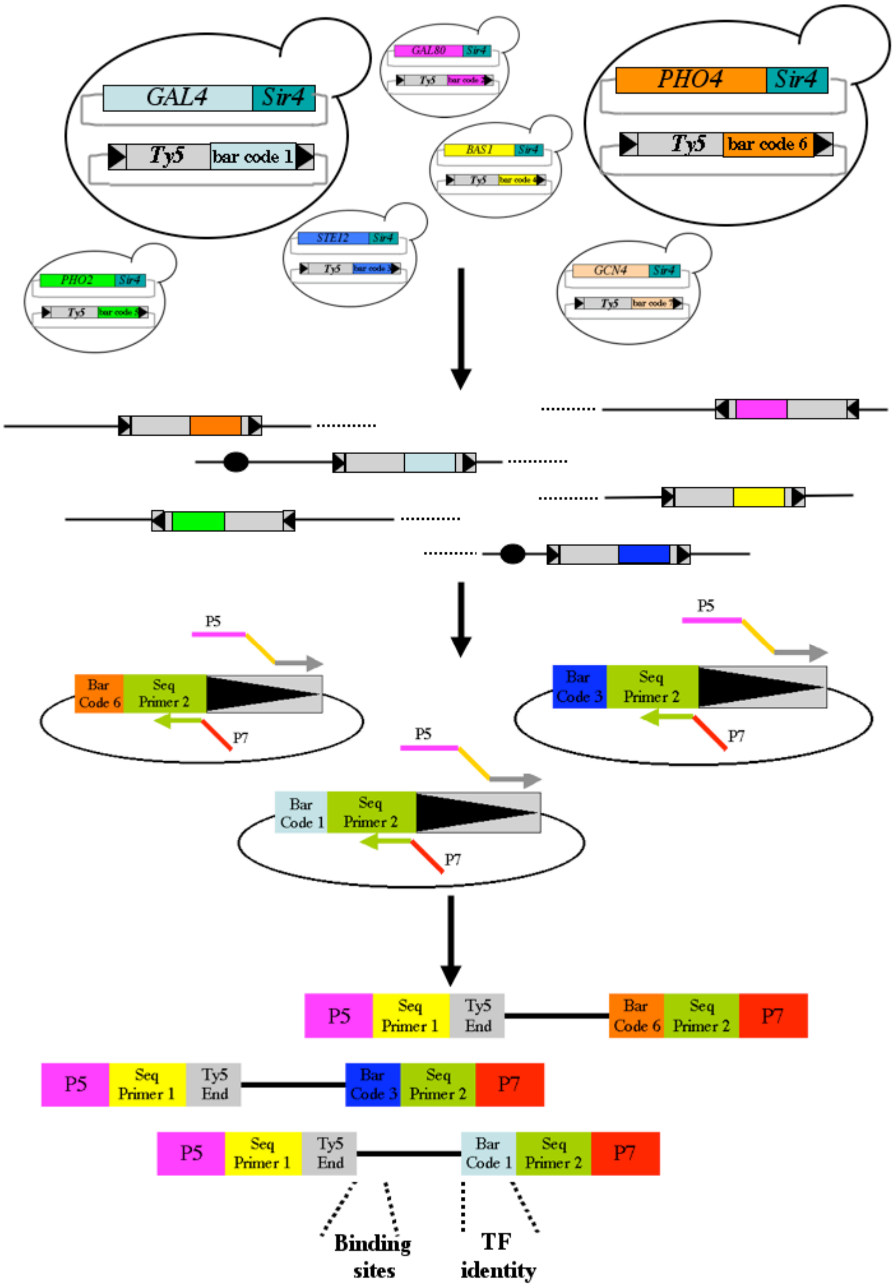


Figure 4.4. Analyzing multiple TFs in one experiment: for each of eight TFs fused to Sir4, a unique 5 bp sequence was cloned into Ty5 to serve as a “molecular bar code”. Each strain was co-transformed with a plasmid encoding a TF-Sir4 fusion and a plasmid carrying its matched bar-coded Ty5 calling card. After transposition, the calling cards were recovered by inverse PCR and sequenced on Illumina GA sequencer with pair-end module. For each paired sequence, we identify Calling Card insertion site and the TF who deposited it there.

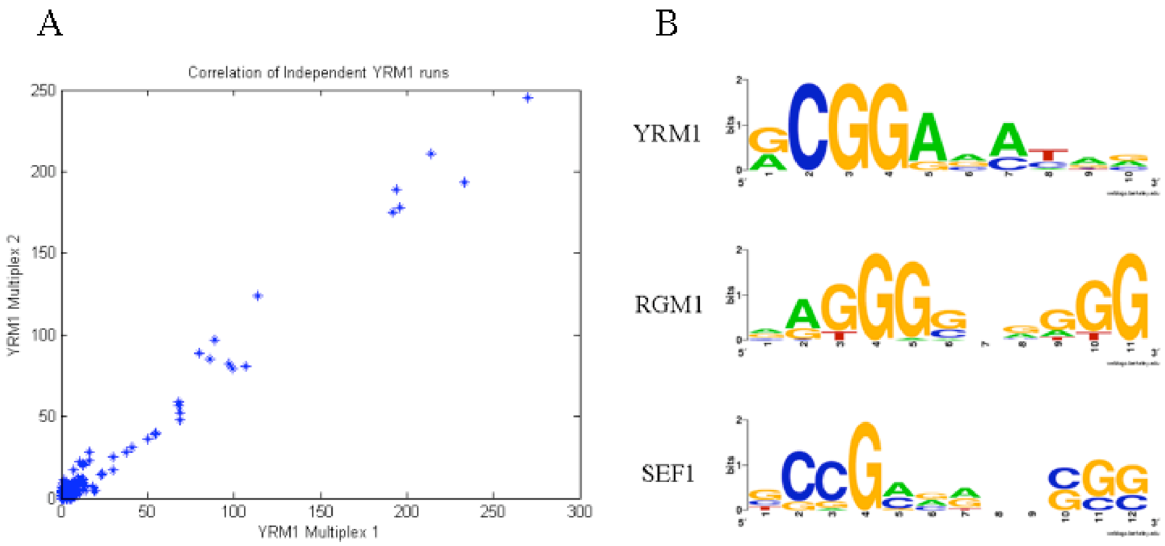


Figure 4.5. Multiplexing experiments are reproducible and productive. (A) Number of independent Calling Cards insertions within each promoter is plotted for two biological replicate experiments multiplexing eight TFs. Data of Yrm1 is shown here. (B) Sequence logos for newly discovered TF binding site motifs.

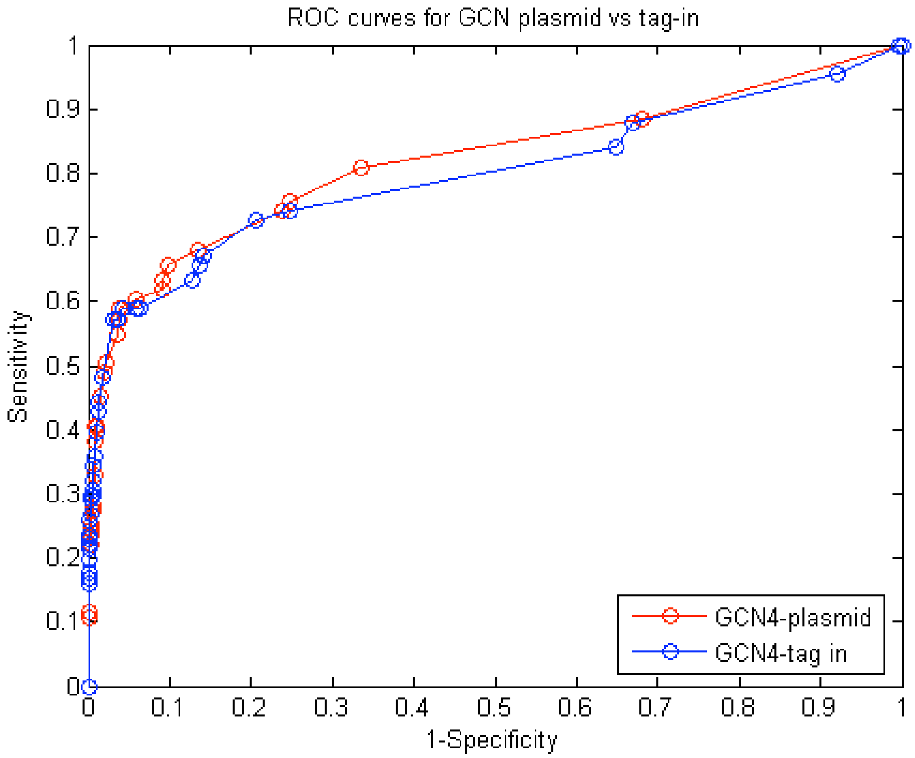
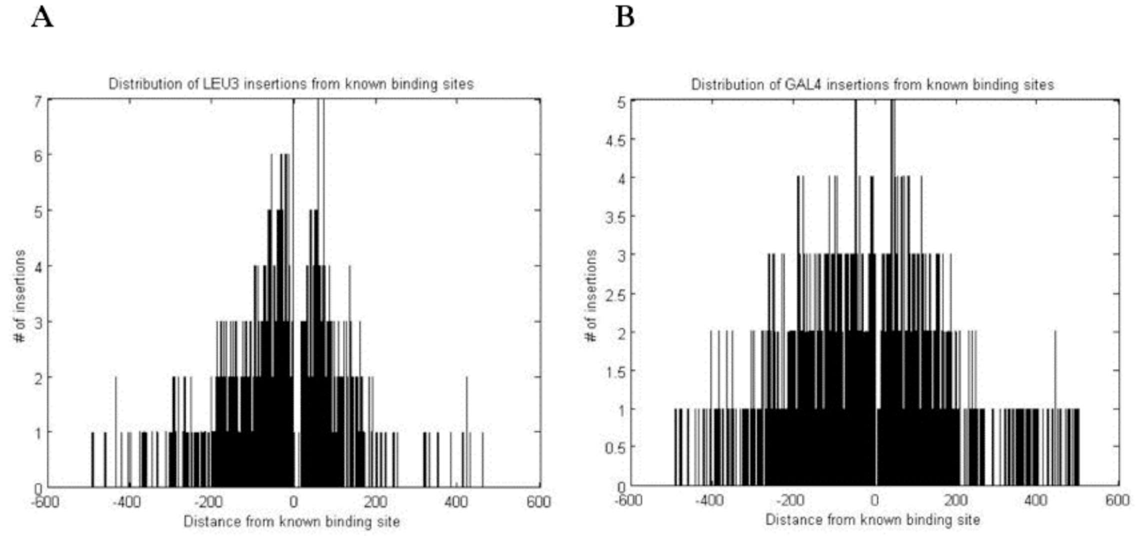


Figure 4.6. Calling Card-seq is fully functional when TF-Sir4 is expressed from native genomic locus. ROC curves are plotted for the Calling Cards data when Gcn4 is expressed from *ADH1* promoter on plasmid (red) and from native genomic locus (blue).

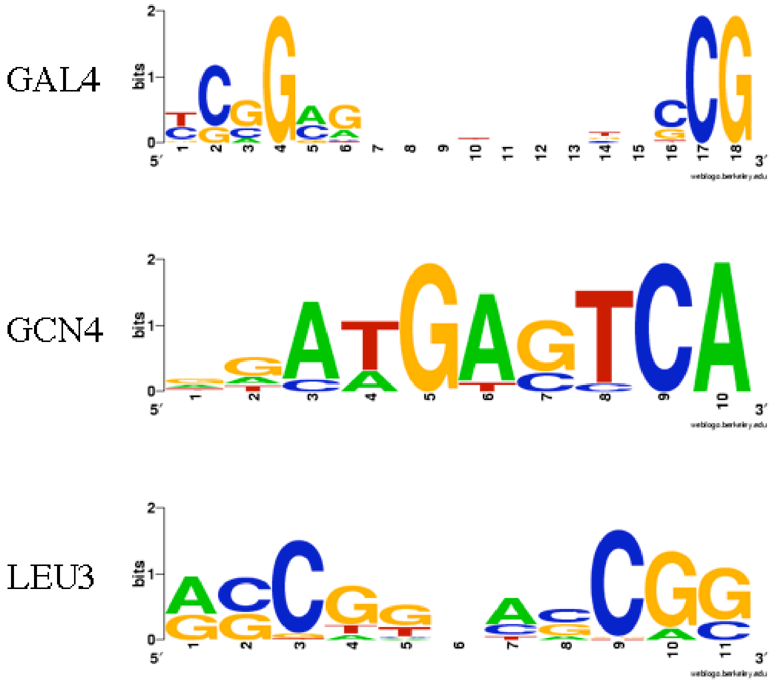
SUPPLEMENTAL FIGURES AND TABLES



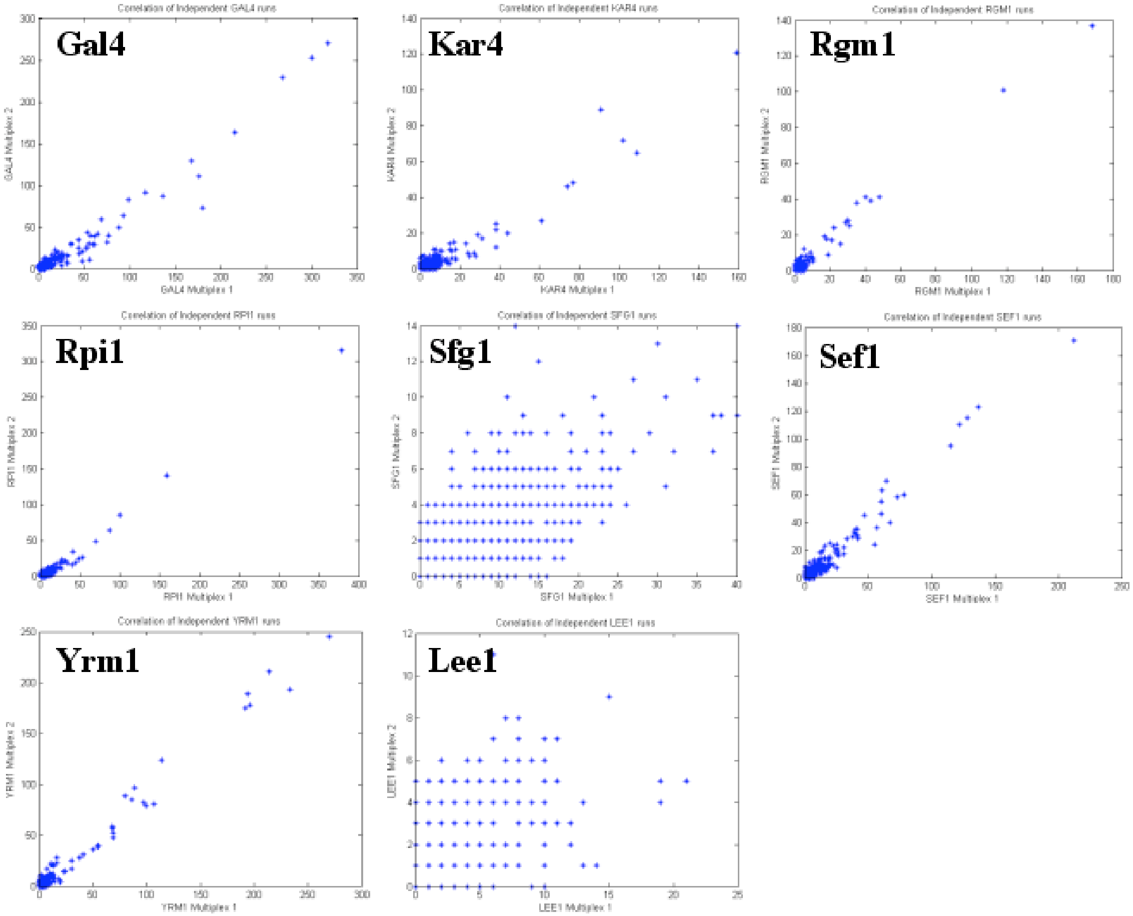
Distribution of LEU3 insertions from known binding sites

Distribution of GAL4 insertions from known binding sites

Supplemental Figure 4.1. The frequency of Calling Card insertions is plotted as a function of distance from known binding sites of Leu3 (A) and Gal4 (B). More than 60% TF-directed insertions occur within 100 base-pairs of known TF binding sites.



Supplemental Figure 4.2. Sequence logos for binding site motifs of Gal4, Gcn4, and Leu3 identified using Calling Card method.



Supplemental Figure 4.3. Number of independent Calling Cards insertions within each promoter is plotted for two biological replicates multiplexing eight TFs.

Supplemental Table 4.1. Potential DNA looping events identified by Calling Cards.

GAL4 Loopers		
Looping Target		Source
YBR021W	FUR4	GAL1/10
GCN4 Loopers		
Looping Target		Source
YOL059W	GPD2	ARG1
YHR020W	YHR020W	ARG4
YHR019C	DED81	ARG4
YER064C	YER064C	ARG5
YER073W	ALD5	ARG5
YOL142W	RRP40	ARG8
YOL143C	RIB4	ARG8
YDR123C	INO2	ARO1
YBR066C	NRG2	BAP2
YBR067C	TIP1	BAP2
YBR069C	TAT1	BAP2
YER088C	DOT6	TRP2
YER054C	GIP2	HIS1
YER056C	FCY2	HIS1
YHR163W	SOL3	YHR162W
YIL118W	RHO3	HIS5
YLR354C	TAL1	ILV5
YIL053W	RHR2	VHR1
YPR035W	GLN1	YPR036W-A
LEU3 Loopers		
Looping Target		Source
YGL008C	PMA1	LEU1
YBR069C	TAT1	BAP2

Supplemental Table 4.2. Target genes identified for each TF using P<0.001 cutoff.

Kar4

Gene	Gene	KAR4_Exp	KAR4_Ins	KAR4_Pval
YCL027W	FUS1	2.03	38	0.00E+00
YCR089W	FIG2	2.7	159	0.00E+00
YDL127W	PCL2	3.39	91	0.00E+00
YDR085C	AFR1	6.77	74	0.00E+00
YDR309C	GIC2	8.13	44	0.00E+00
YHR084W	STE12	2.03	29	0.00E+00
S000006534	tRNA-Asp	5.43	38	0.00E+00
YNL283C	WSC2	5.43	61	0.00E+00
YNL282W	POP3	5.43	61	0.00E+00
YNL280C	ERG24	6.1	102	0.00E+00
YNL279W	PRM1	6.1	102	0.00E+00
S000006612	tRNA-Ile	8.13	77	0.00E+00
YCR009C	RVS161	4.06	38	1.11E-16
YCL075W	YCL075W	26.41	109	1.22E-15
YKL127W	PGM1	4.73	23	2.80E-10
YPL156C	PRM4	2.7	16	4.84E-09
YNL301C	RPL18B	7.46	27	6.87E-09
S000006678	tRNA-Pro	6.77	25	1.47E-08
YNL300W	YNL300W	6.77	24	5.73E-08
YOR344C	TYE7	8.8	28	5.87E-08
YNL192W	CHS1	4.06	18	6.56E-08
YCR084C	TUP1	4.06	17	3.11E-07
YCR086W	CSM1	4.06	17	3.11E-07
YCR087C-A	LUG1	3.39	15	6.18E-07
YCR088W	ABP1	3.39	15	6.18E-07
YLR332W	MID2	3.39	15	6.18E-07
S000006685	tRNA-Pro	3.39	15	6.18E-07
YMR232W	FUS2	2.03	10	9.61E-06
YBR083W	TEC1	14.22	31	3.44E-05
YBR077C	SLM4	6.77	18	8.46E-05
YBR078W	ECM33	6.77	18	8.46E-05
S000006711	tRNA-Arg	2.7	10	1.21E-04
S000006596	tRNA-His	2.03	8	2.66E-04
S000006713	tRNA-Arg	2.03	8	2.66E-04
S000006653	tRNA-Leu	2.03	8	2.66E-04
YPR120C	CLB5	2.03	8	2.66E-04
YMR197C	VTI1	2.7	9	5.05E-04
YMR198W	CIK1	2.7	9	5.05E-04
YPR141C	KAR3	2.7	9	5.05E-04
YPR143W	RRP15	2.7	9	5.05E-04
YPR165W	RHO1	2.7	9	5.05E-04

Rgm1

Gene	Gene	RGM1_Exp	RGM1_Ins	RGM1_Pval
YBL029W	YBL029W	1.25	31	0.00E+00
YPL061W	ALD6	4.12	40	0.00E+00
YIL119C	RPI1	8.95	168	0.00E+00
YNL097C-B	YNL097C-B	5.54	48	0.00E+00
YKL063C	YKL063C	1.61	22	1.11E-16
YBL029C-A	YBL029C-A	1.61	30	1.11E-16
YKL062W	MSN4	1.61	29	1.11E-16
YMR105C	PGM2	1.97	43	1.11E-16
YMR105W-A	YMR105W-A	1.97	43	1.11E-16
YBR126C	TPS1	1.07	19	2.22E-16
YFL052W	YFL052W	0.9	35	2.22E-16
YIL118W	RHO3	5.37	118	6.66E-16
YDR096W	GIS1	3.76	26	7.77E-15
YEL069C	HXT13	1.61	17	1.80E-13
YEL007W	YEL007W	2.68	21	1.85E-13
YFR053C	HXK1	1.97	18	4.97E-13
YEL070W	DSF1	0.54	8	6.28E-09
YDR216W	ADR1	1.25	10	9.34E-08
YGR249W	MGA1	1.25	10	9.34E-08
YOR028C	CIN5	0.9	7	4.67E-06
YDR540C	IRC4	0.9	7	4.67E-06
YDL079C	MRK1	2.15	10	1.58E-05
YDR134C	YDR134C	1.43	8	1.95E-05
YMR244C-A	YMR244C-A	2.51	9	2.83E-04
YMR246W	FAA4	2.51	9	2.83E-04
YOR298C-A	MBF1	1.07	5	8.54E-04
YOR299W	BUD7	1.07	5	8.54E-04
YGR287C	YGR287C	1.07	5	8.54E-04
YHL029C	OCA5	1.07	5	8.54E-04
YHL028W	WSC4	1.07	5	8.54E-04
YLR205C	HMX1	0.71	4	8.56E-04
YLR206W	ENT2	0.71	4	8.56E-04
YKL037W	YKL037W	0.71	4	8.56E-04

Rpi1

Gene	Gene	RPI1_Exp	RPI1_Ins	RPI1_Pval
YEL007W	YEL007W	9.1	378	0.00E+00
YHL029C	OCA5	3.64	159	0.00E+00
YHL028W	WSC4	3.64	159	0.00E+00
YMR194C-B	YMR194C-B	8.5	87	0.00E+00
YMR195W	ICY1	7.88	70	0.00E+00
YBR157C	ICS2	3.64	35	0.00E+00
YGR146C-A	YGR146C-A	4.24	40	0.00E+00
YDR077W	SED1	7.28	41	0.00E+00
YOL031C	SIL1	4.24	100	0.00E+00
YOL030W	GAS5	4.24	100	0.00E+00
YDR186C	YDR186C	4.86	45	0.00E+00
YDR188W	CCT6	4.86	45	0.00E+00
YNL289W	PCL1	3.04	26	2.22E-16
YDR247W	VHS1	3.04	53	2.22E-16
YJR025C	BNA1	3.04	24	4.33E-15
YEL009C	GCN4	1.82	18	1.29E-13
YER088C	DOT6	14.57	48	1.10E-12
YGR108W	CLB1	1.82	17	1.35E-12
YIL123W	SIM1	3.04	20	1.51E-11
YGR250C	YGR250C	7.88	32	2.21E-11
YMR205C	PFK2	4.86	24	9.05E-11
YMR206W	YMR206W	4.86	24	9.05E-11
YKL096W	CWP1	4.24	22	1.84E-10
YGR032W	GSC2	6.06	26	3.70E-10
YGL038C	OCH1	5.46	24	9.40E-10
YGR205W	YGR205W	2.42	15	6.92E-09
YER045C	ACA1	3.64	17	6.41E-08
YAL040C	CLN3	2.42	13	2.89E-07
YPL032C	SVL3	4.24	16	2.46E-06
YJR145C	RPS4A	12.15	30	4.26E-06
YJR147W	HMS2	12.15	30	4.26E-06
YER046W	SPO73	5.46	18	4.89E-06
YMR121C	RPL15B	10.93	27	1.12E-05
YMR122W-A	YMR122W-A	10.93	27	1.12E-05
YPR008W	HAA1	3.04	11	8.09E-05
YOR043W	WHI2	3.04	11	8.09E-05
YBR077C	SLM4	6.06	16	1.97E-04
YBR078W	ECM33	6.06	16	1.97E-04
YMR173W	DDR48	2.42	9	2.16E-04
YBR007C	DSF2	4.24	12	4.72E-04
YCR005C	CIT2	7.28	17	5.69E-04
YOR119C	RIO1	1.82	7	6.05E-04
YOR120W	GCY1	1.82	7	6.05E-04
S000006631	tRNA-Lys	2.42	8	9.17E-04

Sef1

Gene	Gene	SEF1_Exp	SEF1_Ins	SEF1_Pval
YDR216W	ADR1	3.86	40	0.00E+00
YER056C	FCY2	2.2	23	0.00E+00
YDR276C	PMP3	4.96	39	0.00E+00
YMR145C	NDE1	7.17	122	0.00E+00
YPR063C	YPR063C	5.51	73	0.00E+00
YKL217W	JEN1	9.38	57	0.00E+00
YLR153C	ACS2	2.76	31	0.00E+00
YKL110C	KTI12	8.27	47	0.00E+00
YKL085W	MDH1	2.2	24	0.00E+00
YGR067C	YGR067C	2.2	60	0.00E+00
YPR065W	ROX1	4.96	78	0.00E+00
YLR304C	ACO1	3.86	115	0.00E+00
YNR001C	CIT1	6.07	41	0.00E+00
YMR070W	MOT3	6.07	40	0.00E+00
YLL028W	TPO1	8.27	42	0.00E+00
YOR136W	IDH2	2.76	60	0.00E+00
YKL109W	HAP4	6.07	212	0.00E+00
YLR055C	SPT8	4.96	42	0.00E+00
YLR056W	ERG3	4.96	42	0.00E+00
YLR174W	IDP2	2.2	27	0.00E+00
YPR148C	YPR148C	9.38	64	0.00E+00
YMR102C	YMR102C	11.04	128	0.00E+00
YPR149W	NCE102	8.82	61	0.00E+00
YLR297W	YLR297W	6.62	67	1.11E-16
YPL058C	PDR12	12.69	137	2.22E-15
YPL263C	KEL3	4.42	27	5.43E-14
YPL262W	FUM1	4.42	27	5.43E-14
YCR024C-A	PMP1	7.17	33	3.92E-13
YNL118C	DCP2	4.42	25	2.13E-12
YNL117W	MLS1	4.42	25	2.13E-12
YDR275W	BSC2	9.93	38	2.40E-12
YPR036W-A	YPR036W-A	19.86	55	2.47E-11
YBL043W	ECM13	3.31	20	6.92E-11
YDR524C-B	YDR524C-B	8.27	31	3.00E-10
YDR525W-A	SNA2	8.27	31	3.00E-10
YNL037C	IDH1	3.31	19	4.43E-10
YNL036W	NCE103	3.31	19	4.43E-10
YML091C	RPM2	6.07	25	1.72E-09
YCR005C	CIT2	6.62	25	9.58E-09
YGR250C	YGR250C	7.17	26	1.17E-08
YOR043W	WHI2	2.76	15	4.16E-08
YOL126C	MDH2	7.73	25	1.87E-07
YOL125W	TRM13	7.73	25	1.87E-07
YPL202C	AFT2	2.76	14	2.44E-07
YDR508C	GNP1	5.51	20	3.86E-07

YDR510W	SMT3	5.51	20	3.86E-07
YML100W-A	YML100W-A	3.31	15	4.48E-07
YDL174C	DLD1	2.76	13	1.34E-06
YDL173W	YDL173W	2.76	13	1.34E-06
YBR069C	TAT1	6.07	20	1.75E-06
YGL045W	RIM8	3.31	14	2.20E-06
YDR536W	STL1	4.42	16	4.16E-06
YOR274W	MOD5	6.62	20	6.40E-06
YOR119C	RIO1	1.65	9	9.52E-06
YOR120W	GCY1	1.65	9	9.52E-06
YKL096W	CWP1	3.86	14	1.31E-05
YER053C	PIC2	2.76	11	3.30E-05
YOR084W	YOR084W	2.76	11	3.30E-05
YLL027W	ISA1	3.31	12	4.37E-05
YJL116C	NCA3	3.86	13	5.21E-05
YHL007C	STE20	1.65	8	5.85E-05
YOR315W	SFG1	4.96	15	6.36E-05
YNL160W	YGP1	9.38	22	1.22E-04
YLR295C	ATP14	8.27	20	1.48E-04
YNL241C	ZWF1	7.73	19	1.63E-04
YDL047W	SIT4	7.17	18	1.74E-04
YDR033W	MRH1	3.86	12	1.94E-04
YML101C	CUE4	3.86	12	1.94E-04
YGL056C	SDS23	6.07	16	2.01E-04
YGL055W	OLE1	6.07	16	2.01E-04
YJL115W	ASF1	4.96	14	2.10E-04
YDL048C	STP4	8.82	20	3.40E-04
YPR157W	YPR157W	8.27	19	3.85E-04
YNR016C	ACC1	7.17	17	4.73E-04
YFR034C	PHO4	2.76	9	5.98E-04
YMR199W	CLN1	3.31	10	6.53E-04
YMR318C	ADH6	4.42	12	6.83E-04

Yrm1

Gene	Gene	YRM1_Exp	YRM1_Ins	YRM1_Pval
YDR073W	SNF11	6.05	68	0.00E+00
YDR072C	IPT1	6.05	68	0.00E+00
YOR274W	MOD5	6.05	54	0.00E+00
YLR354C	TAL1	6.55	38	0.00E+00
YNL231C	PDR16	6.05	99	0.00E+00
YLR046C	YLR046C	1.51	24	0.00E+00
YPL171C	OYE3	4.54	80	0.00E+00
YPL170W	DAP1	4.54	80	0.00E+00
YGR223C	HSV2	4.54	270	0.00E+00
YGR224W	AZR1	4.54	270	0.00E+00
YMR102C	YMR102C	10.1	194	0.00E+00
YIL056W	VHR1	23.2	214	0.00E+00
YPL137C	GIP3	7.57	69	0.00E+00
YPL135W	ISU1	7.57	69	0.00E+00
YBR150C	TBS1	5.04	50	0.00E+00
YBR151W	APD1	5.04	50	0.00E+00
YKL052C	ASK1	13.62	233	0.00E+00
YKL051W	SFK1	13.62	233	0.00E+00
YDR011W	SNQ2	2.53	69	1.11E-16
YGR197C	SNG1	3.53	114	1.11E-16
YGR198W	YPP1	3.53	114	1.11E-16
YOR064C	YNG1	2.53	41	1.11E-16
YOR065W	CYT1	2.53	41	1.11E-16
YOR273C	TPO4	8.07	55	1.11E-16
YHR048W	YHR048W	2.53	30	1.11E-16
YDR061W	YDR061W	2.53	89	1.11E-16
YGR281W	YOR1	8.07	196	1.11E-16
YOR342C	YOR342C	9.08	97	1.33E-15
YGR280C	PXR1	9.08	192	1.33E-15
YPR036W-A	YPR036W-A	18.16	107	2.66E-15
YOR348C	PUT4	25.73	86	4.55E-15
YOR349W	CIN1	25.73	86	4.55E-15
YLL056C	YLL056C	3.03	23	3.10E-14
YLR191W	PEX13	1.51	16	7.75E-13
YGR035C	YGR035C	1.51	16	7.75E-13
YDR247W	VHS1	2.53	19	4.23E-12
YNL241C	ZWF1	7.07	30	2.83E-11
YEL029C	BUD16	1.51	12	8.67E-09
YCR107W	AAD3	2.53	13	4.78E-07
YBR149W	ARA1	2.53	13	4.78E-07
YGR279C	SCW4	3.03	14	7.48E-07
YCR005C	CIT2	6.05	20	1.67E-06
YDR129C	SAC6	1.51	9	4.43E-06
S000006699	tRNA-Gln	3.53	13	2.02E-05
YMR008C	PLB1	2.53	10	6.79E-05

YKL053C-A	MDM35	2.01	8	2.49E-04
YGR221C	TOS2	3.03	10	3.15E-04
YGR222W	PET54	3.03	10	3.15E-04
YHR028C	DAP2	1.51	6	9.75E-04
YMR220W	ERG8	1.51	6	9.75E-04

Sfg1

Gene	Gene	SFG1_Exp	SFG1_Ins	SFG1_Pval
YIL123W	SIM1	7.82	26	6.72E-08
YDR073W	SNF11	18.74	40	5.89E-06
YDL127W	PCL2	7.82	21	2.41E-05
YDR072C	IPT1	18.74	37	6.08E-05
YAL040C	CLN3	6.23	17	9.06E-05
YOR119C	RIO1	4.68	14	1.14E-04
YOR120W	GCY1	4.68	14	1.14E-04
YGR108W	CLB1	4.68	14	1.14E-04
YMR199W	CLN1	9.37	22	1.19E-04
YMR135C	GID8	17.19	32	4.53E-04
YER064C	YER064C	23.42	40	6.30E-04
YKL096W	CWP1	10.92	22	9.44E-04

Chapter 5: "Calling Cards" for DNA-binding proteins in mammalian cells

Haoyi Wang, David Mayhew, Xuhua Chen, Mark Johnston, and Robi David Mitra

Department of Genetics, Washington University, School of Medicine, 4444 Forest Park Parkway, St. Louis, MO 63108

"Calling Cards" are permanent marks in the genome that record TF binding events. This feature makes them especially useful for studying development of mammalian cells, since the Calling Cards can be used to record protein-DNA interactions in progenitor cells, and the information can be recovered after differentiation. I tested the feasibility of employing the *piggyBac* transposon as a Calling Card in a human cell line. Here I report my preliminary data and propose ways to overcome potential technical difficulties I identified.

I designed the experiments in collaboration with Mark Johnston and Rob Mitra. Xuhua Chen and I performed all the experiments. David Mayhew did all the computational analyses.

ABSTRACT

Transcription regulation is central to mammalian development. Mapping the transcriptional networks that control cell fate decisions has been difficult because no method is available for recording transcription factor binding events throughout cellular lineages. This makes it nearly impossible to correlate transcription factor binding events in progenitor cells to the final fates of their progeny cells. To ameliorate this situation I

endowed transcription factors with the ability to mark the places in the genome they visit with a transposon “Calling Card”. Here I show that this method enables precise mapping of target genes of the transcription factors TP53 and REST of humans. This method will allow us to trace transcription factor binding throughout cellular and organismal development to dissect gene expression networks that control cell-fate commitment in a way that has heretofore not been possible.

INTRODUCTION

Much of mammalian development is transcriptionally regulated. Consequently, considerable effort has been expended on understanding the gene expression networks that control cell division, differentiation, and migration. But mapping the transcriptional networks that control cell fate decisions has proven difficult because existing tools are inadequate. Methods like ChIP-chip (Boyer et al. 2005) or ChIP-seq (Johnson et al. 2007; Robertson et al. 2007) provide a snapshot of transcription factor (TF) binding, but they are unable to record transcription factor binding events throughout development and along different cellular lineages. This makes it nearly impossible to correlate transcription factor binding events in progenitor cells to the final fates of their progeny cells. To solve this problem, we developed transposon “Calling Cards”. The central idea of the Calling Card method is to attach the transposase of a transposon to a TF, thereby endowing the TF with the ability to direct insertion of the transposon into the genome near to where it binds (Wang et al. 2007). The transposon becomes a “Calling Card” that permanently marks the transcription factor’s visit to a particular genomic location. By harvesting the transposon Calling Cards along with their flanking genomic DNA, a genome-wide map of transcription factor binding can be obtained.

We harnessed the *piggyBac* (PB) transposon as a Calling Card (Cary et al. 1989; Ding et al. 2005), because of its attractive features. Its transposition efficiency is significantly higher than commonly used transposons such as *sleeping beauty*, Tol2, and Mos1 in almost all mammalian cell lines tested (Wilson et al. 2007; Wu et al. 2006). PB transposase is amenable to N-terminal DNA-binding domain addition (Wilson et al. 2007; Wu et al. 2006), which is essential to our experimental design. PB transposition lacks overproduction inhibition, and its transposition efficiency can reach ~30% of cells (Wang et al. 2008b; Wilson et al. 2007).

Here I describe the implementation of *piggyBac* Calling Cards and development of a high throughput sequencing method to map Calling Card insertions. I demonstrate that PB Calling Cards can be used to map the target genes of p53 and REST.

RESULTS

Mapping genome-wide *piggyBac* transposition

The "Calling Card" method requires mapping a large number of transposition events, followed by identification of enrichment of insertions close to TF binding sites (Wang et al. 2007). I developed a protocol for high throughput identification of PB insertions in the human genome. I used this protocol to characterize the genome-wide insertion pattern of *piggyBac*. I transfected PB "helper" (encodes PB transposase) and Calling Card "donor" (contains hygromycin (hyg) resistance gene flanked by PB terminal repeats) plasmids (Fig 5.1A) into cell line HCT-116, selecting for hygromycin resistance.

After selection, all colonies contain PB insertions in their genomes. I extracted genomic DNA in a pool and digested it with three restriction enzymes that cut transposon DNA as well as the flanking genomic DNA (Fig. 5.1B). The digested DNA was diluted and circularized by ligation in dilute solution and then amplified by an inverse PCR (Fig. 5.1B). The DNA sequences of the PCR products were determined on an Illumina Genome Analyzer, using a sequencing primer that is designed to read the genomic DNA flanking the PB Calling Card.

In this way we identified ~15,000 independent PB integrations in three biological replicates. We measured the frequency of PB integration into defined genomic locations (Table 5.1). We observed a higher frequency of integration into RefSeq genes and a 10kb window around transcriptional start sites than in randomly selected genomic integration sites, consistent with the available genome-wide PB mapping data (Wilson et al. 2007). Our data set, which is 25-fold larger than the available data set (Wilson et al. 2007), provides a more thorough picture of PB insertions across the human genome, enabling us to identify 18 potential hot spots of PB insertions, using the criterion of more than 5 independent insertions within 10 kb region (Supplemental Table 5.1).

Using PB Calling Cards to map p53 and REST binding

To prove the principle of the PB calling card method, I fused p53 (full length, or from amino acid 100 (which includes the DNA-binding domain and tetramerization domains)) to the N terminus of PB transposase (Fig 5.1A). Attachment of full-length p53 to the PB transposase reduces the efficiency of transposition, but transposase with a shortened p53 (100-end) catalyzes transposition with an efficiency similar to the wild type PB transposase (Fig 5.2A). I transfected TP53 (100-end)-PB transposase and PB "donor" plasmids into HCT-116 cells and grew them in media containing hygromycin (100 μg/ml) for two weeks. The cells were collected and the locations of about 15,000 independent PB Calling Cards were mapped. Many Calling Cards landed within promoters of genes known to be targets of TP53. Transposition events that occurred near the known p53 target gene *GML* are shown in Fig 5.2B. We observed 8 insertions within 2 kilobases of the putative p53 binding site upstream of *GML* and only 3 additional insertions in 150 kb of flanking DNA. This is in contrast to the results obtained with cells with the normal (i.e. not fused to p53) PB transpoase: no transposon insertions were seen within 300kb of the *GML* locus (Fig 5.2C). We identified genes containing more than 3 independent PB integrations within 25kb of their promoter (defined as the 25 kb window from 20 kb 5' to 5 kb 3' of the transcription start site of a gene) as p53 target genes. Our preliminary list of p53 targets (504 genes) includes 22 of the 122 TP53 target genes identified by the ChIP-PET method (Wei et al. 2006). The highly significant overlap of these two data sets ($P= 4.3 \times 10^{-13}$) suggests that TP53-PB transposase directs PB integrations specifically into the promoters of TP53 target genes. In fact, the p53 targets identified with Calling Cards display better concordance with the ChIP-PET study than was observed between that study and previous work (as collated in the A*STAR p53 Knowledgebase).

To verify that this method can be applied to another mammalian TF, I used it to identify the target genes of REST, a TF with a well-defined binding motif and several known target genes (Johnson et al. 2007). Similar to p53, we found Calling Cards deposited by REST to be enriched in promoters of known REST target genes (Fig 5.3).

Out of the 2198 genes called by Calling Cards, 15 are in the 53 target genes identified by ChIP-QPCR (Mortazavi et al. 2006). This overlap is significant (P= 2.2 x 10^{-4}), even though our experiment was performed with a different cell type (HCT-116 cells versus Jarket T cells), and even though our algorithm for calling target genes is unsophisticated. These results suggest that the Calling Card method is capable of identifying target genes of DNA-binding proteins in mammalian cells.

DISCUSSION

The piggyBac transposase is functional when fused to a variety of mammalian transcription factors. I constructed TF-PBase fusions for the TP53, REST, and Sp1, transcription factors, and the E2C artificial Zinc-finger. All four of the fusion proteins efficiently catalyzed transposition of a PB transposon in HCT116 cells. These results suggest that the Calling Card method can be applied to many transcription factors.

This method holds great potential for analysis of mammalian cells. Many TFs are known to be involved in cell fate decisions. It has been difficult to identify the genes that are targets of TFs during differentiation of cells along different lineages. The ability of Calling Cards to "record" TF visits to the genome will allow us to view TF-DNA interactions that happened in progenitor cells. The current PB Calling Card protocol based on transient transfection is ready to be applied for this purpose to study cell differentiation *in vitro*. Further, we expect to be able to use Calling Cards to map the protein-DNA interactions that happen in early development in mice by recovering from fully-grown animals the Calling Cards deposited during their development.

The Calling Card method is still in its infancy. I was able to detect only about 20%

of target genes predicted from other studies. I believe this is partially due to differences in experimental procedures: Wei et al. (2006) induced p53 activation by treating HCT-116 cells with 5-FU for six hours before ChIP; I performed the calling card experiment in the same cell line without 5-FU treatment. ChIP-seq data for REST were generated with a Jakart T cell line (Johnson et al. 2007); I mapped REST binding in HCT-116 cells. We are repeating these experiments in the same cell lines with same treatment, to obtain data that can be directly compared to these published results.

There are many ways to improve this method. First, sample sizes can be increased. Mapping 5,000 Ty5 insertions gave good sensitivity for identifying TF binding in the yeast genome, but the human genome is more than two hundred times larger than the yeast genome, and we have been mapping only 15,000 PB insertions. Assuming PB Calling Cards in mammalian cells perform similar to Ty5 Calling Cards in yeast, we need to map 10-50 fold more PB Calling Cards to achieve good sensitivity. This can easily be accomplished by expanding cell culture and colony collection protocols ten fold. We will also develop improved protocols to ensure efficient identification of all PB insertions by Illumina sequencing. We have not been able to map back all sequencing reads efficiently and accurately: more than half of the millions of 30 bp sequence reads we analyzed were not able to be mapped back uniquely to the human genome. A single base pair error from either the PCR or DNA sequencing could map it back to a wrong locus. There is no way to discriminate between a true insertion with a low number of reads from a false insertion that appears as the result of sequencing errors from other insertions with high read numbers. Therefore, we have been arbitrarily cutting off of the number of reads analyzed. For example, we demand that every mapped sequence have more than 25 copies to be considered real. Inserting the Illumina sequencing primer 2 into the PB donor should improve the efficiency and accuracy of mapping of reads because it will allow us to perform pair-end sequencing of PB samples. That will provide both the 30bp genomic region adjacent to the insertion site as well as the 30bp sequence next to the closest restriction site. With such pairs of sequence tags, the majority of the reads are expected to map to a unique place in the human genome, and sequences with errors are less likely to be mapped to the wrong locus. Given that PB can only insert into TTAA sequences, a limited number of TTAA sites are available for PB integration, which will likely lead to multiple independent insertions in the same site. To improve resolution of the mapping, I have cloned unique 5bp DNA sequence bar codes into different PB donors. Using multiple bar-coded PB donors will enable discrimination of multiple independent insertions into one site.

We also need to develop better algorithms for mapping short DNA sequence reads to the genome. With additional information obtained from paired end sequence reads, we can allow mismatches and map most reads uniquely to the genome. We must also develop methods to align identified target genomic regions and identify TF specific recognition motifs.

The “Calling Card” method applied to mammalian cells promises to help reveal why cell fate decisions are nearly always stochastic processes. I believe that the Calling Card method, in its current form with no further improvements, can be used productively to attack this problem.

METHODS

Cell culture

Human colon adenocarcinoma cell line HCT116 (ATCC) was maintained in McCoy 5A Media Modified (Gibco) supplemented with 10% fetal bovine serum (Gibco). To select for PB transposition, hygromycin was added to the media to a final concentration 100 μg/ml. Cells were cultured at 37 ºC in the presence of 5% CO_2.

Construction of plasmids

Plasmids pcDNA3.1Δneo- *piggyBAC* (PB helper), pcDNA3.1Δneo-Gal4DBD-*piggyBAC* (Gal4-PB helper), and PB donor (Wu et al. 2006) were obtained from Stefan Moisyadi (Hawaii University). To use the "Gap Repair" method (Ma et al. 1987; Wach et al. 1994) for engineering this plasmid, I turned all three plasmids into yeast vectors by inserting into their NaeI site a fragment containing *CEN6*, ARS, and *TRP1*, amplified from pRS314 (Strathern and Higgins 1991) using primers OM8191 and OM8192, to make pBM5209, and in Gal4DBD-PB helper to make pBM5210. A fragment containing *CEN6*, ARS, and *URA3* sequences, amplified from pRS316 (Strathern and Higgins 1991) using primers OM8193 and OM8194, was cloned into the AflIII site of PB donor plasmid to make pBM5211.

All TF-PB helper constructs were built by "Gap Repair" of pBM5210 linearized by digestion with BsrGI. TP53 coding sequences were amplified using OM8420 and OM8422. The 5'-truncated TP53 (100 aa to end) was amplified using OM8501 and OM8422. REST coding sequences were amplified using OM8747 and OM8748. The sequences encoding the E2C artificial Zinc-finger was amplified from pMal-c2-E2C (Tan et al. 2006) using OM8601 and OM8602. Yeast cells were co-transformed with each PCR product and linearized pBM5210 selecting for Trp$^+$ colonies. DNA extracted from yeast colonies was introduced into E. coli and the plasmid was isolated. Each construct was confirmed by Sanger sequencing.

Transfection of cells and transposition of *piggyBac*

All plasmids used for transfection of cells were prepared using EndoFree Plasmid Maxi Kit (Qiagen) following the manufacturer's protocol.

HCT-116 cells were grown to confluency in a 25cm flask then dispersed by adding 1ml Trypcin-EDTA and incubating for 5 minutes at 37 °C. 9 ml of media was added to the flask to resuspend cells thoroughly (10^6 cells/ml). 0.5 ml cell suspension was added into one well of a six-well plate. Cells were grown in a total of 3 ml of media for two days until they reached 50% to 80% confluency. A total of 1 μg of DNA (0.33 μg helper and 0.66 μg donor) transfected into cells with FuGENE 6 (Roche), following the manufacturer's protocol. After 12 hrs cells were trypsinized and resuspended in 2.3 ml of media. For selection of cells in which *piggyBac* transposed, several 400 μl aliquots of cells were plated into one 10cm dish with 10 ml media containing hygromycin (100 μg/ml), resulting in 5 plates for each transfection. For colony counting, 50 μl cells were plated into one 10cm dish with 10 ml media containing hygromycin (100 μg/ml) for each transfection. After 14 days of selection in media containing hygromycin, colonies from all five plates were harvested and pooled for DNA extraction. Cells were fixed for counting with PBS containing 4% paraformaldehyde for one hour and then stained with 0.2% methylene blue overnight.

Inverse PCR

Genomic DNA was extracted from each sample using DNeasy Blood & Tissue Kit (Qiagen) following the manufacturer's protocol. Each DNA sample was divided into three 2 μg aliquots, each digested by MspI or Csp6I or TaqI individually. Digested DNA was ligated overnight at 15°C in dilute solution to encourage self-ligation. After ethanol precipitation, self-ligated DNA was resuspended in 30 μl ddH_2O and used as template in an inverse PCR. Primers that anneal to PB donor sequences (OM8721 and OM8722) were used to amplify the genomic regions flanking PB, as well as adding adapter sequences that allow the PCR products to be sequenced on the Illumina GA analyzer. The PCR products were purified using the QIAquick PCR Purification Kit (Qiagen) and diluted into 10nM concentration. For each sample, the same amount of PCR product from digestion with each restriction endonuclease was pooled and submitted for Illumina sequencing.

Sequence map back and gene calling

Reads were filtered by requiring that the first four bases of the read match the four bases of the end of the *piggyBac* terminal repeat, and that the four following bases are TTAA, the sequence into which *piggyBac* inserts. The first 30 bp of the sequence read corresponding to the genomic DNA adjacent to site of insertion were mapped using a hash table of all possible 30 bp sequences from the Human genome. If a read mapped uniquely to the genome and at least 25 reads mapped back to a specific site then the site was considered real.

Target genes were determined by requiring a minimum number of independent insertions with at least 25 reads in the promoter of a gene, defined as the 25 kb window from 20 kb 5’ to 5 kb 3’ of the transcription start site of a gene.

ACKNOWLEDGEMENTS

We are especially grateful to Stefan Moisyadi (Hawaii University) and Carlos F. Barbas (The Scripps Research Institute) for generously providing reagents and advice, and to Jessica Hoisington-Lopez, Laura Langton, Brian Koebbe, and Jim Dover (Washington University) for their expert assistance with the use of Illumina Genome Analyzer. This work was supported by NIH grants R21RR023960 and 5P50HG003170-03, and funds provided by the James S. McDonnell Foundation.

TABLES

Table 5.1. Frequency of *piggyBac* insertions within intragenic regions of human cells.

Genomic Location[a]	Random	PiggyBac Rep 1	PiggyBac Rep 2	PiggyBac Rep 3
In RefSeq genes	**40.7**	**51.4**	**53.1**	**53.2**
± 5 kb txn start site	**10.5**	**17.7**	**17.6**	**18.4**
± 5kb from CpG islands	**19.1**	**30.5**	**29.0**	**28.8**
± 1kb from CpG islands	**5.7**	**16.4**	**15.2**	**15.1**

[a] The coordinates of RefSeq genes and CpG islands for the 2006 human genome was downloaded from the UCSC genome project website (www.genome.ucsc.edu).

FIGURES

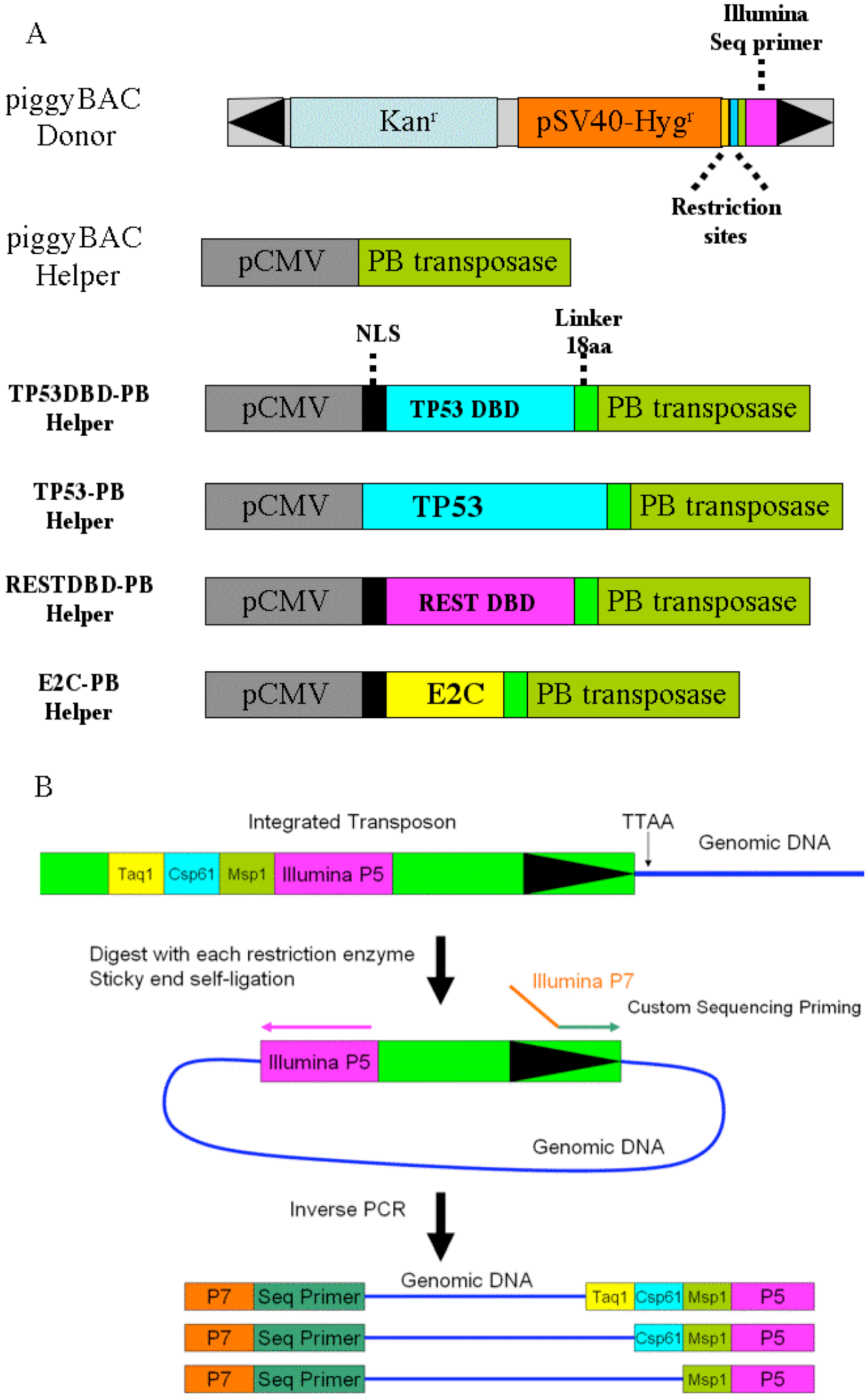


Figure 5.1. *piggyBac* Calling Cards. (A) Constructs used in this study. (B) Mapping PB transposition using Illumina sequencing.

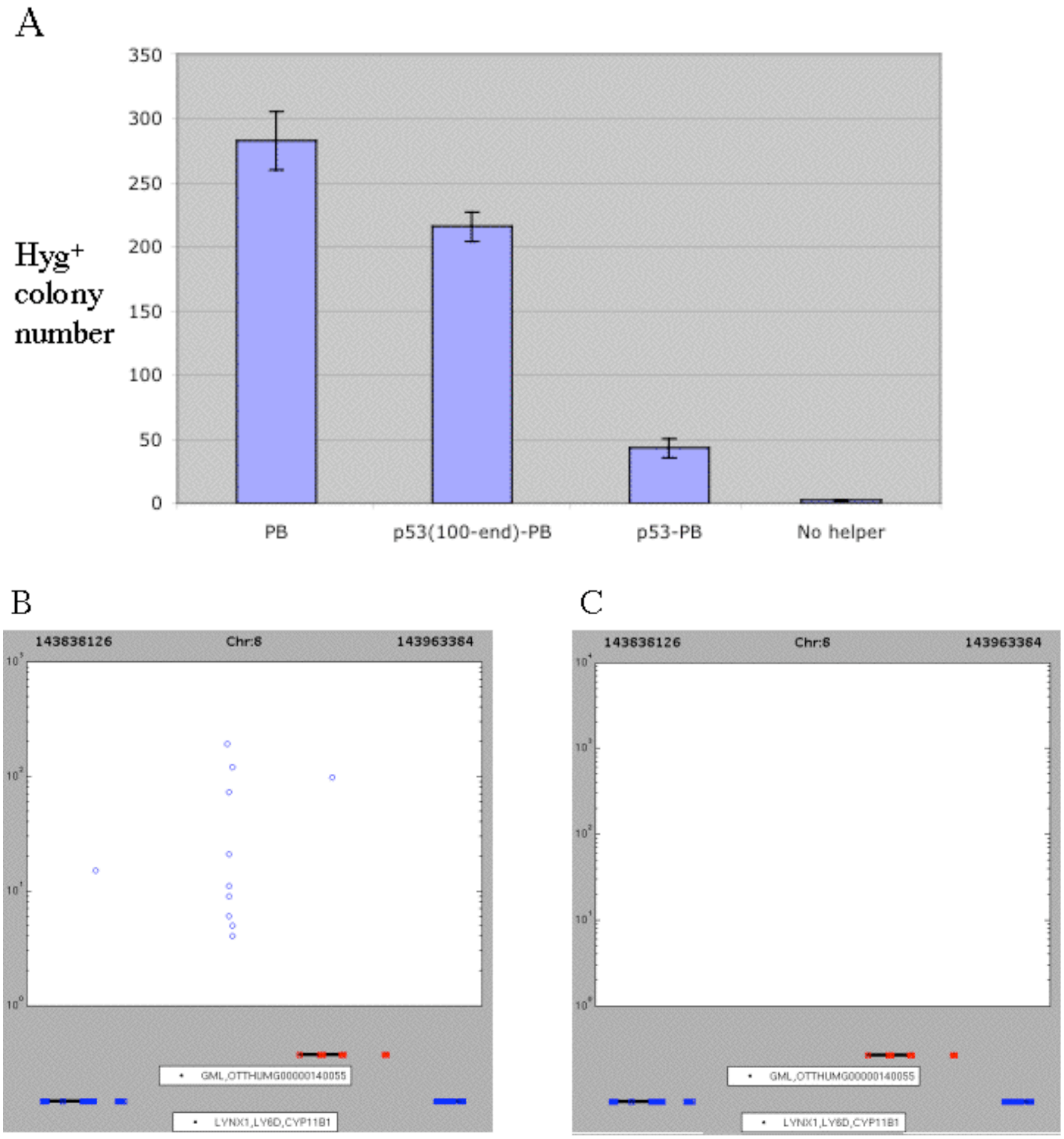


Figure 5.2. Using PB Calling Cards to identify P53 target genes. (A) P53-PB transposase can catalyze PB transposition efficiently. (B) Transposed by P53(100-end)-PB transposase, multiple PB integrations were within promoter of GML. Each blue circle represents an independent transposition event. (C) Using PB transposase, no PB insertion was in promoter of GML.

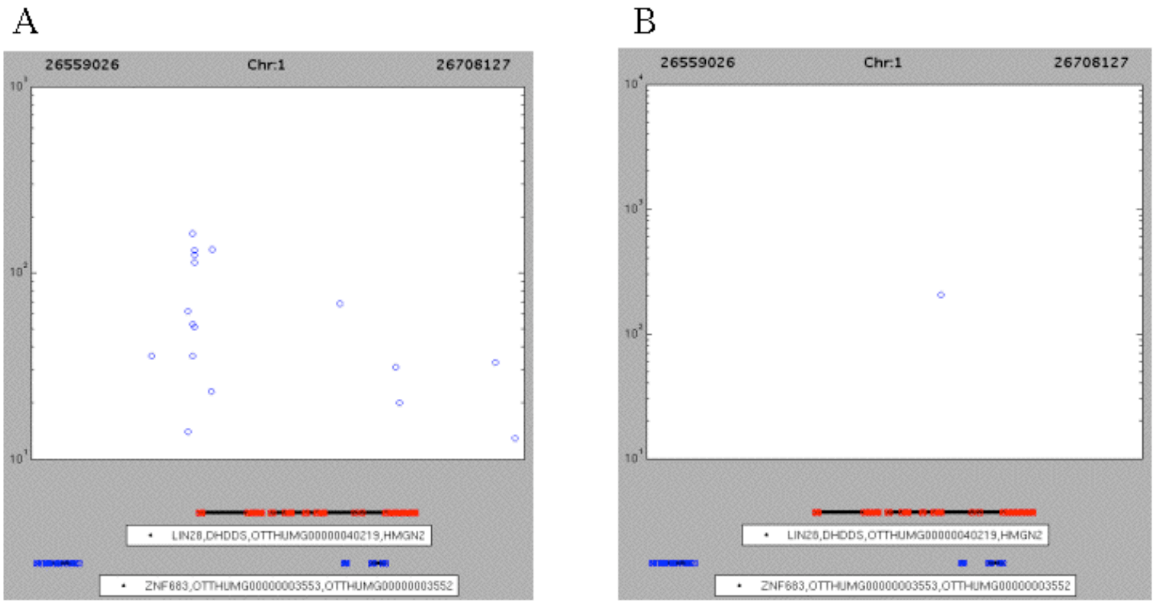


Figure 5.3 PB Calling Cards within promoter of REST target genes. (A) Transposed by RESTDBD-PB transposase, multiple PB integrations were within promoter of LIN28. Each blue circle represents an independent transposition event. (B) Using PB transposase, no PB insertion was in promoter of LIN28.

SUPPLEMENTAL FIGURES AND TABLES

Supplemental Table 5.1. Potential *piggyBac* hotspots in human genome.

Chromosome	Position[a]	Integrations[b]	Gene1[c]	Gene2[c]
11	107992853	7		
1	8103938	6		
1	8171271	6		
1	66534013	6	ENSG00000118473	
1	84540078	6	ENSG00000137976	
1	156386005	6	ENSG00000158477	
1	234133244	6		
3	99740897	6	ENSG00000080819	
3	100026707	6	ENSG00000144810	ENSG00000057019
4	30460764	6	ENSG00000215248	
8	62840849	6		
8	94996868	6		
10	101533607	6	ENSG00000107554	ENSG00000023839
12	64185201	6		
12	74383019	6		
12	87848739	6		
13	98771359	6	ENSG00000125245	
18	3638354	6	ENSG00000177337	

[a] The center of all PB integrations in 10kb window.
[b] The number of independent PB integrations within 10kb window.
[c] The genes within 20kb of the center of integrations.

Chapter 6: Conclusion

Summary of results and their implications

I have described my work developing the Calling Card method for identifying protein-DNA interactions *in vivo.* The central idea of each of the several versions of this method I developed is the same: to endow DNA-binding proteins with the ability to deposit transposons into genomic DNA. These transposons act as "Calling Cards" that permanently mark the transcription factor's visit to that site in the genome. The locations of these Calling Cards can be mapped either by microarray or by second-generation DNA sequencing.

As described in Chapter 2, I first implemented the "Calling Card" idea by isolating individual clones containing Ty5 transposition events, recovering their Calling Cards by inverse PCR, and determining their location in the genome by Sanger sequencing. This method was labor-intensive, but it enabled mapping of hundreds of independent Ty5 tranposition events directed by the Gal4-Sir4 and Gcn4-Sir4 fusion proteins. Of the 96 Gal4-directed insertions that I analyzed, 37 occurred in the *GAL1-10* promoter, 34 in the *GAL7* promoter, and one in each of the *GCY1* and *FUR4* promoters. Of the 80 Gcn4 directed transposition events, 17 occurred in the promoters of known Gcn4 target genes. While these results clearly constituted a "proof-of-principle" for the Calling Card method, I was not able to analyze the large number of transposition events (~2000-5000) needed to identify transcription factor target genes with high sensitivity. Therefore, I developed a protocol to map transposon insertion events by hybridization of the Calling Cards and the DNA adjacent to them to a DNA microarray. I performed inverse-PCR on DNA extracted from a pool of thousands of colonies containing TF-directed Ty5 insertions in their genomes, and hybridized the PCR products to a yeast tilling microarray. This greatly improved the throughput of the method and achieved a sensitivity and specificity comparable to the ChIP-chip method.

The Calling Card method could then be used to answer the question: where in the genome does a particular TF bind? Using a modified protocol, I also showed that the Calling Cards method could also be used to answer the converse question: what TFs bind to a particular genomic region (see chapter 2)? However, the microarray readout does not allow us to answer both of these questions simultaneously. If such a thing could be accomplished, it would allow me to determine the genome-wide binding patterns of multiple (potentially all) yeast transcription factors under a single growth condition in one experiment. To achieve this goal, I decided to incorporate second-generation DNA sequencing technology into the Calling Card method.

I began developing protocols for both the Illumina and polony sequencing (Shendure et al. 2005) platforms, since at that time it was not clear to which technology I would have access. However, in late 2007, I obtained access to an Illumina GA sequencer, so I focused my efforts on that platform and over the next year I worked out ways to map the locations of Calling Cards by determining the sequence of the genomic DNA flanking them. It seemed like a good idea to use the "standard" Illumina sequencing primer for this, so I cloned this sequence into the Ty5 transposon. This allowed me (after enzymatic digestion, self-ligation, and inverse PCR – see chapter 3) to sequence into the yeast genome at the restriction site closest to inserted transposon. This protocol was successful, but it did not provide the exact site of the transposon insertion. In fact, the resolution of this method was no better than the microarray method. To overcome this limitation, I engineered a Ty5 transposon with MmeI restriction sites at the ends of each LTR and developed an inverse PCR protocol that utilized an MmeI digestion as described in Chapter 3. This experimental design is perfect in principle: MmeI cuts 18 bp into the genomic DNA immediately adjacent to theTy5 insertion site. So, after the circularization step, the Illumina sequencing primer and the barcode sequence engineered into the Ty5 transposon would be just upstream of this 18 bp genomic fragment. This allows determination of the precise location of the Ty5 insertion and the bar code sequence from a single 36-bp Illumina sequencing read. In practice, however, the results were far from perfect. As described in Chapter 3, we were able to recover only a few hundred Ty5 transposition events in each experiment. This low "hopping" efficiency made it impractical to multiplex samples, since we were barely able to collect enough transposition events in each experiment to analyze even a single TF.

To improve the transposition efficiency of our engineered Ty5, I designed a method that allows us to perform Illumina sequencing without modifying the sequence of the Ty5 LTR. This protocol utilized the "paired-end" read capability of the Illumina platform, allowing us to obtain a DNA sequence from each end of the amplified fragment: one read maps the insertion site; the other read reveals the DNA sequence of the "bar code" in the Calling Card. This approach enables multiplexing of samples. I futher improved the transposition efficiency by replacing the *His3AI* gene in the transposon with the normal *HIS3* gene (and modifying the selection protocol accordingly). These improvements allowed me to multiplex the Calling Card protocol, so that tens of TFs can be analyzed in a single experiment (Chapter 4). This was gratifying, but the "Calling Cards" method was still not ready to be used to study protein-DNA interactions under different conditions because it involves over-expressing the TF-Sir4 fusion proteins from the strong *ADH1* promoter on a plasmid. Since environmental signals can directly regulate the transcription of some TFs, overexpressing the fusion protein from a heterologous promoter confounds measurement of native TF binding in different conditions. Therefore, I created several yeast strains that expressed the TF-Sir4 fusion proteins from the TFs' promoter at its native genomic location. This genomic tagging strategy worked well, and we were able to achieve performances of the Calling Card method very similar to that achieved in experiments in which the TF-Sir4 on a plasmid was overexpressed.

The current "Calling Card-seq" method is now ready to be applied to study *in vivo* protein-DNA interactions under many different conditions. With it we identified DNA sequence recognition motifs and target genes for six poorly characterized TFs in a single experiment.

Future directions

There are several ways the Calling Card method can be improved. First, we need to further improve transposition efficiency. A ten-fold improvement would allow us to multiplex all 200 yeast TFs in a single experiment. This would also reduce the amount of time needed to induce transposition. The current protocol requires three days, which limits our ability to study the dynamics of gene regulatory networks. A higher transposition efficiency may allow us to reduce the induction time to several hours, enabling us to study transcriptional responses to changing environments. Second, a different method for induction of Ty5 transposition is needed. In our current system, Ty5 is driven by the *GAL1-10* promoter, so Ty5 transposition is induced by growing cells in media containing galactose as the primary carbon source. Since this is not the optimal growth condition for yeast, the cells activate their stress response pathways, making it difficult to test certain conditions. Placing the Ty5 integrase under the control of a tetracycline inducible promoter (TetON) is one way around this problem (Belli et al. 1998). Since doxycycline (a tetracycline analogue) doesn't affect yeast gene expression (Wishart et al. 2005), this would be an ideal method of induction of transposition. Third, a complete library of yeast TF-Sir4 fusion proteins, each expressed from its native locus, is needed. This "tag-in" collection should be made in diploid cells, to produce a diploid with one copy of the wild-type TF and one copy of the TF-Sir4 fusion. This design should largely eliminate phenotypic variance between strains that might be caused by tagging the TFs with the Sir4 fragment.

These improvements will enable rapid characterization of the genome-wide binding patterns of all yeast TFs under many different culture conditions. It would be best to start analyzing growth conditions that lead to the largest global gene expression differences, based on published microarray based gene expression profiling data (Gasch et al. 2000; Roberts et al. 2000). Analysis of all predicted yeast TFs in this way under those conditionsshould provide a catalogue of important protein-DNA interactions. When coupled with gene expression data, the information gathered by these experiments will provide a detailed blueprint of the yeast transcriptional regulatory network. Another exciting application is to characterize protein-DNA interactions in yeast strains in which different genes have been deleted. For example, it would be interesting to compare patterns of transcription factor binding when different genes known to be important for chromatin structure are knocked out (e.g. *SNF2* and *SWI1*). This would allow us to study the specific roles of chromatin structure in regulating transcription factor binding.

Another important future aim is to determine if the Calling Card method can provide a quantitative readout of transcription factor binding. To date, we have been making qualitative predictions (binding or no binding) based on arbitrary cutoffs of significance. However, it may be possible to use the Calling Card method to accurately measure protein-DNA affinities *in vivo*. The total number of sequencing reads that we observe for each insertion is not likely to correlate with binding affinity because the inverse PCR does not amplify all sequences equally well. However, the number of independent Ty5 integrations within a defined genomic region may accurately report binding affinity. We have consistently observed higher numbers of insertions in stronger promoters. By adding a 4-6 bp DNA sequence bar code into each Ty5 element, we can distinguish independent insertions at the same nucleotide position. This will allow us to accurately count the number of independent Ty5 insertions across the yeast genome. By correlating the number of independent Ty5 insertions with binding affinity as measured by Chromatin-IP followed by real-time PCR, the ability of the Calling Card method to provide quantitative results can be determined.

Since we were able to identify DNA sequence recognition motifs and target genes for six poorly characterized TFs (Chapter 4), I am confident that we can produce similar high quality data for the remaining 62 poorly characterized TFs of yeast (Chpater 3). I expect we will find that a few of these putative TFs are not actually DNA-binding proteins (as we found for Lee1p -- see Chapter 4). For the remaining TFs, we will probably be able to predict target genes, and for some of these, identify their sequence recognition motifs. TFs for which we can identify gene targets, but for which we can not identify binding motifs, are particularly interesting. We have already seen several examples of this (e.g. Kar4, Sfg1, Rpi1, chapter 4). An important future direction is to try to understand why we are unable to recover binding motifs for these TFs. Application of all major motif finding algorithms (Harbison et al. 2004; MacIsaac et al. 2006) using less stringent criteria may identify highly degenerate consensus sequences for these TFs. For TFs that resist this analysis, we will need to test several hypotheses using both bioinformatics and experimental methods. We will first look for enrichment of known TF binding motifs within the same promoters; such TFs could serve as a cofactor to facilitate the other TFs' binding. If that is not fruitful, the search window should be enlarged to look for enrichment of known TF binding motifs within neighboring genes. Correlation of the binding locations of these transcription factors with published maps of chromatin and DNA modifications (Pokholok et al. 2005) may be informative. These experiments should allow us to formulate specific hypotheses about individual TFs that can be experimentally tested. By paying close attention to these "outliers", we may be able to reveal novel mechanisms by which TFs bind to their genomic targets.

An interesting observation is that Calling Cards are sometimes deposited into promoters that neighbor strong TF targets (see chapter 4). These promoters often do not contain a binding site for the TF. I imagine the neighboring promoters are brought into close proximity through DNA looping. To test this possibility, I will select several of the best candidate loci for looping and test this hypothesis by deleting the TF binding sites in the strong promoter and determining if Calling Cards are no longer depositedin the neighboring promoter. I could also apply the "chromosome confirmation capture" (3C) method to directly detect DNA looping at these loci. If the Calling Card method is capable of identifying local DNA looping events, application of this technology to study all protein-DNA interactions under different conditions may reveal new features of transcriptional regulation.

Although the Calling Card method is well-established in yeast, the mammalian version of the method is still in its infancy. As described in Chapter 5, I am in the process of making several improvements to the PB Calling Cards method. This improved method should allow us to address interesting questions about the process of development in cell culture systems and, ultimately, in multicellular organisms. For example, Olig2, a basic helix-loop-helix transcription factor, promotes the differentiation of mouse ES cells into motoneurons (Du et al. 2006). We hypothesize that Olig2 binds a different set of target genes in ES cells that become motoneurons than in ES cells that do not acquire that fate. I plan to exploit the ability of of Calling Cards to record TF visits to the genome to test this hypothesis by correlating Olig2 binding in ES cells to the cell fates of their progeny. To analyze the binding of Olig2 during motoneuron differentiation, I will deliver the *piggyBac* transposon and the Olig2:PBase gene fusion to the cells on plasmids. The transformed ES cells will be induced to differentiate into motoneurons by adding sonic hedgehog and retinoic acid to the culture medium (Bain et al. 1995). After differentiation, I will sort motoneurons from other cell types and harvest Calling Cards from the two populations of cells and map the genomic locations of Calling Cards deposited by Olig2 in each population of cells.

Although we can deliver all of the components necessary for PB Calling Cards by transfecting cells with plasmids, we also plan to engineer a mouse ES cell line with multiple PB transposons and a TF:PBase gene fusion integrated into its genome. This is expected to improve transposition efficiency and increase the range of biological questions that can be investigated. The PB Calling Card method can be further improved by engineering cell lines with TF:PBase gene fusions integrated at the transcription factors' native genes to ensure that the transcription factors are expressed at their normal levels. We could also fuse the TF:PBase to the ligand binding domain of the estrogen receptor (ERT2), which will improve the temporal resolution of Calling Cards by allowing us to rapidly and reversibly induce transposition by pulsing the cells with 4-hydroxytamoxifen (Cadinanos and Bradley 2007). This will allow us to record all transcription factor binding events that occur in a short period of time.

Since the *piggyBac* transposon is active in a variety of different organisms, including yeast (Mitra et al. 2008), insects (Handler 2002), and mice (Ding et al. 2005), it is likely that the PB Calling Card method can be applied to the study of all of those organisms. Implementing the Calling Card method in Drosophila is particularly appealing since it is relatively simple to create transgenic flies, large numbers of flies can be grown easily and cheaply, and many powerful genetic tools are available for experimentation on this model organism. An understanding of the DNA-protein interactions that direct fly development should also provide insights into the transcriptional networks that control mammalian development, as several important developmental genes and pathways present in this organism are also present in mammals. (for example, the genes involved in body plan specification (Castanon and Baylies 2002) and tumor formation (Potter et al. 2000) are conserved in both mammals and flies)

In summary, my thesis has focused on developing Calling Card method to study protein-DNA interactions *in vivo*. I have established this method for use with yeast, which promises to deliver a nearly complete list of target genes for all yeast TFs under many different conditions. I have also successfully implemented Calling Cards in mammalian cells. The unique nature of this method will allow us to study protein-DNA interactions in a way that has not previously been possible.

Appendix: Protocol: "Calling Cards" method for high-throughput identification of targets of yeast DNA-binding proteins

Haoyi Wang, Michael Heinz, Seth Crosby, Mark Johnston, and Robi David Mitra
Department of Genetics, Washington University, School of Medicine, 4444 Forest Park Parkway, St. Louis, MO 63108

This protocol is a reproduction of a manuscript published in *Nature Protocol* in 2008. This work was done in collaboration with Michael Heinz, Seth Crosby, Mark Johnston and Robi Mitra. Robi Mitra, Mark Johnston and I designed the protocol and wrote the paper. Michael Heinz and Seth Crosby did the microarray experiments.

ABSTRACT

We present a protocol for a novel method for identifying the targets of DNA-binding proteins in the genome of the yeast *S. cerevisiae*. This is accomplished by engineering a DNA-binding protein so that it leaves behind in the genome a permanent mark—a "Calling Card"—that provides a record of that protein's visit to that region of the genome. The calling card is the yeast Ty5 retrotransposon, whose integrase interacts with the Sir4 protein. If Sir4 is fused to a DNA-binding protein, it recruits the Ty5 integrase, which directs insertion of a Ty5 calling card into the genome. The calling card along with the flanking genomic DNA is harvested by inverse PCR and its genomic location is determined by hybridization of the product to a DNA microarray. This method provides a straightforward alternative to the "ChIP-Chip" method for determining the targets of DNA-binding proteins. This protocol takes about two weeks to complete.

INTRODUCTION

Transcription factors bind to specific sites in the genome and control gene transcription. Identification of the genomic sites bound by all DNA-binding proteins will provide a detailed map of the transcriptional networks that direct different cellular processes and provide a framework for understanding how a cell controls global patterns of gene expression. Here we describe the "Calling Card" method, a tool to provide this information.

Overview of the Calling Card method

The Calling Card method exploits the Ty5 retrotransposon of bakers' yeast. Ty5 mRNA is converted by reverse transcriptase into a double-stranded cDNA that the Ty5 integrase carries to the nucleus and inserts into the genome (Ozcan and Johnston 1999). The Ty5 integrase interacts with the Sir4 heterochromatin protein (Xie et al. 2001). Therefore, any DNA-binding protein can be made to recruit the Ty5 integrase by attaching to it the fragment of the Sir4 protein that interacts with the integrase (Zhu et al. 2003). Consequently, the engineered DNA-binding protein directs insertion of Ty5 into DNA near to where it is bound, leaving behind a permanent mark—a "calling card"—of its visit to that region of the genome. We have exploited this property of Ty5 to develop a method for identifying the genomic targets of DNA binding proteins.

The TF-Sir4 fusions are made by joining the transcription factor of interest to a fragment of Sir4 (amino acids 951 to 1200) that includes the Ty5 integrase-interacting domain (Xie et al. 2001; Zhu et al. 2003). We have been fusing the Sir4 fragment to the C-terminus of the transcription factors, but it may be preferable to fuse it to the N-terminus in some cases (e.g., if the DNA-binding domain of the transcription factor is near the C terminus). Based on our experiments to date, no linker is necessary between TF and Sir4 protein. The TF-*SIR4* fusions can be constructed in yeast by the "gap repair" method (Ma et al. 1987; Wach et al. 1994). See Reagent Setup for details on primer design for obtaining TF coding sequence DNA by PCR.

The "Calling Card" protocol, summarized in Figure 1 and Figure 2, can be divided into five stages: (1) Construction of a yeast strain carrying a plasmid encoding the desired transcription factor (TF)-Sir4 chimera and a plasmid carrying Ty5, (2) induction of Ty5 transposition, (3) selection of cells that have undergone transposition of Ty5, (4) recovery of the Ty5 calling cards from genomic DNA by inverse PCR, and (5) identification of the flanking genomic DNA sequence by hybridization of the inverse PCR product to a DNA microarray.

All the experiments should be done in a *sir4* deletion strain (e.g. YM7635), otherwise wild-type Sir4 protein will compete with TF-Sir4 for binding to Ty5 integrase, causing transposition into telomeres. There are three controls that should be used in any calling card experiment. First, one should analyze Ty5 transposition in the *sir4* deletion working strain, without any TF-Sir4 fusion construct (Box 1). This controls for background transposition that is not directed by the TF-Sir4 chimera. We have found the patterns of transposition to be significantly different between strains without the TF-Sir4 construct and strains expressing a TF-Sir4 fusion protein. A similar pattern of transposition in both strains is a clear indication that something is wrong with the TF-Sir4 construct. Second, to control for the variation in hybridization efficiency across different probes on the microarray, we label genomic DNA and use this as a hybridization control. The inverse PCR samples are labeled with cy5 and the genomic DNA is labeled with cy3. Both labeled samples are hybridized to the same microarray. The intensity values in the control channel (the green, or cy3 channel) are used to estimate the hybridization efficiency of each probe, which allows us to accurately quantify the amount of DNA hybridized in the experimental channel (the red, or cy5 channel). Finally, as a positive control, it is useful to analyze Ty5 transposition in yeast expressing a Gcn4-Sir4 fusion protein. This control, which only needs to be included the first time a calling card experiment is performed, is important for the analysis of the microarray hybridization because it determines the intensity cutoff that separates transposition events from hybridization noise.

Advantages of the Calling Card method

We believe this technology will prove useful for the study of DNA-binding proteins because it is relatively easy to employ, and is in principle orthogonal to the ChIP-Chip method. Even if the calling card technology does not prove to be a substitute for the ChIP-Chip method (Horak and Snyder 2002; Ren et al. 2000), it is likely to complement that well-established method because it can identify targets of proteins that may be refractory to analysis by chromatin IP, and can be used to verify results obtained with the ChIP-Chip method(Wang et al. 2007). In addition, there are opportunities for multiplexing the calling card technology (using DNA barcodes), offering the possibility of identifying the targets of many DNA-binding proteins in a single experiment (Wang et al. 2007). In this protocol, we focus on mapping genome-wide binding of a single transcription factor, but this procedure can easily be extended using modifications detailed in our earlier paper (Wang et al. 2007) to determine all transcription factors that bind to a single promoter. We are in the process of coupling calling card technology with Illumina 1G sequencing to analyze the genome-wide binding of multiple transcription factors in a single experiment.

Limitations of the Calling Card method

There are several limitations of this method in its current state: first, the transposition efficiency of Ty5 is fairly low (~ 10^{-5}), which makes it difficult to sample more than a few thousand transposition events. Second, each transcription factor is driven by the *ADH1* promoter, so its expression level is not native. Third, expression of the Ty5 calling card from the *GAL1* promoter limits the conditions that can be tested. Finally, Ty5 transposition is influenced by host factors (Gao et al. 2002), so the implementation of calling cards in an organism other than *S. cerevisiae* will probably require the use of a different transposon.

Other Methods for identifying target sites of DNA-binding proteins

ChIP-based methods: The “ChIP-chip” method combines chromatin immunoprecipitation (ChIP) with DNA microarrays (chip): DNA is co-precipitated with a DNA-binding protein by ChIP, and then identified by hybridization to a DNA microarray (Horak and Snyder 2002; Ren et al. 2000). ChIP is also now being combined with "next generation” DNA sequencing (Johnson et al. 2007; Robertson et al. 2007). ChIP-chip has been successfully applied to map the target genes of transcription factors in yeast (Harbison et al. 2004; Lee et al. 2002) and other organisms (Boyer et al. 2005; Zeitlinger et al. 2007). The related ChIC and ChEC methods are tailored to the analysis of insoluble proteins, such as the scaffolding components of chromatin (Schmid et al. 2004). These methods are similar to ChIP in that DNA binding proteins are crosslinked to DNA, but they employ a micrococcal nuclease that is tethered to an antibody (ChIC) or the DNA binding protein itself (ChEC), to introduce double stranded breaks in unbound DNA.

The ChIP based methods are powerful because they are highly flexible – they can be used to analyze a wide variety of DNA binding proteins in a number of model systems. One weakness is that the results of ChIP-type experiments often depends on the quality of the antibody employed, although this can be somewhat alleviated by expressing DNA binding proteins with peptide tags. Also, some DNA binding proteins appear to be recalcitrant to ChIP-chip and related methods (Harbison et al. 2004).

Yeast-1 hybrid: A one-hybrid screen can also be used to identify the transcription factors that bind to a specific genomic locus (Wilson et al. 1991). In this method, a query sequence is cloned in front of a reporter gene, and a library of transcription factor-activation domain fusion constructs are screened (Deplancke et al. 2004). This method has been used to reveal the architecture of regulatory networks in *C.elegans* (Deplancke et al. 2006; Vermeirssen et al. 2007). A strength of this method is that it is easily automatable, allowing for high-throughput analysis of many loci and proteins. It has the disadvantage that transcription factor binding is not queried at the native locus.

DamID: Another method for the identification of DNA loci bound by transcription factors is DamID (van Steensel et al. 2001; van Steensel and Henikoff 2000). In this method, a DNA adenine methyltransferase (Dam) is fused to a transcription factor, which targets DNA methylation to adenines that are close to binding sites. This method can be used to analyze proteins that are resistant to ChIP-chip. One possible weakness is that non-specific methylation often occurs, although this can be addressed with the appropriate controls.

MATERIALS

REAGENTS

- Diploid yeast strain with *sir4* deletion, YM7635 (*MATa /MATalpha his3Δ1/ his3Δ1 leu2Δ0/ leu2Δ0 ura3Δ0/ ura3Δ0 met15Δ0/MET15 lys2Δ0/LYS2 sir4::Kan/ sir4::Kan trp1::Hyg/ trp1::Hyg*)

- Plasmid pBM4607 (contains the Gal4DB-*SIR4* fusion with *TRP1* as the selectable marker) (Sequence has been submitted to addgene.org, Plasmid 18795)
- Plasmid pBM5218 (encodes the Ty5 transposon with *URA3* as the selectable marker) (Sequence has been submitted to addgene.org, Plasmid 18796)
- Restriction enzymes: XhoI, HinP1I, HpaII, TaqI (New England Biolabs)
- QIAquick Gel Extraction Kit (Qiagen; Cat. no.: 28704)
- Phusion DNA Polymerase (with manufacturer's buffers) (New England Biolabs; Cat. no.: F-530S)
- 10 mM dNTP mix (ROCHE; Cat. no.: 12779120)
- Yeast growth media and plates (see Reagent Setup):

  YPD.

  SC Glucose –Trp; used to select for the plasmid pBM4607, which contains a *TRP1* marker.

  SC Glucose –Ura; used to select for the plasmid pBM5218, which contains a *URA3* marker.

  SC Glucose –Ura –Trp; used to select for pBM5218 and pBM4607.

  SC Galactose –Ura; used to select for pBM5218, and to activate the GAL1-10 promoter.

  SC Galactose –Ura –Trp; used to select for pBM5218 and pBM4607, and to activate the GAL1-10 promoter.

SC Glucose –His; used to select for cells with a Ty5 transposition event.

SC Glu –His 5-FOA used to select for cells with a Ty5 transposition event and to select against the Ty5 donor plasmid pBM5218.

- Yeast lysis buffer (see Reagent Setup)
- Phenol/chloroform/iso-amyl alcohol (25:24:1) (ROCHE; Cat. no.: 03117979001)

  CAUTION Phenol is toxic when in contact with skin or if swallowed. Chloroform is harmful if inhaled or swallowed.
- 0.5mm glass beads (Biospec Products, Cat. no.:11079105)
- 3M NaOAc (pH 5.2-6.0)
- 70% and 100% EtOH
- E. coli transformation competent cells (GC 10 cells; GeneChoice; Cat. no.: D-7L)
- LB plates containing 100 ug/ml ampicillin (see Reagent Setup)
- QIAprep Spin Miniprep Kit (Qiagen; Cat. no.: 27104)
- Sequencing Primers (IDT), see Reagent Setup:

  OM6189: CACAATATTTCAAGCTATACC;

  OM6373: CTCATCAACCAACGAAACGG;
- T4 Polynucleotide Kinase (New England Biolabs; Cat. no.: M0201)
- T4 DNA ligase (with manufacturer’s 10x ligation buffer) (ROCHE; Cat. no.: 10481220001)

- Inverse PCR primers (IDT), see Reagent Setup and Fig. 1:

  OM6609: CTTTTGGGTTATCACATTCAAC

  OM6610: ATCGTAATTCACTACGTCAAC

  OM6456: CCCATAACTGAATACGCATG

  OM6458: AGGTTATGAGCCCTGAGAG

- REDTaq DNA polymerase (with manufacturer's buffers) (Sigma; Cat. no.: D4309)
- QIAquick PCR Purification Kit (Qiagen; Cat. no.: 28104)
- BioPrime Array CGH Genomic Labeling Module (Invitrogen; Cat. no.: 18095-011)
- aCGH hyb buffer (Agilent; Cat. no.: 5188-5220)
- Yeast Whole Genome 4 x 44K ChIP-on-chip Microarray Kit, (Agilent; Cat. no.: G4493A, Design ID 014810)

EQUIPMENT

- Centrifuge (e.g. Eppendorf Centrifuge 5417R)
- Roller drum
- Bio-Rad E. coli pulser
- NanoDrop ND-1000 spectrophotometer
- Eppendorf Biophotometer
- Thermal cycler (e.g. MJ Research PTC100)
- 30°C and 37°C incubator
- Liquid nitrogen and appropriate container

- Sonicator (e.g. Ultrasonic Processor XL2020)

REAGENT SETUP

Yeast lysis buffer: 20 ml of 10% Triton (vol/vol), 10 ml of 10% SDS (wt/vol), 0.58 g NaCl, 1 ml of 1M Tris (PH 8.0), and 200 µl of 0.5 M EDTA. Add water to 100 ml, filter sterilize.

LB ampicilin medium and plates: Mix 10 g Tryptone, 5 g Yeast Extract, 5 g NaCl, (for plates, add 20 g agar) add water to one liter, autoclave, cool and add 100 mg ampicillin.

Yeast growth media and plates

- **YPD**: Mix 10 g Yeast Extract, 20 g Peptone, and 20 g Glucose, (for plates, add 20 g agar) dissolve in one liter water, autoclave.
- **Synthetic Complete (SC)**: Mix 1.7 g Yeast Nitrogen base (Difco cat. No. 233520), 5 g ammonium sulfate, 20 g glucose or galactose, various nutrient "drop-out" mixes (-His, -Ura, -Trp, -Ura -Trp; US Biologicals, use according to the manufacturer's instructions) (for plates, add 20 g agar), add water to one liter, autoclave. For SC Glucose –His 5-FOA plates, after autoclave, add 1 g 5-FOA. Different SC plates are used to select for the markers on transformed plasmid. For example, SC Glucose –Ura plates are used to select yeast cells transformed with plasmid containing a *URA3* marker.

Primers: Although we provide the sequences for inverse PCR and sequencing primers (see Reagents list), oligos could also be designed by the user. All DNA primers should be synthesized at the 25 nmol scale. No purification other than standard desalting is necessary.

Primer design for cloning TF-Sir4 fusion construct: Each primer (forward and reverse) is composed of two distinct sequences: the first (5') 39 base pairs of each primer have the sequence of the regions flanking the Gal4DBD in the plasmid pBM4607, which are necessary to enable homologous recombination for cloning in yeast cells by gap repair (Ma et al. 1987; Wach et al. 1994). The next ~20bp of each primer is a gene specific sequence designed to amplify the ORF so that it is intact and in frame with the Sir4-encoding sequences in pBM4607. We generally design primers to the first and last 18-22 bps of the ORF. If the transcription factor is to be fused to the N-terminus of the Sir4 fragment (as we usually do), make sure to exclude the stop codon. To fuse the TF to the N-terminus of Sir4 in plasmid pBM4607, the 39 bp homologous sequences are as follows (5' to 3'):

Forward: ATACAATCAACTCCAAGCTTGAAGCAAGCCTCCTGAAAG

Reverse: TTTGGGTTTGCTAGAATTAGTATCACTATGCGACACTCT

PROCEDURE

Construct TF-*SIR4* fusion Timing: 5 to 7 days

1. Design primers to amplify the coding sequence of the transcription factor of interest, as described in Reagent Setup.

2. Amplify the coding sequence of the transcription factor with Phusion DNA polymerase (or other high fidelity DNA polymerase) following the manufacturer's protocol. The PCR mix is made according to the table below, and should be prepared on ice.

Component	Amount (per reaction)	Final amount/concentration
5 x Phusion HF buffer	5 µl	0.5 x
5 x Phusion GC buffer	5 µl	0.5 x
10 mM dNTP mix	1 µl	0.2 mM of each
forward primer 25 µM	1 µl	0.5 µM
reverse primer 25 µM	1 µl	0.5 µM
Yeast genomic DNA	1 µl	10-100 ng
Phusion DNA Polymerase	0.5 µl	1 unit
ddH$_2$O	35.5 µl	
TOTAL volume	50 µl	

CRITICAL STEP To avoid introducing mutations into transcription factor coding sequence by PCR, always use high fidelity DNA polymerase.

3. Program the thermocycler as follows:

Step	Temperature	Time	Cycles
1	98 °C	30 sec	1
2	98 °C	10 sec	
3	60 °C (variable depending on primer design)	30 sec	
4	72 °C	15-30 sec/kb	Go to step 2 for 35

			cycles
5	72 °C	5 min	1
6	4 °C	Indefinitely	1

4. Digest 1μg of pBM4607 (contains the Gal4DB-*SIR4* fusion with *TRP1* as the selectable marker) with 10 units of XhoI at 37 °C for 1 hour.
5. Purify the linearized plasmid by gel electrophoresis. Run XhoI-digested pBM4607 on a 0.7% agarose gel (wt/vol) (containing 10 mg/ml ethidium bromide) at 130V for 1 hour. Cut out the DNA (should be in one band on the gel) and purify using the QIAquick Gel Extraction Kit (following the manufacturer's protocol).
6. Co-transform a *trp1* yeast strain with 10 to 30 ng of the linearized pBM4607 from step 5 and all of the PCR product (usually more than 3 μg) from step 3 as described in Box 2.
7. After two days incubation at 30 °C, multiple yeast colonies should be observed on SC Glucose –Trp plate. Pool 8 Trp$^+$ colonies in 200 μl yeast lysis buffer. Extract DNA as described in Box 3. Resuspend DNA pellet in 100 μl ddH$_2$O.
8. Transform 1 μl extracted DNA into competent *E. coli* cells using Bio-Rad E. coli Pulser following manufacturer's protocol and plate on LB + ampicillin plates.
9. Incubate plates at 37 °C overnight.
10. Pick 4 to 8 *E. coli* colonies from the LB + ampicillin plate and culture each in 1 ml LB + ampicillin media. Incubate on a roller drum at 37 °C overnight.

11. Purify plasmid DNA from each culture using QIAprep Spin Miniprep Kit (follow the manufacturer's protocol), and determine the DNA sequence of the TF-*SIR4* junction using sequencing primers OM6189 and OM6373 (see Reagents).

PAUSE POINT: Transform confirmed constructs in *E. coli* and store as glycerol stocks in -80 °C freezer, which can be kept for years.

Induction and selection of Ty5 transposition Timing: 10 days

12. Co-transform the plasmids containing the TF-*SIR4* fusion (with *TRP1* marker) and the Ty5 transposon (with *URA3* marker) into yeast strain YM7635 as described in Box 2. Use 0.1-0.5 μg of each plasmid for transformation. Remember to carry out a control experiment in parallel, as described in Box 1. TROUBLESHOOTING
13. Incubate for two days at 30 ºC, after which time multiple yeast colonies containing both plasmids should be observed on SC Glucose –Ura –Trp plate.
14. From each plate, pick one colony and culture overnight in 5ml SC Glucose –Ura –Trp media at 30 ºC. Once the culture reaches an OD_{600} of 1 or higher, plate 500 μl of cells on each of 10 SC Glucose –Ura –Trp plates.
15. Grow at 30 ºC for 1 day until a confluent lawn is formed.
16. Replica plate the cells onto SC Galactose –Ura –Trp plates to induce Ty5 transposition. Galactose will activate the GAL1-10 promoter that drives the expression of Ty5. Keep plates at room temperature (22-25 ºC) for 3 days.
17. Select for cells with Ty5 transpositions by replica plating onto SC Glucose –His plates. The integrated Ty5 transposon has a functional His marker, so only cells

with transpositions will grow. Incubate plates for two days at 30 ºC. TROUBLESHOOTING

18. Select colonies that have lost the Ty5-containing plasmid by replica plating onto SC Glucose –His 5-FOA plates. 5-FOA will counter-select the cells containing URA3 gene. Incubate for two days at 30 ºC. TROUBLESHOOTING
19. Harvest the cells from the SC Glucose –His 5-FOA plates with transposed Ty5 by adding 1ml of YPD to each plate. Suspend the cells using a spreader and pipette the liquid into a 15 ml falcon tube. Pool the cells from all 10 plates into one 15 ml falcon tube, which will yield about 8 ml cells in YPD.
20. Aliquot 50 μl cell pellet and extract genomic DNA as described in Box 3.

    PAUSE POINT: Freeze the remaining cells in liquid nitrogen and store in -80 ºC freezer for as long as needed. Extracted DNA can be stored at -20 ºC for several months before proceeding with next step.

Enzyme digestion, DNA fragment circularization, and amplification Timing: 1.5 days

21. Digest 1 μg genomic DNA from step 20 with: TaqI; HinP1I; and HpaII independently. For each digestion: add 2 ml 10 x NEB buffer, 2 ml 10 x BSA, 10 unit restriction enzyme, 1 μg genomic DNA, and add ddH_2O to total 20 ml. For HinP1I and HpaII digestions, incubate at 37 ºC for 1 hr. For TaqI digestion, incubate at 65 ºC for 1 hr.
22. Run 2 ul of each reaction on a 0.7% agarose gel (wt/vol) to confirm DNA digestion. A 200 bp to 5 kb smear should be observed.

23. Purify the DNA from each reaction using the QIAquick PCR Purification Kit, following the manufacturer's protocol. To elute the DNA from the column, apply 30 μl ddH_2O at the center of the column, let sit on bench for 1 min, and spin the column at 18000 x g for 1 min. Measure the DNA concentration using the Nanodrop apparatus.

24. To circularize the digested fragments, prepare the following ligation reaction on ice, and incubate at 15 ºC overnight.

Component	Amount	Final
Digested DNA	50-100 ng	50-100 ng
10 x T4 ligation buffer	10 μl	1x
T4 DNA ligase	1 unit	1 unit
ddH_2O	To 100 μl	

CRITICAL STEP: Do not use more than 100 ng digested DNA in ligation reaction, or inter-molecular ligations will be favored over the desired intra-molecular circularization.

25. For each digested and circularized sample, amplify the ligated products from 5 μl of the ligation reaction by PCR. Set up separate reactions with one pair of primers to amplify the genomic regions on the left side (primers OM6609 and OM6458, see Reagents) and with another pair of primers to amplify the right side (primers OM6610 and OM6456, see Reagents) of Ty5 (Fig. 1). Set up the reactions by mixing the following components on ice:

Component	Amount (per reaction)	Final amount/concentration

10 x RedTaq buffer	5 μl	0.5 x
10 mM dNTP mix	1 μl	0.2 mM of each
5 M Betaine	10 μl	1 M
Forward primer 25 μM	1 μl	0.5 μM
Reverse primer 25 μM	1 μl	0.5 μM
Ligation mix from step 24	5 μl	2.5-5 ng
RedTaq DNA Polymerase	2 μl	2 units
ddH_2O	25 μl	
TOTAL volume	50 μl	

26. Program the thermocycler as follows:

Step	Temperature	Time	Cycles
1	93 ºC	2 min	1
2	93 ºC	30 sec	
3	60 ºC	6 min	Go to step 2 for 28-30 cycles
6	4 ºC	Indefinitely	1

Run 5μl of the PCR products on a 0.7% agarose gel (wt/vol) and a 200 bp to 2 kb smear should be observed.

27. Purify each PCR product with the QIAquick PCR Purification Kit (following the manufacturer's protocol) and measure the DNA concentration using a Nanodrop apparatus. For each transcription factor, pool the same amount of DNA from each PCR product (total six PCR products from three different digested and self-ligated samples).

28. Prepare the control sample: shear 10 μg of yeast genomic DNA using sonicator. Using Ultrasonic Processor XL2020, shear DNA for one minute at full power

(level 10). CRITICAL STEP: Keep the sample in ice-water bath during sonication; use a clamp to hold the tube in the ice-water bath so that the bottom of the tube sits 0.5-1.0 cm above the sonicator probe.

29. Run a portion of the sheared genomic DNA on a 0.7% agarose gel (wt/vol) to confirm DNA shearing. A 200 bp to 2 kb smear should be observed.

    PAUSE POINT: Sample and control DNA could be stored at –20 ºC for several weeks before microarray hybridization.

Microarray hybridization and data analysis Timing: 3 days

30. Label both the PCR products (test DNA) and the sheared genomic (control) DNAs with Invitrogen’s BioPrime Array CGH Genomic Labeling Module, using a different fluorophore (cy3 or cy5) for each. Follow manufacturer' protocol with the following exceptions/ specifications: Input mass for genomic DNA = 1.6 μg DNA/ fluorophore/ array; Input mass for PCR products = 2.0 μg DNA/fluorophore/array.

    CRITICAL STEP: Because this method can be adapted to different microarray platforms, the protocol for hybridization and data analysis may vary. Here, we provide a general overview of the protocol that we employ.

31. Co-hybridize labeled DNAs to Agilent Yeast WGA 4x44K microarrays in Agilent aCGH hyb buffer; characterize each experimental condition in triplicate, using three microarrays. Follow Agilent’s aCGH hybridization protocol with the following exceptions/ specifications: Hybridization overnight (16-20hrs) at 65 ºC at oven rotation of 20 rpm; Washing: B1. Wash 1= 6xSSPE/ 0.005% N-

lauroylsarcosine (wt/vol); B2. Wash 2= 0.06X SSPE; B3. Used Agilent Stabilization and Drying Solution (cat# 5185-5979).

32. Scan the microarrays on Genepix 4000B Microarray scanner (Molecular Devices) to detect cy3 and cy5 fluorescence.

33. Analyze images using the Genepix, v6.0 software package to obtain fluorescent intensities for each feature on the microarray. Use the ratio of the mean fluorescent intensities of the test over control channel to estimate the extent of enrichment of loci present in the test DNA, then rank the loci based on this mean ratio. Next, use the Gcn4-TF positive control to select the appropriate intensity cutoff. We typically choose a cutoff that maximizes the true Gcn4 positives at a 2.5% false positive rate. A list of true Gcn4 targets, as well as a list of genes that are not targeted by Gcn4 can be found in the supplementary material of Pokholok et. al (Pokholok et al. 2005). For a more detailed description of data analysis, please see Wang et al. (Wang et al. 2007) and accompanying Supplemental Information (**http://www.genome.org/cgi/data/gr.6510207/DC1/1**).

TROUBLESHOOTING

Box 1: Experimental control: Ty5 transposition without TF-Sir4 (perform in parallel to the main protocol step 12-16)

1. Transform the plasmid pBM5218 containing the Ty5 transposon (with *URA3* marker) into yeast strain YM7635 as described in Box 2. Use 0.1-0.5 μg of plasmid DNA for transformation. Select transformants on SC Glucose –Ura plates.
2. After two days incubation at 30 °C, multiple yeast colonies should be observed on SC Glucose –Ura plate. From each plate, pick one colony and culture overnight in 5ml SC Glucose –Ura media at 30 °C.
3. Once the culture reaches an OD_{600} of 1 or higher, plate 500 μl of cells on each of 10 SC Glucose –Ura plates.
4. Grow at 30 °C for 1 day until a confluent lawn is formed.
5. Replica plate the cells onto SC Galactose –Ura plates to induce Ty5 transposition. Keep plates at room temperature (22-25 °C) for 3 days.
6. Continue with the main protocol from step 17.

END OF BOX 1

Box 2: Yeast transformation (Gietz and Woods 2006) Timing: 2 days

1. Start a 5 ml YPD culture of the working strain one day earlier. Incubate overnight on a roller drum at 200rpm and 30 °C.
2. The next day, pipette 100 μl cell suspension into 1 ml of water in a spectrophotometer cuvette and measure OD at 600 nm using Eppendorf

Biophotometer. For most yeast strains, culture containing 1x 10^6 cells / ml will give OD_{600} of 0.1.

3. Add 2.5 x 10^8 cells into 50 ml fresh YPD in a culture flask to give 5 x 10^6 cells / ml. Shake the culture at 30 °C and 200rpm for 3 to 5 hours, until the cell density reaches about 2 x 10^7 cells / ml. CRITICAL STEP: Optimal cell density is critical to the transformation efficiency, do not use over-grown cells.
4. Spin down cells at 3000g for 5 min, and wash them with 10 ml sterile water. These cells are sufficient for ten transformations.
5. Aliquot cells for each transformation into 1.5 ml microcentrifuge tube. Spin down cells at 20000 x g for 30 sec and discard the supernatant. Make the total transformation mix first and then add 360 μl of the mix to each tube.

Component	Amount (per reaction)	Final amount/concentration
PEG 3500 50% (wt/vol)	240 μl	33.3%
LiAc 1.0 M	36 μl	0.1 M
Denatured (by boiling) SS-carrier DNA (10 mg/ml)	10 μl	27.8 ng/μl
DNA plus ddH_2O	74 μl	
TOTAL volume	360 μl	

CRITICAL STEP: Be careful to pipette the correct volume of PEG, which is viscous.

6. Vortex the mixture vigorously and incubate the tube in a 42°C water bath for 40 min.
7. Spin down the cells at 20000 x g for 30 sec and discard the supernatant. Add 100 μl ddH_2O into each tube and stir the pellet with pipette tip. Plate appropriate

dilution of the cell suspension onto SC selection media. For example, if a plasmid containing *TRP1* marker was transformed, plate cells on SC Glucose –Trp media.

8. Incubate the plates at 30 °C for two to three days.

END OF BOX 2

Box 3 Yeast Genomic DNA extraction Timing: 1.5 hours

1. Add 200 μl yeast lysis buffer, 200 μl phenol/chloroform/iso-amyl alcohol (25:24:1), and 200 μl 0.5mm glass beads to 50 μl cell pellet. Vortex for 5 to 10 min. CAUTION: Phenol is toxic when in contact with skin or if swallowed. Chloroform is harmful if inhaled or swallowed.
2. Spin the tubes in a microcentrifuge at 20000 x g for 10 min. Transfer the supernatant into a new 1.5ml microcentrifuge tube.

   CRITICAL STEP: avoid the transfer of debris from the interface to reduce the contamination of protein in the extracted DNA.
3. Add 200 μl chloroform to the tube, vortex well, spin at 20000 x g for 5 min. Transfer the supernatant into a new 1.5 ml microcentrifuge tube. CAUTION: Chloroform is harmful if inhaled or swallowed.
4. Add 1/10 volume 3M NaOAc (pH 5.2-6.0) and 2.5 volume of 100% EtOH. Vortex vigorously and put at -80 °C for 30 min.
5. Spin the tube in a microcentrifuge at 20000 x g for 10 min. A pellet of DNA should be visible.
6. Decant the ethanol and add 1ml 70% EtOH to the DNA pellet. Invert the tube several times, spin at 20000 x g for 5 min.

7. Decant the 70% EtOH, vacuum dry the DNA pellet and resuspend DNA in 100 μl TE or ddH_2O. The DNA concentration should be approximately 200 ng/μl.

END OF BOX 3

TROUBLESHOOTING

Troubleshooting advice can be found in Table 1.

Table 1: Troubleshooting

Step	**Problem**	**Reason**	**Solution**
12	Few or no colonies after transformation	Co-transformation of two plasmids is inefficient.	Use more competent cells and plasmid DNA for transformation
16,17,18	Bacterial contamination on plates	Plates can easily become contaminated during replica plating.	Important to clean bench with ethanol before replica plating. Autoclave velvets wrapped in foil (no more than 10 per package) thoroughly before use.
17	No colonies growing on SC Glucose –His plates (selecting for cells with Ty5 transposition)	Homologous recombination between 5'and 3' Ty5 LTRs in the plasmid with the calling card results in the deletion of Ty5.	Before inducing transposition, verify that the strain carries an intact Ty5 calling card by a PCR assay using a pair of primers that amplify a region within Ty5.
33	Results from the strain with TF-Sir4 are similar to results from the control strain without TF-Sir4.	This TF-Sir4 fusion construct is non-functional.	Determine the sequence of the entire coding sequence of TF-Sir4 to ensure there are no significant mutations. Also, expression of the fusion protein can be confirmed by Western blotting with anti-Myc antibody (the Myc tag is fused to Sir4 in pBM4607). If the TF-Sir4 coding sequence and protein expression are fine, perhaps the TF interferes with the function of the Ty5 integrase. Try fusing only the DNA-binding domain to Sir4.

ANTICIPATED RESULTS

In order to define a set of genomic regions that have a high probability of being adjacent to a 'calling card', we used the calling cards to identify targets of the well-characterized transcription factor Gcn4 and empirically chose a cutoff that minimizes the rate of false negatives at a false positive rate of 2.5% (Wang et al. 2007). (A list of genes known to be regulated by Gcn4 and a list of genes that are not regulated by Gcn4 was provided by Pokholok et al. (Pokholok et al. 2005)) For each experiment, we performed three technical replicates. Probes with fluorescence ratios above the cutoff in at least two of the three measurements were considered significant. We ignored data from probes that cover telomere regions because Ty5 can insert into these regions of the genome due to homologous recombination with Ty5 elements that reside there. We also excluded *HIS3* probes because *HIS3* sequences from the Ty5 calling cards are present in the inverse PCR product.

Gal4 and Gcn4 provide good positive controls for the method. Gal4-Sir4 leaves calling cards at *GAL1-10*, *GAL7*, *GAL3*, *GAL2*, *FUR4*, *GCY1*, and *PCL10*, approximately in that order of abundance[3]. Since a large number of calling cards are deposited upstream of *GAL1-10* and *GAL7*, the probes for these regions are often saturated in the test channel on the microarray. Gcn4 has more targets than Gal4, and consequently Gcn4-Sir4 leaves calling cards at a larger number of places in the genome[3]. A list of real and false Gcn4 targets can be found at Pokholok et al. (Pokholok et al. 2005). For both Gal4-Sir4 and Gcn4-Sir4, the false negatives should be around 49% at a false positive frequency of 2.5%. The negative control strains (i.e. no TF-Sir4 fusion) will contain transpositions that localize largely to the telomeres, although we also observe some background transposition in regions of open chromatin. We generally observe very different patterns of transposition in the negative control than in samples with TF-Sir4 fusions. It appears that background transposition (e.g. to the telomeres) is largely suppressed when the Sir4 protein is tethered to a transcription factor.

An example of the raw data from microarray hybridization experiments of Gal4 and Gcn4 is shown in Table 2. For a strong target (*GAL1-10*), an intermediate target (*GAL2*), and a non-target (*ACT1*) of Gal4, two known targets (*CPA2* and *HIS5*) and a non-target (*ACT1*) of Gcn4, the top three probes on the microarray of each promoter are listed. The exact cy5/cy3 ratio often varied between biological replicates, but the relative ranking of target genes in each experiment remained largely the same.

ACKNOWLEDGMENTS

We are grateful to Dan Voytas (Iowa State University) for generously providing reagents and advice. This work was supported by NIH grants R21RR023960 and 5P50HG003170-03, and funds provided by the James S. McDonnell Foundation.

COMPETING INTERESTS STATEMENTS

The authors declare no competing financial interests.

TABLES

Table A.1. Raw data from microarray hybridization

	GeneName	SystematicName	cy5/cy3 ratio
GAL4			
Strong Target	GAL10_GAL1	chr2:278552-278611	2113.88144
Strong Target	GAL10_GAL1	chr2:278210-278269	1913.56122
Strong Target	GAL10_GAL1	chr2:278766-278825	739.809151
Intermediate Target	GAL2	chr12:290045-290104	242.987142
Intermediate Target	GAL2	chr12:289916-289975	150.780762
Intermediate Target	GAL2	chr12:289271-289330	19.8380465
Non-Target	ACT1	chr6:54741-54800	0.0035058
Non-Target	ACT1	chr6:53476-53535	0.01081764
Non-Target	ACT1	chr6:54282-54341	0.01021794
GCN4			
Known Target	CPA2_YMR1	chr10:632975-633034	146.171857
Known Target	CPA2_YMR1	chr10:633354-633413	136.359583
Known Target	CPA2_YMR1	chr10:633184-633243	75.1000869
Known Target	PRM5_HIS5	chr9:142513-142572	37.9003228
Known Target	PRM5_HIS5	chr9:142382-142441	37.6663545
Known Target	PRM5_HIS5	chr9:142799-142858	18.3060131
Non-Target	ACT1	chr6:55296-55355	1.27141814
Non-Target	ACT1	chr6:55068-55127	0.69365667
Non-Target	ACT1	chr6:54741-54800	0.22661866

FIGURES

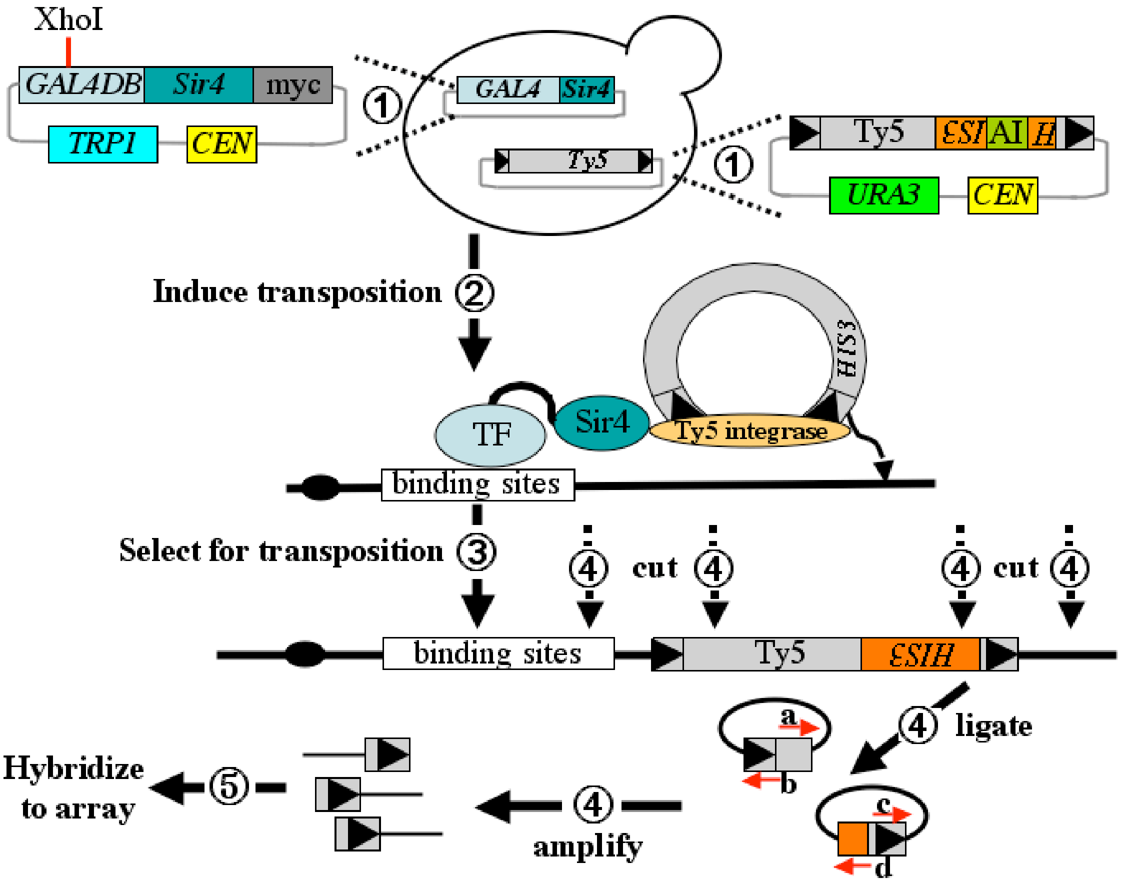


Figure A.1. The five stages of the "Calling cards" protocol: (1) Construction of a yeast strain carrying a plasmid encoding the desired transcription factor (TF)-Sir4 fusion and a plasmid carrying Ty5. A *HIS3* marker is inserted into Ty5 in the opposite direction, within which lies an artificial intron (AI). The two black triangles at the ends of Ty5 represent the long terminal repeats (LTR) of Ty5. (2) Induction of Ty5 transposition. (3) Selection of cells that have undergone transposition of Ty5. (4) Recovery of the Ty5 calling cards from genomic DNA by inverse PCR. Primer a is OM6458; primer b is OM6609; primer c is OM6610; primer d is OM6456. (5) Identification of the flanking genomic DNA sequence by hybridization of the inverse PCR product to a DNA microarray. Modified from Fig.1 in Wang et al. (Wang et al. 2007)

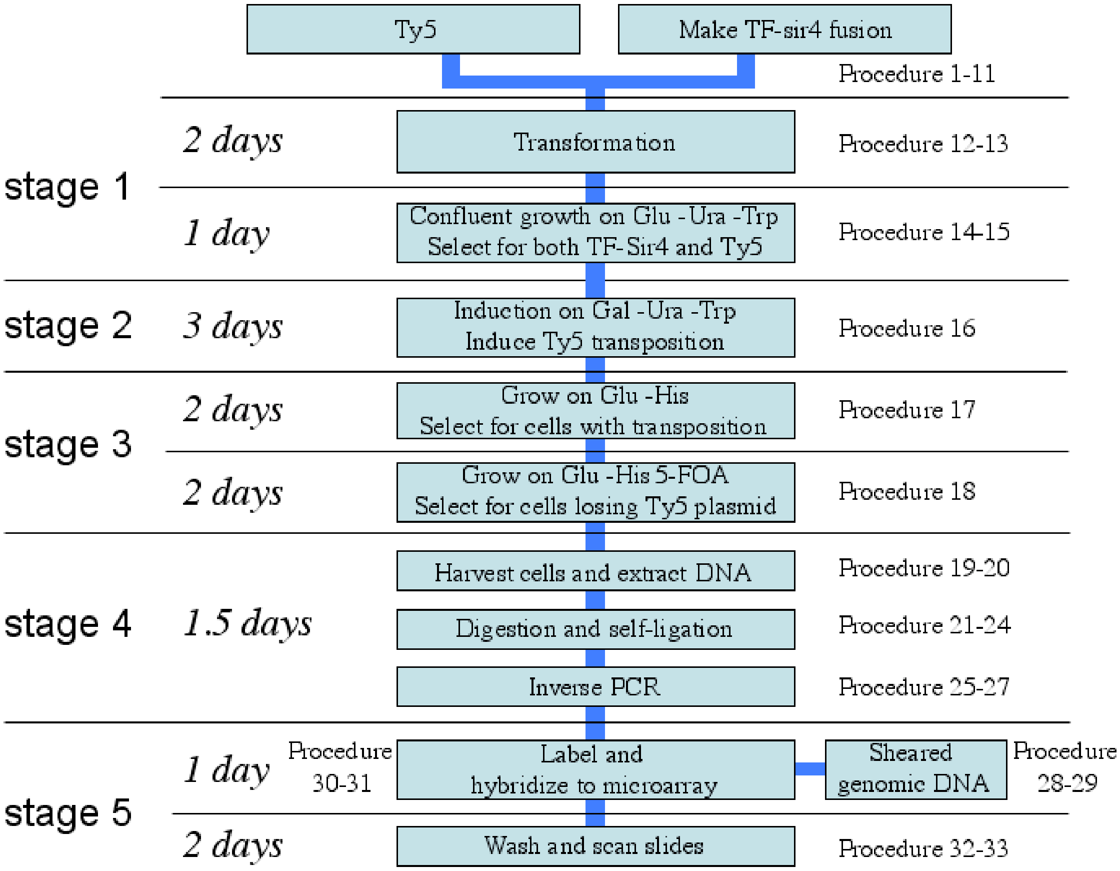


Figure A.2. A sample timeline for the “calling cards” protocol.

REFERENCES

Adryan, B. and Teichmann, S.A. 2006. FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. *Bioinformatics* 22: 1532-1533.

Agalioti, T., Lomvardas, S., Parekh, B., Yie, J., Maniatis, T., and Thanos, D. 2000. Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell* 103: 667-678.

Ayer, D.E. 1999. Histone deacetylases: transcriptional repression with SINers and NuRDs. *Trends Cell Biol* 9: 193-198.

Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C.D., Gossett, A.J., Hasinoff, M.J., Warren, C.L. et al. 2008. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* 32: 878-887.

Bae, K.H., Kwon, Y.D., Shin, H.C., Hwang, M.S., Ryu, E.H., Park, K.S., Yang, H.Y., Lee, D.K., Lee, Y., Park, J. et al. 2003. Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nat Biotechnol* 21: 275-280.

Bain, G., Kitchens, D., Yao, M., Huettner, J.E., and Gottlieb, D.I. 1995. Embryonic stem cells express neuronal properties in vitro. *Dev Biol* 168: 342-357.

Barbaric, S., Munsterkotter, M., Svaren, J., and Horz, W. 1996. The homeodomain protein Pho2 and the basic-helix-loop-helix protein Pho4 bind DNA cooperatively at the yeast PHO5 promoter. *Nucleic Acids Res* 24: 4479-4486.

Belli, G., Gari, E., Piedrafita, L., Aldea, M., and Herrero, E. 1998. An activator/repressor dual system allows tight tetracycline-regulated gene expression in budding yeast. *Nucleic Acids Res* 26: 942-947.

Berens, C. and Hillen, W. 2003. Gene regulation by tetracyclines. Constraints of resistance regulation in bacteria shape TetR for application in eukaryotes. *Eur J Biochem* 270: 3109-3121.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G. et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122: 947-956.

Brachmann, C.B., Davies, A., Cost, G.J., Caputo, E., Li, J., Hieter, P., and Boeke, J.D. 1998. Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14: 115-132.

Brenowitz, M., Senear, D.F., Shea, M.A., and Ackers, G.K. 1986. "Footprint" titrations yield valid thermodynamic isotherms. *Proc Natl Acad Sci U S A* 83: 8462-8466.

Briggs, M.R., Kadonaga, J.T., Bell, S.P., and Tjian, R. 1986. Purification and biochemical characterization of the promoter-specific transcription factor, Sp1. *Science* 234: 47-52.

Cadinanos, J. and Bradley, A. 2007. Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Res* 35: e87.

Carballo, E., Lai, W.S., and Blackshear, P.J. 1998. Feedback inhibition of macrophage tumor necrosis factor-alpha production by tristetraprolin. *Science* 281: 1001-1005.

Carthew, R.W., Chodosh, L.A., and Sharp, P.A. 1985. An RNA polymerase II transcription factor binds to an upstream element in the adenovirus major late promoter. *Cell* 43: 439-448.

Cary, L.C., Goebel, M., Corsaro, B.G., Wang, H.G., Rosen, E., and Fraser, M.J. 1989. Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology* 172: 156-169.

Castanon, I. and Baylies, M.K. 2002. A Twist in fate: evolutionary comparison of Twist structure and function. *Gene* 287: 11-22.

Chandler, V.L., Maler, B.A., and Yamamoto, K.R. 1983. DNA sequences bound specifically by glucocorticoid receptor in vitro render a heterologous promoter hormone responsive in vivo. *Cell* 33: 489-499.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M. et al. 1998. SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 26: 73-79.

Chodosh, L.A., Carthew, R.W., and Sharp, P.A. 1986. A single polypeptide possesses the binding and transcription activities of the adenovirus major late transcription factor. *Mol Cell Biol* 6: 4723-4733.

Cohen, B.A., Mitra, R.D., Hughes, J.D., and Church, G.M. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* 26: 183-186.

Curcio, M.J. and Garfinkel, D.J. 1991. Single-step selection for Ty1 element retrotransposition. *Proc Natl Acad Sci U S A* 88: 936-940.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. 2002. Capturing chromosome conformation. *Science* 295: 1306-1311.

Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A.J. 2004. A gateway-compatible yeast one-hybrid system. *Genome Res* 14: 2093-2101.

Deplancke, B., Mukhopadhyay, A., Ao, W., Elewa, A.M., Grove, C.A., Martinez, N.J., Sequerra, R., Doucette-Stamm, L., Reece-Hoyes, J.S., Hope, I.A. et al. 2006. A gene-centered C. elegans protein-DNA interaction network. *Cell* 125: 1193-1205.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. 2005. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* 122: 473-483.

Du, Z.W., Li, X.J., Nguyen, G.D., and Zhang, S.C. 2006. Induced expression of Olig2 is sufficient for oligodendrocyte specification but not for motoneuron specification and astrocyte repression. *Mol Cell Neurosci* 33: 371-380.

Fields, S. and Song, O. 1989. A novel genetic system to detect protein-protein interactions. *Nature* 340: 245-246.

Fried, M. and Crothers, D.M. 1981. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res* 9: 6505-6525.

Gabriel, A., Dapprich, J., Kunkel, M., Gresham, D., Pratt, S.C., and Dunham, M.J. 2006. Global Mapping of Transposon Location. *PLoS Genet* 2: e212.

Galas, D.J. and Schmitz, A. 1978. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* 5: 3157-3170.

Gao, X., Rowley, D.J., Gai, X., and Voytas, D.F. 2002. Ty5 gag mutations increase retrotransposition and suggest a role for hydrogen bonding in the function of the nucleocapsid zinc finger. *J Virol* 76: 3240-3247.

Garner, M.M. and Revzin, A. 1981. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res* 9: 3047-3060.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11: 4241-4257.

Gelperin, D.M., White, M.A., Wilkinson, M.L., Kon, Y., Kung, L.A., Wise, K.J., Lopez-Hoyo, N., Jiang, L., Piccirillo, S., Yu, H. et al. 2005. Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. *Genes Dev* 19: 2816-2826.

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

Gietz, R.D. and Woods, R.A. 2006. Yeast transformation by the LiAc/SS Carrier DNA/PEG method. *Methods Mol Biol* 313: 107-120.

Gilmour, D.S. and Lis, J.T. 1984. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc Natl Acad Sci U S A* 81: 4275-4279.

Granek, J.A. and Clarke, N.D. 2005. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol* 6: R87.

Griffiths, A.D. and Tawfik, D.S. 2006. Miniaturising the laboratory in emulsion droplets. *Trends Biotechnol* 24: 395-402.

Guarente, L., Yocum, R.R., and Gifford, P. 1982. A GAL10-CYC1 hybrid yeast promoter identifies the GAL4 regulatory region as an upstream site. *Proc Natl Acad Sci U S A* 79: 7410-7414.

Handler, A.M. 2002. Use of the piggyBac transposon for germ-line transformation of insects. *Insect Biochem Mol Biol* 32: 1211-1220.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J. et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431: 99-104.

He, Y., Imhoff, R., Sahu, A., and Radhakrishnan, I. 2009. Solution structure of a novel zinc finger motif in the SAP30 polypeptide of the Sin3 corepressor complex and its potential role in nucleic acid recognition. *Nucleic Acids Res*.

Ho, S.W., Jona, G., Chen, C.T., Johnston, M., and Snyder, M. 2006. Linking DNA-binding proteins to their recognition sequences by using protein microarrays. *Proc Natl Acad Sci U S A* 103: 9940-9945.

Hoeffler, J.P., Meyer, T.E., Yun, Y., Jameson, J.L., and Habener, J.F. 1988. Cyclic AMP-responsive DNA-binding protein: structure based on a cloned placental cDNA. *Science* 242: 1430-1433.

Horak, C.E. and Snyder, M. 2002. ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol* 350: 469-483.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D. et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102: 109-126.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292: 929-934.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316: 1497-1502.

Johnston, M. and Davis, R.W. 1984. Sequences that regulate the divergent GAL1-GAL10 promoter in Saccharomyces cerevisiae. *Mol Cell Biol* 4: 1440-1448.

Kadonaga, J.T. 2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 116: 247-257.

Kadonaga, J.T. and Tjian, R. 1986. Affinity purification of sequence-specific DNA binding proteins. *Proc Natl Acad Sci U S A* 83: 5889-5893.

Ke, N., Gao, X., Keeney, J.B., Boeke, J.D., and Voytas, D.F. 1999. The yeast retrotransposon Ty5 uses the anticodon stem-loop of the initiator methionine tRNA as a primer for reverse transcription. *Rna* 5: 929-938.

Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 26: 1351-1359.

Kim, J.L., Nikolov, D.B., and Burley, S.K. 1993a. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* 365: 520-527.

Kim, S.Y., Chung, H.J., and Thomas, T.L. 1997. Isolation of a novel class of bZIP transcription factors that interact with ABA-responsive and embryo-specification elements in the Dc3 promoter using a modified yeast one-hybrid system. *Plant J* 11: 1237-1251.

Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. 1993b. Crystal structure of a yeast TBP/TATA-box complex. *Nature* 365: 512-520.

Kristie, T.M. and Roizman, B. 1986. Alpha 4, the major regulatory protein of herpes simplex virus type 1, is stably and specifically associated with promoter-regulatory domains of alpha genes and of selected other viral genes. *Proc Natl Acad Sci U S A* 83: 3218-3222.

Kurihara, L.J., Stewart, B.G., Gammie, A.E., and Rose, M.D. 1996. Kar4p, a karyogamy-specific component of the yeast pheromone response pathway. *Mol Cell Biol* 16: 3990-4002.

Lahav, R., Gammie, A., Tavazoie, S., and Rose, M.D. 2007. Role of transcription factor Kar4 in regulating downstream events in the Saccharomyces cerevisiae pheromone response pathway. *Mol Cell Biol* 27: 818-829.

Lai, W.S., Carballo, E., Strum, J.R., Kennington, E.A., Phillips, R.S., and Blackshear, P.J. 1999. Evidence that tristetraprolin binds to AU-rich elements and promotes the deadenylation and destabilization of tumor necrosis factor alpha mRNA. *Mol Cell Biol* 19: 4311-4323.

Latchman, D.S. 1997. Transcription factors: an overview. *Int J Biochem Cell Biol* 29: 1305-1312.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I. et al. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298: 799-804.

Lee, T.I. and Young, R.A. 2000. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* 34: 77-137.

Lee, W., Mitchell, P., and Tjian, R. 1987. Purified transcription factor AP-1 interacts with TPA-inducible enhancer elements. *Cell* 49: 741-752.

Leuther, K.K. and Johnston, S.A. 1992. Nondissociation of GAL4 and GAL80 in vivo after galactose induction. *Science* 256: 1333-1335.

Li, J.J. and Herskowitz, I. 1993. Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system. *Science* 262: 1870-1874.

Lucau-Danila, A., Delaveau, T., Lelandais, G., Devaux, F., and Jacq, C. 2003. Competitive promoter occupancy by two yeast paralogous transcription factors controlling the multidrug resistance phenomenon. *J Biol Chem* 278: 52641-52650.

Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. 2000. An overview of the structures of protein-DNA complexes. *Genome Biol* 1: REVIEWS001.

Lusted, L.B. 1971. Decision-making studies in patient management. *N Engl J Med* 284: 416-424.

Ma, H., Kunes, S., Schatz, P.J., and Botstein, D. 1987. Plasmid construction by homologous recombination in yeast. *Gene* 58: 201-216.

MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D., and Fraenkel, E. 2006. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* 7: 113.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.

McKnight, S.L. and Kingsbury, R. 1982. Transcriptional control signals of a eukaryotic protein-coding gene. *Science* 217: 316-324.

Meijer, A.H., Ouwerkerk, P.B., and Hoge, J.H. 1998. Vectors for transcription factor cloning and target site identification by means of genetic selection in yeast. *Yeast* 14: 1407-1415.

Mitra, R., Fain-Thornton, J., and Craig, N.L. 2008. piggyBac can bypass DNA synthesis during cut and paste transposition. *Embo J* 27: 1097-1109.

Mortazavi, A., Leeper Thompson, E.C., Garcia, S.T., Myers, R.M., and Wold, B. 2006. Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire. *Genome Res* 16: 1208-1221.

Natarajan, K., Meyer, M.R., Jackson, B.M., Slade, D., Roberts, C., Hinnebusch, A.G., and Marton, M.J. 2001. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Mol Cell Biol* 21: 4347-4368.

Nutt, S.L., Heavey, B., Rolink, A.G., and Busslinger, M. 1999. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature* 401: 556-562.

Ochman, H., Gerber, A.S., and Hartl, D.L. 1988. Genetic applications of an inverse polymerase chain reaction. *Genetics* 120: 621-623.

Oliphant, A.R., Brandl, C.J., and Struhl, K. 1989. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol Cell Biol* 9: 2944-2949.

Orlando, V. 2000. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* 25: 99-104.

Oshima, Y., Ogawa, N., and Harashima, S. 1996. Regulation of phosphatase synthesis in Saccharomyces cerevisiae--a review. *Gene* 179: 171-177.

Ozcan, S. and Johnston, M. 1999. Function and regulation of yeast hexose transporters. *Microbiol Mol Biol Rev* 63: 554-569.

Pelham, H.R. 1982. A regulatory upstream promoter element in the Drosophila hsp 70 heat-shock gene. *Cell* 30: 517-528.

Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N.M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P.A., Herbolsheimer, E. et al. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 122: 517-527.

Potter, C.J., Turenchalk, G.S., and Xu, T. 2000. Drosophila in cancer research. An expanding role. *Trends Genet* 16: 33-39.

Ptashne, M. and Gann, A. 1997. Transcriptional activation by recruitment. *Nature* 386: 569-577.

Reece-Hoyes, J.S., Deplancke, B., Shingles, J., Grove, C.A., Hope, I.A., and Walhout, A.J. 2005. A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol* 6: R110.

Reid, J.L., Iyer, V.R., Brown, P.O., and Struhl, K. 2000. Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol Cell* 6: 1297-1307.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E. et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290: 2306-2309.

Roberts, C.J., Nelson, B., Marton, M.J., Stoughton, R., Meyer, M.R., Bennett, H.A., He, Y.D., Dai, H., Walker, W.L., Hughes, T.R. et al. 2000. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287: 873-880.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A. et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4: 651-657.

Santiago, Y., Chan, E., Liu, P.Q., Orlando, S., Zhang, L., Urnov, F.D., Holmes, M.C., Guschin, D., Waite, A., Miller, J.C. et al. 2008. Targeted gene knockout in mammalian cells by using engineered zinc-finger nucleases. *Proc Natl Acad Sci U S A* 105: 5809-5814.

Schmid, M., Durussel, T., and Laemmli, U.K. 2004. ChIC and ChEC; genomic mapping of chromatin proteins. *Mol Cell* 16: 147-157.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728-1732.

Sikorski, R.S. and Hieter, P. 1989. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics* 122: 19-27.

Singh, H., LeBowitz, J.H., Baldwin, A.S., Jr., and Sharp, P.A. 1988. Molecular cloning of an enhancer binding protein: isolation by screening of an expression library with a recognition site DNA. *Cell* 52: 415-423.

Solomon, M.J., Larsen, P.L., and Varshavsky, A. 1988. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 53: 937-947.

Solomon, M.J. and Varshavsky, A. 1985. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A* 82: 6470-6474.

Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S.G., Cyert, M., Hughes, T.R. et al. 2006. Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* 21: 319-330.

Spilianakis, C.G. and Flavell, R.A. 2004. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nat Immunol* 5: 1017-1027.

Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R., and Flavell, R.A. 2005. Interchromosomal associations between alternatively expressed loci. *Nature* 435: 637-645.

Strathern, J.N. and Higgins, D.R. 1991. Recovery of plasmids from yeast into Escherichia coli: shuttle vectors. *Methods Enzymol* 194: 319-329.

Struhl, K. 1981. Deletion mapping a eukaryotic promoter. *Proc Natl Acad Sci U S A* 78: 4461-4465.

Struhl, K. 1995. Yeast transcriptional regulatory mechanisms. *Annu Rev Genet* 29: 651-674.

Takahashi, K. and Yamanaka, S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126: 663-676.

Tan, W., Dong, Z., Wilkinson, T.A., Barbas, C.F., 3rd, and Chow, S.A. 2006. Human immunodeficiency virus type 1 incorporated with fusion proteins consisting of integrase and the designed polydactyl zinc finger protein E2C can bias integration of viral DNA into a predetermined chromosomal region in human cells. *J Virol* 80: 1939-1948.

Tice-Baldwin, K., Fink, G.R., and Arndt, K.T. 1989. BAS1 has a Myb motif and activates HIS4 transcription only in combination with BAS2. *Science* 246: 931-935.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403: 623-627.

Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M.A., Groudine, M., Weiss, M.J., Dekker, J., and Blobel, G.A. 2005. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol Cell* 17: 453-462.

van Steensel, B., Delrow, J., and Henikoff, S. 2001. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat Genet* 27: 304-308.

van Steensel, B. and Henikoff, S. 2000. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol* 18: 424-428.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. 2009. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 10: 252-263.

Vermeirssen, V., Barrasa, M.I., Hidalgo, C.A., Babon, J.A., Sequerra, R., Doucette-Stamm, L., Barabasi, A.L., and Walhout, A.J. 2007. Transcription factor modularity in a gene-centered C. elegans core neuronal protein-DNA interaction network. *Genome Res* 17: 1061-1071.

Wach, A., Brachat, A., Pohlmann, R., and Philippsen, P. 1994. New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. *Yeast* 10: 1793-1808.

Wang, H., Heinz, M.E., Crosby, S.D., Johnston, M., and Mitra, R.D. 2008a. 'Calling Cards' method for high-throughput identification of targets of yeast DNA-binding proteins. *Nat Protoc* 3: 1569-1577.

Wang, H., Johnston, M., and Mitra, R.D. 2007. Calling cards for DNA-binding proteins. *Genome Res* 17: 1202-1209.

Wang, M.M. and Reed, R.R. 1993. Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast. *Nature* 364: 121-126.

Wang, W., Lin, C., Lu, D., Ning, Z., Cox, T., Melvin, D., Wang, X., Bradley, A., and Liu, P. 2008b. Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proc Natl Acad Sci U S A* 105: 9290-9295.

Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z. et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124: 207-219.

Weintraub, H., Tapscott, S.J., Davis, R.L., Thayer, M.J., Adam, M.A., Lassar, A.B., and Miller, A.D. 1989. Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc Natl Acad Sci U S A* 86: 5434-5438.

West, R.W., Jr., Yocum, R.R., and Ptashne, M. 1984. Saccharomyces cerevisiae GAL1-GAL10 divergent promoter region: location and function of the upstream activating sequence UASG. *Mol Cell Biol* 4: 2467-2478.

Wheelan, S.J., Scheifele, L.Z., Martinez-Murillo, F., Irizarry, R.A., and Boeke, J.D. 2006. Transposon insertion site profiling chip (TIP-chip). *Proc Natl Acad Sci U S A* 103: 17632-17637.

Wilson, D., Charoensawan, V., Kummerfeld, S.K., and Teichmann, S.A. 2008. DBD--taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* 36: D88-92.

Wilson, M.H., Coates, C.J., and George, A.L., Jr. 2007. PiggyBac transposon-mediated gene transfer in human cells. *Mol Ther* 15: 139-145.

Wilson, T.E., Fahrner, T.J., Johnston, M., and Milbrandt, J. 1991. Identification of the DNA binding site for NGFI-B by genetic selection in yeast. *Science* 252: 1296-1300.

Wishart, J.A., Hayes, A., Wardleworth, L., Zhang, N., and Oliver, S.G. 2005. Doxycycline, the drug used to control the tet-regulatable promoter system, has no effect on global gene expression in Saccharomyces cerevisiae. *Yeast* 22: 565-569.

Wu, C., Wilson, S., Walker, B., Dawid, I., Paisley, T., Zimarino, V., and Ueda, H. 1987. Purification and properties of Drosophila heat shock activator protein. *Science* 238: 1247-1253.

Wu, S.C., Meir, Y.J., Coates, C.J., Handler, A.M., Pelczar, P., Moisyadi, S., and Kaminski, J.M. 2006. piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. *Proc Natl Acad Sci U S A* 103: 15008-15013.

Xie, W., Gai, X., Zhu, Y., Zappulla, D.C., Sternglanz, R., and Voytas, D.F. 2001. Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. *Mol Cell Biol* 21: 6606-6614.

Yuan, D.S., Pan, X., Ooi, S.L., Peyser, B.D., Spencer, F.A., Irizarry, R.A., and Boeke, J.D. 2005. Improved microarray methods for profiling the Yeast Knockout strain collection. *Nucleic Acids Res* 33: e103.

Zeitlinger, J., Zinzen, R.P., Stark, A., Kellis, M., Zhang, H., Young, R.A., and Levine, M. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev* 21: 385-390.

Zhu, C., Byers, K.J., McCord, R.P., Shi, Z., Berger, M.F., Newburger, D.E., Saulrieta, K., Smith, Z., Shah, M.V., Radhakrishnan, M. et al. 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* 19: 556-566.

Zhu, Y., Dai, J., Fuerst, P.G., and Voytas, D.F. 2003. Controlling integration specificity of a yeast retrotransposon. *Proc Natl Acad Sci U S A* 100: 5891-5895.

Zhu, Y., Zou, S., Wright, D.A., and Voytas, D.F. 1999. Tagging chromatin with retrotransposons: target specificity of the Saccharomyces Ty5 retrotransposon changes with the chromosomal localization of Sir3p and Sir4p. *Genes Dev* 13: 2738-2749.

Zou, S., Ke, N., Kim, J.M., and Voytas, D.F. 1996. The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. *Genes Dev* 10: 634-645.

Zou, S., Wright, D.A., and Voytas, D.F. 1995. The Saccharomyces Ty5 retrotransposon family is associated with origins of DNA replication at the telomeres and the silent mating locus HMR. *Proc Natl Acad Sci U S A* 92: 920-924.