Chapter 8

Summary

Transcription factors (TFs) are an essential part of gene regulation. Mutations in TFs, and their binding sites (TFBSs), can result in muscle diseases such as myotonic dystrophy, rhabdomyosarcoma, Waardenburg syndrome type 2, congenital myasthenia, and diseases related to muscle regeneration, (overview in Martin 2003 (26)). 50% of tumors, resulting from loss of cell cycle control, carry a mutation in the TF p53 (18). Therefore, to cure these diseases and many more it is essential we better understand TFs, their genomic targets, and function.

One research area that has improved greatly with modern sequencing technologies is the study of TFs. Chromatin-immunoprecipitation (ChIP) provides a means to isolate stretches of DNA bound by a protein, such as TFs. These fragments of DNA can be sequenced, resulting in genome wide identification of TFBSs. Previously polymerase-chain reactions (PCRs) and micro-array technologies mostly only looked at target regions, such as promoters, missing many binding regions. Currently, next-generation sequencing of ChIP DNA provides full genome (target free) results, at a lower cost, higher reproducibility, and with the ability to detect low-affinity binding sites.

We have focused on two biological processes of interest: the cell cycle and muscle differentiation. Both are essential processes: cell cycle for replication/division and myogenesis for muscle development and regeneration. Defects in the cell cycle result in death and cancer, whereas muscular dystrophies may result from impaired myogenesis. TFs, such as p300 or CBP for cell cycle control, or MyoD or Myogenin for myogenesis, have been identified to regulate their parent processes, though full details of their binding locations and regulation have not been eluted. Still, many other TFs have not been discovered or related to either process. We have made an approach to further broaden our knowledge of cell cycle and myogenic control through known TFs.

In chapter 2 we present a web application called CORE_TF (Conserved and Over-REpresented Transcription Factor binding sites) that identifies TFs that occur more often in an experimental set of sequences compared to a random set of sequences. It also has the ability to identify TFBSs that are conserved across different organisms. Initially this was developed to identify TFs that potentially regulate co-expressed genes from micro-array studies. However, CORE_TF can also be used with next-generation sequencing expression studies and to identify co-regulators from
Summary

micro-array or next-generation sequencing of ChIP samples.

We expanded on CORE_TF’s principle of identifying over-represented TFBSs in Chapter 3. Instead of using CORE_TF’s Match to identify binding sites, we used a novel tool called Sunflower. Sunflower models competition between TFs for the same nucleotide sequences. This is closer to the actual biological state. After identifying potential TFBSs with Sunflower, we used the same statistical test as CORE_TF to identify TFBSs that are enriched in an experimental set compared to a random set. This process is not as user friendly or fast as CORE_TF, but gives improved results.

As we began to implement our own wet-lab work with a next-generation sequencing platform (Illumina’s Genome Analyzer) we realized we needed a general pipeline to begin analysis of our data. Therefore, we developed GAPSS (General Analysis Pipeline for Second-generation Sequencers). As discussed in chapter 4, GAPSS gives us the possibility to quickly edit for contaminating linker sequences, align our data to the genome, make it viewable in a genome browser, and present this data as defined regions.

In chapter 5 we investigated cell cycle control, as regulated by the TFs CBP and p300. ChIP-seq was performed in a model cell line. The data was analyzed initially with the GAPSS pipeline described in chapter 4. By using CORE_TF (chapter 2) we managed to identify TFs that work as partners with CBP and p300. Though these TFs are highly similar and seem to regulate similar genes, we were able to identify targets specific to each TF and potential regulatory partners (e.g. AP-1, AP-2, SP1, and SRF).

Our work in chapters 2 and 3 often relied on analyzing promoter regions. However, often alternative (or previously unannotated) promoters are used during particular processes, in different tissues, and at distinct time points. In chapter 6 we used CAGE and SAGE techniques coupled with next-generation sequencing to provide a better look into the promoters and genes that differed between proliferating and differentiating mouse myoblasts. This used the GAPSS pipeline of chapter 4 for initial data analysis. To prove these novel promoter regions were muscle specific we searched for and found over-representation of muscle specific TFs.

This thesis demonstrates techniques to identify TFs regulating a process, both with novel in silico and modern wet lab techniques, such as next-generation sequencing of ChIP DNA. We elucidated the roles of myogenic and cell cycle control TFs, specifically MyoD, Myogenin, CBP, and p300, but these techniques could be applied to transcriptional control of any other biological process.