In silico and wet lab approaches to study transcriptional regulation

Hestand, M.S.

Chapter 1

Introduction

Today we have the technology to quickly sequence entire genomes, but annotating those sequences is still a daunting task. Discerning their function is even a more massive challenge. With only four nucleotides, the genome encodes tens of thousands of genes. Sequence content also determines regulation, providing sites for regulatory elements to control gene transcription. Regulatory elements that bind to genomic DNA can be in the form of proteins termed transcription factors (TFs). However, regulation goes beyond just sequence, encompassing epigenetic factors, from methylation to chromatin remodeling. To even further complicate the picture, regulation can occur at the RNA level by microRNAs, degradation, and alternative splicing. Translational control and post-translational modifications may also further determine the final gene product (a protein for many genes). The comprehensive picture is extremely complicated and too large for one individual to master. This thesis is devoted to one fraction of this picture: TFs and their target binding sites. We have studied two biological processes: the cell cycle (control) and myogenesis. By using a combination of *in silico* and wet lab work, including next-generation sequencing technology, we can better understand the TFs involved in transcriptional regulation of these processes, as outlined in this thesis.

1.1 Biological Background Information

Genetics and Genomics

The genomic code is embedded in our DNA, which is composed of a double helix of strands of nucleotides. There are four nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). DNA can be transcribed into RNA, composed of the same nucleotides other than T becoming uracil (U). RNA synthesis occurs in what is termed a $5'$ to $3'$ direction, using the DNA as a template. RNA in turn can be directly functional or translated into amino acids, the building blocks of proteins. Genetics is the study of genes. Classically genes were considered the portions of DNA that are transcribed into RNA, which is spliced in higher organisms. The portions of RNA (and corrosponding original DNA template sequence) that is retained after splicing are called exons and the portions removed are called introns. Most of the

spliced RNA is then translated into amino acids. What is not translated is called the untranslated region (UTR).

All the nucleotides in a person's DNA make up their genome. Traditionally, focus was only on all the genes in an organism. However, as our knowledge expanded to comprise regulatory elements not within genes the full genome became an interest of study. Genomics is the study of the genome. This includes how much of a gene is transcribed into RNA, termed gene expression, or transcriptomics. The number of genes in the human genome is difficult to know for sure. One explanation is that in one annotation transcripts may overlap constituting one gene and in another annotation the overlap may not be found indicating two separate genes. The initial sequencing of the human genome estimated 30,000 to 40,000 protein-coding genes (1) and a year later the number of genes was estimated to be closer to the low end of 30,000 protein-coding genes (2). Currently, as annotated by Ensembl v53 (3), there are 37,435 total genes (Biomart query, including non-protein coding genes (3; 4)). As more and more high-throughput datasets become available this number should become more reliable.

TFs and Promoters

TFs are regulatory proteins, or protein complexes, that bind to DNA, and positively or negatively influence gene expression. Pattern finding algorithms have been developed to identify TF binding sites (TFBSs) that are presumed to occur in a group of nucleotide sequences. A group of target nucleotide sequences could be known promoters. Promoters are typically a variable amount of base pairs (bp) surrounding the transcription start site (TSS) at the $5'$ end of a gene. For promoters, pattern finding is based on the presumption that promoters with similar regulation/expression have common regulators, and therefore similar TFBSs in their sequences. These regulating TFBSs should therefore have a high occurrence in similarly regulated/expressed genes' promoters.

TFs may also bind other TFs, and are then termed coactivators. There is evidence that some TFs may preferentially bind one strand of the DNA (5). Traditionally, the binding sites of TFs were looked for in the promoter region. One early example of promoter binding is Sp1, which binds the promoter region of beta globin genes (6), as well as 1641 promoters in an additional study(7).

Properly defining the promoter region of genes has been difficult. The promoter is often considered around the TSS, so the first exons of currently annotated genes indicate potential promoters. Many traditional annotation approaches have been results of sequencing RNA and aligning those sequences to a reference genome to infer exon locations. This was often done from the $3'$ end and the process was often considered complete when a full coding sequence was determined. Therefore, many exons with non-protein coding sequence were not annotated. In addition, genes do not necessarily have a single transcript per gene. Often, genes have multiple transcripts, comprised of different combinations of exons. This is a process that contributes to cells being different in one tissue than another. Due to these alternative transcripts, the promoter being used in a specific cell may be around a different exon than the annotated first exon of a gene.

Promoters may also be divided into several classes. Some promoters, such as those containing a TATA box (the target sequence of the polymerase II complex),

have TSSs with very specific locations, whereas others may contain broad TSSs with multiple positions of transcription initiation (8). The first class of promoters tend to be tissue specific, whereas the second class is more likely to be associated with house keeping genes (8). The latter promoters also tend to contain a higher number of GC dinucleotides than expected, termed CpG islands (8; 9; 10). CpG island promoters encompass a majority of mammalian promoters, whereas a minority of promoters are CpG poor (8). These issues are important to keep in mind since TFs may have a preference for one promoter type over another.

However, TFs may bind regions other than the promoter. When looking at some TFs, such as p53, TFBSs may also be located in introns and 3′ regions (11). TFs may actually bind far from a gene. These regions, which also regulate gene expression, are termed enhancers. The difference between promoters and enhancers is that both are regulatory regions, but promoters also contains the sites that basic transcriptional machinery binds to.

Whether a TF can bind its target DNA or not can also be regulated by the accessibility of the DNA. Open chromatin, accessible to the transcriptional machinery and associated with active gene expression, is termed euchromatin. Many epigenetic factors (not encoded by the DNA) are associated with euchromatin, including hypomethylation of CpG islands, multiple histone modifications and variants, and chromatin remodeling complexes (12). All of these factors can therefore have an influence on whether a TF can bind its target DNA or not. Whole regions of the chromosome, potentially containing multiple genes, may be regulated by what are termed locus control regions. One example is that of the locus control region for beta globin genes where binding of proteins to the locus control region play a critical role in multiple (up to 80 kb away) genes' activation (reviewed in (13)).

Better understanding TFs will give us greater knowledge into how the genome is regulated. In a larger view it may help us to even define what makes us human. With the high concordance between coding DNA in the human and chimpanzee (>99% at the protein level) it has long been believed that what largely makes us human is not the genes themselves, but the regulation of their transcriptome (14).

The Cell Cycle

One of the hallmarks of living cells is the process of cell duplication. This so-called cell (division) cycle is a tightly regulated process due to the expression and activation of stage-specific proteins that control the different cell cycle transitions (G1/S, S, and G2/M phases; reviewed by Satyanarayana and Kaldis, 2009 (15) and Malumbres and Barbacid, 2009 (16)). Loss of control of the cell cycle can lead to increased cell proliferation, resulting in tumors. By better understanding the regulators of the cell cycle scientists hope to guide research into cures for diseases such as cancer. Chapter 5 of this thesis involves a study of TFs which play a role in the cell cycle.

Many factors contribute to cell cycle regulation, including hormones, growth factors, cytokines, cyclin-dependent protein kinases, cyclins, the retinoblastoma (RB) protein, bcl-2 protein, myc protein, bax protein, the E2F family of TFs, and the TF p53 (17; 18). The tumor suppressor p53 is a crucial cell cycle regulator, with an estimated 50% of tumors carrying a mutation in the p53 encoding gene (18). In Chapter 3, based on *in silico* predictions, we identify TFs that potentially cooperate with p53.

p53 itself can be regulated by coactivators such as p300 and CBP (19). Besides by interactions with TFs, these acetyltransferases also regulate gene expression by altering chromatin accessibility via the acetylation of proximal nucleosomal histones. Despite their high levels of homology, the coactivators are not able to substitute for each other during embryogenesis as was shown by mouse knockout experiments (20; 21). Thus, in chapter 5 we selected these two coactivators for study.

Myogenesis

Several other parts of this thesis (chapters 2, 3,and 6) aim at elucidating the roles of TFs regulating myogenesis. Myogenesis is the process of muscle formation and development. The process of myogenesis may be divided into two parts: embryonic and adult. During embryogenesis somites develop into mesodermal precursor cells (22; 23). These mesodermal precursor cells are pushed towards a myogenic lineage by two primary myogenic TFs: MyoD and Myf5 (22). These resulting cells are termed myoblasts, which further differentiate into primary and secondary myofibers.

We focus on the process of adult myogenesis, through which myoblasts cease proliferating and fuse together to form multinucleated myofibers. In skeletal muscle this was traditionally and simplistically believed to be controlled by four major TFs: MyoD, Myf5, Myogenin, and MRF4, with the first two functioning in early differentiation and latter two in late differentiation (24). However, as our knowledge of biological pathways and processes expands it is becoming apparent that many TFs and other elements are responsible for the regulation of myogenesis. Besides the four major TFs, Charge *et al.* 2004 review many molecules, including other TFs (including Pax7, Pax3, Slug, myocyte nuclear factor (MNF), and Msx1) that contribute to myogenesis (22). A year later an initial blueprint of myogenic differentiation was published including MyoD, Myogenin, and MEF2 targeting a large number of additional TFs, with connections being made to TEAD4/TEF-3, ARNT, Copeb/KLF6, NFE2l2/NRF2, and ATF4 (25). As genetics moves forward it is likely more and more TFs will be identified that play a role in myogenesis.

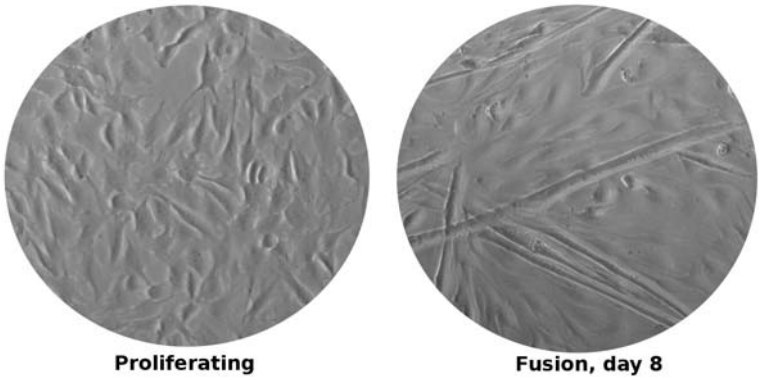

Figure 1.1: Proliferating and Differentiating Mouse C2C12 cells

Several systems exist to study myogenesis in the laboratory. These includes patient

samples, mouse strains, and cell lines. This thesis primarily uses a mouse cell line termed C2C12. These cells proliferate with serum, but when serum deprived stop proliferating and begin to differentiate and fuse into myotubes (Figure 1.1). This process typically takes seven to nine days.

Defects in myogenic regulation (via a TF mutation or alteration of its target) result in a multitude of diseases, including myotonic dystrophy, rhabdomyosarcomas, Waardenburg syndrome type 2, congenital myasthenia, and diseases related to muscle regeneration (overview in Martin 2003 (26)). By gaining a better understanding of the genetic architecture of late myogenesis we hope to aid researchers towards developing cures for such illnesses.

1.2 Conventional Wet-lab Methods

RNA Expression

Many kits and techniques now exist for the isolation of RNA. A traditional method for over twenty years is an extraction with guanidinium thiocyanate, Phenol-chloroform, and sodium acetate, followed by isoproponal precipitation clean-up (27).

Serial Analysis of Gene Expression (SAGE) (28) (Figure 1.2) and Cap Analysis of Gene Expression (CAGE) (29) (Figure 1.3) are two methods to isolate small parts at either end of mRNAs. These were classically concatenated and cloned into libraries and then sequenced. With next generation sequencing technology (see below) it is possible to directly sequence the SAGE/CAGE sequences (termed DeepSAGE (30) and DeepCAGE (31)).

SAGE is a method developed to quantify all the transcripts expressed in a genome (28). This commonly works by isolating RNA poly-A tails with oligo(dT) beads, converting into cDNA, performing a first restriction digest (NlaIII which cuts at CATG's), retaining the 3′ most fragments, adding a linker to the 5′ end with a restriction site, then using an additional enzyme that recogizes the linker site (such as MmeI) to cut a certain number of bp from the 5′ end each fragment, typically 14-20, adding a second linker to the 3′ end, and finally cloned and sequenced (32) (Figure 1.2).

CAGE is a technique to sequence the 5′ end of transcripts and therefore better annotate TSSs, which can be used to provide better promoter annotation (29). CAGE works first by creating single strand cDNA and then capturing the 5′ cap, present on all mRNAs, with an antibody or biotinylated cap-trapper (Figure 1.3)(29). A linker sequence is then added to the 5′ end which contains sequence to bind to the sequencer's glass slide (for Illumina next-generation sequencing), a sequencing primer, and a restriction enzyme site. Double strand cDNA synthesis is then performed and a restriction enzyme actually cuts a number of bp downstream of the linker restriction enzyme site, providing approximately 20-26 bp of the original 5′ end of the transcript. A final linker is added for the sequencing protocol and in current protocols the library is run through a next-generation sequencing machine.

In contrast to 5′ or 3′-end focused methods, true whole transcriptome sequencing, also called mRNA-seq, is a method by which cDNA generated on the total RNA by random priming is amplified, sheared, and sequenced (33). This method therefore provides a more complete picture of RNAs, but can be more complicated for expression

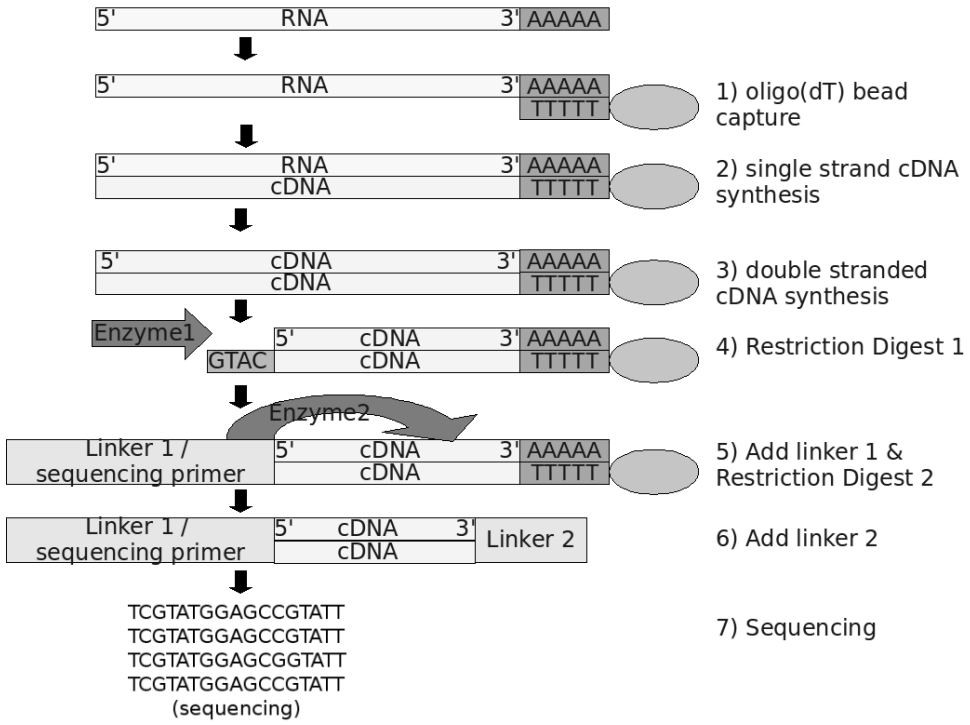

Figure 1.2: DeepSAGE: Sequencing of serial analysis of gene expression libraries starts by capturing the poly-A tail of mRNAs with oligo-dT beads. RNA is converted to cDNA and made double stranded, followed by a first restriction digest. The 3′ most fragments are retained, sequencing specific linkers adapted with a restriction site, and a second restriction digest performed that cuts downstream of the introduced restriction site. A second linker sequence is adapted and next-generation sequencing can then be performed.

analysis since one transcript may be represented by a greater diversity of tags. Having multiple random tags per transcript also reduces the quantity of total transcripts detected, reducing statistical power for calling differential expression levels. This is increasingly offset by the major increases in sequencing depth.

Isolating TF bound DNA

Chromatin immunoprecipitation (ChIP), is a wet lab technique to identify the targets of a specific TF (Figure 1.4). In general, this technique begins by formaldehyde fixing cells so that the TFs are fixed to the DNA. The cells and nucleus are then lysed, often with detergents, and the chromatin (DNA bound by RNA and protein) is isolated and cleaned up. This chromatin is then fragmented with chemicals or sonication. TF bound fragments of chromatin are then immunoprecipitated using an antibody targeting the TF of choice. This isolated pool of TF bound chromatin fragments

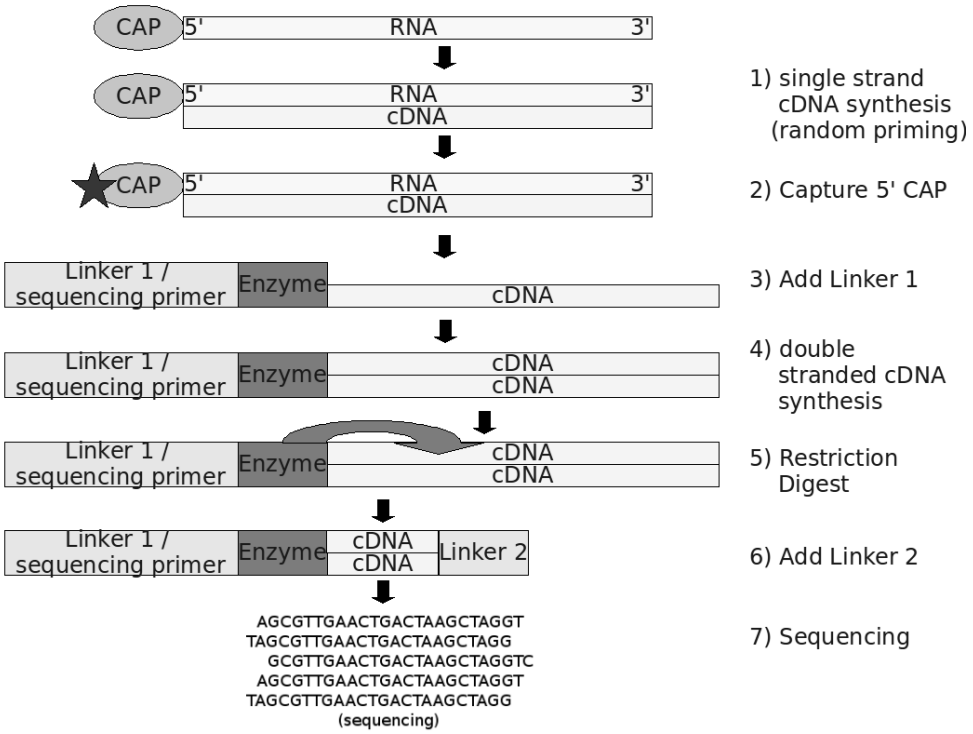

Figure 1.3: DeepCAGE: Sequencing of cap analysis of gene expression libraries starts with random priming for single strand cDNA synthesis and then capturing RNAs by their 5′ cap. A linker with a restriction site and sequencing linker is ligated and double stranded cDNA synthesized. A restriction enzyme is used that cuts downstream of the restriction site. The 5′ fragments are retained and a second linker ligated to the 3′ end of the fragment. These linker adapted sequences can then be applied to next-generation sequencers.

are then reverse cross-linked and cleaned up to leave only DNA fragments that were originally bound by the TF of interest.

The ChIP wet-lab method can be coupled with several genomic technologies to analyze ChIP target sequences genome-wide. When ChIP sequences are hybridized to a microarray (see below) it is termed ChIP-chip (or ChIP-on-chip) (34). An alternate approach is massive parallel sequencing, either with a paired-end ditag approach (ChIP-PET) (11), or directly using a next-generation sequencer (ChIP-seq) (35), as addressed below. These methods both start with ChIP, resulting in a pool of TF bound DNA. In ChIP-PET these are cloned into a plasmid vector, converted to concatenated and cloned PETS, and then sequenced (11). ChIP-seq is less laborious, omitting the cloning and concatenation steps, by just directly ligating linkers and sequencing the ChIP DNA.

Only several ChIP-seq experiments have been published at the time of this thesis, though large numbers of ChIP-chip studies have been published. ChIP-seq is expected

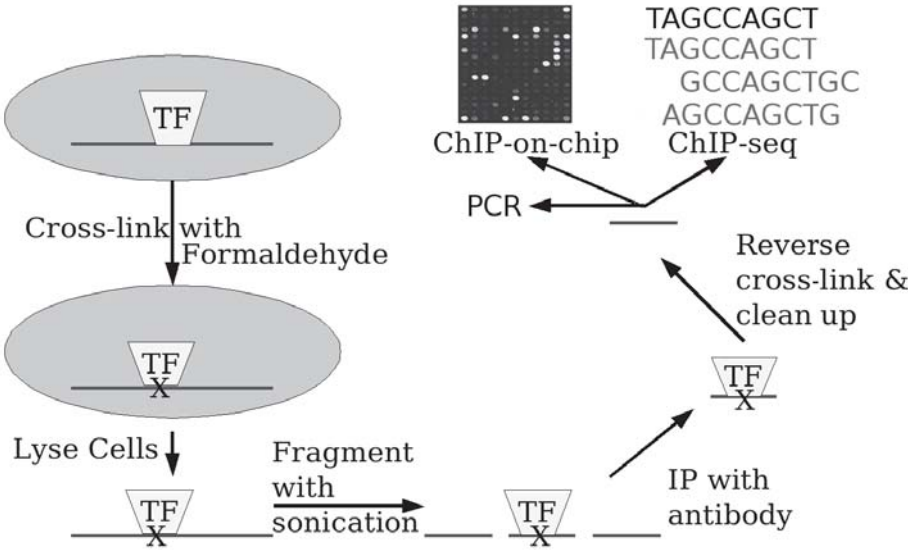

Figure 1.4: ChIP techniques: Chromatin immunoprecipitation (ChIP) works by cross-linking TFs to the DNA with formaldehyde, lysing cells, fragmenting chromatin with sonication, immunoprecipitating TF bound DNA fragments with an antibody, reverse cross-linking to remove TFs, and cleaning up the final pool of DNA fragments (originally bound by the TF of interest). These pools of DNA fragments can be analyzed by PCR, microarray (ChIP-(on)-chip), or next-generation sequencers (ChIP-seq).

to be an up and coming technology. It has the advantage, in comparison to ChIP-chip, of requiring less input material, the potential to identify TFBSs with low affinity, not being limited to target regions (*i.e.* probes on a microarray), not having hybridization errors, and is less costly for whole genome analysis (35). This method will rewrite the books on how TFs bind genome wide, identifying many TFBSs in intragenic regions that were not studied previously or were bound at too low a concentration to be detected by microarrays. This additional wealth of data will provide more sequences to mine for position weight matrices (PWMs, see below) and improve upon existing PWMs, resulting in improved *in silico* predictions.

Single Target Readout methods

The polymerase chain reaction (PCR) is a method to amplify a stretch (up to several kilobases) of DNA. DNA regions are targeted using primers specific to the DNA region. A polymerase is used to read the DNA and replicate it. The simplest use of this is to see if the DNA stretch is present in the genome. After amplification the product can be viewed on an agarose gel, and if appropriate markers are included the size can be estimated. The intensity of a band (compared with a control sample) on this gel can represent the relative quantity of DNA in the original sample, but to be more precise an adaptation of PCR is used. Quantitative real-time PCR, termed qPCR, uses fluorescent dyes or probes to quantify the amount of target DNA. RNA

can also be converted to cDNA with a reverse transcriptase and qPCR performed, termed RT-qPCR. This is especially useful and cost-effective to determine expression levels of RNA because of simplicity and high sensitivity.

Other methods also exist to detect TF bound DNA. This includes luciferase assays, deletion constructs, gel shift assays, and the TransFactor kit. Luciferase assays are a technology in which a promoter from a gene is cloned in front of a gene encoding a luciferase gene. When activating TFs bind to this promoter they activate the luciferase gene, causing the cell or organism to produce light under proper conditions. Deletion constructs are a means of eliminating a portion of a gene's promoter, then observing the effect.

Gel shift assays, involves running DNA through a gel. If a stretch of DNA has a TF bound to it, the sequence will run out slower on a gel. This is a relatively faster method than the previous two, but only indicates binding and no regulatory function. The TransFactor kit works on a similar level, determining binding of a TF to a target DNA sequence using a TF specific antibody, a secondary antibody, and colorimetry.

High-throughput Readout methods

Microarrays were one of the first technologies to study genetics at a genomic scale in a single test. Microarrays traditionally consist of a glass slide with thousands or millions of probes attached to it. These probes have sequences that bind target sequences. The target sequences are labeled with a dye that cause bound probes to give a fluorescent signal. Therefore, spots, consisting of clusters of probes, give a signal relative to the quantity of their target sequences in the sample analyzed. The most common use of microarrays involves hybridizing RNA to study gene expression levels.

Alternatively, microarrays used in conjunction with ChIP can search for a large number of TF targets. Promoter based and whole genome tiling arrays also exist to analyze the afore mentioned ChIP samples. These arrays consist of probes that are "tiled" (spaced) across promoters, or the entire genome. These can provide ideal target regions to study ChIP.

In the past few years several new technology platforms have emerged that perform DNA sequencing on a massive scale at a fraction of the speed and cost of traditional sequencing technologies. The primary three systems are the 454 by Roche, the Illumina Genome Analyzer (formerly Solexa) by Illumina, and the SOLiD by Applied Biosystems. Our department has two Illumina Genome Analyzers so this thesis's next-generation sequencing (NGS) has been performed on this system.

Though all classified as second or next-generation sequencers, these platforms have very different mechanics. The 454 is based on attaching DNA fragments to beads (one fragment to one bead), emulsion PCR amplification of the fragments on the beads, and loaded onto a PicoTiterPlate (one bead per well) for sequencing (*www.454.com*). Sequencing is performed by sequentially adding complementary nucleotides that emit a florescent signal, detected by a camera (*www.454.com*). SOLiD also uses beads and emulsion PCR, but then the amplified products are applied to a glass slide (*www.appliedbiosystems.com*). Several series of ligations are performed in which fluorescently labeled di-base probes are used for detection (*www.appliedbiosystems.com*). This system differs in that a fluorescent signal does not reflect the addition of an exact

nucleotide, but a pair (which is termed colorspace) (*www.appliedbiosystems.com*). Illumina differs in that no beads or emulsion PCR are used. Adapter ligated sequences are first attached to a slide, and then bridge amplification is performed on the slide (*www.illumina.com*). Nucleotides are then sequentially added which emit a different fluorescent signal for each of the four nucleotides, which is recorded by a camera (*www.illumina.com*).

Table 1.1: Next-Generation Sequencing System Specifications

Company	Applied Biosystems	Roche	Illumina	Applied Biosystems
Machine	traditional sequencing (3730xl DNA Analyzer)	FLX Titanium	Genome Analyzer IIx	SOLiD 3 Plus System
read length	up to 900 bp	400-500* bp	35-100 bp	35-100* bp
# reads per run	96 or 384 x 16 plates	∼1 million	∼150-200 million*	∼200 million*
run time	0.5-3 hours	10 hours	2-9.5 days**	3.5-14 days**
reference	*www.applied biosystems.com*	*www.454.com*	*www.illumina.com*	*www.applied biosystems.com*

*Numbers from website adapted based on personal experience. **Run times depend on the number of cycles (bp sequenced per read). Machine details are based on website specifications in February 2010.

These systems can produce vast amounts of data, however the read length, total bp sequenced, and sequence time vary between instruments (Table 1.1). The read length and total bp sequenced are also continuously increasing with advancements in chemistry and mechanics. It has been shown that next-generation sequencers outperform microarrays in precision, reproducibility, and sensitivity, likely by avoiding the problems associated with hybridization techniques (36). NGS (also called deep-sequencing or second-generation sequencing) also escapes the limitation of only looking at the targets that have been spotted on a microarray, *i.e.* performing a "content-limited" analysis.

Typically NGS analysis begins by converting data to sequences and filtering for quality. For the Illumina Genome Analyzer, this means converting image files and filtering on quality with their pipeline. For most NGS applications the next step is to align to a reference genome. Traditionally for longer reads alignments could be done with BLAST (37) or BLAT (38), but these algorithms do not perform well with large numbers of short reads, such as those provided by the Illumina Genome Analyzer. To align short reads many different alignment algorithms have been developed in the past years, including Eland (part of the Illumina GA Analysis Pipeline: fast, but only good for reads $\leq$ 32bp), Maq (39), Rmap (slow, but accurate) (40), Cloudburst (fast and accurate, but large system requirements) (41), Bowtie (fast) (42), and BWA (fast) (43). When a reference genome is not available, sequences are often built into contigs with the tool Velvet (44). From here analysis is very dependent on the application being analyzed.

1.3 *In silico* Prediction of TFs and TFBSs

Pattern Finders

As mentioned earlier, pattern finding algorithms can be used to identify TFBSs in sets of TF bound DNA sequences. Modern pattern finders include MEME (45; 46) and Gibbs samplers (47; 48; 49), which can find one or more variable patterns in DNA or protein sequences.

Position Weight Matrices

One method to identify TFBSs for known TFs is using PWMs (50). These matrices summarize experimental information on the sequential preference of a TF (Figure 1.5). The two leading databases of experimentally determined PWMs are TRANSFAC (51; 52) and JASPAR (53; 54). TRANSFAC has the advantage of more PWMs (834 matrices (release 11.4, December 2007)) (52) compared to JASPAR (123 matrices) (54). However, to use the larger TRANSFAC Professional (there is also a smaller public version free to all non-commercial users) a paid license is required, whereas JASPAR is free. These PWMs are used by programs like Match (51; 55) or Sunflower (56) to identify TFBSs in a nucleotide sequence by evaluating the nucleotide similarity of the PWM with the sequence.

	C	A	T	G
Nucleotide 1:	0	0	10	0
Nucleotide 2:	1	4	4	1
Nucleotide 3:	0	0	5	5
Nucleotide 4:	8	0	1	1
Nucleotide 5:	1	1	7	1

Figure 1.5: A Theoretical Position Weight Matrix (PWM): At the top is a theoretical chart of a 5 nucleotide PWM made up from 10 experiments. For each nucleotide is a count of how many experiments found that nucleotide. Below is shown a visual representation of the chart information.

Over-Representation of TFBSs

However, even with PWMs, identifying TFBSs is a difficult task, considering genomes may be in the billions of base pairs and TFBSs may be only 12-14 bp in size (49).

One method to improve upon TFBS predictions in a set of genes is to look for over-representation of TFBSs in the promoters of co-regulated/co-expressed genes. Using a similar presumption as described for pattern finders, it is presumed that

similarly regulated/expressed genes' promoters contain common regulators. Therefore, target TFBSs identified through PWMs should occur more often in a similarly regulated/expressed set of genes' promoters than in a random set of genes' promoters. This method has been developed to include work on complex organisms such as human (57). This method relies on using proper target sequences. Therefore, good gene/promoter annotation is critical, such as that provided by CAGE techniques.

Conservation of TFBSs

Another method to look for *de novo* TFBSs is by searching for conservation between orthologous promoters (58). This method is based on the presumption that functional elements are evolutionarily conserved and mutations in these elements could therefore be detrimental to the organism (58; 59). Programs that use conservation to determine TFBSs include oPOSSUM (60) and ConTra (61).

1.4 Thesis Overview

This thesis looks at TFs and TFBSs discovery first through *in silico* predictions based on previous ChIP and expression data, then wet lab work with *in silico* confirmation. Chapter two focuses on CORE_TF, a web site developed to identify over-represented and cross-species conserved TFBSs in a set of similarly regulated genomic regions, such as up-regulated genes' promoters from a microarray study. The third chapter achieves a similar goal to chapter two to identify over-represented TFBSs, but also models competition between TFs, which better models the true biological system and, thus, improves results. Chapter four presents a pipeline, titled GAPSS, to analyze NGS data that was used for data analysis of chapters five and six. Chapter five focuses on ChIP-seq wet-lab work and data-analysis, including GAPSS and CORE_TF, to better understand the role of CBP and p300 in cell cycle control. The sixth chapter primarily focuses on using CAGE to better annotate muscle specific TSSs which should improve promoter based TFBS predictions. Chapters seven to nine wrap up this work, explaining how a combination of multiple *in silico* and wet lab techniques lead to a better understanding of the transcriptional control of genes.