ABSTRACT

Title of Dissertation MODEL BASED APPROACHES TO CHARACTERIZE HETEROGENEITY IN GENE REGULATION ACROSS CELLS AND DISEASE TYPES

Mahfuza Sharmin, Doctor of Philosophy, 2017

Directed by Professor Héctor Corrada Bravo
Department of Computer Science
&
Professor Sridhar Hannenhalli
Department of Cell Biology & Molecular Genetics

Access to large genome-wide biological datasets has now enabled computational researchers to tackle long-standing questions in Biomedicine through the lens of Machine Learning (ML) and Artificial Intelligence (AI). The potential benefits of such computational approaches to biological research are immense. For example, efficient, and yet interpretable, machine learning models of disease/drug response/phenotype can impact our life at both personal and social levels. However, heterogeneity is found at multiple scales in biology, manifested as the context-specificity of biological processes. This context-specific heterogeneity poses a major challenge to ML models. Even though context-specific models are often trained, this is mostly done without the benefit of mechanistic insights about the biological processes being modeled, and as such do not help improve our biological understanding.

This dissertation addresses these challenges and their limitations by: a) designing appropriate features and ML models motivated by the current biological hypothesis at hand, b) building pipelines to analyze multiple context-specific models together, and c) developing data integration and imputation methods to address the problems of insufficient and missing data.

The first project studies loss of methylation or hypo-methylation in large blocks causing aberrant gene activity, a well-known phenomenon in cancer. To find the associated markers, I designed a classification model of hypo-methylated block boundaries and non-boundaries in colon cancer.

The second project models binding of transcription factor (TF) to specific DNA element to the genome, one of the principal components of gene regulation. Since condition specificity of TF binding is not yet well understood, this dissertation examines a design of cell type-specific models for transcription factor (TF) binding using ChIPSeq data. A meta-analysis pipeline, called TRISECT, is applied for multiple TF binding models to understand heterogeneity of cell specificity across those models.

Next, models for breast cancer metastasis using gene expression data are discussed. In breast cancer metastasis, the affinity towards distant tissues called secondary tissues has not been comprehended. Therefore, going beyond mere discriminatory models, I propose another meta-analysis pipeline, MONTAGE intending to understand the organotropism of breast cancer metastasis across secondary tissues.

Building ML models can be hindered by the data size, specially, for rare diseases. Therefore, by necessity, molecular data have been merged across multiple studies, and across multiple technical platforms which has vulnerability of so called batch effects diluting the actual biological signal. Existing methods are not capable of removing multi-variate confounding artifacts leading to inaccurate models. To circumvent this issue, this dissertation examines a deep learning based technique (deepSavior) which 'translates' the gene expression profile from samples of one technical platform to another platform.

To summarize, this dissertation makes three distinct contributions, a) designing effective ML model to explore the determinants of cancer-associated hypomethlation, b) designing meta-analysis pipelines to compare multiple related but context-specific ML models to understand heterogeneous relations among biological processes, and b) developing new method to overcome the data integration and imputation challenges.

MODEL BASED APPROACHES TO CHARACTERIZE HETEROGENEITY IN GENE REGULATION ACROSS CELLS AND DISEASE TYPES

By

Mahfuza Sharmin

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Héctor Corrada Bravo, Chair/Advisor
Professor Sridhar Hannenhalli, Co-Chair/Co-Advisor
Professor Hal Daumé III
Professor Stephen M. Mount
Professor Eytan Ruppin

Preface

Portions of the material presented in this dissertation have either been published in peer-reviewed journals or are being prepared for submission to peer-reviewed journals. Contents from Chapter 2 and 2 have been published. Contents from Chapter 4 and 5 are in preparation for submission.

A list of the papers that constitute my dissertation and a list of other papers where I have contributed to thus far.

Chapter 2:

- Sharmin, M, Bravo, HC, Hannenhalli, S. Distinct genomic and epigenomic features demarcate hypo-methylated blocks in colon cancer. BMC Cancer, 16(1), 88.

Chapter 3:

- Sharmin, M, Bravo, HC., Hannenhalli, S. Heterogeneity of transcription factor binding specificity models within and across cell lines. Genome Research, 26(8), 1110–1123.

Chapter 4:

- Sharmin, M, Bravo, H. C., Hannenhalli, S. Modelling metastasis and organotropism for breast cancer using gene expression. (Unpublished).

Chapter 5:

- Sharmin, M, Hasio, J, Vyas, Y, Bravo, HC, Hannenhalli, S. Overcoming cross-platform batches using multi-modal auto-encoder. (Unpublished).

A selection of papers that I have contributed to (being a co-author):

- Basu, M, Sharmin, M, Das, A, Nair, NU, Lee, JS, Chang, YC, Ruppin, E, Hannenhalli, S. Pan-tissue transcriptomic and genetic analyses of hypertension reveals patients subgroups differing in key clinical phenotypes. (Under review in Genetics).
- Magen, A, Das, A, Lee, JS. Sharmin, M, Gutkind, S, Ruppin, E, Hannenhalli, S. Beyond synthetic lethality: multiple other gene interaction types play an important and comparable functional predictive role in cancer. (Unpublished).

Dedication

To Monjura and to my “real” well wishers…

Acknowledgments

In the name of Allah, the most beneficial, the most merciful

I wholeheartedly express my gratitude to my advisors, Prof. Héctor Corrada Bravo and Prof. Sridhar Hannenhalli for their continuous support and guidance. I feel privileged to be part of their research groups. They were incredibly patient with my naivete and lack of experience during the early days of my research. So, I had a chance to grow and today, can see myself as an independent researcher. If not for the second chance to stay in their labs, I would not have been able to turn into who I am today. I also express gratitude to the rest of the dissertation committee (Prof. Eytan Ruppin, Prof. Hal Daumé III, and Prof. Stephen M. Mount) for their guidance and suggestions. Without their direction, it would have been impossible to complete this dissertation. I am fortunate to be mentored by Mihai Pop who was not only a Teaching Mentor for me, but also a person who showed both empathy and care when I needed them the most, without expecting anything.

I feel blessed to have an awesome family, caring cousins, encouraging childhood friends from Bangladesh. Their affection and trust in me did not let me stop looking forward and hoping for the best. Specially, the blind support and mad love from my younger sister Monjura and my mother were the energy to stay strong during every part of my hardships. I really feel lucky to have Jillet, Joyce, Jennifer, Hanan, Heba and Mounica as my roommates. Over the past couple of years, they, one by one, have made my apartment in graduate campus a perfect "home away from home".

I am happy to have met all my colleagues from HCBravo lab, specially Joyce, Florin, Faezeh and Kwame for their support during the hardest times of my graduate life. At times, they were more than any relatives, more than any native speakers of my own home country. I am also happy to have colleagues from Hannenhalli lab: Shrutii, Justin, Kun, Avinash, Hiren, to name a few, for their constant encouragement and assistance. From CBCB, I want to thank many others for their friendliness and warmth: Keith, Justin, Jayaram, Sushant, Mahashewta, Nishant, Joo, Nidhi, Jay, Mohammad, Nate, Nick and others. All of them contributed to make CBCB a second home in the campus. I also thank my 1st and 2nd year friends/seniors/juniors Awalin, Sagar, Jason, Kent, Arun, Garrett, Mohit, Varun, Faezeh, Victoria, Meethu, Sudha, Samet, Sarthak, Ramakrishna, Andres, Aishwarya, Snigdha, Teng, Milad and Melika who definitely made my campus life more exciting and enjoyable.

I want to thank Prof. Ashok Agarwala, Prof. Amr Baz and Prof. Teng Li from the Future Faculty Program, Prof. Bahram Momen from Environmental Science Department and many other Professors from Computer Science Department for their suggestions and inspirations which reshaped my outlook about the path of a

scientific career. I am grateful for receiving not only official assistance from Jodie Gray and Fatima Bangura in AVW but also feel special for their care and love. I also appreciate for all the assistance I have received from Christine Bogan, Jennifer Story, Arlene Schenk, Denise Cross, and other stuffs from UMICAS.

Last, but not least, I thank my physicians, Dr. Susie Lew, Dr. Abraham Dabela, Dr. Meryl Waldman and the assistants Evelyn and Margarita for showing positive attitude and great care for last couple of years, for sharing their knowledge with patience and providing me the opportunity to participate in bio-marker studies. This eventually sparked ideas and spirit inside me to serve humanity by staying in the field of genomics.

List of Figures

List of Tables

1 Background

1.1 Motivation and Contribution

Machine learning (ML) has become mainstream in several domains, including language translation, facial and speech recognition, spam detection, and marketing. The revolution of Machine Learning has also made its way into genomics, especially due to continued technological advances that incessantly increase our ability to comprehensively measure a variety of molecular phenomena over large populations. In fact, access to large genome-wide biological datasets now enable computational researchers to tackle long-standing questions in Biomedicine through the lens of Machine Learning (ML) and Artificial Intelligence (AI). The potential benefits of such computational approaches to biological research are immense, for example, modeling any disease/drug/phenotype would significantly impact our life at both personal and social levels.

Currently, the opportunities for personalized medicine applications are challenged by, a) the complexity of biological systems, and b) the size and complexity of the available datasets to probe biological systems. The former challenge necessitates computational approaches to generate and prioritize hypotheses and the latter demands techniques to fill gaps of missing data and data integration. One way of generating and prioritizing hypotheses is to design effective but interpretable machine learning models. However, heterogeneity is prevalent at multiple scales in biology, manifested as the context-specificity of biological processes and functional effects of individual genes. Such heterogeneity poses additional challenges to computational and statistical modeling. Nonetheless, context-specific models are often built in the presence of such heterogeneity but are mostly used without the benefit of mechanistic insights about the processes being modeled. As such, these models do not help improve our understanding of these biological processes.

This dissertation addresses the above challenges and limitations by: a) designing appropriate features and ML models motivated by the current biological hypotheses at hand, b) by building pipelines to analyze multiple context-specific models together, and c) the development of novel data integration methods.

Loss of methylation or hypo-methylation in large blocks is a very well known phenomena in cancer. Such hypo-methylation leads to aberrant gene activity in cancer. First, I designed a model to identify biological determinants of hypo-methylated block boundaries in colon cancer. The design of this model was motivated from the following observations. Nucleosome and heterochromatin lie near the methylation block boundary of the cell when it is at normal state. However, they shift away from the boundary when the cell goes to cancer state.

Based on the above, I hypothesized that the genetic and epigenetic features of methylation boundaries might explain whether these boundaries have distinct properties compared to the non-boundaries regions and whether the relevant features are responsible for the formation of large block of hypo-methylation and hyper-variability of genes. Based on the model and downstream analysis I found that boundaries have distinct properties, they act like pseudo-promoter even though they are not promoter and the genetic features of methylation boundaries interact with chromatin modifying enzymes.

Second, I designed cell type-specific models for transcription factor (TF) binding. Binding of transcription factor (TF) to specific DNA element to the genome is one of the principal components of gene regulation. However, condition specificity of TF is not yet well understood and we are interested in finding the determinants of TF binding specificity. The TF models used here are built using the sequence features taking the binding information from ChIPSeq (Chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing). Our ensemble based TF binding models (EMT) perform favorably compared to previously published models.

Going beyond mere discriminatory models, I designed and applied a meta-analysis pipeline TRISECT (Ensemble model of TF Binding and Clustering) for a set of tissue specific TF binding models. TRISECT aims to understand heterogeneity of multiple cell specific TF binding models. Using TRISECT, I demonstrated that TF can have both ubiquitous and cell type-specific functions. The rules that govern binding of a TF to DNA can exhibit different levels of heterogeneity, contributed by interaction partners and such binding rules can transcend cell types, and are informative of the function of the gene targets.

Third, I built models for breast cancer metastasis in distant organ specific fashion using gene expression data of primary tissue (breast). The target organ-specific metastasis models showed 70-90% AUC-ROC (Area Under Receiver Operating Curve). No models for this task were previously reported. It is well-known that when cancer cells spread to a distant organ, it does so with more affinity towards certain tissues than others. Such affinity, called organotropism, is not well understood. To this end, I applied another meta-analysis pipeline, MONTAGE (Models of organotropism and metastasis using gene expression). MONTAGE intends to cheracterize distant tissue affinity of metastatic cancer cells and patient heterogeneity.

Building ML models can be hindered by the data size, specially, for rare diseases. Therefore, by necessity, gene expression data are commonly integrated across multiple studies, and across multiple technical platforms. However, integrating data across studies/platform has vulnerability of having so called batch effects that often overshadow the actual biological signal we are interested in. To date, batch correction methods either remove confounding principal components along technical batches or explicitly model the batches as bias for each molecular feature (e.g., a gene) independently. However, as ML models include non-linear interactions among multiple genes, the multi-variate confounding artifacts misrepresent the models, even when they are built using data that is batch corrected data by existing methods. To circumvent this issue, borrowing techniques from natural language translation, I propose a deep learning based technique (deepSavior) which can project the expression data into a smaller non-linear space and then regenerate the gene expression profile from samples of one technical platform to another platform.

To summarize, this dissertation contains three kinds of contributions, a) designing effective ML models to test the biological hypotheses at hand, b) designing meta-analysis pipelines to compare multiple related but context-specific ML models to understand heterogeneous relations among biological processes, and b) developing new method to overcome the data integration challenges.

The rest of the dissertation is organized as follows. The following subsections introduces a) the basic biology of transcriptional regulation by genetics and epigenetics, b) basics of cancer metastasis, c) the ML models used in this dissertation, d) prior available methods on batch corrections, and e) basics of neural networks. Chapter 2 presents the models of methylation block boundaries and the downstream analysis. EMT and TRISECT are described in Chapter 3. MONTAGE pipeline, the findings of cancer heterogeneity are presented in Chapter 4. Chapter 4 also sets the premise for the necessity of new method development for batch correction and data imputation. Chapter 5 introduces the deep learning method (deepSavior) and the performance on both single cell expression data and bulk-Seq expression data.

In particular, the contributions of each chapter are shown below.

- Chapter 2. *H.C.B. and S.H. conceived and designed the project. M.S. performed all the analyses. All authors helped write the manuscript. All authors read and approved the final manuscript.*
- Chapter 3. *S.H. conceived the project. S.H. and M.S. designed the analyses in consultation with H.C.B. M.S. performed the analyses. S.H. and M.S. wrote the manuscript with help from H.C.B.*
- Chapter 4. *M.S. conceived the project. M.S., S.H., and H.C.B. designed the analyses. M.S. performed the analyses. Everybody participated in writing the manuscript.*
- Chapter 5. *M.S. conceived the project. H.C.B and M.S. designed the analyses in consultation with S.H. M.S. performed the analyses. M.S., H.C.B. and S.H. wrote the manuscript. J.H. and Y.V. helped M.S. with technical issues.*

1.2 Transcriptional regulation by genetics and epigenetics

Cells are basic structural and functional building blocks of all living organisms. Both prokaryotic (without nucleus) and eukaryotic (with nucleus) cell contain cytoplasm carrying proteins and organelles encapsulated by cellular membrane [1], [2]. Nearly all living cells carry DNA (and RNA) which is the genetic material containing hereditary information. DNA resides in the cytoplasm for prokaryotic cells and is protected and separated by the nuclear membrane for eukaryotic cells [3]. DNA holds all the instructions for life of an organism in the form of functional segments, called genes, which encode for protein molecules, as well as other non-protein-coding genes such as tRNAs, ribosomal-RNAs, micro-RNAs, pi-RNAs, etc [4], [5].

The information in DNA is stored as a code consisting of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T) [6], [7]. The order, or sequence, of these bases determines the information available for building and maintaining an organism. Together, a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. Human DNA consists of about 3 billion bases, and more than 99.9 percent of those bases are the same in all people. An important property of DNA is that it can replicate itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the mother cell.

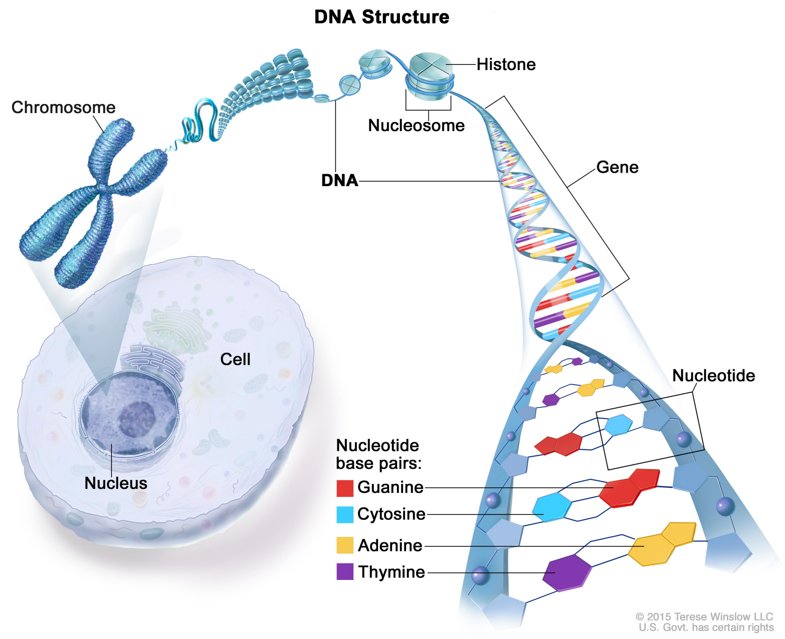


Figure 1.1 Illustration of cell and DNA structure

DNA is organized in one or more molecules or chromosomes. Chromosomal DNA is packaged inside the nucleus with the help of histone proteins: the DNA-protein complex is called chromatin. These positively-charged proteins strongly adhere to negatively-charged DNA to form complexes called nucleosomes. DNA is wrapped around the eight histone proteins of each nucleosome. Nucleosomes fold up to form chromatin fiber, which forms loops averaging 300 nanometers in length. The 300 nm fibers are again compressed and folded to produce a wider fiber, which is tightly coiled into the chromatid of a chromosome. When DNA is lightly packed, it is called euchromatin (unfolded or unwind DNA) or open chromatin or accessible state and otherwise it is called heterochromatin or closed chromatin or inaccessible state [8].

Proteins carry out all essential processes necessary to maintain life, including development, cellular, tissue, and organismal functions, and reproduction. The availability of proteins determines what bio-chemical reactions and thus functions are going to be carried out by the cell. According to the central dogma of molecular biology, the protein production is instructed by the gene in DNA: DNA produces RNA which goes out of cytoplasm to be turned into a protein [9]. The 1st phase of this process is called transcription and 2nd phase is called translation. In many organisms, the translated protein can be further modified by various enzymes. This process, referred to as post-translation modification, is not covered by the central dogma [10].

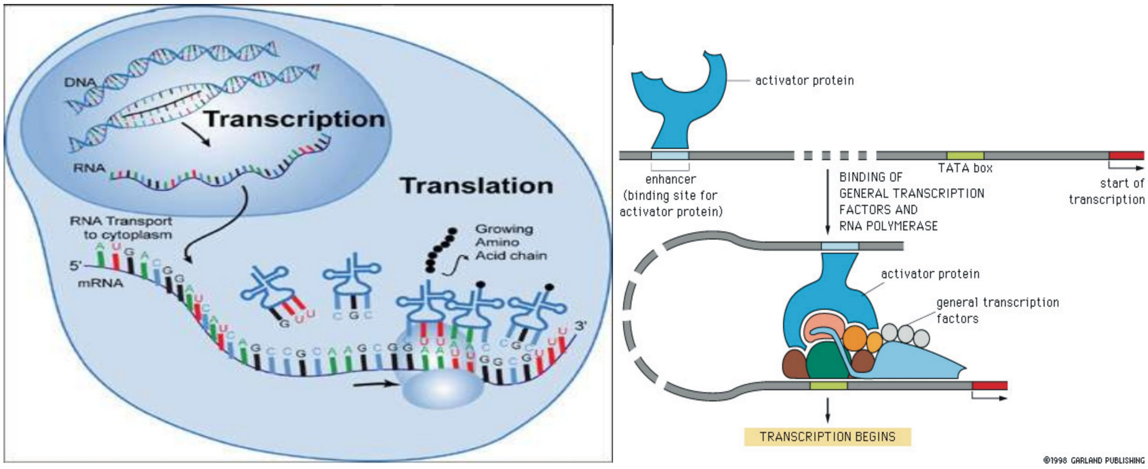


Figure 1.2 Illustrations of transcription and translation.

In eukaryotic cells, the transcription process first generates primary transcript mRNA (pre-mRNA), which is then spliced into the final product - the mature mRNA molecule. During translation, a protein complex called ribosome reads the mRNA according to the genetic code, where each mRNA triplet codon encodes for an amino acid. Thus, mRNA is used as a template to assemble a chain of amino acids that form the final protein product. In eukaryotic cells, transcription occurs in the nucleus while translation occurs in cytoplasm, therefore mRNA are transported out of the nucleus to the cytoplasm [11].

The transcription process is controlled by a class of proteins called Transcription Factors (TF). They bind to the DNA, in the promoter (upstream genomic region) as well as other distal regulatory regions of a gene, using DNA binding domains which recognize a 6-20 base-pair sequence signature or motif. A promoter (and regulatory region) contains a specific set of motifs, also called transcription factor binding sites (TFBS), which allow specific set of TFs to bind and modulate expression of the target gene and in turn the amount of protein produced. For transcription, a promoter needs to be unwound from histones (i.e. accessible) so that TFs can bind and a pre-initiation complex can be formed by RNA polymerase to read the DNA [12].

Many TFs are activators, while others are repressor of genes. Gene regulation can happen not only by binding to the promoter but also by binding to a distal genomic region which can reach up to 1Mbp away from the transcription start site (TSS). TFs involved in such distal regulation are referred to as enhancers. An enhancer physically interacts with the gene-promoter by forming a chromatin loop, whereas the regular TF-gene-promoter interactions are mostly linear. The presence or absence of TF determines which genes are going to be on and which genes are going to be off. In sum, the combinations of genes, thereby availability of certain proteins can determine the functionality carried out by the cell; in another words, TFs and enhancers are the crucial determinants of cell identity [13].

A set of chemical modifications to the DNA and to the histones can change the local accessibility of DNA for TF binding and therefore can modulate gene expression [14]. DNA methylation is the modification to DNA that silences gene expression by not letting any TF to bind. H3K4me3 and H3K27me3 are histone modifications where the former activates the gene by making the promoter accessible to TFs and RNA polymerase, and the latter represses the gene. Histone modifications and DNA methylation are also known to be inherited during cell division and therefore are collectively called epigenetics. Epigenetics, in summary, modulates how transcription machinery reads the genetic instruction from DNA in a cell. It is also widely known that undesirable epigenetic changes cause many human diseases [15].

Each cell type expresses a unique subset of genes. Conversely, the set of the genes expressed in a cell determines its identity. For example, the set of genes that is expressed in blood cells is different from those in immune cells or in neurons. That's the reason for all the cell types to look and act differently even though they contain same DNA sequence. Cancer cells also activate sets of genes that are different from any normal cell, thus acting differently from any normal cell. During cellular differentiation, a daughter cell acquires the capability to express different set of genes than the parent cell. Understanding cellular differentiation has significant impact both in the understanding of biology and clinical applications [16].

1.3 Cancer and Metastasis

Cancer is among the leading cause of death worldwide and in the US. In 2012, around 15 million new cases of cancer and 8.2 million deaths were reported [17], and in 2015 about 90.5 million people were reported to have cancer [18]. It is expected that the number of cancer cases will increase by 70% in the next two decades. Among all diseases National Institute of Health allocates the highest amount of its budget to the cancer research.

Cancer is a unique genetic disorder where the transcription machinery and other cellular processes are hijacked to allow cancer to proliferate and migrate. Existing cellular processes and regulatory networks are reprogrammed in systematic manner to adapt the need of such proliferation and migration. In order for a normal cell to transform into a malignant cancer cell, a series of genetic and transcriptomic alterations need to occur to the genes controlling cell growth and differentiation. The genetic alterations can be divided into two broad categories: alterations of oncogenes and alterations of tumor suppressor genes [19]. The former promotes cell growth and division, while the latter inhibit cell proliferation. Genetic changes can occur at different genomic levels and by different mechanisms: gain/loss of an entire chromosome, mutations, insertions, deletions. Epigenetic alterations also occur frequently in cancers. Epigenetic alterations refer to functionally relevant modifications to the genome that do not change the nucleotide sequence. Large blocks of hypo- and hyper-methylation, histone modifications and changes in chromosome architecture are common phenomena in cancer [20], [21]. All epigenetic alterations regulate gene expression without changing the underlying DNA sequence and may last through cell divisions for multiple generations.

Hanahan et. al. [22] suggested several essential alterations in cells required to transform into a tumor: Self-sufficiency of growth signal, antigrowth, apoptosis, limitless potential of replication, angiogenesis, invasion and metastasis.

Self-sufficiency of growth signal: Normal cells require specific growth signal (GS) from extracellular signaling molecules to proliferate. Tumor cells, in contrast, show a greatly reduced dependence on the external growth stimulation by mimicking growth signals or by permanently activating the pathways that respond to the GSs [23].

Antigrowth or Insensitivity to growth-inhibitory signals: Uncontrolled proliferation is blocked by many antigrowth signals through trans-membrane signaling receptors and intracellular signaling pathways in normal cells. The signals either force a cell out of the proliferation or permanently switch off the proliferation potential of a cell. Cancer escapes these antigrowth factor signals to keep proliferating uncontrollably. Such insensitivity to antigrowth signals can be achieved by disruption of tumor suppressor genes that primarily control those signals and pRB pathway responsible for blocking antigrowth signals [24].

Apoptosis or Avoidance of programmed cell death: Programmed cell death, known as apoptosis, is a major mechanism by which uncontrolled growth is controlled in the normal cells. The acquired resistance to the apoptosis is a hallmark of all cancer types. Cancer acquires the apoptosis resistance through a variety of strategies: mutation of p53 (tumor suppressor gene regulating apoptosis), overexpression of anti-apoptic signals e.g. AKT/PKB pathway, increased capability to detect DNA damage or abnormalities etc [25].

Limitless number of cell divisions: Three acquired capabilities - independence of the growth signals, insensitivity to antigrowth signals, and resistance to apoptosis - do not suffice in supporting uncontrolled tumor growth and tumorigenesis due to an intrinsic limit on a number of cell divisions allowed. Once cells have achieved a certain number of doubling they stop dividing, a concept termed as senescence. This program is independent of cell signaling. In order for cells to grow in malignant tumor, they must evade this program too. Telomeres located at the ends of chromosomes are the counting devise, which shorten with every cell division. The progressive shortening causes cells to eventually lose their capability to divide further. Telomere maintenance is evident in all types of tumors. In most tumors, their maintenance is mediated by telomerase up-regulation, the enzyme responsible for maintaining telomere length in stem cells [26].

Angiogenesis or Promotion of blood vessel construction: The formation of new blood vessels is referred to as angiogenesis. Nutrients and oxygen are supplied by blood to each cell and are necessary for maintenance and survival. The expanding tumor needs additional routes for blood supply. Cancer hijacks the angiogenesis to ensure adequate oxygenation. This is achieved by disruption of the production of factors that regulate blood vessel formation [27], [28].

Invasion of tissue and Formation of metastasis: Advanced stages of tumors eventually acquire capability to invade adjacent tissue and metastasize to distant sites [29]. Most of cancer types do not lead to patient's death unless they metastasize. In fact, 90% of cancer deaths are due to metastasis.

Metastasis is the spread of cancer from one body site to another, a stage of cancer arrived at by a complex series of steps from single or multiple cancer cells. Cancer cells acquire the ability to break the Extra Cellular Matrix (ECM), leave the original tumor site, migrate to other parts of the body [30]. The migration can occur by the following routes: a) hematogenous spread, b) lymphatic spread, c) transcoelomic and d) transplantation or implantation. For sarcoma and certain types of carcinoma, e.g. renal cell, the common route is hematogenous spread: distribution by blood stream. Because of their thinner walls, veins are more frequently invaded than are arteries, and metastasis tends to follow the pattern of venous flow. Except sarcoma, the most common route of metastasis is lymphatic spread which allows the transport of tumor cells to lymph nodes which drain off the metastatic cells into the systemic venous system and thus these cells can spread through the haematogenous route. Transcoelomic is the spreading via body cavities such as peritoneal, pleural, pericardial, or subarachnoid spaces. Transplantation is the spreading via regional lymph nodes near the primary tumor. Localized spread to regional lymph nodes near the primary tumor is not normally counted as metastasis, although this is a sign of worse prognosis [31].

The location of the metastases is not always random, with different types of cancer tending to spread to particular tissues at a rate that is higher than expected by statistical chance alone. Breast cancer, for example, tends to metastasize to the bones and lungs. The propensity for a metastatic cell to spread to a particular tissue is called 'organotropism'. According to “Seed and soil” theory of Stephen Paget, “it is difficult for cancer cells to survive outside their region of origin, so in order to metastasize they must find a location with similar characteristics. For example, breast tumor cells, which gather calcium ions from breast milk, metastasize to bone tissue, where they can gather calcium ions from bone. Malignant melanoma spreads to the brain, presumably because neural tissue and melanocytes arise from the same cell line in the embryo” [32].

The "seed and soil" theory was challenged by James Ewing proposing that metastasis occurs purely by anatomic and mechanical routes [33]. This hypothesis has been recently utilized to suggest several hypotheses about the life cycle of circulating tumor cells (CTCs) and to postulate that the patterns of spread could be better understood through a 'filter and flow' perspective [34]. However, contemporary evidence indicates that the primary tumor may dictate organotropism by inducing the formation of pre-metastatic niches at distant sites, where incoming metastatic cells may engraft and colonize. Specifically, exosome vesicles secreted by tumors have been shown to home to pre-metastatic sites, where they activate pro-metastatic processes such as angiogenesis and modify the immune contexture, so as to foster a favorable microenvironment for secondary tumor growth.

It is theorized that metastasis always coincides with a primary cancer, and, as such, is a tumor that started from a cancer cell or cells in another part of the body. However, over 10% of patients presenting to oncology units will have metastases without a primary tumor found. In these cases, doctors refer to the primary tumor as "unknown" or "occult," and the patient is said to have cancer of unknown primary origin (CUP) or unknown primary tumors (UPT). It is estimated that 3% of all cancers are of unknown primary origin [35].

1.4 Ensemble Models

A classification problem is to find a function (or a set of functions) that can discriminate a data points membership in one of multiple different classes. On occasion, functions that define membership to a specific class is referred to as a *hypothesis* in this context. Ensemble methods refer to a classifier that itself consists of multiple classifiers; such classifier combinations may outperform non-ensemble classifiers as each classifier in the ensemble may model a specific hypothesis required discrimination. So as a whole, the set of classifiers can capture the diversity of the class-membership pattern. We provide a short discussion of two common ensemble methods in the following.

1.4.1 Random Forest

Random Forest [36] is a combination of bagging [37] and a special case of the random subspace method [38]. Bagging is an ensemble meta-algorithm. In this composite model, each sub-model has equal weight, constructed from multiple independent samples, D_i. Each D_i is constructed from dataset D using uniform random selection and with replacement. Each D_i is then used to train a separate sub-model m_i. The average of these sub-models is considered as the outcome of the final model. Bagging reduces variance and helps to avoid over fitting, in many ensemble techniques bagging is done as pre-step of modeling. On the other hand, in random subspace method, each sub-model is constructed on D_i where the feature set of D_i is a sub-set of original feature set. The feature subset is selected without replacement. For classifying a new observation, the output of all sub-models is combined by majority voting or averaging the posterior probabilities.

Random Forest consists of a set of decision trees, each tree is a sub-model here. Each sub-model is trained on a bootstrap sample and the feature of sample dataset is a subset of original feature set. Typically, the size of the feature subspace is decided as the one-third of the original size, i.e. number of features selected in each bootstrap sample is one-third of the original feature number [39].

1.4.2 Adaboost

Boosting is an iterative method where weak learners are constructed based on the performance of the current classifier [40]. In the basic boosting method, the algorithm gives equal weight for each sub-model. While working with subsequent sub-models, the model puts more emphasize on misclassified examples. In the 1st stage, all the examples have equal weight. All the misclassified examples are given higher weight, and the next model is trained on the newly weighted dataset. The weights of the misclassified examples are updated again based on the combined model and the training step is repeated. The weight update and training is done lml number of times. There are many variation of boosting, the most popular one is Adaboost or adaptive boost [41]–[43].

In adaptive boost not only the weights of the examples are updated but also the weights of sub-models are tweaked as the training progresses. The problem is seen as a minimization of error function which is defined as the error of current model and new weak learner. The new weak learner is weighted in such a say that the total error decreases. Each weak learner produces an output, hypothesis h_j, for each sample in the training set. At each iteration t, a weak learner is selected and assigned a coefficient α_t such that the sum training error E_t of the resulting t-stage boost classifier is minimized. The ensemble of basic boosting is shown by $F_t(x) = \sum_{t=1\ldots T} f_t(x)$, and the ensemble of adaboost is expressed by $F_t(x) = F_{t-1}(x) + \alpha_t h_t(x)$. Here, h_t is new hypothesis and α_t is chosen in such a way that sum of training error is minimized, $E_t = \sum_i E[F_{t-1}(x_i) + \alpha_t h_t(x_i)]$. Because of the good performance, the variations of boosting are applied to adaboost framework [44].

1.5 Batch Correction Methods

Systematic differences and non-biological variations due to experimental and technological conditions in sequencing experiments are called batch effects [45]. Various hybridization (e.g. microarray) and sequencing technologies (e.g. RNASeq), are used to determine gene expression profiles of samples coming from different states (disease, cell cycle, normal). The expression profiles are useful measurement to understand the gene-phenotype relationships. Due to practical reasons, the number of samples processed for sequencing is limited. For example, for rare disease the samples can come from multiple labs and hospitals, the samples can be sequenced using different technologies, array types or platform, even the replicate samples can be generated several days/months apart, experiments can be done by different people, they can be performed under different environmental conditions. All these contribute to the differences in gene expression patterns that are unrelated to the underlying biology of interest. Several techniques have been developed to remove such differences as described below. Among them, the 1st three methods are applicable when the batches are known.

1.5.1 Singular Value Decomposition

SVD is a linear transformation of the expression data from the genes × arrays space to the “eigen genes” × “eigen arrays” space [46]. The new space is of lower dimensional than the original space and in the new space, the data are diagonalized with each eigen gene expressed only in one eigenarray and with the corresponding "eigen expression" level indicating their relative significance. The eigen genes and eigen arrays are unique, and therefore also data-driven, orthonormal super positions of the genes and arrays, respectively. After determining the eigen genes and eigen arrays, those inferred to represent noise or experimental artifacts are filtered out and the rest is normalized. The caveat of SVD method is that it is hard to detect the right eigen gene to remove and the removed eigen gene and eigen array might be combination of both noise and the phenotype of interest and hence not worthy of removing all of the variation across that direction.

1.5.2 Distance Weighted Discrimination

DWD does not remove all information along maximum variance, rather adjusts the mean along the mean discriminating hyperplanes [47]. In particular, DWD finds the separating hyperplanes (DWD direction vector) between two sets of samples. The sub-populations (e.g. respective source subsets) are all projected in that DWD direction, and the sub-population projected means are computed. Each subpopulation is then shifted in the DWD direction, by an appropriate amount, through the subtraction of the DWD direction vector multiplied by each projected mean for each gene. The DWD method can only be applied to two batches at a time. A way around for more than two batch scenarios can be achieved using a stepwise approach. In this approach, the two most similar batches are adjusted first, and then the third against the previous (adjusted) two are compared. Such stepwise method works reasonably well in their three-batch case, but when many more batches are present or when batches are not very similar, the iterative approach could potentially break down.

1.5.3 ComBat

ComBat [48] has two main advantages over previous methods, a) it is robust for small number of samples, e.g. less than 10 whereas SVD/PCA, DWD requires at least 25 samples, b) it removes both linear and non-linear noise. In ComBat, batch effects are modeled out by standardizing means and variances (L/S model parameters) across batches. These adjustments can range from simple gene-wise mean and variance standardization to complex linear or non-linear adjustments across the genes. Specifically, the L/S model parameters that represent the batch effects are estimated by "pooling information" across genes in each batch to "shrink" the batch effect parameter estimates toward the overall mean of the batch effect estimates (across genes). These EB estimates are then used to adjust the data for batch effects, providing more robust adjustments for the batch effect on each gene.

1.5.4 Surrogate Variable Analysis and Limma

Unlike previous methods, SVA identifies and estimates the variation of unknown batches (unmeasured or unmodeled factors of both biological and technical sources) to overcome the problems caused by heterogeneity in expression studies [49]. For example, due to the complexity of our genomes, environment, and demographic features, there are many sources of variation when analyzing gene expression levels. Therefore, to understand the relationship between two variables, such as a drug and its effect on a disease, we might not want the effect of the variation of age and sex on the disease. In SVA, a residual matrix, R is constructed by removing the signal of the primary variable(s) of interest. Signatures of additional heterogeneity is identified by singular value decomposition (SVD) and based on permutation test, those singular vectors are retained that represent significant variation than expected by chance. For each singular vector, the subset of genes are identified who are associated with the variation of the singular vector. Next, for each subset of genes, a surrogate variable is built based on the full expression heterogeneity signature of that subset in the original expression data. After the surrogate variable are detected, they can be treated as other known batches to remove biases using any previous method, e.g. ComBat [48]. However, usually the surrogate variables are used as covariates in a differential expression (DE) analysis so that differentially expressed genes are accounted for the batches.

Limma [50] is used to find differentially expressed genes between case and control by fitting a linear model for each gene considering heteroscedasticity of different genes. Limma offers interface of providing the information of surrogate variables so that the measured differential expression signals are due to strictly case and control not due to on any unmodeled variations.

1.6 Artificial Neural Networks

An artificial neural network (ANN) is a structure of information processing using interconnected processing elements or nodes. This structure is analogous to the vast network of neurons in a brain. Figure 1.3 depicts a general architecture of a neural network [51].

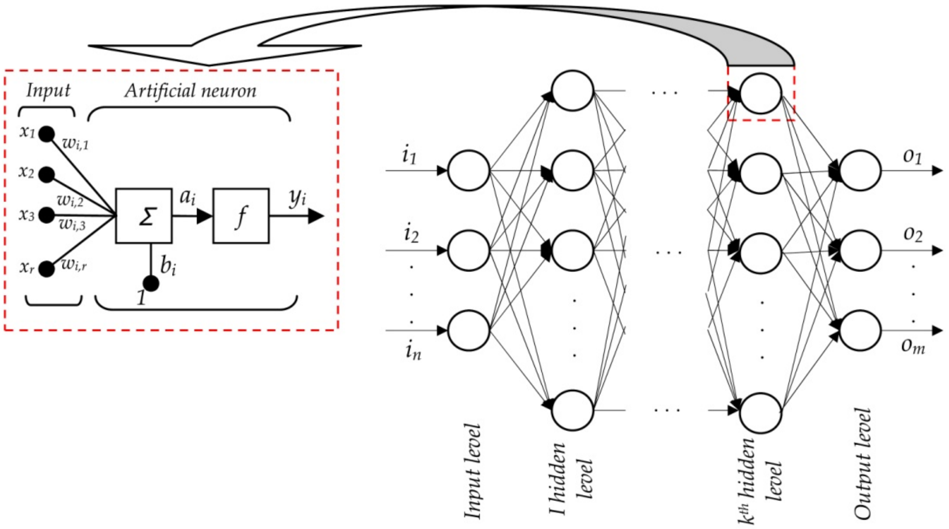


Figure 1.3 Example architecture of Artificial Neural Network

Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one neuron to the input of another. Typically, neurons are connected in layers, and signals travel from the first (input), to the last (output) layer. An ANN is typically defined by three types of parameters: a) the interconnection pattern between the different layers of neurons, b) the weights of the interconnections, which are updated in the learning process and c) the activation function that converts a neuron's weighted input to its output activation. A function defined by a neuron is a composite of all incoming neurons which are also composite function of other incoming neurons. A widely-used type of composition is the nonlinear weighted sum, where $yi = f(x) = K(\sum W[i,j].g[j](x))$, where K (commonly referred to as the activation function) is some predefined function, such as the hyperbolic tangent or sigmoid function. The important characteristic of the activation function is that it provides a smooth transition as input values change, i.e. a small change in input produces a small change in output. It will be convenient for the following to refer to a collection of functions $g[j]$ as simply a vector $g = (g[1], g[2], \ldots g[n])$. Together, an ANN can approximate very complex function. Among many varieties of network architecture, for the sake of relevance to this dissertation, residual network and auto-encoder are discussed below.

1.6.1 Residual Network

Residual neural networks is a recently introduced class of very deep neural nets [52], [53] typically formed by concatenation of many blocks (Figure 1.4), where each block receives an input x (the output of the previous block) and computes output y = x + δ(x), where δ(x) is the residual between original input and distorted input. The advantages of Residual neural networks over other architectures are they can avoid exploding or vanishing gradients during back propagation and thus can grow deeper without determining performance. Since a Residual neural network block consists of a residual term and an identity term, it can easily learn functions close to the identity function, when the weights are initialized close to zero, which is shown to be a valuable property for deep neural nets.

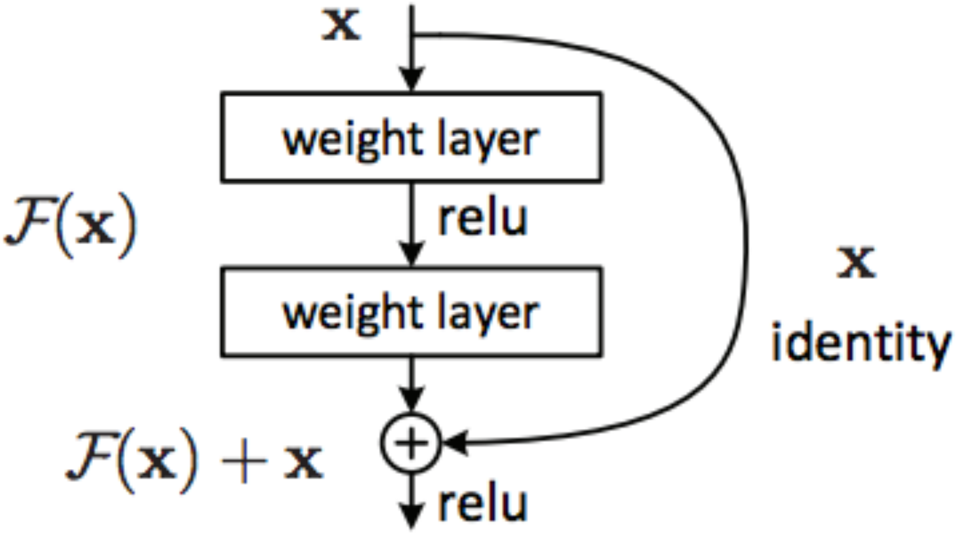

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 1.4 Building block of Residual Network

1.6.2 Auto-encoder

An auto-encoder is an artificial neural network used for unsupervised learning of efficient coding of the input [54]. The aim of an auto-encoder is to learn an encoding for a set of data in, mostly, lower dimensionality reduction. When the encoding is done in higher dimensional space, the corresponding network is called sparse auto-encoder. Architecturally, the simplest form of an auto-encoder is a feedforward, non-recurrent neural network, like the multilayer perceptron (MLP), having an input layer, an output layer and one or more hidden layers connecting them, but with the output layer having the same number of nodes as the input layer, and with the purpose of reconstructing its own inputs. An auto-encoder (Figure 1.5) always consists of two parts, the encoder and the decoder, which can be defined by the following equations.
$z = \sigma(WX + b)$ and $X' = \sigma'(W'z + b')$.

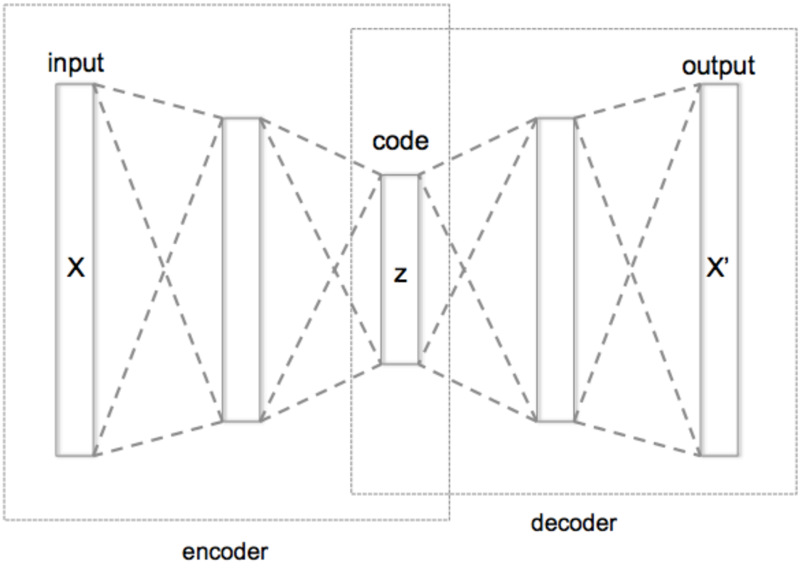


Figure 1.5 Schematic architecture of auto-encoder.

The auto-encoder is trained with a squared loss function between X and X' or KL-divergence of X' and X'. Denoising auto-encoders take a partially corrupted input and is trained to recover the original undistorted input. To train an auto-encoder for denoising data, it is necessary to perform preliminary stochastic mapping from X to X' in order to corrupt the data and use X' as input for a normal auto-encoder and use the $loss(\mathrm{X}, \mathrm{X}')$.

2 Demarcation of hypo-methylated blocks by distinct features in colon cancer

2.1 Background and Related works

Cells in an individual adopt hundreds of distinct phenotypes in their structure and function. This dramatic phenotypic variability through development and disease cannot be explained by genetic differences alone. Phenotypic variability is also partly encoded by the so-called epigenetic variation – varying degrees of chemical modifications of the DNA and nucleosome histones that the genomic DNA is wrapped around [55], [56]. Epigenetic mechanisms are integral to gene regulation; and, their role in cellular differentiation [21], aging [57] and disease [20] are areas under active investigation. DNA methylation is one of the earliest known epigenetic modifications, for which cellular inheritance mechanisms are now well understood [58]. Although a direct relationship between locus-specific DNA methylation and gene expression is well known, a more specific involvement of DNA methylation in various diseases, particularly in cancer, is only beginning to be investigated in a comprehensive manner [20], [59], [60]. Collectively, these studies have identified specific oncogenes that are hypomethylated, and thus activated, in cancer [61]; certain tumor suppressor genes that are hypermethylated, and thus inactivated [62], and additional methylation changes in cancer [59], [60].

A recent study showed well-demarcated, large regions, collectively covering half of the genome, to be differentially methylated in cancer [20]. Moreover, presence of such large cancer-specific differentially methylated regions (cDMRs) was found to be a general epigenomic signature across many cancer types [20]. The cDMRs contain important genes involved in mitotic cell cycle and matrix remodeling and were shown to exhibit extreme gene expression variability. Moreover, cDMRs are highly enriched among regions that are differentially methylated during stem cell reprogramming of induced pluripotent stem cells

[63]. Subsequent investigations revealed that cDMRs significantly overlapped with Lamina Attachment Domains (LAD), Large organized chromatin lysine modifications (LOCK) [64] and Partially Methylated Domains (PMD) in cancer [21]. Additionally, 1kb regions flanking cDMR boundaries were shown to be enriched for DNase hypersensitive sites [65]. Nucleosomes were found to be locally enriched in hypomethylated regions in normal tissue [66]. Collectively, these observations led the authors to postulate a model of cancer progression involving epigenetic instability of well-defined genomic domains [20]. However, investigations of additional genomic and epigenomic correlations of cDMRs, and ultimately the causes of cDMR formation are necessary to gain a better mechanistic understanding of the role of DNA methylation in cancer, and also to harness the full potential of these earlier studies for epigenetic-based cancer diagnostics [67].

Vast majority of large cDMRs are in fact hypomethylated in cancer, i.e. less methylated in cancer tissue than the corresponding normal tissue, and such hypomethylation happens in large contiguous genomic regions called hypomethylated blocks. Here, we focused on previously identified ~13k hypomethylated blocks (HMB) in colon cancer, which encompass approximately half the genome [20]. Given the length of HMBs and their general overlap with chromatin structural features such as LADs and enrichment of DNAse hypersensitive sites at HMB boundaries, it is likely that the genome and the epigenome at HMB boundaries hold the clues to the underlying mechanisms of genome wide hypomethylation with distinct boundaries. We therefore analyzed a number of genomic and epigenomic features at the HMB boundaries including TF binding motifs, epigenomic marks, and three-dimensional chromatin structural features (Figure 2.1).

Our analysis revealed that the classical promoter epigenomic mark – H3K4me3, is highly enriched at HMB boundary in normal colon tissue, and the boundaries that are enriched for promoter marks are also enriched for *in vivo* binding of the insulator protein CTCF in colon cancer. We also found that the HMB boundaries harbor distinct combinations of TF motifs. Our *Random Forest* machine learning model that uses TF motifs as features can distinguish boundaries not only from regions inside and outside HMBs, but surprisingly, from active promoters as well, with very high accuracy (F-measure ~ 0.98). Interestingly, the TFs that preferentially bind at HMB boundaries and their interacting partners are involved in chromatin modification. Finally, we found that HMB boundaries are associated with the boundaries of Topological Associating Domains (TADs), which form the backbone of chromatin structure [68].

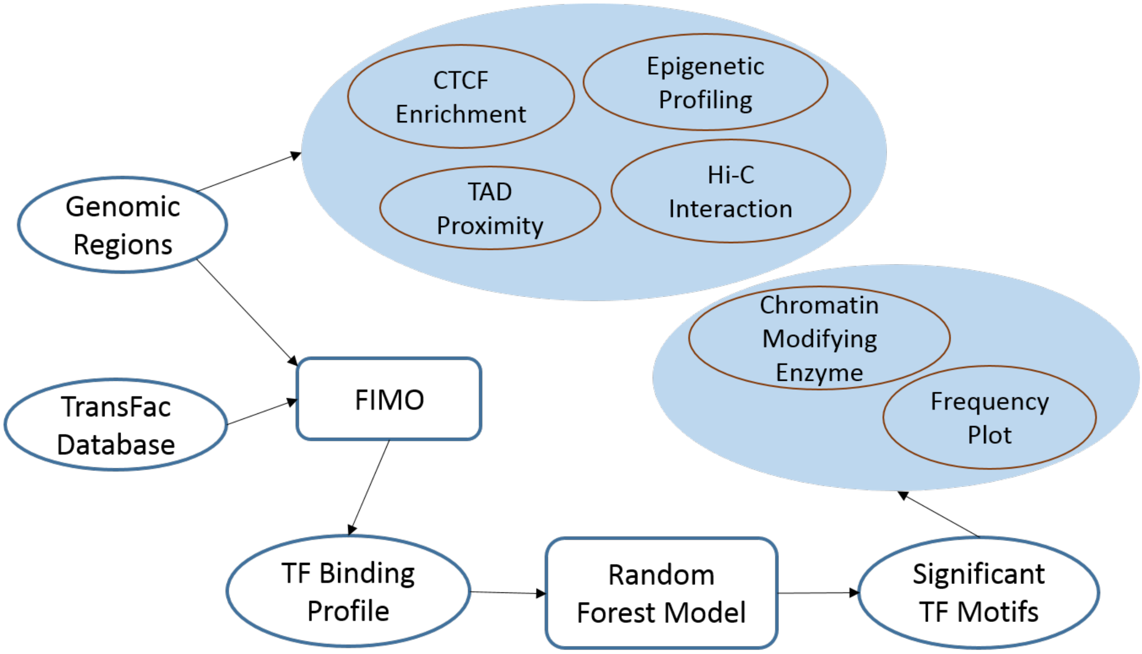


Figure 2.1 Schematic of analysis pipeline for hypo-methylated block boundaries. *Starting with ~13,000 HMBs, we perform a number of tests to assess the association of HMBs and HMB boundaries with Topological Associating Domains, Physical interaction within and across HMBs, profiles of various epigenetic marks, and CTCF binding. In addition, we identified TF motifs enriched at the HMB boundaries relative to various controls and assessed the ability of a random forest model to distinguish HMB boundaries from other domains based on TF binding site motifs. Finally, we assessed the spatial profile and functions of enriched TF motifs and their interacting partners.*

Taken together, our analyses suggest that the overall architecture of HMBs is guided and restricted by pre-existing chromatin architecture, while their creation in cancer may be caused by aberrant activity of promoter-like sequences at the boundary, with a direct chromatin modification activity.

2.2 Results

2.2.1 Overview

Our objective is to characterize genetic and epigenetic features that demarcate hypomethylated blocks in cancer, in order to gain insights into the mechanism and functional implications of these genomic blocks. Our findings are organized as follows: First, we determined and examined epigenomic marks that are enriched at HMB boundaries. Second, we analyzed genomic properties, namely, putative binding sites for all vertebrate transcription factors at HMB boundaries. Third, we showed that many of the motifs enriched at HMB boundaries exhibit specific positional distributions aligned with the HMB boundary. Fourth, we investigated specific transcription factor motifs enriched at HMB boundaries and their links to chromatin modifying enzymes (CMEs), in order to understand the mechanistic link between transcription factor binding and chromatin structure. Fifth, we furthered examined the link between genetic/epigenetic properties of the HMB boundaries and CMEs by analyzing the association between HMB boundaries and topologically associating domains (TAD) boundaries, which define the structural backbone of the chromatin. Finally, we examined at HMB boundaries, the putative sites for CTCF, which acts both as mediator of chromatin loop formation as well as an insulator that restricts the spread of chromatin marks.

2.2.2 Boundaries of hypomethylated blocks are enriched for promoter-associated histone mark H3K4me3.

Previous studies have shown cross-talk between DNA methylation and various histone modifications [69]. Given that HMBs exhibit relatively sharp demarcation of their boundaries [20], we investigated the patterns of various histone marks in normal colon tissue in the vicinity of HMB boundaries. We summarized the signal strength of six histone marks in 20 kbp flanking the HMB boundaries (see Methods) from human colon tissue data downloaded from the Epigenetic Roadmap Website (www.roadmapepigenomics.org). Histone marks H3K4me3 and H3K9ac, known to be associated with active promoters, showed a distinct peak immediately outside the HMBs (Figure 2.2). Patterns for other histone marks (H3K4me1, H3K9me3, H3K27me3 and H3K36me3) did not show noticeable trends at HMB boundaries (Figure 2.3).

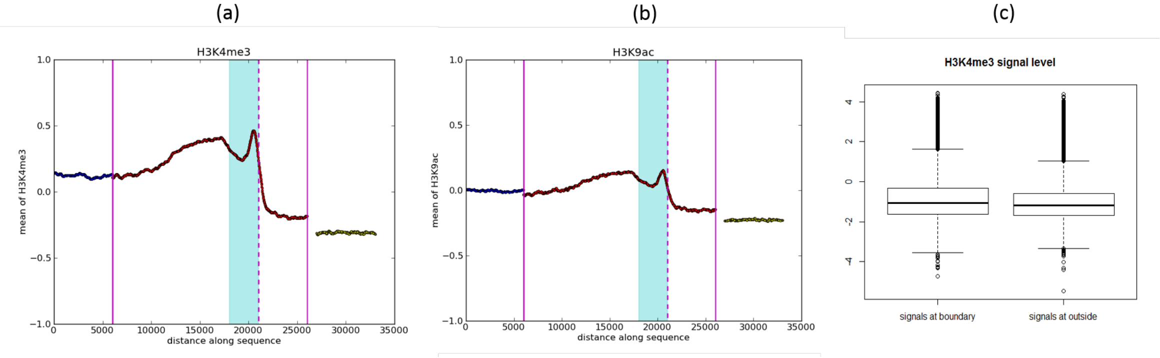


Figure 2.2 Histone modifications enriched near HMB boundaries.

Mean normalized ChIP signal for (a) H3K4me3 and (b) H3K9ac as a function of genomic distance to HMB boundary. The dotted vertical line (pink) depicts the precise location where the HMB starts while the shaded (cyan) region is the 3 kb HMB boundary region as defined in this paper. The solid vertical lines (pink) indicate inside (right) and outside (left) of HMBs. (c) Distribution of normalized H3K4me3 signal in HMB boundary regions and outside HMBs.

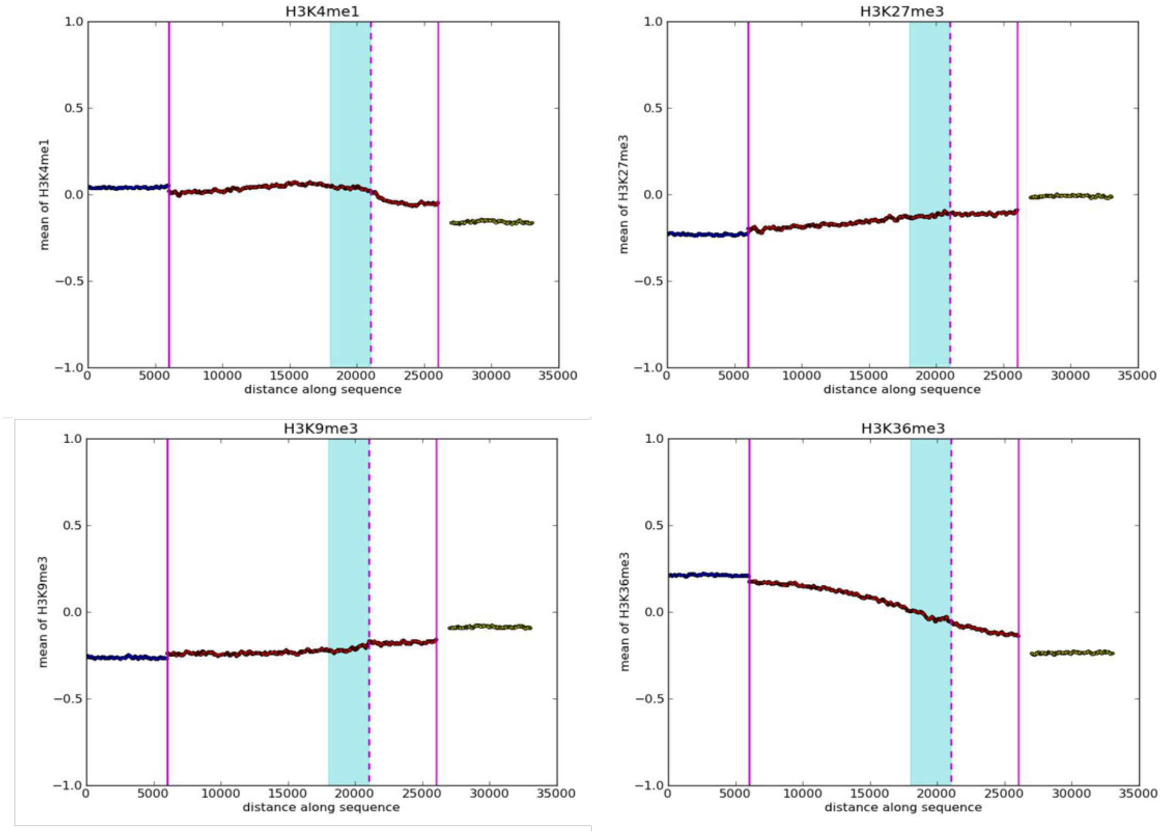


Figure 2.3 Patterns of various histone marks near HMB boundaries.

Given the enrichment for promoter histone marks at the HMB boundaries, we considered the possibility that the HMB boundaries coincide with or are near gene promoters. We excluded the HMB boundaries (5kb outside the HMB and 1 kb inside the HMB) that overlapped with the transcription start site of any gene or pseudogene (including non-coding genes), based on Gencode annotation [70], and repeated the analysis of histone mark pattern. The remaining boundaries still showed a significant, but smaller than previously mentioned peak, at the HMB boundary. For instance, as shown in Figure 2.2c, H3K4me3 signal strength in 3kb *outside HMB*s was lower than that in the regions immediately outside HMBs. The mean of normalized signals (see Methods) at the HMB boundaries was -0.82, while at random 3kb regions outside of HMBs the mean signal was -1.01 (Wilcoxon test p-value = 7.08e-42). This suggests that the observed enrichment of histone modification at HMB boundaries is not entirely due to annotated promoters for genes or pseudogenes.

2.2.3 HMB boundaries harbor distinguishing TF binding motifs.

Given the enrichment for promoter-like histone marks near HMB boundaries, we assessed whether HMB boundaries are distinct from non-boundary regions as well as other known promoters in terms of their TF binding motifs. For this purpose, in addition to the HMB boundary regions we defined three sets of regions of 6kb length (see Methods): (1) *Inside:* regions within HMBs, (2) *Outside:* regions between HMBs, and (3) *Promoters*. All regions were non-overlapping and in each pairwise comparison task, the GC content was similar in the two sets of regions (See Methods). For each 6kb region we constructed a 932-dimensional feature set quantifying the fraction of CpG Island overlaps and the number of motif matches for each of the 931 vertebrate TF motifs from TRANSFAC, v2011 [18], using FIMO [71] as the motif search tool. We then applied Random Forest (RF) classifiers on the feature set to distinguish HMB boundaries from the other genomic region sets under study. We trained the RF using 70% of the data and noted the classification accuracy on the remaining 30% of the data. The classification performances are shown in Table 2.1. Surprisingly, HMB boundaries can be distinguished from even other promoters with very high accuracy (F-measure ~ 0.978); Figure 2.4 shows the ROC curve corresponding to classification between HMB boundaries and promoters (ROC curves for the rest of the classification tasks are presented in Figure 2.5. We were able to recapitulate the RF results of HMB boundary versus promoter classification accuracy using Support Vector Machine (SVM) (F-measure ~0.97) – SVM is a classic tool for learning the combination of features of set of sequences that distinguishes the set from the control set. This suggests that the motif composition at HMB boundaries is distinct from those in promoter regions. We also obtained high discriminative performance when distinguishing HMB boundaries from regions inside HMBs (F-measure ~0.90).

	Sensitivity	Specificity	F-score	AUC	Size of Data Set
Boundary vs. Inside	0.90	0.89	0.90	0.96	41425
Boundary vs. Outside	0.84	0.81	0.83	0.91	41430
Boundary vs. Promoter	0.98	0.97	0.98	0.99	31051
Boundary vs. Promoter (SVM)	0.97	0.97	0.97	0.99	31051

Table 2.1 Performance of Random Forest classifier for HMB boundaries relative to other genomic regions.

'Inside' and 'outside' correspond to regions inside or outside HMBs respectively. Random sampling of these regions was stratified to match the length and CG content of HMB boundaries (see Methods). The last row corresponds to a Support Vector Machine classifier used to replicate the Random Forest result on the HMB boundary vs. Promoter region classification. In all cases, 70% of the data used as training and 30% used for testing. Sensitivity, Specificity and F-score were noted at the optimal F-score.

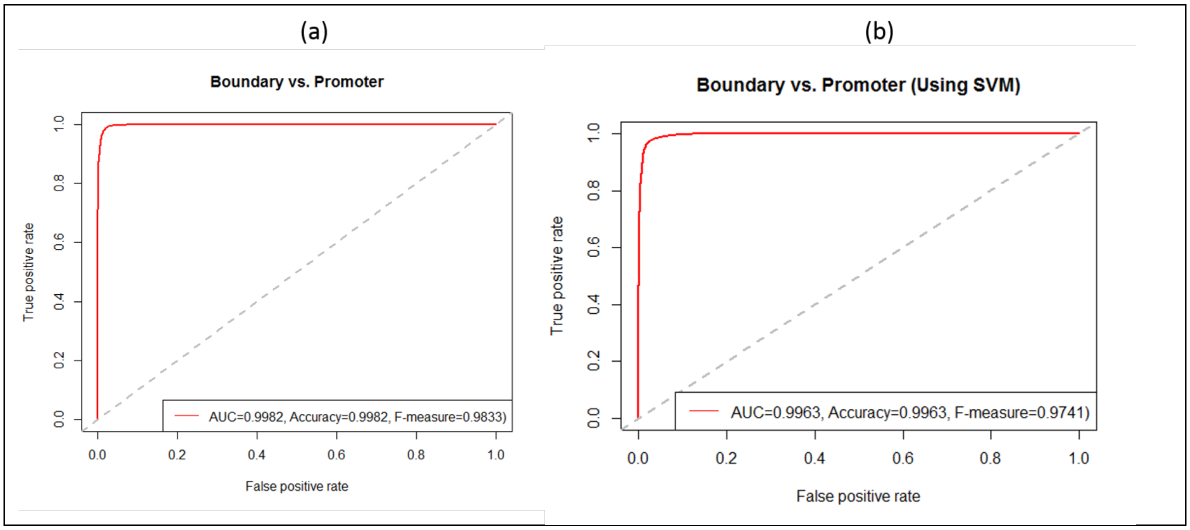


Figure 2.4 ROC curves for classifiers distinguishing HMB boundaries and promoters based on TF binding site motifs.

(a) Using Random Forest classifier, (b) Using Support Vector Machine classifier. Each ROC curve is based on predictions on a held-aside set of genomic regions (see Methods).

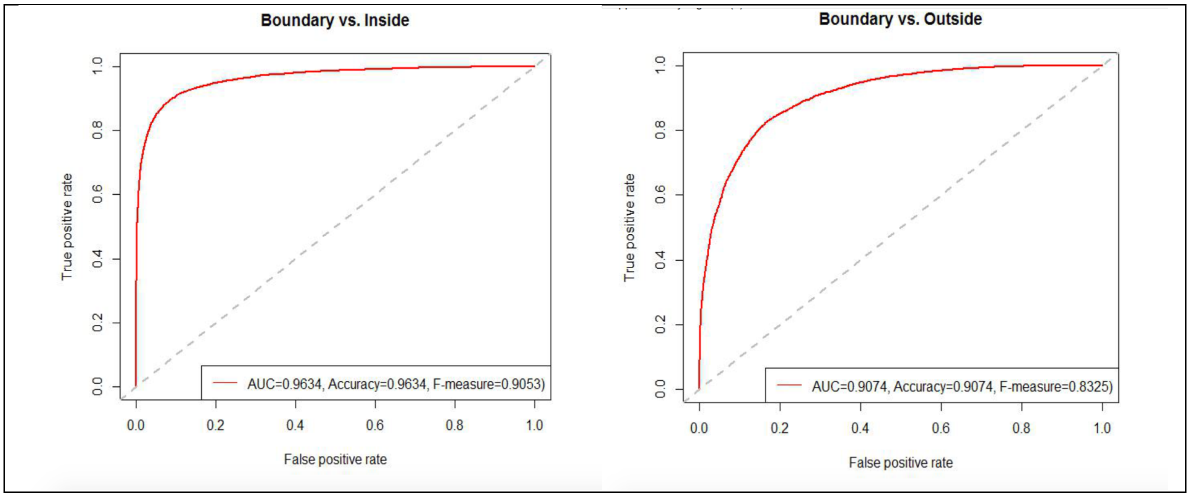


Figure 2.5 ROC curves for classifiers distinguishing HMB boundaries and inside/outside of HMB.

2.2.4 Positional distribution of discriminating motifs.

Next, we assessed whether TF motifs that distinguish HMBs exhibit a positional bias relative to the HMB boundaries. To prioritize the motifs we used the Mean Decrease Accuracy as the measure of a motif’s relevance to a specific discrimination task (see Methods). Table 7.4 (Supplementary Data) lists the top 20 most discriminating motifs in the classification of HMB boundaries against inside-HMB, outside-HMB, and promoters. Also, we only selected 46 motifs that were enriched above a threshold in the boundary (see Methods). For each of the 46 motifs, we plotted the frequency of the motif in 100 bps windows within the 6 kb HMB boundary regions, averaged over all HMB boundaries. Figure 2.6 shows the positional profile for the two most discriminating transcription factors ZFX (TRANSFAC id M01593) and SP1 (TRANSFAC id M00196) as an illustration; the profiles of all other motifs are included in Figure 7.1 in Supplementary Section. We next estimated for each motif the positional bias of binding sites within HMB boundaries by taking the most extreme (high or low) frequency of binding motifs among all 100 bp windows. The extreme frequencies of binding motifs were normalized and converted to Z-scores across all 100 bp windows in the 6kb regions. Z-score provides a standardized measurement of deviation from the mean frequency of binding motifs across the 46 motifs. We found that the majority of extreme frequency was located near the HMB boundaries: within 6k block the median location is 5574 from the outside of the boundary with a standard deviation of 892. Z-scores for all motifs ranged from 2.35 to 5.94 with a mean of 3.48 (See Figure 7.1 in Supplementary Section for all positional profiles, the corresponding Z-score for both boundary and promoter). This suggests that discriminating motifs have a skewed positional distribution that exhibits extreme enrichment very close to the HMB boundaries.

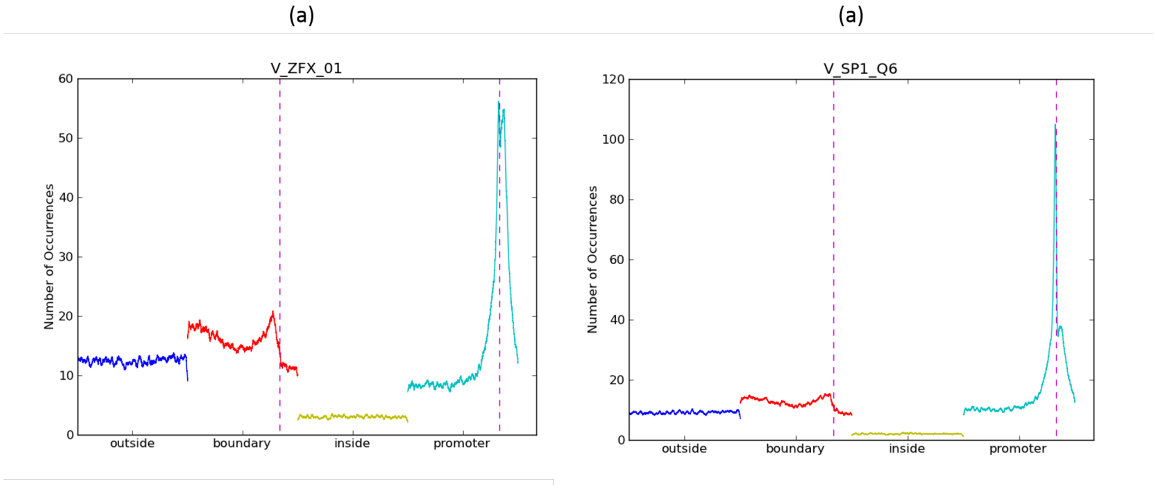


Figure 2.6 Positional profile of binding sites for ZFX and SP1.

Number of occurrences in 100 bp windows as function of genomic distance to HMB or promoter start site for TFs ZFX_01 (a) and SP1_Q6 (b). The dotted vertical line indicates the location of HMB and promoter respectively. 'Outside' and 'inside' correspond to 6 kb sized genomic regions outside or inside HMBs respectively.

2.2.5 Characterization of the most discriminating Transcription Factor motifs.

Some TFs are directly involved in histone modification and some other TFs are known to interact with chromatin modification enzymes [72]. We assessed whether the TFs whose motifs are most discriminative of HMB boundaries are involved in chromatin modification, either directly or by interacting with a chromatin modification enzyme. We first compiled a set of 492 genes annotated as Chromatin Modification Enzymes (CME) from the ENSMBL database. For each of the 931 TRANSFAC motifs, we obtained the Ensemble Gene ID for the corresponding TF protein and then obtained the set of annotated proteins known to interact with the particular TF using the string-db R package, which is based on the STRING database of protein interactions [73]. For each pair of regions compared (say, HMB boundary versus Promoter), we assessed whether the most discriminating motifs and their interacting partners are enriched for CMEs. To do so we obtained the top 20, 25, 40, and 50 motifs according to Mean Decrease Accuracy (see Methods), and compared the prevalence of CMEs among these motifs and their interacting partners against the rest of the available TF proteins as background. For each comparison, we assessed enrichment of CMEs using Fisher's Exact test. We found that the most discriminating TF motifs (Table 7.1-Table 7.3 from Supplementary Data) in HMB boundaries and their interacting partners were enriched for CMEs relative to all other regions (inside HMB, outside HMB, and promoter regions, Table 2.2). Encouragingly, the fold enrichment of CMEs increases monotonically as we restrict ourselves towards more significant TFs, from top 50 to top 20 motifs only. These results suggest that relative to inside and outside regions, the HMB boundaries not only harbor distinct motifs but these motifs could also be responsible for distinct epigenetic profiles at HMB boundaries.

Classification	**Top 20**		**Top 25**		**Top 40**		**Top 50**	
	OR	**P-value**	**OR**	**P-value**	**OR**	**P-value**	**OR**	**P-value**
Boundary-Inside	1.66	2.8e-9	1.57	6.1e-8	1.48	7.5e-7	1.46	7.8e-7
Boundary-Outside	1.61	3.0e-8	1.53	3.5e-7	1.50	2.4e-7	1.45	1.4e-6
Boundary-Promoter	1.64	7.6e-9	1.56	1.2e-7	1.44	3.9e-6	1.45	1.1e-6

Table 2.2 Enrichment of chromatin modification enzymes among the most discriminating TF motifs and their interacting partners.

Odds ratio (OR) and Fisher test P-value for a chromatin modification enzyme enrichment test using the most discriminating (20, 25, 40 or 50) TF binding site motifs for each classification task (as described in Table 2.1).

Table 7.5 (Supplementary Data) lists the 135 CMEs that interact with the top 20 enriched motifs in each of the three comparisons – boundary versus inside, outside, and promoter. Interestingly, these 135 CMEs include two DNA methyltransferases DNMT3A/B, and also P300, which is a well-known marker of regulatory enhancers.

2.2.6 Hypo-methylated blocks may be informed by chromatin structure.

Our analysis so far suggests that the HMB boundary regions possess distinguishing genomic and epigenomic characteristics, which may underlie their role as nucleation or termination of the methylation alteration. In addition, it is likely that the spread and confinement of epigenomic alteration within HMBs may be informed by preexisting chromatin organization and structure. This is suggested by a previous study that showed a significant overlap between cDMRs and LADs [20].

Based on Hi-C assay, which provides quantitative evidence of physical interactions between genomic loci, previous work has identified the so-called Topological Associating Domains (TAD), which are mega-base-sized genomic regions with a much greater interactions within the regions relative to across regions. TADs are relatively conserved across cell lines and species, and thus represent an underlying structural backbone of the chromatin. Based on 3,127 TADs reported in [68], we measured the proximity of each TAD boundaries to the closest HMB boundary, and compared the resulting positional distribution with that for a control set of randomly selected genomic loci. TAD boundaries are significantly closer (~43kb) in genomic distances to a HMB boundary compared with the expected ~71kb (ratio of mean = 3.8, ratio of median = 1.7, Wilcoxon test p-value = 5.4e-55, Figure 2.7a).

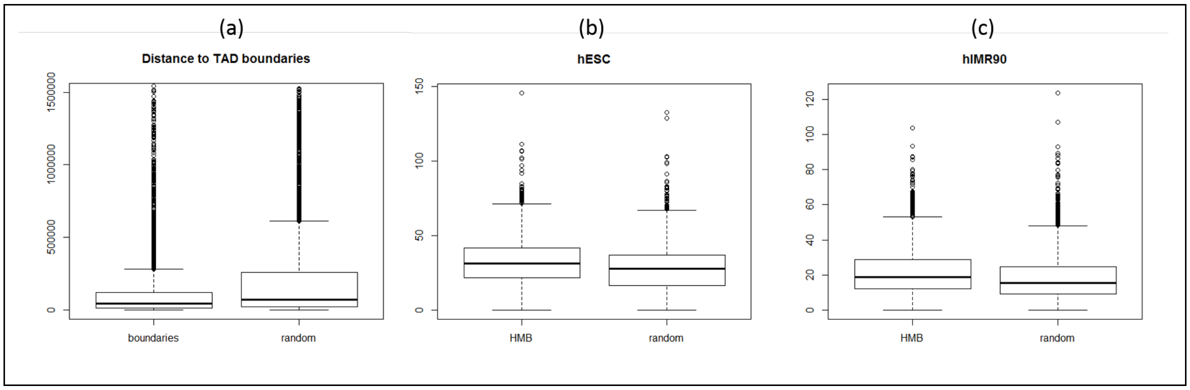


Figure 2.7 Hypo-methylated blocks associate with topological domains in the chromatin structure.

(a) Boxplot of genomic distance (in bases) between TAD boundary (obtained from hESC) to nearest colon cancer HMB boundary. Distances between TADs and random 6 kb genomic regions are included as background. (b) Boxplots of average Hi-C interaction for bins within HMBs in hESC, with average interactions randomly generated genomic regions of similar size and GC content included as background. (c) same as (b) and for IMR90 cell line.

Because TADs were identified based on a statistical overrepresentation of intra-region interaction, we also directly assessed using the Hi-C data, whether HMBs show an enriched intra-block interaction compared to inter-block interactions. Unfortunately, Hi-C data is not available for human colon tissue. Based on the Hi-C data in hESC, and hIMR90 cell line (yuelab.org/hi-c/download.html), as shown in Figure2.7b-c, we found a significantly greater interactions within HMBs compared to within random blocks controlled for length (For hESC: mean_HMB = 32, mean_Random = 27, Wilcoxon test p-value = 3.2e-38. For hIMR90: mean_HMB = 21.8, mean_Random = 18.3, Wilcoxon test p-value = 4.1e-34).

Overall, these analyses suggest that long domains of altered methylation in colon cancer may in part be informed by the underlying chromatin structure of the normal cell.

2.2.7 CTCF binding sites coincide with the H3K4me3 signal in HMB boundaries.

Among its numerous roles, CTCF is known to act as insulator by restricting the spread of heterochromatin, and is also involved in the maintenance of three dimensional chromatin conformation in part by stabilizing long-distance interactions [74]. Consistent with the role of insulator, CTCF binding sites are enriched between TADs [68]. We assessed whether CTCF binding sites are enriched near HMB boundaries. We downloaded the *in vivo* CTCF binding sites for colon cancer tissue from CTCFBSDB 2.0 database (insulatordb.uthsc.edu/). We found that HMBs were often bounded by CTCF binding sites. The frequency of CTCF in the 6 kb HMB boundaries (21%) was significantly higher than random blocks inside (14%) and outside (18%) HMBs, where the total number of regions in each set was ~20k. Moreover and interestingly, the HMB boundaries with a CTCF binding site had significantly higher levels of H3K4me3 signal than the boudaries without a CTCF binding site (ratio of mean = 1.4, Wilcoxon test p-value = 3.7e-24). Overall, this suggests that HMB boundaries are enriched for CTCF, as is expected for structural chromatin domains, but the presence of CTCF is in fact linked to the promoter-like characteristic of HMB boundaries.

2.3 Methods

2.3.1 Data processing: Hypomethylated blocks

We obtained coordinates for 13,540 reported long hypomethylated block (HMB) in colon cancer with an average and median size of 144 kbps and 39.5 kbps, respectively [20]. We define the boundary of an HMB as its 5kb flanking regions outside the HMB plus an additional 1kb inside the HMB. The choice of 5kb for the flanking region is arbitrary and 1kb inside is included to offset a lack of precision in localizing HMB boundary (e.g., Supplementary Figure 10b of [20]).

2.3.2 Random Forest based discrimination of HMB boundaries

We used Random Forest classifiers [75] to distinguish the resulting 27,080 6-kb-long HMB boundary from other genomic regions: (1) *inside HMB* - randomly selected 6kb block from inside of the HMBs, excluding HMB boundaries; (2) *outside HMB* - randomly selected 6kb regions from outside of the HMBs excluding HMB boundaries; (3) *promoter* - randomly selected 6kb promoters for protein-coding genes, including 5 kb upstream and 1 kb downstream of the transcription start site using the Ensembl annotation (www.ensembl.org, version 69). Given two sets of sequences (e.g., HMB and inside-HMB), and a set of characteristics (i.e. features) describing each sequence (e.g., putative binding sites for a set of transcription factors), the Random Forest classifier learns the combinations of features that distinguish one set of sequences from the other. When given an unforeseen sequence and its features, our Random Forest classifier can determine the set to which the sequence belongs to based on its features. The more distinguishing the features of the two sequence sets are (e.g., HMB and inside-HMB), the higher the accuracy with which our classifier can determine the set to which a new sequence belongs. To design the right control while building the Random Forest classifier, in each sequence set we selected the same numbers of regions for each pairwise classification task, while controlling for the GC content. For instance, when classifying between HMB boundaries and promoters, we selected two sets of regions that are non-overlapping and with similar GC content distribution. Finally, each set of sequences were composed of ~20k sequences.

As feature sets in the Random Forest classifiers, 931 motifs corresponding to vertebrate TFs were obtained from TRANSFAC v2011 [76]. Putative motif binding was determined in each 6kb region using the FIMO (Find individual Motif Occurrences) software [71]. Each 6 kb region was represented as a 931-dimensional feature vector where the measurement of each dimension is the count (0 or greater) of binding sites of each corresponding motif within the 6kb region. To build each classifier, we used the implementation from 'randomForest' package [77]; we used the default parameter setting except for the number of features (m) to be sampled randomly at each split of a decision tree. The default value of m is typically one-third of total number of features. However, we choose m=92 after tuning the random forests for optimal parameters. While tuning, the classifier was built with default m, and the out-of-bag error was estimated to update the value of m. In a random forests classifier, each tree was grown to the largest extent possible, i.e. without any pruning and to decide the classification of an unseen sequence the majority vote of the trees was considered.

We assessed the classification accuracy using a 70%-30% split of the data into training and test sets, chosen randomly, for each of the pairwise classification tasks distinguishing HMB boundaries from the three sets of regions: inside HMB, outside HMB, and promoters. The classification accuracies are reported using both area under curve (AUC) of the receiver operating curve and harmonic mean of precision and recall (F-measure). As an additional robustness measure, we also performed the HMB boundary versus promoter classification task using Support Vector Machine (SVM) implemented in R statistical package (www.**r**-project.org), based on 10-fold cross-validation.

2.3.3 CpG island overlap as an additional feature in the Random Forest Classifiers

CpG islands tend to exhibit increased methylation in colon cancer. Consequently, HMBs are frequently `broken' by CpG islands [20], and thus their boundaries frequently overlap CpG islands. Therefore, motifs can be found more frequently in HMB boundaries than inside or outside HMBs simply due to the presence of CpG islands. We used the fraction of the 6kb region that overlaps any of the 28,681 CpG islands annotated in the UCSC genome browser (genome.ucsc.edu) as an additional feature in the classification task, in addition to controlling for GC content in the classification task.

2.3.4 Identifying most discriminating motifs

We determined the importance of each motif in distinguishing between region types using *Mean Decrease Accuracy* obtained from the Random Forest classifier. Mean decrease accuracy of a feature measures the reduction in classification error upon including the corresponding feature in the model, and thus represents the importance of the motif in distinguishing HMB boundaries from a specific control region set; the higher the mean decrease accuracy the more important the feature is. We also determined enrichment of each motif in HMB boundaries relative to each control set (inside, outside, or promoters) using Fisher's exact test. The motif is considered as enriched (depleted) in the HMB boundaries relative to the control when the corresponding odds ratio is greater than 2 (less than 0.5).

2.3.5 Epigenetic data processing

Genome-wide profiles of six histone marks (H3K4me1, H3K4me3, and H3K9ac, H3K9me3, H3K27me3 and H3K36me3) in normal colon mucosa tissue were downloaded from the Roadmap Epigenetics Project website (www.roadmapepigenomics.org/). We calculated average signal for each histone mark (at 20bp resolution as provided by the Roadmap project) within each 6kb region in HMB, inside HMB, outside HMB, and promoter region. ChIP-Input was also obtained for normalization. To get the normalized values, we took the log ratio of methylation levels of histone marks and their corresponding ChIP-Input at the base-pair resolution.

2.3.6 Chromatin interaction measurement in hypomethylated blocks

To obtain the chromatin interaction information, we used Hi-C experimental data, which provides the spatial proximity information between pairs of different genome segments [78]. We obtained Hi-C data for human embryonic stem cell (hESC) and lung fibroblasts (hIMR90) cell lines from [68] as normalized interaction matrices with 40 kb bin size denoting the frequencies of physical contacts among pairs of genomic loci at a genome-wide scale. We mapped those 40 kb bins onto the HMBs and disregarded partially mapped blocks so HMBs smaller than 40kb were excluded from the analysis. We then measured interaction strength within each HMB as the sum of all pairwise bin interactions within the HMB divided by the number of 40 kb bins within the HMB. As a negative control, the same was done for randomly chosen non-overlapping genomic regions with same lengths as HMBs.

2.3.7 Measuring Proximity to Topologically Associating Domains

We downloaded the locations of 3,029 topological associating domains (TADs) from [68] for hESC cell lines. For each boundary of the TAD we obtain the minimum distance to a HMB boundary. As a control, we selected 13k random non-overlapping blocks of same sizes as HMBs. As for real HMBs, we also obtained the minimum distance of each TAD boundary to a random block selected for control.

2.3.8 Fisher's exact test: calculating enrichment/depletion of motif in different regions and finding motif interaction with chromatin modification enzymes (CME).

The contingency table for testing enrichment/depletion of each motif is shown below.

	Positive	Negative
Presence	a	c
Absence	b	d

*a (*respectively *b)* denotes the number of positive examples in which a motif is present (respectively absent). Similarly, *c* (respectively *d*) denotes the number of negative examples in which a motif is present (respectively absent).

The contingency table for testing interaction with CME is shown below.

	Selected Motifs	Other Motifs
Interact with CME	a	c
Do not interact with a CME	b	d

a (respectively *b*) denotes the number of selected motifs that themselves are CMEs or do not interact with a CME (respectively all others). Similarly, *c* (respectively *d*) corresponds to the control for testing CMC interaction using all the other motifs that themselves are not CMEs.

2.4 Discussion

In this study, we have characterized the regulatory landscape of large regions of methylation loss in colon cancer. We have found that the putative binding sites for specific TFs potentially involved in chromatin modification are distinguishing features of the DNA sequence at HMB boundaries. We also found that while activating histone marks common to promoters are enriched in HMB boundaries, HMB boundaries still show a distinct pattern of TF motif profile relative to known promoters. Finally, we found that the specific domains where HMBs occur are reflective of general chromatin organization of the normal cell.

Based on our qualitative assessment, we found that TFs enriched in HMB boundaries include those involved in demethylation, cell proliferation and cell cycle, hallmarks of cancer. For instance, for the most discriminative motif Sp1, high expression of Sp1 is known to disrupt cell cycle. Sp1 deregulation might be beneficial for tumor cells and its overexpression is known to induce apoptosis of untransformed cells [79]. Other members of Sp TF family also play roles in metastasis and growth of different tumor types [80]. In our analysis, multiple TFs from this family were found to be enriched in HMB boundaries. Zfx presents another illustrative example, as it controls the self-renewal of embryonic and adult hematopoietic stem cells [81]. Zfx also controls BCR-induced proliferation and survival of B lymphocytes [82]. Another detected TF, FoxO is central to the integration of growth factor signaling, oxidative stress and inflammation, and is involved in tumor suppression [83] and DNA demethylation process in B-cell development [84]. Finally, TF Zfp281 is known to play a role in cell pluripotency [85], chromatin remodeling [86], and inhibition of nanog auto-repression [87].

Loss of methylation in large domains has been identified as a consistent and stable mark in solid tumors [20], [88]. While the degree of methylation loss increases with tumor progression, intra-sample variability in DNA methylation and gene expression is greater within these domains [88]. These findings point to a general loss of epigenomic and transcriptomic stability that is essential to the normal behavior of the cell. The co-localization of these domains with lamin-associated domains [20], with TADs (as found in this study), and the enrichment of CTCF binding in the boundaries of these domains suggest that a loss of chromatin organization is concomitant with this loss of epigenomic and transcriptomic stability.

We note a few limitations of our analyses. Our analyses are based on 6 kb region flanking the HMB boundary. This choice, while reasoned, is somewhat arbitrary. Although our analyses suggest that HMB formation is associated with specific genomic, epigenomic, and chromatin features, it does not clarify the causality leading from TF binding to hypomethylation and ultimately to the previous observed aberrant gene expression in HMBs. While we observed specific patterns of certain epigenomic marks at HMB boundaries, these may be ultimately a reflection of the genomic characteristics [89]. Moreover, our analysis is based on putative binding site and not based on in vivo binding data for the TFs, which are currently not available for a majority of TF. Nevertheless, our analyses do suggest a potential link between specific genomic marks and HMB boundaries, which require future experimental studies of the underlying mechanisms.

Taken together, our analyses suggest that the overall architecture of HMBs is guided by pre-existing chromatin architecture, while their creation in cancers may be caused by aberrant activity of promoter-like sequences at the boundary. Our results are consistent with a model where a loss of chromatin organization and a concomitant loss of epigenetic stability make previously inaccessible TF binding sites accessible for proteins involved in chromatin modification as well as cellular fate, whose binding sites are enriched within domains of inaccessible chromatin where HMBs reside. The binding of specific DNA binding factors at HMB boundaries may further participate in methylation loss.

3 Heterogeneity of Transcription Factor binding specificity models

3.1 Background and Related works

Transcriptional regulation is mediated by the binding of transcription factors (TF) to specific DNA elements in the genome [90], [91]. While the *in vitro* binding specificity of many human TFs has been determined, it is well-recognized that the *in vitro* binding specificity of a TF is not sufficient to its explain condition-specific *in vivo* binding [92], [93]. This realization has spurred investigations of additional determinants of *in vivo* binding, such as heterogeneity of TF's binding motif [94], broader sequence context and inter-position dependence [95], homotypic clusters of binding sites [96], cooperative binding of the TF with its partners [97], [98], condition-specific chromatin context [98]–[101], and local DNA properties [96], [99]. While, overall, both local genomic and epigenomic features are deemed important in determining *in vivo* occupancy of a TF, recent reports suggest that *in vivo* binding of a TF can be accurately predicted based solely on the genomic signatures near the binding site without relying on the epigenomic context [96], [102]; this is consistent with additional recent reports, showing that the epigenome itself is encoded by the genomic context [89], [103].

Prior models of *in vivo* TF binding have shown that the genomic context of a binding site effectively encodes the condition-specific *in vivo* binding specificity [95], [102]. This can be explained by the substantial plasticity of a TF's interaction with other TFs' and the modular nature of a TF binding co-operativity [104]. The availability of specific combinations of interacting TFs can then guide *in vivo* binding to specific loci where the binding site of the interacting TFs are present in close proximity to each other, along with the availability of corresponding TFs [94].

Previous sequence-based modeling of *in vivo* TF binding was performed in a cell type-specific fashion [95], [102]. These cell type-specific models exhibit substantial *inter*-cell type heterogeneity, which is expected, given the variation in the availability of the potentially interacting TFs. In particular, Arvey et al. 2012 explicitly modeled potential interactions of the primary TFs with multiple additional co-factors, while general sequence properties were used as features in Mathelier & Wasserman 2013. These previous approaches, however, build a single model for a cell type, thus implicitly assuming a homogeneous cell type-specific TF binding model. As such, previous models have not investigated intra-cell type model heterogeneity. Intra-cell type TF binding heterogeneity is expected for the same reasons as inter-cell type heterogeneity. Moreover, in many instances, a binding specificity model trained in one cell type can predict a subset of *in vivo* binding in another cell type [102], suggesting that binding models, or parts thereof, are shared across cell types.

The motivation of the following chapter is to evaluate the heterogeneity of sequence-based cell type-specific *in vivo* TF binding models, and the extent to which binding rules (*sub-models*) are shared across cell types. We have developed an ensemble model-based approach (***TRISECT***) to reveal both cell-specific and cell-independent rules for the *in vivo* TF binding. Application of *TRISECT* to 23 TFs, each with genome-wide *in vivo* binding data in 4 – 12 cell types strongly suggests that the cell type-specific binding rule for a TF consists of multiple sub-models, a subset of which are shared across cell types, and points to shared functional underpinnings. This refinement to our understanding of the genomic context of *in vivo* binding specificity can facilitate future investigations of transcriptional regulation and its genetic determinants.

3.2 Results

3.2.1 TRISECT – Ensemble model of TF binding

An illustration of the *TRISECT* analysis pipeline is presented by Figure 3.1A and a brief description of the pipeline is provided below (for additional details see Methods).

Overview. As the first step, we developed an ensemble model (*EMT*) to discriminate a TF's *in vivo* bound genomic loci (foreground) from non-bound sites (background), balancing model complexity (number of sub-models in the ensemble) against the cross-validation classification accuracy. Given a set of genome-wide loci, bound by a specific TF, we first identified sets of foreground and background (control) sequences. The foreground set consisted of 100 bp sequences centered at the ChIP-seq peak. As a stringent background sequences, as done previously [102], we used 100 bp regions ~200 bp away from the peak location. We considered a variety of feature sets for discrimination (see below). The *EMT* model was trained using the Adaboost method where each sub-model is a decision tree (Figure 3.1B) built from a bootstrap sample [105]–[107]. Next, given a TF's *EMT* models for all cell types, each cell type-specific sub-model was represented by a point in a *d*-dimensional space, with *d* corresponding to the number of relevant features. We constructed clusters of the data points for a TF (representing the sub-models across all cell types), using *k*-Nearest Neighbors algorithm (*k*-NN). The sub-models within a cluster represent binding rules that are similar within or across the cell types.

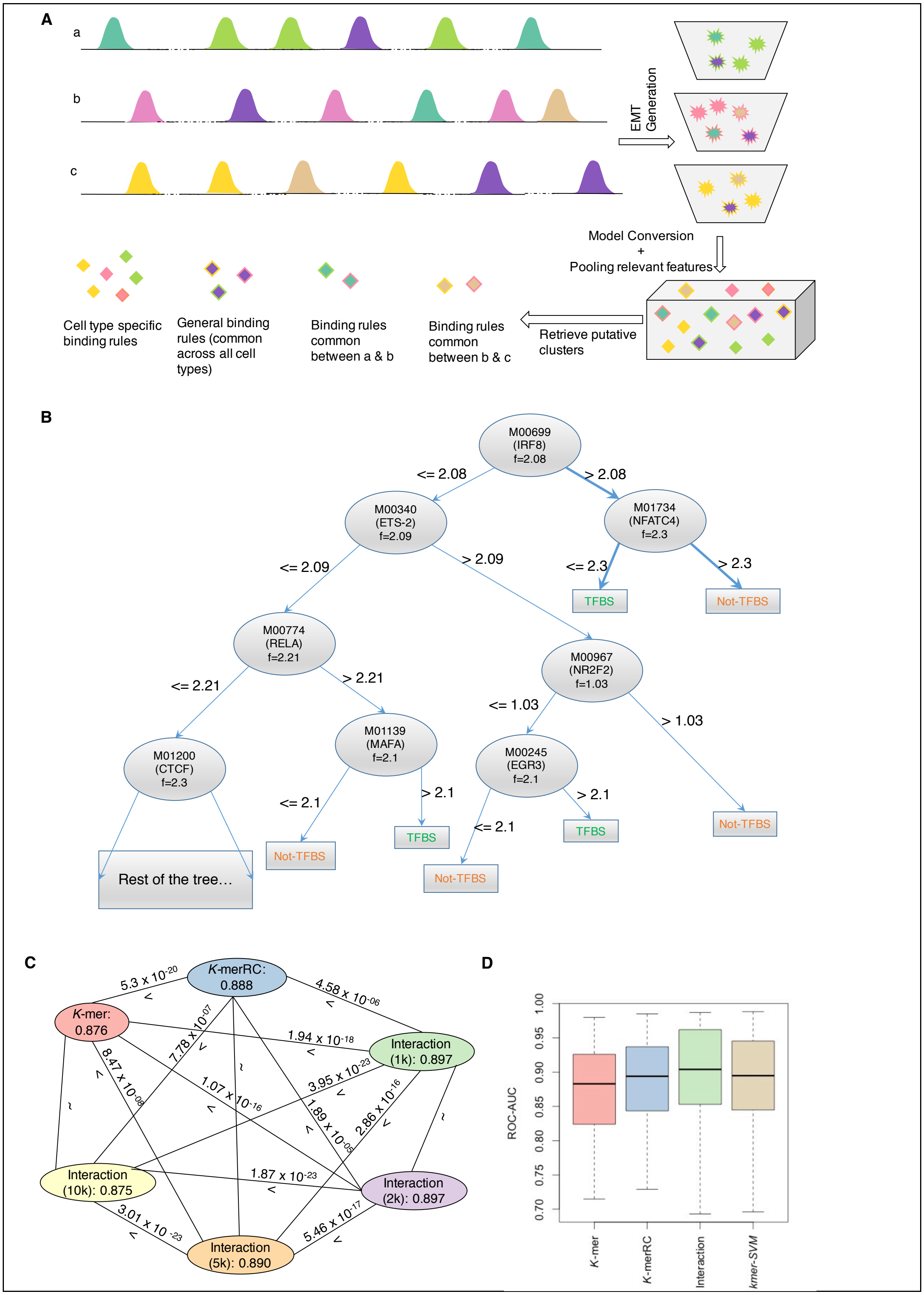


Figure 3.1 Overview and benchmarking.

(A) Schematic of TRISECT pipeline. Color indicate different binding rules or sub-models and rows (a, b, c) represent different cell types. Green, pink and yellow colors indicate cell type-specific sub-models. Each ensamble model (EMT) is represented by a bucket of sub-models (top right). Stars and diamonds with the same color denote corresponding sub-models and data points after transformation into reduced feature space, respectively. Each sub-model is represented by a decision tree. The sub-models across cell types are clustered. Cyan is common between cell types a and b, light-brown is common between cell types b and c, and purple is common across all three cell types. (B). An example sub-model taken from the Interaction model for CEBPB-Gm12878. Each node in the tree is labeled with the TRANSFAC id, corresponding gene name and the threshold at which the feature is split. Two binding rules are highlighted indicating TF binding and no TF binding. In (C) and (D) same color is used to denote the models using the same features. (C). Comparison of accuracy between all pairs of feature-sets. Nodes are labeled with feature type and mean accuracy. Edges are labeled with ">" (greater) or "<" (less) sign and two sided Wilcoxon p-value. (D) Accuracy (ROC-AUC) distribution of EMT for K-mer/K-merRC/Interaction (1k) and those of kmer-SVM models.

EMT Feature sets. We considered three feature sets for the 100 bp foreground and background sequences. The first feature set, *K*-mer, was comprised *of 6*-mer frequencies within each 100 bp sequence (total 4096 features). The second set, *K*-merRC, consisted of unified *6*-mers and their reverse complement frequencies (total 2080 features. The third feature set included the binding scores for 981 vertebrate TF motifs from TRANSFAC 2011 database. We defined the models built from the third feature set as the *Interaction* model, as the features represent potential TFs that might contribute to the binding of the reference TF (The TF for which *EMT* was built). For *Interaction* models, we used four thresholds for motif match in the PWMSCAN tool [108] where a threshold denotes the background match frequency – one hit in every 1kb, 2kb, 5kb, and 10kb.

EMT Training. We applied *TRISECT* to 23 TFs, each with ChIP-seq data in 4 to 12 cell types (a total of 135 TF-cell pair *EMT*s, Table 7.6 from Supplemental Data). A TF was included in this study if (i) the TF has narrow-peak data for at least 4 cell lines with at least 4000 bound sites in each cell line, and (ii) the TF has an established position weight matrix (PWM) in TRANSFAC 2011 database. See Figure 7.2 of Supplemental section for TF web-logos and Table 7.7 from Supplemental Data for other information about each TF including family names. *EMTs* were trained using 75% of the full dataset and performance assessment of *EMT*s was conducted using the remaining 25%. Model details such as the number of sub-models, model size etc. are provided in Table 7.8 from Supplemental Data.

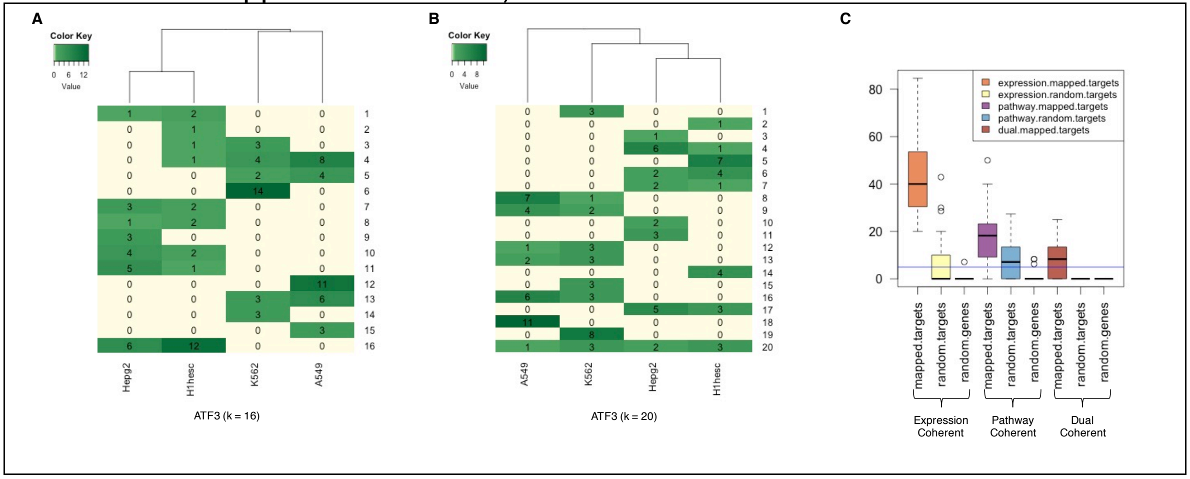
Each *EMT* includes multiple decision trees and each path from root to leaf in an estimated decision tree sub-model captures one binding rule that asserts how a combination of motifs and their binding affinities contribute to the target TF's binding. As an illustrative example, Figure 3.1B shows an arbitrarily selected sub-model of CEBPB in the Gm12878 cell line. Two of the binding rules are "presence of IRF8 with score greater than 2.08 and presence of NFATC4 with score of less than 2.3" - when these rules are met by the reference TF, CEBPB, is likely to bound. Whereas "presence of IRF8 with score greater than 2.08 and presence of NFATC4 with score of greater than 2.3” hinders CEBPB binding. Supplemental Note 1 and Figure 7.3 of Supplemental section include further interpretation of a sample sub-model (decision tree), a summary of how the reference TF’s motifs are distributed among the sub-models, and a discussion of model robustness for various parameter choices.

EMT performance. Model accuracy was quantified using Area Under the Receiver Operating Curve (ROC-AUC) on the 25% test set (Figure 3.1C, Figure 7.3C of Supplemental Data). We compared the model performances, using Wilcoxon test across 135 TF-cell type pairs for the 6 sets of *EMT*s (*K*-mer, *K*-merRC, and *Interaction* at 4 thresholds (namely, *Interaction* (1k), *Interaction* (2k), *Interaction* (5k), *Interaction* (10k)) (Figure 3.1C). We found that *K*-merRC significantly outperforms the *K*-mer model (two sided Wilcoxon p-value 5.3 x 10^{-20}). This is consistent with the fact that TF binding occurs on double-stranded DNA and as such does not have directionality (except in relation with other interacting TFs). Therefore, unifying each *k*-mer with its reverse complement is more representative of the biological determinants of TF binding. Following this line of reasoning, PWMs can provide an even better abstraction of DNA binding specificity and, as expected, the PWM-based models outperform the *k*-mer-based models, two sided p-value 4.58 x 10^{-6} when comparing *K*-merRC to *Interaction*-1k. Therefore, for sub-model clustering and other downstream analyses we selected *Interaction* (1k)-based *EMT* (heretofore referred to as *Interaction* model).

Comparison with previous model. Next, we compared *EMT* model (using *K*-merRC and *Interaction*) with previously published model based on Support Vector Machine (*kmer-SVM*) [102]. In *kmer-SVM*, the authors considered both *k*-mers and their reverse complements of size 8 with minimum matches of size 6. Applying the kmer-SVM pipeline to our dataseset, the resulting ROC-AUC for all the TF-cell pairs are listed in Table 7.9 from Supplemental Data. Figure 3.1D suggests that *Interaction* model performs favorably relative to *kmer-SVM*.

3.2.2 Intra-cell type heterogeneity and inter-cell type sharing of binding rules

Given the favorable performance of EMT, and its architectural differences to *kmer-SVM,* we next assessed whether *EMT* was better able to exploit the heterogeneous binding rules across the genome, as dictated by different combinations of co-occurring and co-regulated (i.e. potentially interacting) TFs. Conceptually, a ‘binding rule’ refers to the specific combination of motifs (along with their importance) aiding in the binding of a reference TF. While a general binding rule may be difficult to state concisely, it can be operationally defined in terms of a collective ensemble of cell type-specific binding rules. Each decision tree (a sub-model) operationally defines a binding rule, in terms of presence of specific motifs above/below a certain binding score. Furthermore, in general, the relative importance of features decrease with increasing depth of the node in the decision tree, with the first few levels contributing a substantial portion of the decision. Although a decision tree represents a statistical model for TF binding, by applying strict thresholds for motif scores and considering only the top few layers, in principal, a concise 'binding rule' can be derived, albeit, at a loss of information. For a specific TF and cell type combination, we captured the binding rules by a set of sub-models (decision trees). Then to investigate commonality and uniqueness of binding rules for a TF across cell types, we pooled all sub-models from cell-specific *EMT*s, represented them by feature importance and clustered them using *k*-NN clustering algorithm. Next, we constructed a cluster-membership matrix mapping the number of sub-models originating from different cell types within each cluster. As an example, Figure3.2A-B shows the cluster-membership matrix for the TF ATF3 for cluster sizes 16 and 20. The matrices show both cell type-specific (Figure3.2A, cluster #6) and ubiquitous (Figure3.2B, cluster #20) clusters. Examining the cluster mapping for all TFs (Figure 7.4 of Supplemental Data), a wide range of patterns emerge. For certain TFs, many clusters tend to map to single cell type, suggesting the cell type-specific binding modalities of these TFs (EP300, JUN), while other TFs have ubiquitously applicable binding rules, such as YY1 and TBP, suggesting the cell type independent binding rules and, presumably, function. Importantly, many clusters consist of sub-models from multiple, but not all, cell types. We ensured that inter-cell type sharing of binding rules is not simply due to the shared binding loci across cell types (Supplemental Note 2 and Figure 7.5 of Supplemental Data). Subsequent analyses are based on *k* = 16; the reason for this choice is discussed in Supplemental Note 3).



Hepg2	H1hesc	K562	A549	
1	2	0	0	1
0	1	0	0	2
0	1	3	0	3
0	1	4	8	4
0	0	2	4	5
0	0	14	0	6
3	2	0	0	7
1	2	0	0	8
3	0	0	0	9
4	2	0	0	10
5	1	0	0	11
0	0	0	11	12
0	0	3	6	13
0	0	3	0	14
0	0	0	3	15
6	12	0	0	16

A549	K562	Hepg2	H1hesc	
0	3	0	0	1
0	0	0	1	2
0	0	1	0	3
0	0	6	1	4
0	0	0	7	5
0	0	2	4	6
0	0	2	1	7
7	1	0	0	8
4	2	0	0	9
0	0	2	0	10
0	0	3	0	11
1	3	0	0	12
2	3	0	0	13
0	0	0	4	14
0	3	0	0	15
6	3	0	0	16
0	0	5	3	17
11	0	0	0	18
0	8	0	0	19
1	3	2	3	20

Figure 3.2 Assessment of TRISECT.

(A&B) Cluster membership matrix using k-Nearest Neighbors algorithm (k-NN) where k=16 in (A) and k=20 in (B). Row represents clusters and columns represent cell types. Each element in the matrix denotes the number of sub-models in the cluster from each cell type. Some clusters consist of sub-models from multiple cells (cluster#20 in B), while some other consist of sub-models from a single cell type (cluster#6 in A). (C) Functional and Expression coherence of sub-model clusters: fraction of multi-cell clusters found to be coherent using k-Nearest Neighbors algorithm (k-NN). Y-axis is the coherence percentage. Among the conditions (X-axis), mapped.targets denotes when genes are assigned to cluster based on

TRISECT pipeline, random.targets indicates the clusters consisting of random genes among all targets and random.genes indicates the cluster consisting of random genes. Here, expression coherence was defined with using an expression threshold of log2CPM>=1, i.e. a gene is considered as ON when the log2CPM>=1. The horizontal line (blue color) denotes the coherence level of 5% of the total multi-clusters.

Previous research [109] showed that so-called ‘zinger’ motifs are enriched in ChIP-seq regions of several unrelated TFs. We conducted additional analysis to ensure that our clustering results are not affected by the zinger motifs (Supplemental Note 4 and Figure 7.6 of Supplemental Data). Moreover, it is possible that *EMT* can falsely yield multiple sub-models, even in the absence of heterogeneity, and those sub-models can be falsely clustered. By looking at the clustering tendency of the sub-models, we examined the heterogeneity across sub-models and found that it is possible to separate the sub-models into distinct clusters. (Supplemental Note 5 and Supplemental Figure 7.7B-C).

Next, we assessed the functional underpinning of shared binding rules across cell types (see Methods for details). Specifically, we assessed whether two co-clustered loci from different cell types (i.e., those obeying similar binding rules) are functionally associated relative to loci from the same cell type, but belonging to different clusters, indicating that they are obeying different binding rules. We measured a cluster-specific score for each binding sequence, and assigned each binding site in each cell type to one or more clusters. As per convention, we assigned each binding site to the nearest gene as a potential transcriptional target; 88% of the target genes were within 50 kb from the binding site (median distance 4.5 kb) (Figure 7.7G of Supplemental Data). To assess functional coherence of clusters, we defined two metrics: expression coherence and pathway coherence. Expression and pathway coherence are measured as the fraction of gene-pairs in a cluster (regardless of cell type) that are respectively co-expressed, or belong to same pathway. We assessed the significance of coherence using two sided Fisher's exact test. As shown in Figure3.2C, ~40% (~18%) multi-cell type clusters show significantly higher (p-value < 0.05) expression-coherence (pathway-coherence) than the background (expectation is 5%) and 5.5% of the clusters show both significant expression and pathway coherence (called dual coherence). Applying a more stringent p-value threshold (< 0.001), these coherent percentages are 35% (expression), 10% (pathway) and 4% (dual). Moreover, the expression and pathway coherence are highly correlated across clusters (spearman correlation=0.56, p-value=0.02). As a negative control, we conducted the same set of tests for random clusters with the same size as the real clusters. In both cases, the coherence was no greater than the null expectation (Figure3.2C).

Taken together, these analyses support the existence of heterogeneous sets of TF binding rules governing the *in vivo* binding and suggests that a subset of rules are shared across cell types with functional implications.

3.2.3 The role of interaction partners in a TF's binding occupancy

By using 981 PWMs for a comprehensive set of vertebrate TFs as the basis for features, *EMT* implicitly incorporates the contributions of interaction partners in predicting *in vivo* binding of the reference TF. To quantify the contribution of putative interacting motifs, we repeated the *EMT* training and testing using only the PWMs corresponding to the reference TF. Individual TFs are represented by multiple motifs in the literature (ranging from 1 to 8, with a median of 3; Table 7.7 from Supplemental Data), many of which differ substantially from each other, suggesting potential functional implications [110], [111], e.g. 75% of the intra TF PWM-pairs have less than 85% PWM-similarity, in contrast to 99% of inter TF PWM-pairs [112]. We refer to these motifs as the *reference motifs*, and, in contrast to the *Interaction* model, the *EMT* model utilizing only the reference motifs are referred as *NonInteraction* model. Figure 7.8 from Supplemental Data shows the prediction accuracies for the *Interaction* and the *NonInteraction models*; the diagonal elements represent the cross-validation accuracies within a cell type, while the off-diagonal elements represent the accuracy when *EMT* is trained on one cell type (row) and tested on another (column). Comparing the within cell type cross-validation accuracy for the *Interaction* and *NonInteration* models (Figure 3.3A, Figure 7.8 of Supplemental Data). The *Interaction* models have higher predictive accuracy than *NonInteraction* models, which is consistent with the expectation that *in vivo* binding of a TF relies on interactions among several TFs.

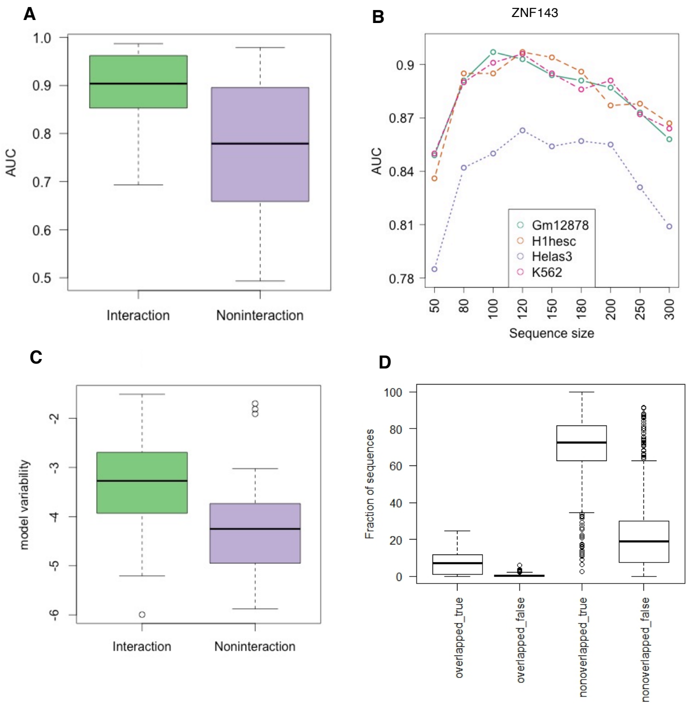


Figure 3.3 Association between the number of interaction partners and model-accuracy.

In plots (A) and (C) Interaction and Noninteraction models are indicated with green and purple respectively. (A) Comparison of cross-validation prediction accuracy for Interaction and Noninteraction models. (B) The trend of model accuracy with increasing sequence size for TF ZNF143 (selected arbitarily for illustration). Models from each cell line are indicated with different colors. (C) Comparison of model variability in log scale (variability of cross-cell type performance for each model) for Interaction and Noninteraction models. (D) Distribution of the fraction of test sequences falling into one of the four categories: Overlapped_true denotes correctly and overlapped_false incorrectly classified sequences having at least 50% overlap between training sequences in one cell type and test sequences in another cell type. Nonoverlapped_true (nonoverlapped_false) denotes correctly (incorrectly) classified sequences that do not overlap with any sequence in the training set.

Next, we conjectured that in the *Interaction* model, allowing for greater numbers of partners enables learning of more complex binding rules, leading to increased binding prediction accuracy. We therefore assessed the effect of the length of the region flanking the binding site on prediction accuracy (see Methods). We note that beyond 100bp, due to narrowing the gap between the foreground and the background region, the discrimination accuracy is expected to decrease. Despite this, in several cases (Figure 3.3B & Figure 7.9 of Supplemental Data), the increase in ROC-AUC beyond 100bp suggests that a larger context may be necessary in these cases to capture the binding rules. Nevertheless, we chose a sequence context of 100bp to make our model comparable to the previously published *kmer-SVM* [102].

For a given TF, we also quantified the variability of the model accuracy in different cell types (see Methods). We define cross-cell type prediction accuracy as the performance of a model from one cell type tested on another cell type. For these performance accuracy of models, we expect greater variability for the models relying on cell type-specific interaction partners than the models only relying on reference motifs. Our analysis supports this expectation, suggesting that the sequence information required for *in vivo* binding is encoded by the TF's own motifs which does not vary substantially across cell types (Figure 3.3C). Conversely, the role of context and interaction-dependences in TF binding varies substantially across cell types (Figure 3.3C). However, the small variability in cross-cell type prediction accuracy when using the *NonInteraction* model is likely due to the heterogeneity of TF binding motif. We quantified the inter-motif divergence for each TF as either the number of annotated motifs, or the motif-divergence (defined over all motifs-pairs) (see Methods). We found that the performance variability of *NonInteraction* models is positively correlated with both measures of motif divergence (Spearman correlation=0.63, 0.67; two sided p-value=1.2 x 10^{-3}, 6.3 x 10^{-4} respectively).

In Figure 7.8 of Supplemental Data, the off-diagonal elements for the *Interaction* model shows higher cross-cell type performance relative to the same elements for *NonInteraction* model. This higher performance suggests that the binding 'rules' are shared between cell types. We ensured that the high cross-cell type performance is not simply due to overlaps in the genomic loci used to train and test the model between cell types, i.e., the genomic loci on which the model was trained in one cell type does not substantially overlap with the loci tested in another cell type. Overall, across TFs and cell type pairs, the fractional overlap in genomic loci ranges from 0 to 10%, with a mean and median of ~4% (Figure 3.3D). This suggests that it is the binding rule, independent of specific sequence instances, that is shared across cell types.

Furthermore, we found that when using the *Interaction* model, the cross-cell type accuracy is symmetric. In other words, a high (low) accuracy in cell type *Y* using *EMT* trained on cell type *X* implies a high (low) accuracy in cell type *X* using the model learned from cell type *Y*. To demonstrate this symmetry, we normalized the off diagonal elements of cross-cell performance matrices by the reference AUC by diving each row by the corresponding diagonal ROC-AUC. Then we showed in Figure 3.4A, the lower and upper diagonal ranks are highly correlated (Spearman correlation of upper and lower triangle of resulting matrices is 0.68, two sided p-value 9.5 x 10^{-53}, Figure 3.4A), supporting our claim that the interaction-dependent (therefore genomic-context dependent) binding rules are shared across cell types. In stark contrast, there is a lack of symmetry in cross-cell prediction accuracy when *NonInteraction* model is used (Spearman correlation = 0.04, two sided p-value 0.4, Figure 3.4B and Figure 7.10 of Supplemental Data).

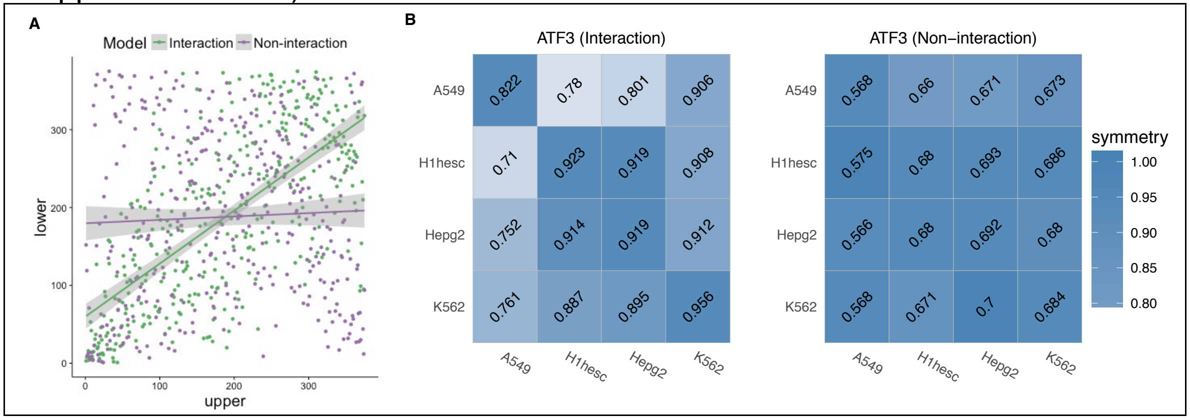


Figure 3.4 Comparing cross-cell type performance matrix of Interaction and Noninteraction models.

(A) Ranks of the normalized symmetry of upper and lower diagonal matrices of cross-cell type performance. Interaction and Noninteraction models are colored green and purple respectively. (B) In each matrix, row represents the cell on which the model is trained and column represents the cell from which the test data is used. Diagonal elements are within cell type performance and each matrix is color coded according to the extent of the non-diagonal element symmetry. The symmetry is calculated by normalizing each row by the reference model (diagonal element).

In summary, our analyses suggest that the cell type-specific TF interactions play critical role in determining the cell type-specific *in vivo* binding, and *EMT* reveals some of the interactions underlying the cell type-specific binding of a reference TF.

3.2.4 Cell-specific biological roles by putative co-factors

Our results so far suggest that cell type-specific co-factors of a TF are a major driver of cross cell-type *in vivo* binding variability. To gain further insights into the functional implications of cell type-specific co-factors, for each reference TF, we identified its cell type-specific co-factors using the feature importance of the corresponding motif as estimated by the model. To minimize redundancy, we excluded motifs with substantially high co-occurrence frequency with at least one of the reference motifs (see Methods). To further minimize false positives, we assessed the enrichment of motif occurrence within the cell-specific ChIP-seq peaks of the reference TF relative to background and retained only those putative co-factor motifs that were significantly enriched (odds ratio > 1.2 and two sided p-value < 0.05, see Methods). The rationale for choosing 1.2 as the odds ratio threshold is discussed in Supplemental Note 6.

Several lines of evidence support *TRISECT* identified cell type-specific TF co-factors, referred to as putative co-factors. First, we showed that there exists an enrichment of protein-protein interactions (PPI) among a reference TF and its corresponding co-factors as compared to the PPI interactions among all motifs (Table 7.12a from Supplemental Data). Additionally, the putative co-factors are enriched for either heterodimerizing TFs or for the TF family that the reference TF belongs to for almost 70% of all TF-cell pair cases (see Methods, Table 7.12b-c from Supplemental Data). The enrichment of same family as that of reference TF is consistent with the fact that TFs form dimer with other TFs preferably from same family [96], [113]. We also performed protein domain enrichment analysis (Table 7.13 from Supplemental Data) using DAVID tool [114], [115], and found that more than 80% of enriched domains are involved in homo- or hetero-dimerization consistent with the findings from Table 7.12 from Supplemental Data.

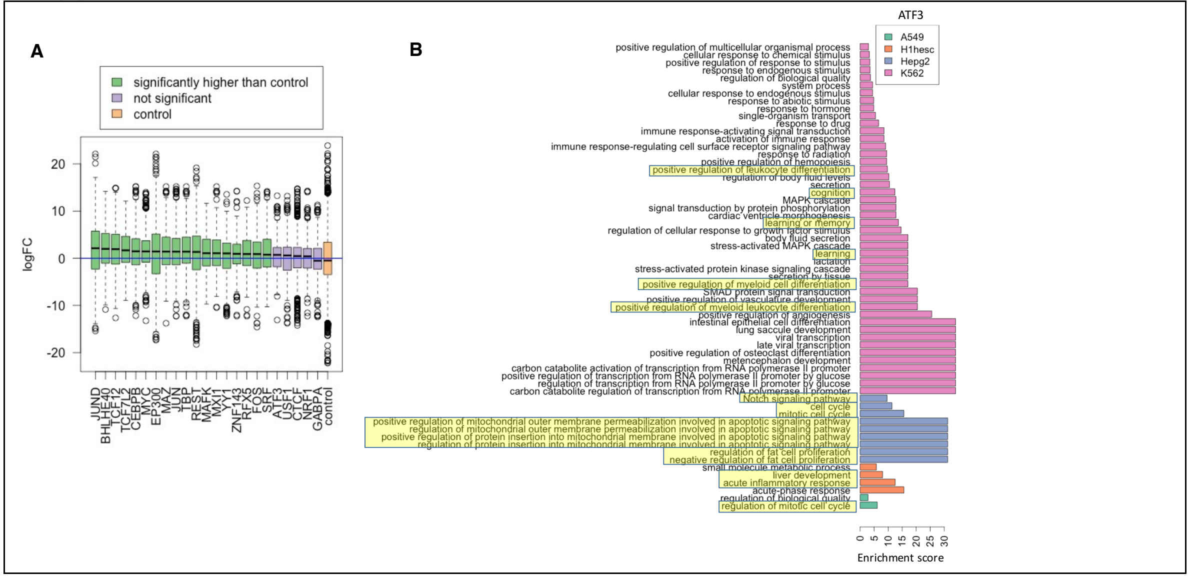


Figure 3.5 Functional validation of putative co-factors.

(A) Each boxplot corresponds to all co-factors of a TF in X-axis and Y-axis denotes the log fold change (logFC) of the expression of co-factors in relevant cell vs. non-relevant cell. The 'blue' horizontal line at Y=0 denotes no fold change. For a TF motif detected as a co-factor in n cell lines, and not in another m cell lines, we calculated log fold change (logFC) in the TF's expression between the two sets of cell lines. Identified co-factors have higher expression in the cell lines they are detected in (relevant cells). (B) Enrichment scores of GO terms obtained from GO analysis of co-factors in four cell types of ATF3 (selected arbitrarily). The known cell type-specific biological roles are highlighted.

Second, we expect higher expression of putative co-factors in the cell types where they are identified as co-factors by our analysis. For each co-factor (excluding ubiquitous co-factors), we determined the log-fold difference in expression between the cell types where it is identified as co-factor relative to cell types where it is not (see Methods). The distributions of log fold changes of the co-factors are compared with a control set of fold ratios as presented in Figure 3.5A. For most TFs, the co-factors show significantly higher expression in the relevant cells. This is not true only in 5 cases: ATF3, USF1, CTCF, NRF1 and GABPA. Among these 5 cases, CTCF is a known cell type-independent TF, GABPA and NRF1 exhibit higher cell independence than other TFs as shown via an independence test.

Third, we assessed whether the relationship between a reference TF and its co-factor is symmetric. For this assessment we limit the analysis to 23 TFs, as for the current study we have models and associated co-factors only for these TFs. Specifically, we assessed whether a reference motif from one TF appears as co-factors in the TFs whose reference motifs are also reported as co-factors in the first TF. For all X-Y TF pairs where one TF is deemed co-factor of the other and both TFs have available ChIP-seq data in the same cell line, we found that the correlation between the enrichment score of motif X in the biding sequences of TF-Y and vice versa is 0.41 (two sided p-value = 5.19 x 10^{-14}). This suggests a degree of co-dependence among TFs for their DNA binding.

Finally, for each TF's cell type-specific co-factors, we performed biological processes (BP) GO term enrichment analysis using the GOrilla tool [116] relative to all 981 motifs. We found significant differences in the assigned BP of a TF's co-factors among cell types. Remarkably, the BP can vary across cell types while still being functionally related to the reference TF. As an example, Figure 3.5B shows the enriched BP (false discovery rate ≤ 10%) for ATF3 in 4 cell types. ATF3 is a stress-inducible TF involved in homeostasis regulating cell-cycle, apoptosis, cell adhesion and signaling [117], [118]. We found that ATF3 co-factors are enriched for cell cycle and proliferation functions in 3 out of 4 cell lines. In the stem cell line, the identified co-factors are involved in liver regeneration and inflammatory response, consistent with previous studies showing a direct link between ATF3 induction to liver injury and regeneration in mice [119], [120]. Furthermore, enrichment of NOTCH and apoptotic signaling among co-factors in the Hepg2 cell line is consistent with ATF3's role in glucose homeostasis and other primary liver functions [117]. Surprisingly, we find enrichment of cognition, learning and memory among the TF co-factors in the leukemia cell line. Since leukemia is a cancerous cell line, non-native gene expression is not unexpected [121], [122]. While, ATF3 is not known to play a direct role in neuronal function, a functionally and structurally related protein CREB has a well documented role in neuronal activity and long-term memory formation in brain [123]. This raises the possibility that either ATF3 has a unknown role in cognition, or the same set of co-factors are involved in memory formation in conjunction with other TFs.

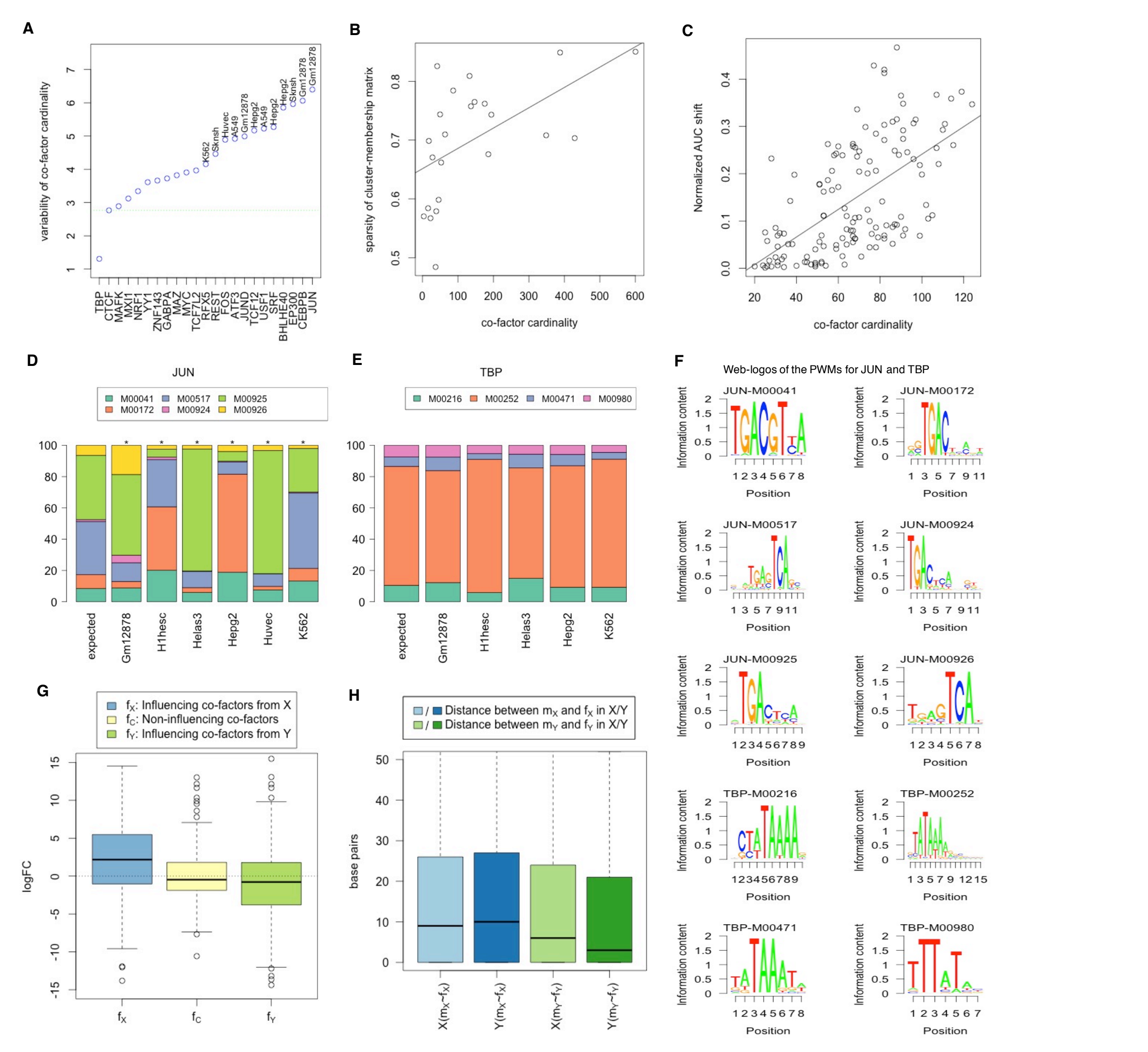


Figure 3.6 EMT model heterogeneity is associated with cell type-specificity of co-factors.

(A) The plot shows for each TF the variability of co-factor cardinality across cell types. Each point is labeled by cell type where the relevant TF has specific usage, based on the literature and has the largest number of co-factors. TBP and CTCF are the most ubiquitous TFs. The 'green' dotted horizontal line denotes the variability of cardinality for CTCF co-factors. (B) Sparsity of cell-membership matrix correlates with co-factor cardinality. (C) Normalized ROC-AUC difference of Interaction and NonInteraction models for a specific TF-cell type pair correlates with co-factor cardinality. (D-E) Motif usage for the reference TF in the NonInteraction models of different cells, for JUN and TBP as two extreme examples. Y-axis denotes the feature importance of motif usage in the NonInteraction model. The

sequence logos for the corresponding reference PWMs are presented in (F). In (G) and (H) fX (fY) denotes the influencing co-factors of mX (mY) in cell line, X (Y) (G) Left: Log fold change (logFC) between relevant and non-relevant cell type for influencing co-factors of mX; Middle: logFC for non-influencing co-factors; Right: logFC between non-relevant and relevant cell type for influencing co-factors of mY. (H) Genomic proximity of the motif-specific interaction partner with the motif. mX~fX denotes the nearest genomic distances (in base pairs) from mX motif to any co-factors in the set of fX and so on.

For other TFs, the enriched GO-terms are listed in Table 7.14 from Supplemental Data (enrichment scores ranges from 1.22 to 93.75 with a median of 7.44, false discovery rate cutoff of 10%). The corresponding discussion based on a review of the literature is provided in Supplemental Note 7 and Supplemental Note 8 includes example co-factors in various cells. This can serve as a resource for further investigation into the cell type-specific binding and function of a broad array of TFs.

We noted substantial variability in the number of detected co-factors across cell types for a TF. Interestingly, a literature survey suggests that the cell types for which the reference TF has specific known function, the number of co-factors in that cell type is comparatively higher. For example, REST has well-known neuronal functions and its binding sites in neurons exhibit lack of cognate RE1 motifs [124], suggesting co-factor dependence. Consistently, Sknsh (brain cancer cell line) has the highest co-factor cardinality for REST. Similarly, JUN plays a specific role in hematopoetic differentiation and we found that Gm12878 (normal blood cell line) has the largest number of co-factors [125]. We reasoned that a TF with greater cell type-specific roles would exhibit greater variability in co-factor cardinality. For each TF, we measured the variability of its co-factor cardinality across cell types. As shown in Figure 3.6A, interestingly, TFs with ubiquitous and invariant roles such as TBP and CTCF have the least variable co-factor cardinality. Based on the trend shown in Figure 3.6A, we use the variability of co-factor cardinality as a proxy for the TF's *cell type-specificity*. As an additional support, this proxy also correlates with the *Sparsity* measure of cluster-membership matrix. Specifically, for each TF we computed the *sparsity* of its cluster-membership matrix (presented in Figure3.2A-B & Figure 7.4 of Supplemental Data) using *Gini index* [126], [127]. Figure 3.6B shows that *sparsity* is positively correlated with the variability of co-factor cardinality (Spearman correlation = 0.66, two sided p-value = 9.2 x 10^{-4} using *k*-NN).

We also assessed whether differences in prediction accuracy achieved by the *Interaction* model and the *NonInteraction* model for a particular TF-cell type pair may reflect the TF's co-factor dependence. We compared co-factor cardinality to the normalized distance between *Interaction* and *NonInteraction* model performance (*AUC shift*). As shown in Figure 3.6C, the A*UC shift* is positively correlated with co-factor cardinality (Spearman correlation = 0.65, two sided p-value = 2.7 x 10^{-17}).

Previous studies have found that the DNA sequence specificity of a TF can be influenced by its interactions with co-factors [128], [129]. Interestingly, a close inspection of the feature importance estimated by the *NonInteraction EMT* model shows that for different cell types the composition of utilized reference motifs varies. Figure 3.6D-E presents all cell type-specific usage of JUN & TBP (see Figure 7.11 of Supplemental Data for other TFs); JUN shows significantly different binding specificity from the expected usage in different cell types (marked with asterisk, see Methods), while TBP does not. Notably, such diverse usage is observed using *NonInteraction* models, suggesting a cell type-specific motif preference. In Figure 3.6D, M00925 (JUN) and the reverse complement of M00926 (JUN) are almost identical, yet they show very different usage. Even though both PWM have very similar distributions of scores over the same genomic regions, in most cases M00925 yields slightly higher score than M00926 and once M00925 is selected by a model, M00926 is deemed as redundant and not considered as important further. Hence, they show dissimilar importance. However, in our downstream analysis of assessing contribution of cell-specific usage, none of them are selected as having cell-specific influence and thus has no impact on the analysis.

We further investigated the potential contribution of cell type-specific co-factors in modulating the cell type-specific motif usage for the reference TF. In this regard, we identified pairs of reference motifs (m_X & m_Y) having the most differential usage in cell types X and Y respectively. For each such pair we selected a set of candidate co-factors (f_X & f_Y) which could potentially aid the TF for cell type-specific binding; we call them (f_X & f_Y) influencing co-factors of m_X and m_Y respectively. Next comparing the log fold change (logFC) of f_X & f_Y in cell type X versus Y (Figure 3.6G), shows that the influencing co-factors have higher expression in relevant cell types. Moreover, the influencing co-factors are more proximal to the influenced motif in the relevant cell type (Figure 3.6H, see Methods for details).

Taken together, cell type-specific co-factors revealed by TRISECT are consistent with their cell type-specific expression and function which may be critical in modulating a TF's cell type-specific biological function.

3.3 Methods

3.3.1 Data Processing

We downloaded the ChIP-seq peaks for 23 TFs from ENCODE [130] (Table 7.6 from Supplemental Data). For each TF we selected only those cell lines for which narrow-peak data was available. We chose the more stringent of the two criteria – top 5000 most significant peaks, or FDR q-values<0.2 to select the binding sites. The criteria are reasoned by the availability of enough data to build a model and the backward compatibility of the previous method [102]. Notably, not all ENCODE datasets provide q-values and in that situation we generate the list of q-values from the given p-values [131]. Relative to the center of ChIP-seq peaks, the DNA regions of length 100bp were identified as the foreground. As negative control, we sampled flanking regions of 100bp from 200bp away from the positive sequences. Again, the choice for the size and location of foreground and background can be rationalized by the backward compatibility. In fact, choosing control sequences from near the foreground makes the modeling problem harder than when they are chosen from arbitrary locations in the genome. Moreover, control sequences overlapping with any peak were excluded. Due to the proximity of the negative examples, both foreground and background are expected to have similar GC-composition [102] and chromatin accessibility. However, we explicitly controlled for the GC composition using sequence set balancing technique when comparing the foreground and the background [89]. In the sequence set balancing, the GC percentage is divided into N bins (e.g. we choose N=100). Then for both foreground (*F*) and background (*B*) sets, the number of sequences falling into each bin are enumerated: *F[i]* & *B[i]* where *i=1 to N*. Finally, in each bin *min(F[i], B[i])* sequences are selected randomly from foreground and background set. This way each set of sequences will have similar distribution of GC-composition. After sequence set balancing, we discarded any cell line resulting in fewer than 4000 sites. In our list of TFs, EP300 is non-sequence specific. Even so, EP300 is localized to the chromatin by interacting with other motifs. Like Arvey et al. we include EP300, specifically to reveal those putative interactions.

In addition to the 100bp foreground and background, we also extracted another 6 sets of foreground and background of size 120, 150, 180 200, 250 and 300 base pairs. We keep increasing the size of foreground to check how much additional information was added to the model by the increased sequence size. Note that for all sequence sizes the middle point of the background does not vary; so as the sequence size is increased the gap between foreground and background decreases.

3.3.2 Learning EMT (Ensemble model of TF binding)

We considered three types of feature set for the sequence specificity model: (1) *K*-mers - frequencies of 4096 6-mers in the 100bp sequence, (2) *K*-merRC - frequencies of 2080 *k*-mer (*k*=6) groups equating a *k*-mer and its reverse complement, and (3) *Interaction* (***Lk***) – we obtained all 981 vertebrate positional frequency matrices (PFM) from TRANSFAC 2011 as the features. Each PFM was converted into positional weight matrices (PWMs), which is a log-likelihood matrix, by, (1) adding a pseudocount of 0.2 of ‘C’, ‘G’, and 0.3 for ‘A’,’T’ in line with genome composition, (2) normalizing the frequencies to get probabilities for each base, (3) dividing each base probability by the background probabilities (0.2 of ‘C’, ‘G’, and 0.3 for ‘A’,’T’), and (4) taking the log of the probability ratio. The resulting PWMs were then used to get the motif matches using PWMSCAN [108]. Here, ***Lk*** refers to the PWM hit threshold (hit expected every *L* kb on average in the genome); we used *L* = 1, 2, 5, or 10. In particular, we use log(1/L k) as the threshold value to call a PWM ‘match’. For instance, at L=1, the expected frequency of matches is once every 1kb, corresponding to a 20% chance of a match in a 100 bp region or its reverse complement. Previous research showed that clusters of homotypic 'weak' binding sites are prevalent in regulatory regions [132] and such presence of multiple weak binding sites, called homotypic cluster of binding sites, are preferred to single strong binding actual binding [133]. To mimic this binding affinity, from the output of PWMSCAN, we decided to use the sum of PWM-score (-log(match score)) for all matches as the feature value. However, we also collected the 'maximum score' and 'average score' of the bindings for each training sequences and measured their correlation with our feature value. The high correlations (0.8 and 0.87 respectively) suggest a minimal effect on downstream analysis and overall conclusions. Finally, we used the log sum of PWM-score to compensate for the skewed distribution of the number of binding sites for individual TFs.

We found that the model performance was better for the 1k than the 2k thresholds, and at much higher stringency the model performance significantly deteriorates due to the sparsity of the matches (Figure 7.3C of Supplemental Data). Further, we determined the feature importance of the motifs for each TF-cell pair at those four thresholds. For each TF-cell pair, we calculated the correlation of the feature importance based on 1k threshold with those based on other thresholds, i.e., three correlation values. Thus in total, we calculated 405 correlation measures for 135 TF-cell pairs. We found that 90% of those correlations are significant, ranging from 0.21 to 0.81 with a median of 0.52. Considering the relative performance of the *Interaction* (1k) model, in the subsequent analysis we use them as the representative *Interaction* model, and refer to it as such.

We chose Adaptive boosting [106], [107] as our composite model where each sub-model within the ensemble is a decision tree and each decision tree is constructed based on a bootstrap sample. We used the Adaboost framework implemented in R gbm package [134]. In the framework, Huber loss function is selected to reduce over-fitting. We estimated the classification accuracy of the model based on 25% held out data set, while 75% data were used to build the cell-specific models. In Supplemental Note 1, we summarize the interpretation of a model and parameter choices.

3.3.3 Model conversion, Duda-Hart test and Hopkins statistics

Each sub-model is represented by a point in a *d*-dimensional space. Each dimension denotes a feature and the value along the dimension indicates the importance of the feature for the sub-model. Therefore, each model (consisting of multiple sub-models) can be represented as a set of points in a *d*-dimensional space where $d \leq$ number of features (981). For a model, the feature importance was measured using the prediction performance improvement for out-of-bag sample predictions. We modified the gbm package [134] implementation of feature-importance to accommodate the calculation for single tree or the sub-model in question. In other words, we determined the contribution of a single tree (sub-model) in prediction performance improvement using the same out-of-bag samples. We disregard the features which do not contribute to any sub-model. We conducted *Duda-Hart* test to show that whether the sub-models belong to one or multiple clusters. We measured *Duda-Hart* or dh-ratio (ratio of within-cluster sum of squares and overall sum of squares) for all cluster pairs, based on either cell type-specific set of sub-models, or the pooled set of sub-models across all cell types for a TF [135]. While calculating dh-ratio, *k*-Nearest Neighbors algorithm (*k*-NN) was used for clustering. Since the final output of *k*-NN depends on initial random set of centers, the dh-ratio calculation was repeated 1000 times to ascertain robustness. We noted that all test results were significant (p-value < 0.001).

Hopkins statistics (H) was measured to check clustering tendency of the sub-models. To measure Hopkins statistics (H), the sub-models are again represented as a set of points. H is defined by the following.

$$H = \frac{\sum_{j=1\ldots m} U_j^d}{\sum_{j=1\ldots m} U_j^d + \sum_{j=1\ldots m} W_j^d}$$

W_j are the nearest-neighbor distances of *m* randomly chosen points (sub-models), which demarcate the sampling window. $U_{\llcorner}$ are the minimum distances of the sub-models from *m* random points in the sampling window. To define the sampling window, we either took 25 to 75 percentile of the feature values or from δ to max.value-δ along each dimension, where δ denotes the standard deviation of the feature value [136]–[138]. To estimate p-value, we repeat the above procedure 1000 times and measured the H value. The p-values range from 0.026 to less than 0.001.

3.3.4 Clustering sub-models

For a TF, we obtained the sub-models from all cell types, and then clustered all sub-models using *k*-Nearest Neighbors algorithm (*k*-NN), where each sub-model is an instance and the features of the instances are individual feature-importance obtained in the context of respective cell-specific model. Before feeding into the *k*-NN, we remove all the features whose cumulative importance over all sub-models is zero. To check robustness, the sub-models are also clustered using XY-fused version of self-organizing map [139] from kohonen R package [140]. To make it comparable to *k*-NN, sub-models were clustered without preexisting sub-model cell labels, i.e. we assumed 100% weight for X map.

3.3.5 Assignment of sequences and target genes to the clusters

A cluster of sub-models can be viewed as a new ensemble. Therefore, for each cluster, we built a gbm object by treating the cluster as an ensemble and used it the same way an original *Interaction* model would score a sequence. Thus, we scored each binding site sequence against each cluster, and a sequence is assigned to a cluster when it is scored above a threshold (of 1) by the cluster. The choice of the threshold was based on the rationale that the intercept (bias of the model [134]) of cell-specific models are ~1, and for a high-confidence positive sequence, the model-score should be greater than the intercept. Each bound sequence (from all cell lines) is mapped to a set of clusters. For each bound sequence, the nearest gene on the genome is considered to be its putative target, as per convention [141]. Hence, each cluster corresponds to a set of target genes coming from different cells.

3.3.6 Measuring pathway and expression coherence

To measure the functional coherence, we determined the target gene array of size M-by-N for M clusters and N cell types. The *M*-by-*N* array thus includes a set of genes corresponding to each cluster in a particular cell type. We compared gene-pairs from the same row across columns (same cluster, different cells) to a background of gene-pairs along columns from different rows (same cell, different cluster). Then we apply the Fisher's exact test in a cluster-centric fashion by comparing the fraction of co-clustered gene-pairs in the foreground as compared to the background. The measure is named as expression coherence: whether targets gene pairs from same cluster but different cell lines are more co-expressed than those from different clusters but same cell line. A gene-pair is considered co-expressed if both of the genes are turned on (RNA-seq log2CPM > 1) in their respective cells; CPM stands for Counts per Million. CPM, instead of the standard FPKM measure to quantify gene expression suffices for our purpose as we only compare a gene’s expression across samples, and not with other genes in the same sample. We showed similar trend of expression coherence with different expression threshold (log2CPM>=5) (Figure 7.7E-F of Supplemental Data).

Pathway coherence is also assessed in similar fashion: whether the target genes from different cell lines that are assigned to the same cluster are more functionally related (i.e. in the same pathway) than the target genes coming from the same cell but from different clusters. Pathway data was downloaded from KEGG pathway database (www.genome.jp/kegg).

3.3.7 Robustness of EMT and sub-model clustering

While building *EMT* using gbm R package, we used the default parameter settings except maximum depth of variable interaction (interaction.depth), minimum number of observations in the trees terminal nodes (n.minobsinnode) and learning rate (shrinkage). Our parameter choices are the following – interaction.depth: 15, n.minobsinnode: 30, shrinkage: 0.05. To check model and pipeline robustness, we build models with different values of these three parameters and compared the performance and model size (number of learned sub-models). We found that performance and model size becomes stable after interaction depth of 15 (Figure 7.3D-E of Supplemental Data), performance and model size do not vary much with the change of n.minobsinnode from 25 to 45 (Figure 7.3G-H of Supplemental Data), and performance does not change which shrinkage from 0.1 to 0.5 (Figure 7.3I of Supplemental Data). However, model size varies with the shrinkage parameter setting because with lower learning rate, it takes longer to reach an optimum, and it results an increase in the model size (Figure 7.3J of Supplemental Data). Therefore, for different shrinkage parameters, we measured the clustering consistency. To this end, we took the models built with shrinkage=0.05 as the reference models and we compared the clustering pattern of reference models with the set of models built using different shrinkage value. More specifically, we determined if a pair of sequences fall into same cluster for the reference model, does it also fall in same cluster for a different shrinkage value. We found that on average 96% of the sequence-pairs fall in the same clusters regardless of shrinkage (Figure 7.3K of Supplemental Data).

3.3.8 Model variability, and Motif-divergence

Model variability is defined by its normalized-predictability across cell lines. For each model, n ROC-AUC values are obtained using the held-out dataset of n cell-lines. Cross-ROC-AUC values are normalized by self-ROC-AUC value. Mathematically,

$$var_{model_i} = \frac{\sum_{j \neq i, j \epsilon cells} rocauc_j}{rocauc_i}.$$

Motif-divergence is defined by the following equation. $motif.div._{pwms} = \sum_{i,j \in pwms} \frac{dist_{i,j}}{IC_i + IC_j}$. Here, $dist_{i,j} = 1/similarity_{i,j}$ and IC_i is the information content of the ith motif. Similarity between two PWMs is calculated following the normalized version of the sum of column correlations [142].

3.3.9 Identification of co-factors

EMT provides importance of all features in discriminating the foreground from the background. We retained all features with nonzero importance. From the initial set, we removed any motif that has 60% PWM-similarity (consensus overlap) for at least 50% of the binding site locations with any of the reference motifs. Next, we calculated an enrichment score (i.e. odds ratio) of the motif in the foreground binding sites relative to control sites. We retained the motifs with greater than 1.2-fold enrichment and two sided p-value < 0.05. The resulting motifs were considered as co-factors. For further analysis, we considered cell-specific co-factors by removing common motifs across cells. In particular, we excluded all co-factors that are common between any two cell-lines. The functional cell-specificity measure for a TF is determined using the variability of co-factor cardinality of such unique co-factors.

3.3.10 Validation of co-factors using PPI and TF family

We obtained protein-protein interaction (PPI) data from STRING v10 [143]. Using the TRANSFAC 2011 database, we determined the mapping from motifs to ENSEMBL protein id and the number of motif pairs having PPI. Using hyper-geometric test we calculated the enrichment of PPI between reference TF and each set of cell-specific co-factors. The test summary indicated that 81% of the TF-cell cases have higher PPI enrichment among the interactions involving reference TF and their co-factor (Table 7.12a from Supplemental Data).

We compiled each PWM's family and the list of heterodimerizing PWMs from TRANSFAC 2011 database. To identify heterodimerizing TFs, we looked for the presence of keyword 'heterodimer' and absence of 'no' or 'not' in the description of the motif. Table 7.11 from Supplemental Data shows the heterodimerizing PWMs. Detailed manual inspection of a random subsample suggests that this automated criterion may result in ~5% false positives. We also noted that occasional use of the term 'dimer' instead of 'heterodimer' may lead to ~20% false negatives. For the hyper-geometric test of family-enrichment, we compared how many co-factors belong to the family of reference motif relative to the 981 motifs. Heterodimer enrichment was tested similarly. The enrichments scores (odds ratios) and p-values are reported in the Table 7.12b-c from Supplemental Data. The Table shows that 70% of the model-co-factors are either enriched for heterodimerizing TFs or TFs coming from same family.

3.3.11 Gene expression and differential gene expression

For gene expression, we used RNA-seq data downloaded from ENCODE (Table 7.10 from Supplemental Data). For each cell, we obtained between 2 and 4 RNA-seq samples depending on the availability and obtained the number of reads aligned to the gene. We corrected for batch effect using sva R package [144]. To estimate the differential expression between two sets of cell lines (those in which a TF is deemed a co-factor, and those where it is not), we used the linear model implemented in limma package of R [145].

For each co-factor, we determined all possible relevant & non-relevant cell pairs and took the log fold change (logFC) of the expression in those cells. To determine the control gene expression, we considered the same sets of cell pairs but took the logFC of an arbitrary gene instead of the co-factor. In both cases, we considered only significant differential expressions (logFC values with p-value < 0.05) provided by the limma package [145].

3.3.12 Cell-specific PWM for the reference TF

We obtained relative feature importance of the reference motifs from the *Noninteraction* models and compared them with random expectation. To calculate the random expectation, 1000 *Noninteraction* models are learned based on randomly sampled 4000 sites from all binding sites across cell-lines. From 1000 models, 1000 relative feature importance was calculated. Each set of relative importance was assumed a point in p-dimensional space where p is the number of reference motifs. We considered the relative importance vectors as data points from multivariate normal distribution and for each vector we calculated the Mahalanobis distances from the centroid which follows a chi-square distribution [146]. The degrees of freedom (d) for the chi-squared distribution was determined using maximum likelihood estimate and a p-value was generated from a chi-square distribution function of d degrees of freedom.

3.3.13 Influencing co-factors, proximity to the influenced motif, and expression in the most used cell

We identified the influencing co-factor set in the cell where one motif is used much more frequently than the others. More specifically, for a TF, we identified pairs of motifs and cell types where there is a maximal differential in cell type usage of the two motifs (i.e. one of the motifs has the highest usage in one cell type and the lowest usage in another, and vice versa). For such pairs of cell types X, Y, and corresponding reference motifs m_X & m_Y, we determined the candidate motif-specific co-factors f_X and f_Y as follows. We first separated the sequences from cell types X and Y where m_X and m_Y matches are found, respectively. Next, we assessed each putative co-factor's motif enrichment in each sequence set relative to the other sequence set. If the putative co-factor is enriched in X relative to Y we consider it as a putative influencing co-factor for m_x, and likewise for m_y. All other co-factors (f_c) are considered non-influencing, and serve as negative control.

We measured the fold change (logFC) of all influencing and non-influencing co-factors in X vs. Y using limma package [145]. To demonstrate the genomic proximity between influenced motif and influencing co-factors, we chose the nearest distance between them among potentially multiple motif matches.

3.3.14 Ubiquitous vs. cell-specific sub-models

We designated a cluster as cell type-specific if all member sub-models (at least 5) came from the same cell type. We then estimated skewness for each multi-cell type [147] based on the numbers of sub-models contributed to the cluster by various cell types. If the skewness was less than 25%, we designated the cluster as ubiquitous. For each cluster, we counted the number of relevant features (i.e., with non-zero importance). Among the relevant features, we retained only those which were deemed as putative co-factors for at least one of the cell-specific models in our earlier analysis. The retained co-factors are designated ubiquitous or cell type-specific based on the label of the cluster they belong to. Any common features from the two sets are removed. For each feature, we collect the expression across cell types in question and measure the skewness of gene expression [147].

3.4 Discussion

In this study, we have presented a novel ensemble-based framework –*TRISECT*, to investigate intra-cell type heterogeneity and inter-cell type commonality of *in vivo* TF binding rules. To the best of our knowledge, this is the first study to comprehensively demonstrate that *in vivo* binding specificity rules are composed of multiple components, or sub-models, many of which are shared across multiple cell types. Importantly, non-orthologous targets of binding sites across cell types governed by a shared binding sub-model exhibit a greater functional and expression coherence than targets of binding sites in the same cell type that are governed by different binding rules. For each TF, *TRISECT* identified cell type-specific co-factors that are supported by gene expression data and literature studies supporting their cell type-specific function.

We chose Adaboost as our ensemble model due to its architectural advantages with respect to our ultimate goal of analyzing common and distinct binding rules, or sub-models, across ensembles learned for each cell type. Boosting ensemble methods, including Adaboost, are designed to learn optimal tree sub-models for successive reweighted bootstrap samples. This is in contrast to other ensemble methods, including the popular Random Forest (RF) approach which seeks to increase variability of sub-models by estimating weak sub-models from un-weighted bootstrap samples. Since our primary goal is to reveal model heterogeneity, we chose to cluster sub-models generated by Adaboost rather than Random Forest's weak learners.

In terms of prediction accuracy, *EMT* compared favorably to the previously reported sequence-based discriminative model (*kmer-SVM*) [102]. Apart from the modeling approach, our study differs from Arvey et al. 2012 in several other aspects. The previous study compared the cell type-specific models for only two cell types – GM12878 and K562, while we have investigated in-depth the cell type-specificity of *TRISECT* across 4-12 cell types for each TF. While the previous work primarily discusses cell type-specificity and ubiquity of their models, by clustering the cell type-specific sub-models, our work investigates the extent of shared binding rules; cell type-specificity and ubiquity are extreme cases thereof. In addition to the cell type-specific variability in proximal co-factors, we investigated in much greater depth the cross-cell type variability in the preferred motif for the reference TF. Together, these novel aspects of our study adds to the knowledge of sequence information that specify a TF's *in vivo* binding in various cell types.

Another recent study [96] aimed at deciphering the determinants of *in vivo* occupancy of a TF showed that TF binding specificity is influenced by nearby homotypic sites (for the reference TF), the local nucleotide composition, and certain DNA physical properties. Moreover, the preferred *in vivo* binding in homotypic clusters was related to a preferred nucleotide composition, e.g. GC-rich for zinc finger TFs and AT-rich for homeodomain reference TFs, in the binding site flanking region. These previous findings are consistent with the fact that the co-factors identified by *TRISECT* are enriched for same family of TFs as the reference TF and thus have similar preference for nucleotide composition to the reference TF. In the previous work [96], the accuracy in discriminating bound vs. unbound sequences after controlling for the presence of a putative site for the reference TF was modest (ROC-AUC ~ 0.6). In contrast, we have shown that the motifs for the reference TF alone can discriminate bound sites from unbound control sites with ROC-AUC ~ 0.78, suggesting that the reference TF is the most informative determinant of *in vivo* binding, which is indeed expected, and was also observed by Pique-Regi et al [148]. The additional power of discrimination comes from either the presence of co-factor motifs, as suggested before [94], [102], or from nucleotide composition and other DNA physical properties [96]. Interestingly, DNA flexibility measured by propeller twist [149] is highly dependent on GC-content [150], which in turn is related to motif composition, as we have noted. Overall, the three properties, nucleotide composition, DNA physical properties, and motif composition are interrelated. The specific advantage of an ensemble model based on motif composition is that, apart from achieving favorable accuracy, it is functionally more interpretable and can provide insight into a TF's cell type-specific functions.

Context-dependent function of a *cis* regulatory region requires binding of a specific combination of TFs. This modularity contributes to morphological evolution through changes in cis elements controlling transcription, while avoiding the pleiotropic effects of TF gene's expression change [151]. Shared sub-models of TF binding rules across cell types, as revealed by *TRISECT*, may suggest shared history of cell types.

The ability of a TF to bind to diverse reference motifs and in conjunction, interact with diverse combinations of co-factors serves to enhance its functional repertoire across contexts [102], [152]. Our analyses reveal a cell type-specific preference for the reference motif as well as the cell type-specific interaction partners of a TF. We found the expression of cell type-specific interaction partners to be higher in the cell types where they are expected to interact with the TF, and their function is consistent with the context based on the literature. Thus, our study provides further support for a TF's cell type-specific functions, and more importantly, enables further investigation into the mechanisms underlying a TF's diverse cell-specific functions.

4 Heterogeneity of breast cancer metastasis

4.1 Background and Related works

Metastasis is the spread of a cancer from the primary oncogenic site to a different secondary organ. Current data suggest that metastasis from a primary organ to secondary organs is biased, that tumors from a primary tissue tend to spread to a secondary organ more often than other tissues, and the mechanisms underlying this biased 'organotropism' is not fully understood. Previous efforts toward this have been limited to mouse model and cell lines breast cancer metastasis signature to a specific secondary tissue [260], [261], or molecular characterizations of various breast cancer sub types [262], [263]. A detailed molecular characterization of organotropism has not been reported.

An important requirement of characterizing organotropism is recognizing that every tumor is composed of multiple clonal populations with distinct mutational and transcription profiles. This molecular heterogeneity presents a major hurdle toward developing effective cancer therapies. Such heterogeneities occur in biological systems at several levels, from rules that govern molecular interactions to cellular identity. Thus, the characterization of heterogeneities is fundamental to effective modeling of biological systems.

The motivation of the current study is to model the heterogeneity of models of breast cancer metastasis to characterize components of the model that are unique to specific secondary organs and those that are shared among them. Toward this, we have developed an ensemble model-based approach (MONTAGE) to reveal both tissue-specific and tissue-independent rules of gene interactions for breast cancer metastasis.

4.2 Results

4.2.1 MONTAGE – Models of organotropism and metastasis using Gene Expression

Overview. An illustration of the *MONTAGE* analysis pipeline is presented in Figure 4.1 and a brief description of the pipeline is provided below (for additional details see Methods).

As the first step, we developed an ensemble model of metastasis (*EMM*) to discriminate patients' gene expression profile of secondary metastasis (foreground) from the profile of primary cancer (background). The data was collected from Harrell et al, who integrated expression data of 4 cohorts sequenced in 3 platforms (see Methods for details). Given this expression data with distant breast cancer metastasis to bone, brain, liver and lung, and primary breast cancer covering (Table 4.1), we first selected the genes/features with high variability across patients regardless of disease status. Then, depending on single or non-single metastasis destination, we built two kinds of *EMM* models and three kinds of tissue specific EMM models as described in Table 4.2. The EMM models are: a) unique metastasis model and b) non-unique metastasis model; and the tissue-specific EMM models are: a) tissue specific unique models, b) tissue specific non-unique models and c) tissue specific metastasis models. The 1st sets of models were built to gauge the general metastasis signature and the 2nd set of models we built to assess secondary tissue specificity. In each set, we divided them into unique and non-unique model to differentiate the heterogeneity of patients for tissue exclusive and non-exclusive way. Each *EMM* model was trained using the Adaboost method where each sub-model is a decision tree built from a bootstrap sample [105]–[107]. Each tissue-specific model is composed of an ensemble of sub-models. Next, given the tissue specific *EMM* models for all secondary tissues, each tissue-specific sub-model was represented by a point in a *d*-dimensional space, with *d* corresponding to the number of relevant features. We constructed clusters of the data points for breast cancer metastasis (representing the sub-models across all tissues), using *k*-Nearest Neighbors algorithm (*k*-NN). The sub-models within a cluster represent similarity in gene expression combinations (or 'rules') within or across the tissues.

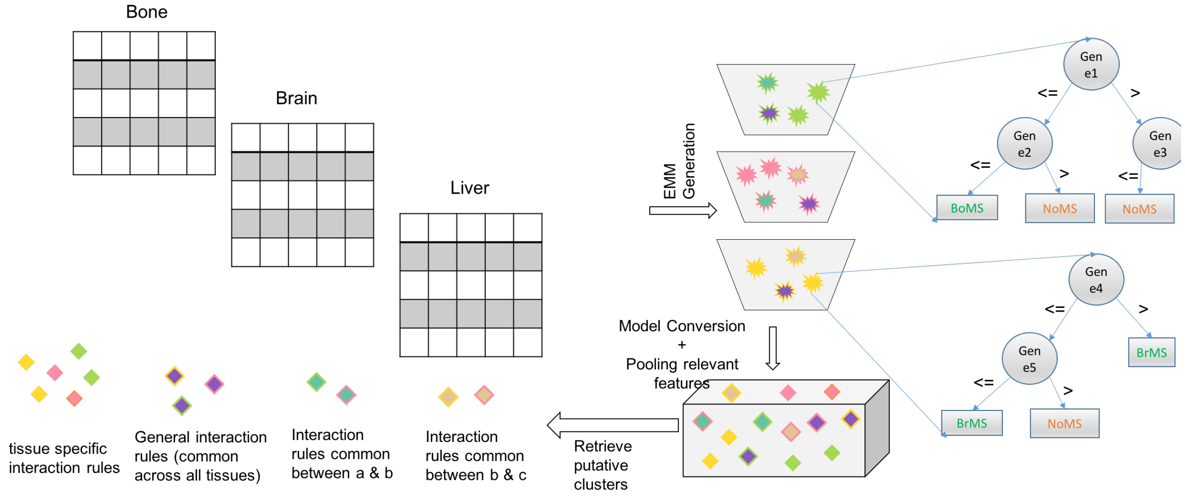

Bone
Brain
Liver
EMM Generation
Model Conversion
+
Pooling relevant features
Retrieve putative clusters
tissue specific interaction rules
General interaction rules (common across all tissues)
Interaction rules common between a & b
Interaction rules common between b & c
Gen e1
<=
>
Gen e2
Gen e3
<=
>
<=
BoMS
NoMS
NoMS
Gen e4
<=
>
Gen e5
BrMS
<=
>
BrMS
NoMS

Figure 4.1 MONTAGE pipeline.

Metastasis to non-unique destination				
noMS	BoMS	BrMS	LiMS	LuMS
211	238	49	107	101
Metastasis to unique destination				
noMS	BoMS	BrMS	LiMS	LuMS
211	138	14	28	42

Table 4.1 Number of samples with primary breast tumor (noMS), breast to bone metastasis (BoMS), breast to brain metastasis (BrMS), breast to liver metastasis (LiMS) and breast to lung metastasis (LuMS). Unique (non-unique) destination refers to the samples which has been metastasized to only one (one or more) distant organ(s).

	Foreground	Background
Unique metastasis models (model.u)	Metastasis samples of unique destination	noMS samples
Non-unique metastasis model (model.nu)	Metastasis samples of non-unique destination	noMS samples
Tissue specific unique model (e.g. BoMS or model.tu)	Secondary Tissue specific samples with unique destination	noMS samples
Tissue specific non-unique model (e.g. BoMS or model.tnu)	Secondary Tissue specific samples with non-unique destination	noMS samples
Tissue specific metastasis model (e.g. BoMS or model.tm)	Secondary Tissue specific samples with unique destination	Other MS samples

Table 4.2 Model description

EMM Feature Selection. Considering the relatively small number of samples used for each model, the total number of features is too large. Therefore, we performed feature selection in the overall data. To make the models comparable to each other, we built a universal set of features to be used for all models. As universal features, we chose genes having cross-sample expression variance more than K% of the mean variance of housekeeping genes. The housekeeping genes are collected from (http://www.stat.berkeley.edu/~johann/ruv/) and used as control as they are expected to have less variability than non-housekeeping genes. We varied K=85 and K=95 yielding 3102 and 920 features respectively.

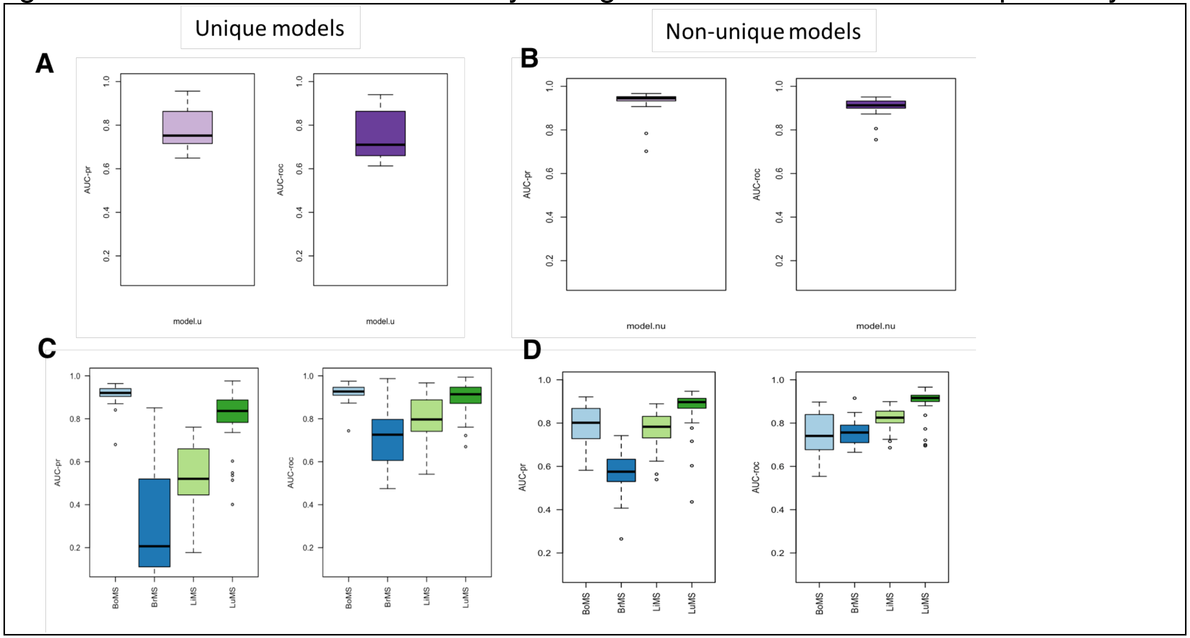


Figure 4.2 Performance of metastasis models.

(A) metastasis model using unique samples, B) metastasis model using non-unique models, C) Tissue specific metastasis models using unique samples, d) Tissue specific metastasis models using non-unique samples.

EMM performance. Model accuracy was quantified using Area Under the Receiver Operating Curve (AUC-roc) and Area Under Precision Recall Curve (AUC-pr) on test set of 4-fold cross validation (Figure 4.2). According to Figure 4.2, non-unique models have higher accuracy than unique models, most likely due to the higher number of samples used for building non-unique models. For the same reasoning, the BrMS has lower accuracy than other tissue specific unique/non-unique/metastasis models. Here we measured both AUC-roc and AUC-pr to ensure that our observed accuracies (70-90% AUC-roc) are not biased due to unequal number of foreground and background samples. More specifically, AUC-pr measure is not biased when the foreground and background sample-counts are very different. High (low) AUC-roc and with high (low) AUC-pr denotes that AUC-roc measures are not biased. Additionally, we checked that our feature selection is not causing any overfitting of the models and they are not learning any noise (Supplementary Note 1 & 2).

4.2.2 Intra- and inter-tissue heterogeneity as revealed by MONTAGE

Given the performance of *EMM*, and its architectural properties*,* we next assessed whether *EMM* can exploit the heterogeneous rules of genetic interactions, as manifested by different combinations of genes. Conceptually, a ‘genetic interaction rule’ refers to the specific combination of gene expression values leading to the metastasis state. Each decision tree (a sub-model) operationally defines a set of interaction rules, in terms of activation of specific genes above/below a certain expression threshold. Furthermore, in general, the relative importance of features decrease with increasing depth of the node in the decision tree, with the first few levels contributing a substantial portion of the decision. Although a decision tree represents a statistical model for metastasis, by applying strict thresholds for gene expression and considering only the top few layers, in principal, a concise ‘genetic interaction rule’ can be derived, albeit, at a loss of information. For a specific disease state and tissue combination, we captured the genetic interaction rules by a set of sub-models (decision trees). Then to investigate commonality and uniqueness of interaction rules for a metastasis state across tissues, we pooled all sub-models from tissue-specific *EMM*s, represented them by feature importance and clustered them using *k*-NN clustering algorithm. Next, we constructed a cluster-membership matrix mapping the number of sub-models originating from different tissues within each cluster. As an example, Figure 4.3A-B shows the cluster-membership matrix for the unique and non-unique tissue specific metastasis models. The matrices show both tissue-specific (Figure 4.3A, cluster #4) and ubiquitous (Figure 4.3B, cluster #2) clusters. Examining the cluster mapping it is apparent that, unique models tend to map to single tissue, suggesting the tissue-specific behavior, while non-unique models have more widely distributed clusters suggesting the tissue independent interaction rules and, presumably, function. Importantly, many clusters consist of sub-models from multiple, but not all, tissues.

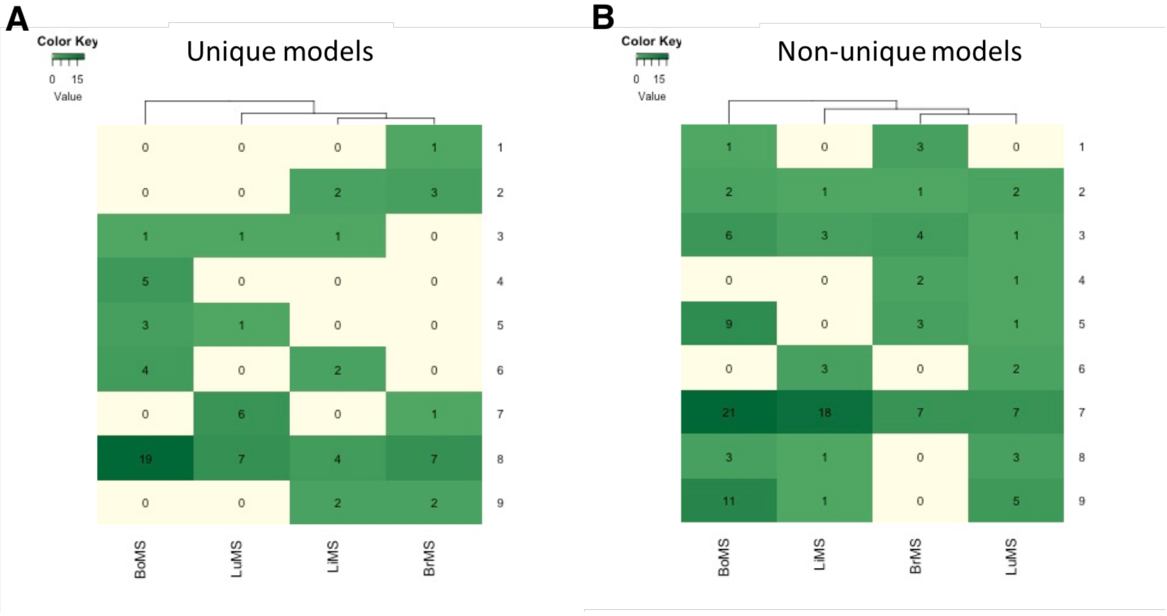


Figure 4.3 Cluster membership matrix of tissue specific metastasis models.

Given the existence of heterogeneity across metastasis models we determined the shared and unshared markers of tissue specific metastasis (Method). The markers are model-specific important genes that are unique to specific model and common across models. For example, we identified 152 genes as unique to liver metastasis, 122 genes as unique to lung metastasis and 12 gene as common to both liver and lung metastasis. Then, we assessed their enrichment in the gene sets of each pathway from KEGG database. For 1st set of genes only “Cell Adhesion Molecules (CAMs)” pathway is found as significantly enriched, for 2nd set of genes only “Proteoglycans in cancer” and for 3rd set of gene only “Metabolism of xenobiotics by cytochrome P450” pathway are found as significantly enriched. Literature review says that, EpCAM is highly expressed in breast to liver metastasis, but not lung metastasis, proteoglycan carrier is active in breast to lung metastasis but not in liver metastasis and cytochrome associated genes are involved in both breast to lung/liver metastasis.

In sum, the above clustering of sub-models support the existence of heterogeneous sets of genetic interactions rules governing the metastasis state and suggests that a subset of rules are shared across tissues.

4.2.3 Platform-associated markers confound MONTAGE

MONTAGE successfully revealed the markers for breast to liver and breast to lung metastasis, but not for other secondary organs. Our data is compiled from multiple technical platforms to quantify gene expression. Following the best practices, we corrected our data for batch effects, which in our cases primarily consists of multiple platforms. We found that even though we explicitly correct for batches, MONTAGE seems to learn features that distinguish the two platforms. Specifically, since the data has been corrected for batches across platform and according to the principal component analysis (as shown in Figure 4.6 from Supplementary Data), there should not be any bias left across batches and cohorts. However, unfortunately, we can detect batches or platform with very high accuracy (Figure 4.4A) on the corrected dataset regardless of the batch correction methods. According to Table 4.3, which shows the data size built for unique EMM, it is clearly a problem of experimental design, rather any batch correction problem. Because all the background samples are coming from only rosetta platform, the EMMs are, in principle, capturing the platform-associated markers. However, Table 4.4 shows that the data size used for building non-unique EMMs are not biased to single platform, yet the platform models have very high accuracy (Figure 4.4B). It can be argued that the separating hyperplane for platform detection and the separating hyperplane for metastasis detection are different and thus should not be affected by each other. To assess their independence, we determined the set of importance genes of both metastasis models and platform models. We then designed a model by training on platform status and then assessed the model's accuracy in distinguishing metastatic from primary samples. The high accuracy of this model indicates that the platform detection and metastasis detection are not independent (Figure 4.4C). This suggests that the Harrell et al. dataset is not suitable for metastasis signature detection and analysis of heterogeneity. We also confirmed the inefficacy and inadequacy of the current dataset, by showing that the set of important genes derived from metastasis models and platform models has 60% similarity (data not shown).

Platform	gpl96	gpl570	rosetta
NoMS	0	0	159
MS	126	95	68

Table 4.3 Number of foreground and background samples used for unique EMM models.

Platform	gpl96	gpl570	rosetta
NoMS	13	143	159
MS	126	95	68

Table 4.4 Number of foreground and background samples used for Non-unique EMM models.

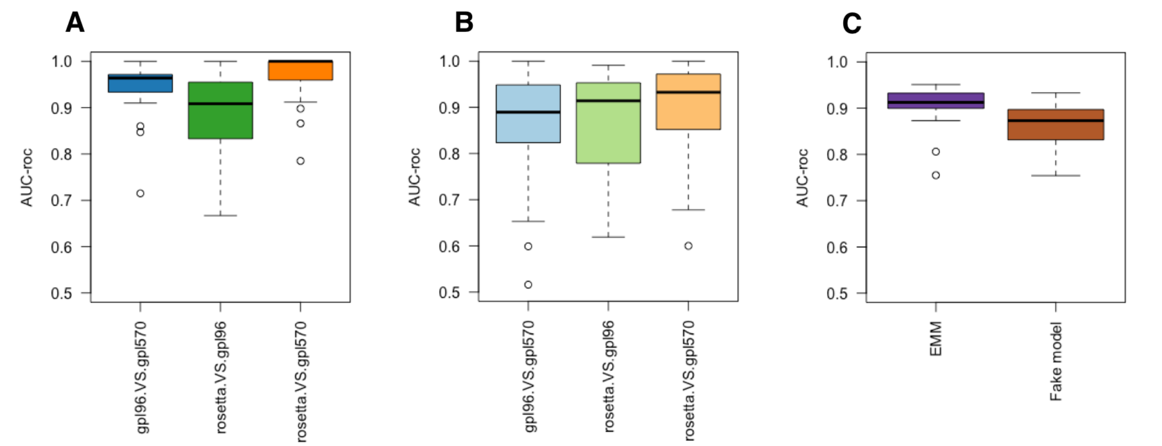


Figure 4.4 Platform detectability and performance of Fake model.

Unfortunately, in more than 50 research articles, the same dataset has been used either as primary data to test hypothesis about molecular characterization of breast cancer subtypes or as independent data to validate research findings. One of the main contributions of Harrell et al. [262] is the correlation of tissue-specific signatures with certain subtype by PAM50 genes [264], [265], a 50-geneset identified by PAM (Prediction Analysis Microarray) algorithm to determine breast cancer subtypes: Basal, Claudin, Her2, LumA and LumB. However, the differential expression of PAM50 genes across platforms casts doubt on this mode's ability to correctly determine cancer-subtypes. Moreover, the accuracy of platform models, metastasis models and fake model (labeled as pam50) are found to be very high using PAM50 genes (Figure 4.5A). One can argue that a hypothetical scenario in which a specific cancer subtype is sampled exclusively from a specific platform would give the high accuracy in Figure 4.5A. We nullified such argument by showing that metastasis prediction after training only on platform show high accuracy even if we restrict the analysis to a single cancer type (labeled by LumA in Figure 4.5A). In addition, Table 4.5 shows that the different cancer subtypes are not biased toward any platform. Collectively, these observations question the credibility of correct expression of PAM50 genes and the conclusion reached by using their values. Another article [263], used the same dataset to verify their findings of 208 Irf7 genes being involved in metastasis prognosis from mouse data. However, as shown in Figure 4.5B, even using those 208 genes we can achieve high metastasis prediction accuracy after training the model simply based on platform.

	Gpl570	Gpl96	Rosetta
Basal	28	22	38
Claudin	16	14	25
Her2	31	23	45
LumA	47	22	90
LumB	24	26	63
Normal	7	12	27

Table 4.5 Number of samples in different cancer subtypes in 3 platforms.

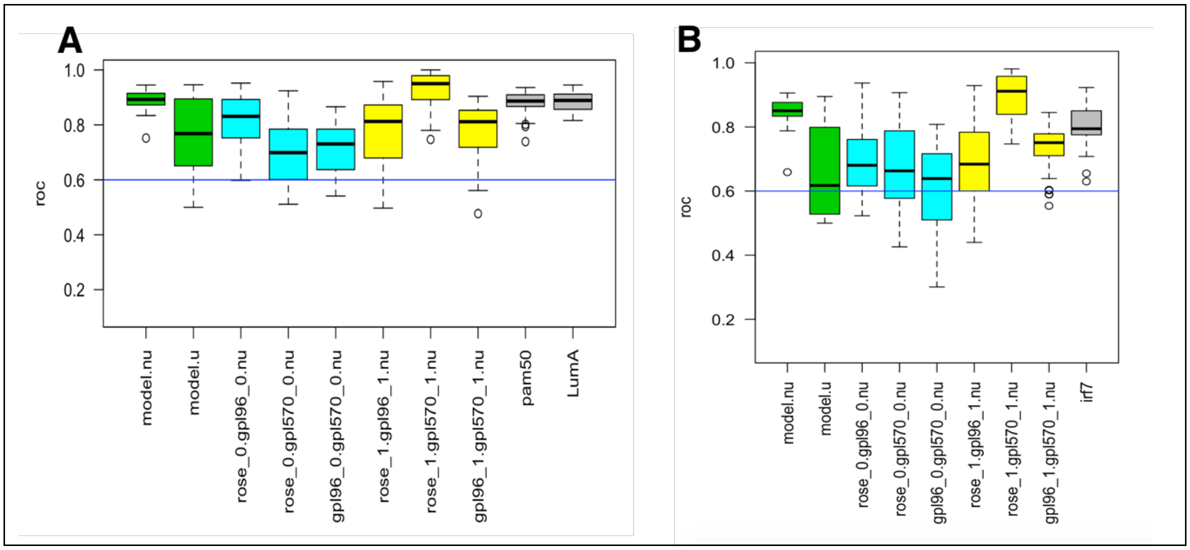


Figure 4.5 Performance of platform models and fake model A) using PAM50 genes and B) using IRF7 signaling pathway genes.

Taken together, these results indicate that the batch-corrected dataset is highly noisy and confounded such that these analyses not only yield incorrect conclusion about metastasis heterogeneity, but also casts serious doubts on the previous findings about cancer metastasis signatures.

4.3 Methods

4.3.1 Data Processing

We downloaded the gene expression data from Harrell et al. [262] where the samples were integrated from 4 cohorts and the measurements were sequenced in 3 different microarray platforms. The data set was already corrected for batches across platform using Distance Weighted Discrimination (DWD) method [47]. For sanity, we conducted principal component analysis and confirmed that the data are not biased across platforms or across cohort (Figure 4.6 from Supplementary Information).

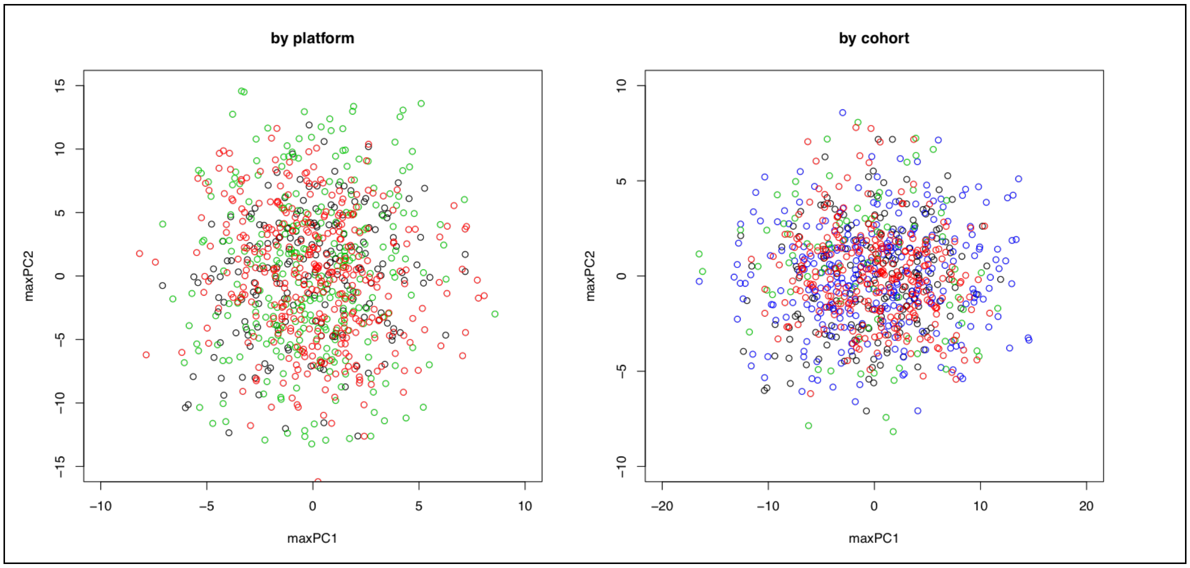


Figure 4.6 Principal components of the expression data.

Because of high number of features as compared to the number of available samples, we conducted universal feature selection. According to the universal selection, we disregarded any metastasis status of the samples and measured the variance of the expression of each feature independently across all samples. Next, we retained the features which shows variance of less than 5% of the housekeeping gene variance. The cut-off of such variance measure is shown in Figure 4.7 from Supplementary Information.

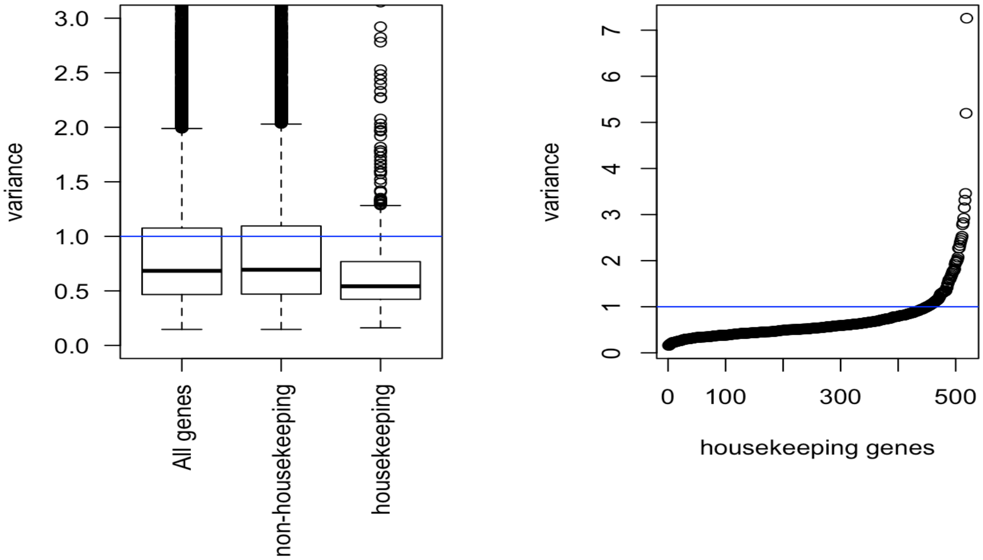


Figure 4.7 Expression variance of housekeeping and non-housekeeping genes. The blue horizontal line denotes the measurement for less than 5% of housekeeping genes.

We chose Adaptive boosting [106], [107] as our composite model where each sub-model within the ensemble is a decision tree and each decision tree is constructed based on a bootstrap sample. We used the Adaboost framework implemented in R gbm package [134]. In the framework, Huber loss function is selected to reduce over-fitting. We estimated the classification accuracy of the model based on 4-fold cross validation. As the bootstrap sample sets can be different in each run of the Adaboost model, we repeated the building of each model 50 times to account for any random bias.

4.3.2 Clustering sub-models

For each tissue specific metastasis model, we obtained the sub-models from all secondary tissue, and then clustered all sub-models using *k*-Nearest Neighbors algorithm (*k*-NN), where each sub-model is an instance and the features of the instances are individual feature-importance obtained in the context of respective secondary tissue specific model. Before feeding into the *k*-NN, we remove all the features whose cumulative importance over all sub-models is zero.

4.4 Discussion

In this study, we have presented a novel ensemble-based framework – *MONTAGE*, to investigate intra-tissue heterogeneity and inter-tissue commonality of genetic interaction rules in the context of breast cancer metastasis. To the best of our knowledge, this is the first study to comprehensively identify genetic interaction rules, each rule composed of more than 2 genes, many of which are shared across multiple tissue specific conditions. Additionally, we showed that why such study can be challenging using current data and given that challenge can also nullify some previous findings of cancer metastasis.

We chose Adaboost as our ensemble model due to its architectural advantages with respect to our ultimate goal of analyzing common and distinct binding rules, or sub-models, across ensembles learned for each cell type. Boosting ensemble methods, including Adaboost, are designed to learn optimal tree sub-models for successive reweighted bootstrap samples. This is, in contrast to other ensemble methods, including the popular Random Forest (RF) approach which seeks to increase variability of sub-models by estimating weak sub-models from un-weighted bootstrap samples. Since our primary goal is to reveal model heterogeneity, we chose to cluster sub-models generated by Adaboost rather than Random Forest’s weak learners.

5 Auto-encoder based non-linear batch correction

5.1 Background and Related works

High throughput gene expression profiling is ubiquitous in all of biomedical research. However, using gene expression profiles for such studies are not straightforward. Before effective usage of expression profile, the measurements need to be free of biases which are incurred due to many non-biological relevant sources: experiments done by different lab in different ozone level, experiments done by different personnel, experiments done at different time points, and different conditions. Collectively, these biases are called batch effects. In addition to that, the technology for sequencing the samples change with the advances in biotechnologies. It is likely that new patient samples are sequenced in different technology, and concurrent use of multiple technologies is common. All these technological differences mandate either to use the expression data separately causing smaller sample size and reduction of statistical power or to aggregate data across multiple sequencing technology after correcting for batches across platform.

Previous research addressed batch affects by variety of techniques each of which has their respective advantages and disadvantages. Singular value Decomposition (or SVD [46]) corrects for batches by directly removing the singular vectors (termed as eigengene and eigen array) which have any non-biological information along that vector. Distance Weighted Discrimination (DWD, [47]), on the other hand, removes the artifacts indirectly by projecting the expression on to mean separating hyperplane of two batches. Both SVD and DWD necessitates large number of samples. ComBat addresses the small sample size problem [48] by modeling the batch as additive and multiplicative noise for each gene independently. ComBat is often used along with SVA [49] which identifies unknown sources of noise unlike all the other previous methods. Limma [50] has also been used to incorporate batch information while finding the differentially expressed genes. Notably, neither SVA nor Limma correct for batches explicitly.

A common caveat of all previous methods is that none of them consider the interaction of multiple features in non-linear fashion, a necessity for both biological and practical reasons. Phenotypes are affected by simultaneous interactions of multiple genes. Common machine learning models built using gene expression data also incorporate non-linear interaction of multiple genes. Hence, a method employing both multi-variate non-linearity are essential for effective batch correction. This motivates our proposed deep learning based method, *deepSavior*, offering the afore-mentioned characteristics.

Recently, two deep learning based methods, ResNet and ADAGE, have been proposed for reducing noise in expression data [266], [267]. In the former method, the authors learnt the batches using residual network and the latter method utilizes auto-encoder to remove noise and find relevant group of genes while reducing noise. While *deepSavior* utilizes auto-encoder based technique, it learns the expression translation of one technical platform to another technical platform in multi-modal fashion [268]. In our study, we explicitly show why such multi-modal learning is critical. We also show the efficacy of our method in single cell protein and gene expression data.

5.2 Results

5.2.1 Exemplifying the essentiality of deepSavior

Prat et al. combined a dataset (Table 5.1) from 4 cohorts sequenced in 2 microarray platforms: gpl96 and gpl570. As the data were coming from multiple sources, it calls for batch correction while integrating the data.

Cancer subtype / Platform	gpl96	gpl570
ER+	120	125
ER-	217	190

Table 5.1 Data size of Prat et al.

Figure 5.1 shows the principal component analysis of the batch corrected data using DWD technique. According to the plots, the data are not biased by cohort or platform but retains the information of ER status of the samples which is the biological variable of interest.

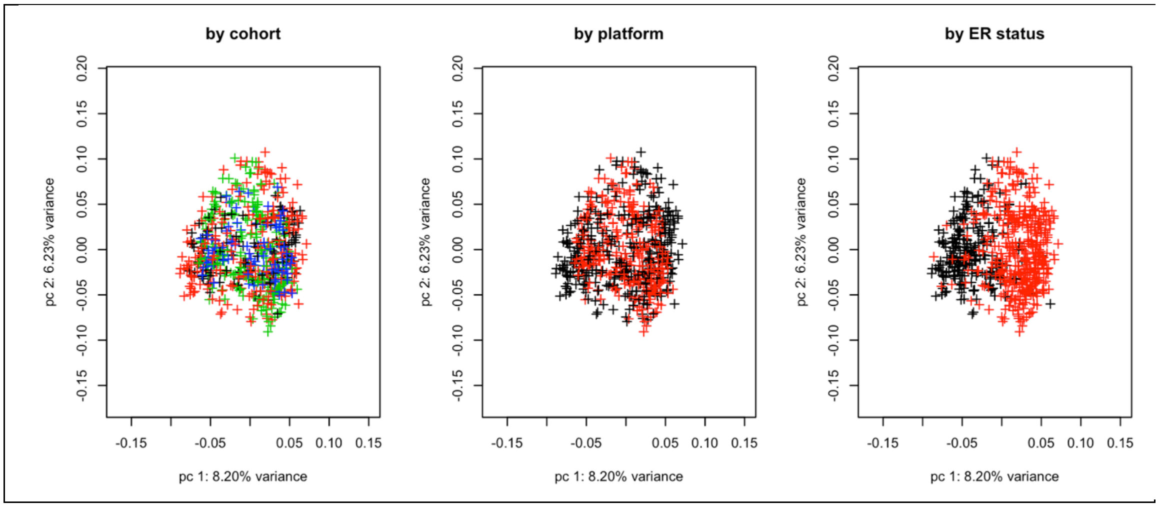


Figure 5.1 Principal components after batch correction.

Next, we selected features with high variability regardless of their ER/platform status and build machine learning models using Adaboost method on the batch corrected data. The models were built to detect ER status, both within each platform independently (ERD96 and ERD570 in Figure 5.2A) and pooling the two platforms (ERD in Figure 5.2A). The models performed with high accuracy in 4-fold cross validation fashion. We also build models on batch corrected data to detect platform, both within samples of same ER status (PD+ and PD- in Figure 5.2A) and pooling both ER status (PD in Figure 5.2A). The performance of the platform detection models is also measured in 4-fold cross validation fashion and shows high AUC-roc (Area Under Receiver Operating Curve). High accuracy in separating the two platforms is surprising given that the data was corrected for platform. Nevertheless, the argument supporting batch correction can be that the hyperplane separating the two platforms and the hyperplane separating the ER status are different and thus do not interfere with each other. We give a counter argument by building two Fake models, a) train a model to detect platform and test it for ER status detection (PERD) and b) train a model to detect ER status and test it for classifying platform (ERPD). As shown in Figure 5.2A, both Fake models show high performance indicating that the two above-mentioned hyperplanes are related. In addition, we repeated the above analysis using ~1k most variable genes as features from batch corrected data by ComBat method and using the principal components of batch corrected data by the same method. In both cases, we arrived at the same conclusion about platform detection, ER status detection and prediction by fake models (data not shown).

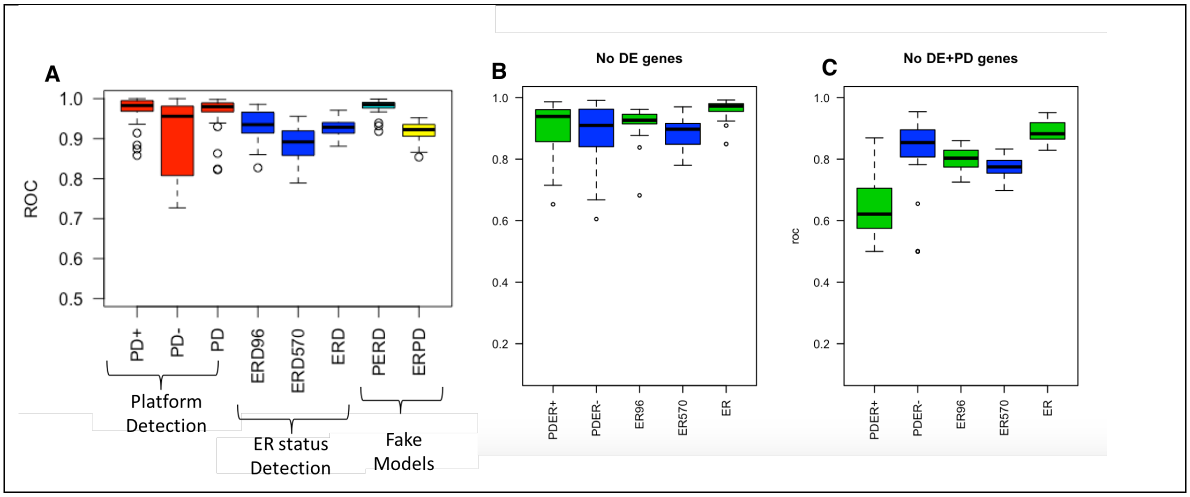


Figure 5.2 Performance of platform models, ER status models and Fake models.

Additionally, in the feature selection stage we removed the features which are found as differentially expressed (DE) between two platforms. However, as the removal of feature selection is done for each feature/gene independently, the removal of DE genes does not have any effect on the platform detection models (Figure 5.2B). On the other hand, removal of platform detectable genes from the platform model reduces the accuracy of all model (Figure 5.2C). Machine learning models, like Adaboost, Random forest, SVM with non-linear kernel (e.g. Gaussian or RBF kernel) introduces non-linear interaction across multiple features. On the other hand, any noise reduction at the batch correction stage either use linear method or use non-linear method for each feature independently. Hence none of the current batch correction method are applicable for expression data, especially when we want to reuse the corrected data for building any ML model.

deepSavior is a method which offers both multi-variate and nonlinear interaction of features. ResNet and ADAGE are also deep learning based method which have the above two properties. However, ResNet assumes the identity map between input and output while learning the batches as additional noise. This assumption applies for batches across same technology or platform, but does not apply for samples coming from different technology. For example, the probe sets from one platform can give ~14k Ensembl gene ids and the probe sets from another platform can generate ~17K Ensembl gene ids. The total common set of Ensembl ids could be ~12K and all previous methods work on taking those ~12K ids and the identity assumption does not hold here. In addition, the same set of ids coming from two different technology, the associated noise or difference in expression is far more complex than simple noise due to non-platform batches. Same reasoning applies to the efficacy of ADAGE which showed success on reducing noise on all expression data from GPL84 platform. In addition to that, while removing batches with any existing method, one must remove all the unmapped genes of different platform. The architecture of *deepSavior*, as described in next section, provides an opportunity for not leaving any information out of the dataset.

5.2.2 deepSavior – deep learning architecture to tackle biases across technical platform

deepSavior is multi-modal learning based neural network. The general network architecture of *deepSavior* is presented by Figure 5.3A and a brief description of the architecture is provided below (for additional details of activation and loss function see Methods).

As the first step, we take separately preprocessed (log-normalized and scaled) expression data from two platforms. Without loss of generality, one of them is termed as ‘left’ expression and the other one is termed as ‘right’ expression. The number of input node is the number of features in each side (left or right), i.e. they do not need to be equal. The input layer is mapped to a smaller dimensional space (Dimension Reduction Layer - DRL) and both left and right DRL have same number of nodes. Then, the left and right DRL are mapped to a shared layer. The Decode and Reconstruction layer are simply reverse transform of original mapping. Notably, reconstruction layer is going to generate output in the same range of the activation functions used for each node of the layer. Therefore, an additional Linear Transform Layer (LTL) is added so that the tail of the distributions generated by each output layer is not truncated at [-1, 1] or [0, 1]. The characteristics of input data is shown in Figure 5.3B, a set of samples with expression data from both left (CL) and right side (CR), a set of samples with only left side (UL) and another set of samples with only right side (UR).

While training the architecture with CL and CR, all 4 parameters sets ($\{w_l\}$, $\{w_r\}$, $\{w_l^{prime}\}$, $\{w_r^{prime}\}$) are updated and while training with UL/UR only the parameters from relevant sides are updated. After training, given a new sample expression from left (or right), both left and right expression data can be reconstructed. Given such reconstruction, we claim, *deepSavior* is not necessarily reducing any noise, rather it is learning how to translate expression measurements from one platform to expression measurements of a different platform. Figure 5.3C-D depicts this fundamental difference between previous batch correction methods and *deepSavior*.

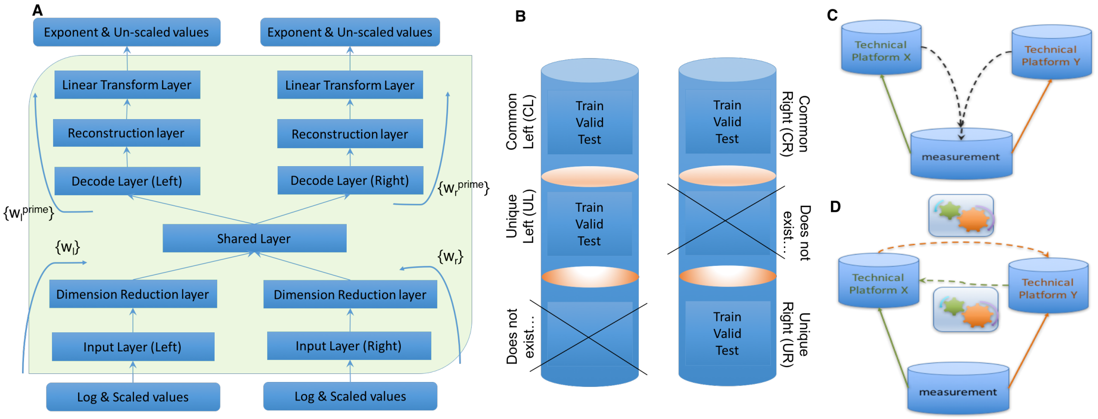


Figure 5.3 A) *deepSavior* architecture, B) General input format and C) Noise reduction by existing methods, D) Learning expression translation by *deepSavior*.

5.2.3 Efficacy of deepSavior in simulated data

We first assessed the efficacy of deepSavior on simulated data. We simulate data for both Left and Right, but to simulate UL and UR, we simply discard the data from the other side to reflect the real-world scenario (cross marked in Figure 5.3B). However, as in simulated data the counterpart of UL/UR exists, against which the reconstructed output can be verified. For simulation, we generated two sets of expression data of 10k samples, each with 25 genes. Each gene is assumed to follow a Gaussian mixture of two components (to represent on and off state) and the data was generated following a 50 by 50 correlation matrix with the assumption that each gene's expression is correlated with that of all other genes in and the left and right side. As an illustration, 4 input features of left and right are presented in Figure 5.4A.

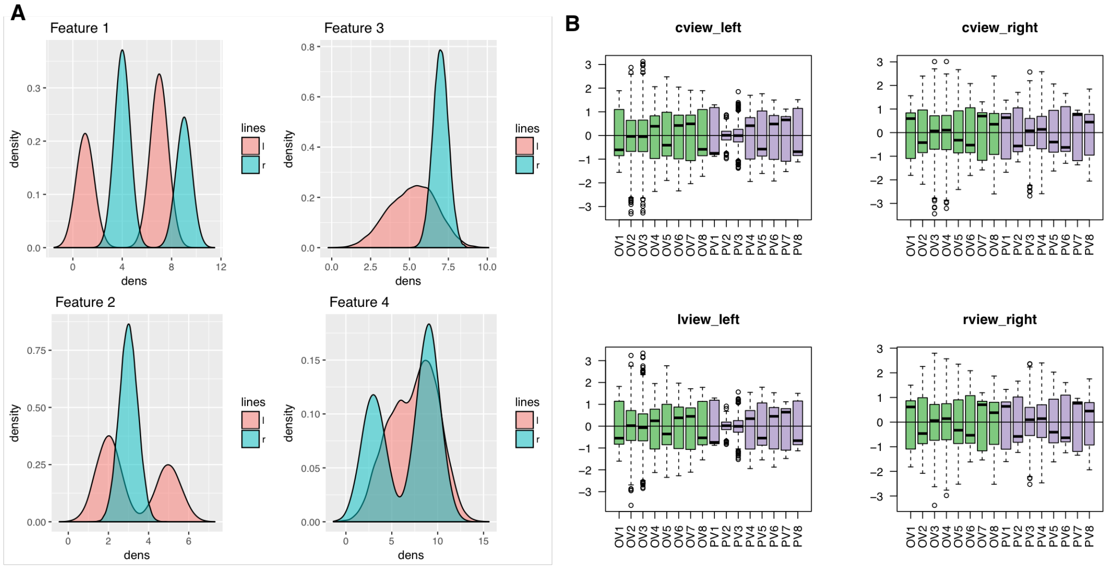


Figure 5.4 A) input feature distribution, B) Distribution of input and output from *deepSavior*. In B, “cview_left” (“cview_right”) denotes the left (right) input expression data from CL (CR) and “lview_left” (“rview_right”) denotes the left (right) input expression data from UL (UR). Notably, when the input expression is from left (right), the distribution is compared between corresponding test right (left) unseen by the model and predicted right (left).

The simulation scheme is illustrated in Figure 5.3B and training, validation and test dataset were taken from each part or CL/CR, and UL/UR. The *deepSavior* model was trained with training dataset for n number of iterations, where n was chosen based on improvement of loss function for validation set at least by \delta amount from previous iteration. After training, the model is used for prediction of the both left and right expression given only one side of data, e.g. given the left expression data we measured the predicted left and predicted right. As an illustration, we only show the distribution of predicted and actual test right for 8 features when the input is from left side and vice versa. In Figure 5.4A, the “cview_left” (“cview_right”) denotes the left (right) input expression data from common section i.e. CL (CR) and “lview_left” (“rview_right”) denotes the left (right) input expression data from unique section, i.e. UL (UR). The rationale for showing common and unique section is to show the similarity of performance measurements regardless of the input data coming from common or unique sections. The similar performance measure indicates that the performance found for common test data is not because the model has seen the data from both sides, rather it applies for both common and unique data section. Therefore, for real data we restrict the performance assessment based on common data.

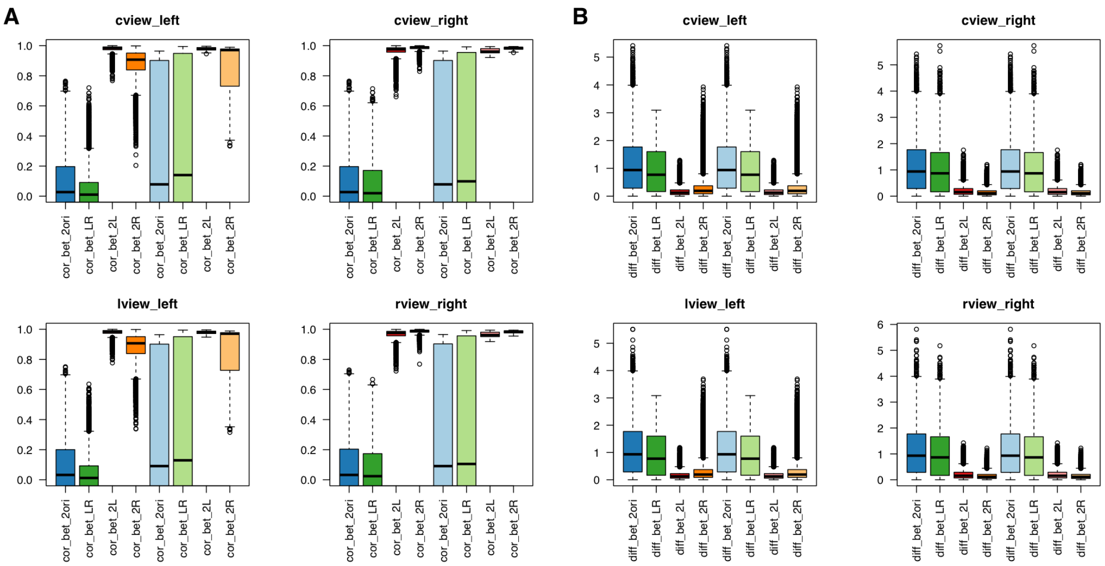


Figure 5.5 A) Correlation and B) Square loss of test data. In each plot, 1st 4 boxplots are measured along rows, i.e. across gene and 2nd 4 boxplots are measured along columns, i.e. across people. “cview_left” (“cview_right”) denotes the left (right) input expression data from CL (CR) and “lview_left” (“rview_right”) denotes the left (right) input expression data from UL (UR). “2ori” denotes the two original input expression, “LR” stands for predicted/output left and right, and “2L” (“2R”) stands for input left (right) and predicted/output left (right).

Figure 5.5 shows high correlation in predicted data and test data, similar correlation and square losses in original inputs and predicted outputs, and low square losses in predicted data and test data are apparent from Figure 5.5. As argued previously, the correlation and square loss show similar behavior both for common data section and unique data section.

In sum, *deepSavior* is capable of reconstructing expression data given only one side of data. The challenge for real data is the lack of availability of large samples which can be ignored for single cell data and can be worked around by simulating bulkSeq data.

5.2.4 Application of deepSavior in CyTOF data

CyTOF is a mass spectrometry machine to measure protein abundance. Previous method, batch correction method ResNet measured the efficacy of their model using CyTOF single cell data. In this dataset, there are 4 sets of single cell expression data with 25 features: Person1_baseline Day1 & Day2, Person1_3months Day1 & Day2, Person2_baseline Day1 & Day2 and Person2_3months Day1 & Day2. In each set, the batches are considered between Day1 & Day2. In this section, we compare *deepSavior* with ResNet to point out that in addition to platform or technological translation, *deepSavior* translates the expression in the same way across regular batches. Here, left expression is taken from Day2 and right expression is taken from Day1. For illustration, we have used only the 1st dataset: Person1_baseline Day1 & Day2.

Notably, single cell data are special in the sense that there is expression of each cell in two batches and we do not have any cell to cell correspondence between two batches. Thus, there is no direct common/shared dataset in two sides. However, in special cases, we might have information about cell-type in each batch which are measured by various markers. CyTOF have cell-type information and that's why we can utilize this information to make an artificial common/shared dataset (CL as shown in Figure 5.3B). Since cell-type information is an expensive measurement and many other single cell data might not have such information, we also try to make the artificial CL dataset without using the cell-type information directly. We refer the former method as "with cell-type" and the latter as "without cell-type".

The summary of expression datasets is presented in Table 5.2. More specifically, in "with cell-type", we matched two expressions as left and right counterpart based on same cell type and in "without cell-type", we matched two expression based on MMD (maximum mean discrepancy), membership of a cell expression according to SOM (Self-Organizing Map) cluster and without any information of cell-type (see Method for details). After training the model, we predict the translated expression using both left and right side data together because unlike platform batches, for every left expression there exists a corresponding right expression data and vice versa.

	Person1_baseline (with cell-type)	Person1_baseline (without cell-type)
Common Left/Right (CL/CR)	1315	452
Unique Left (UL)	135	1008
Unique Right (UR)	135	1008

Table 5.2 Number of cells in each data section.

We used the following criteria to compare deepSavior with ResNet: MMD, Frobenius Norm, PCA plots. MMD is similarity of two data sets or two distributions in RBF kernel space. The lower the MMD the better the correction and the expression translation. Since MMD is measure taking a subset of data points from each dataset, the MMD of same dataset, e.g. MMD(Day2, Day2) is not going to be zero, but very small number which we call as 'baseline'. According to Table 5.3, the baseline MMD measure is 0.1271 and before any correction the measure is 0.6627. After correction by ResNet, the MMD goes down to 0.2702. Notably, *deepSavior* can translate either towards left or towards right side and hence two entries of MMD measure is being shown (~0.2). *deepSavior* translates the expression data minimizing the MMD further than ResNet.

	Person1_baseline (with cell-type)	Person1_baseline (without cell-type)
(Day1, Day2)	0.6627	0.6627
(Day2, Day2)	0.1271	0.1274
(ResNet(Day1), Day2)	0.2702	0.2815
(deepSavior_left, left)	0.2119	0.3656
(deepSavior_right, right)	0.1969	0.4017

Table 5.3 MMD between two sets of expression data.

Frobenius Norm is the difference between two correlation matrices. We measured the correlation matrix of source/left and target/right. The lower the norm, the less discrepancy between two expression data and the better. We took the ratio of two norms after and before correction/translation is applied. The lower the ratio, the better the correction is. Table 5.4 presents the results for ResNet and *deepSavior* indicating that with respect to the ratio of Frobenius Norm, *deepSavior* performs either comparable or better than ResNet.

	Person1_baseline (with cell-type)	Person1_baseline (without cell-type)
ResNet	0.4217	
deepSavior (left)	0.4481	0.6397
deepSavior (right)	0.1688	0.7253

Table 5.4 Ratio of Frobenius norm between before and after method (ResNet and deepSavior).

We then checked the cumulative distribution function (CDF) of features before applying any method, after calibration by ResNet, and after translating by *deepSavior*. As an illustration, we picked 4 arbitrary features in Figure 5.6. According to the CDF, *deepSavior* performs comparable to ResNet (average KS statistics are presented in Table 5.5).

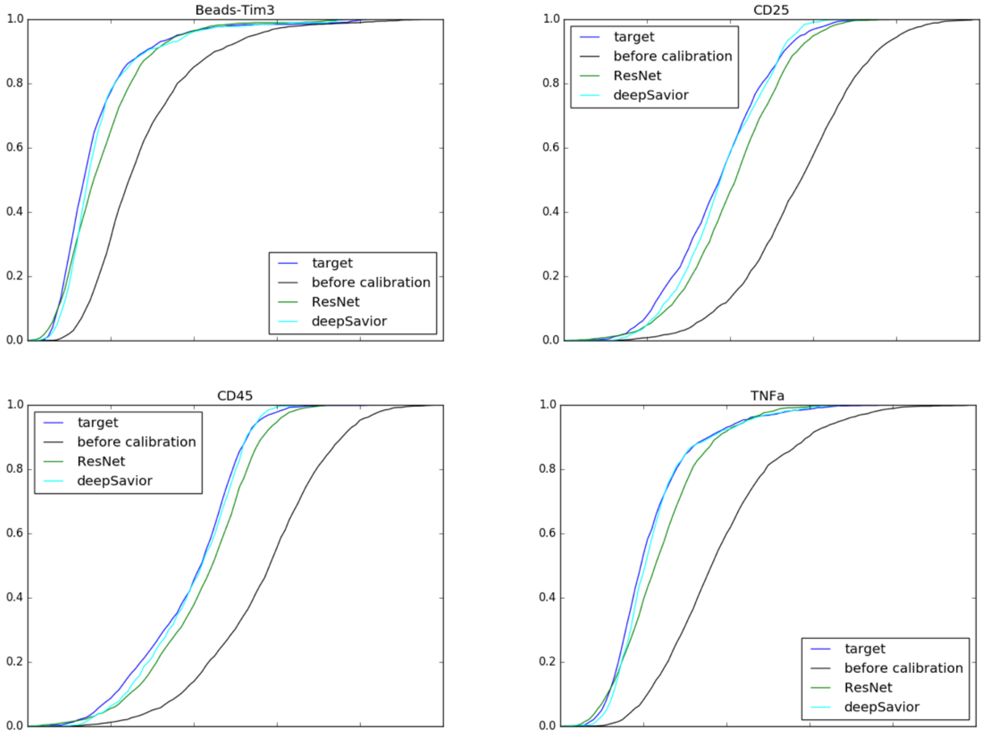
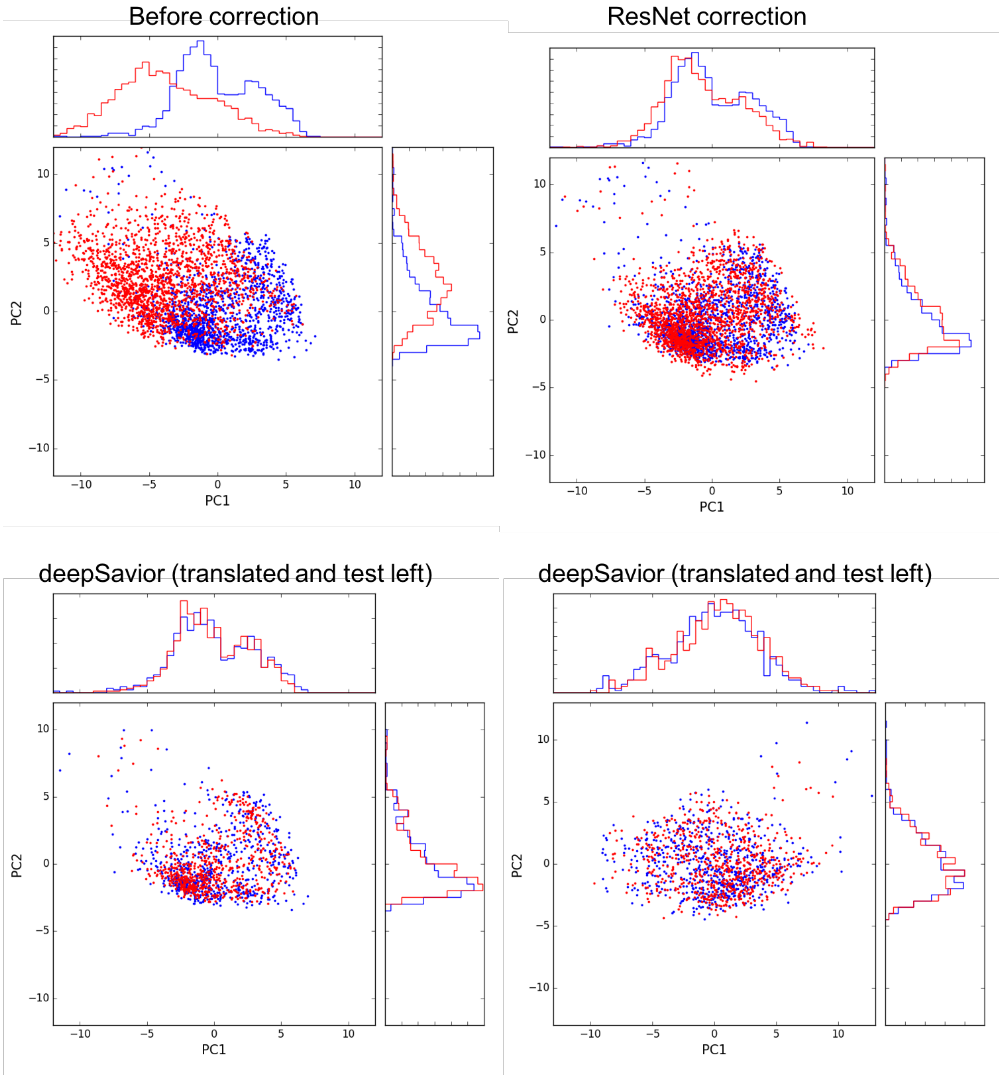


Figure 5.6 Cumulative distribution function of features.

	Person1_baseline
Before correction/translation	0.3934
ResNet	0.1068
deepSavior (with cell-type)	0.0411
deepSavior (without cell-type)	0.1458

Table 5.5 KS statistics of the CDF presented in Figure 6.

Next, we assessed the principal component analysis of the expression data both at the population (Figure 5.7) and sub-population (Figure 5.8) level. Population level denotes all expression data regardless of cell-type and sub-population level indicates the data taking from only one cell type to ensure that the correction/translation are not interfering the cell-type information.



Before correction

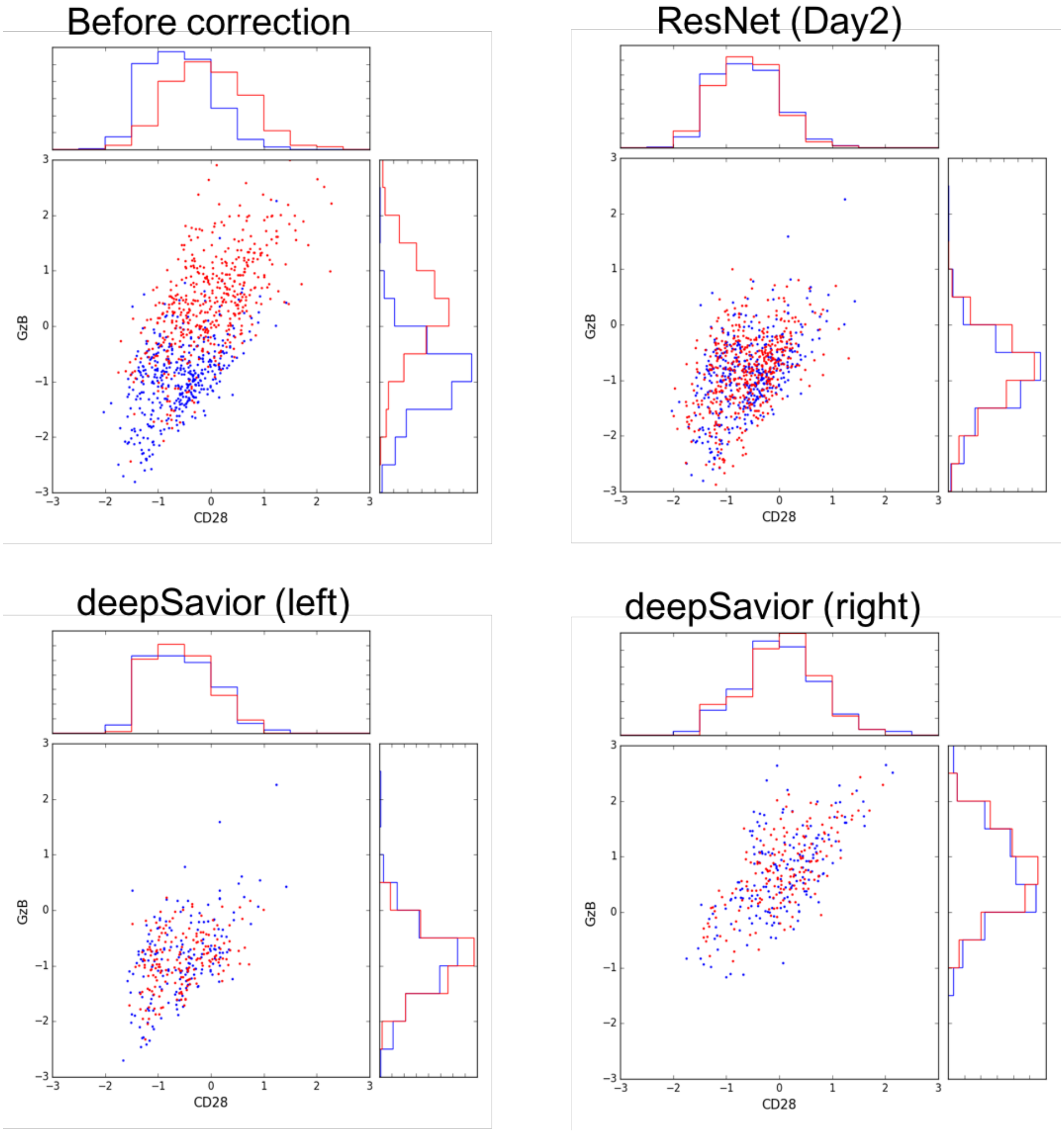
ResNet correction

deepSavior (translated and test left)

deepSavior (translated and test left)

Figure 5.7 Principal component analysis for all cell types together.

According to principal component analysis (Figure 5.7 and Figure 5.8), it is apparent that *deepSavior* performs as good as ResNet which is not surprising, as both models considers multi-variate and non-linearity of the features. However, the superiority of *deepSavior*, as illustrated above, might stem from the fact that the translation is done on the test data which the model never saw which is not the case for ResNet. Moreover, for the same platform batch the existence of both batches of data are possible, but for different platform expression data, the model needs to learn from only one batch, possibly with different number of features, which is not facilitated by ResNet.



Before correction

ResNet (Day2)

deepSavior (left)

deepSavior (right)

Figure 5.8 Principal component analysis for cell-type 1.

	With cell-type		Without cell-type	
	Batch model	Cell-type model	Batch model	Cell-type model
Before anything	0.93	0.98	0.93	0.98
ResNet Correction	0.77	0.98	0.77	0.98
deepSavior (left)	0.77	0.99	0.81	0.96
deepSavior (right)	0.75	0.98	0.79	0.89

deepSavior (outputs)	0.99	0.94	0.90	0.95

Table 5.6 AUC-roc of SVM batch and SVM cell-type models.

Finally, we measured whether we can translate the expression data across batches by preserving the cell-type information. To this end, we built batch model and cell-type model using Support Vector Machine (SVM) to predict batch and cell-type respectively. Based on Table 5.6 , both ResNet and *deepSavior* reduced the stark difference between two batches while retaining the difference between cell-types. Additionally, the last row of Table 5.5, confirms that *deepSavior* retains the batch and cell-type information in the predicted output as well.

In sum, *deepSavior* is well suited to be applicable for any batch correction done by ResNet.

5.2.5 Application of deepSavior in single cell RNASeq data

Next, we assessed the applicability of deepSavior to single-cell mRNA expression levels generated by DropSeq. We utilized the same dataset as in Shaham et al., which has two batches of seven replicates to study bipolar cells of mouse retina. We obtained the preprocessed data from Shaham et al., as according to Shekhar et al. (2016) most of the signal is captured by the leading 37 principal components. Therefore, in mouse_retina dataset we have 37 features for both left and right expression. In mouse_retina data, unlike CyTOF, the differences due to batches are very subtle (Figure 7.16 from Supplementary Information), and there is no cell-type information. As shown by Shaham et al., t-SNE plot demonstrates clusters of cells might be representative of various cell-types (Figure 5.9A and Figure 5.9B). We presented similar non-linear visualization after *deepSavior* translation (Figure 5.9C and Figure 5.9D). Based on the t-SNe plot, similar to ResNet, *deepSavior* also learns the expression translation while retaining the cell type information. Notably, the plots for *deepSavior* is sparser than ResNet as the former method is discarding some data points during training and testing.

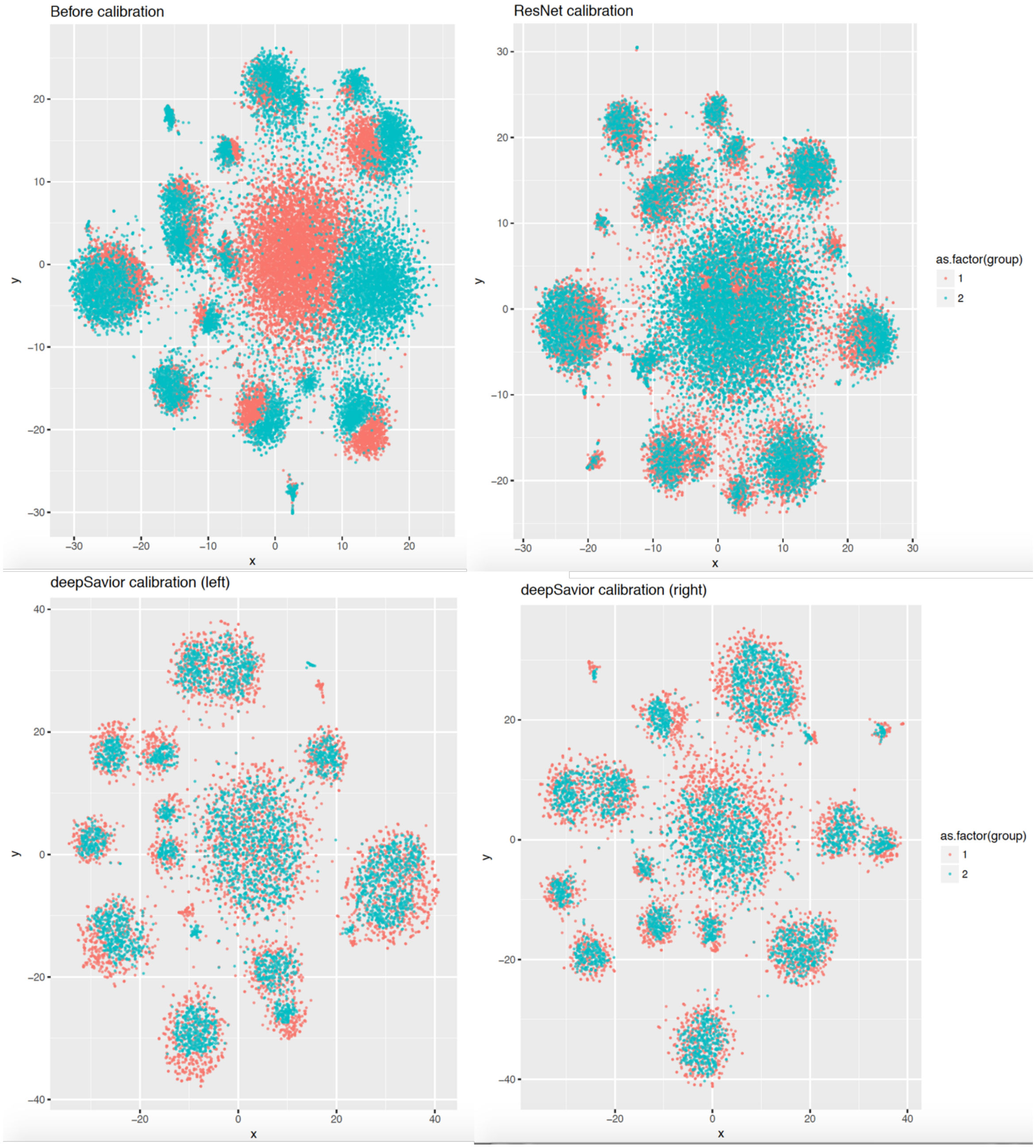


Figure 5.9 t-SNE plot for two batches of mouse_retina data.

Finally, we presented 3 features where deepSavior outperforms ResNet: the features are selected based on KS statistics measure. For most features the KS statistics between left and right (source and target) are very close to each other after correction/translation, and the measure for correction by ResNet and translation by deepSavior (mean difference 0.02). However, for the three features, as presented in Figure 5.10, the difference of two KS statistics are more than 0.2.

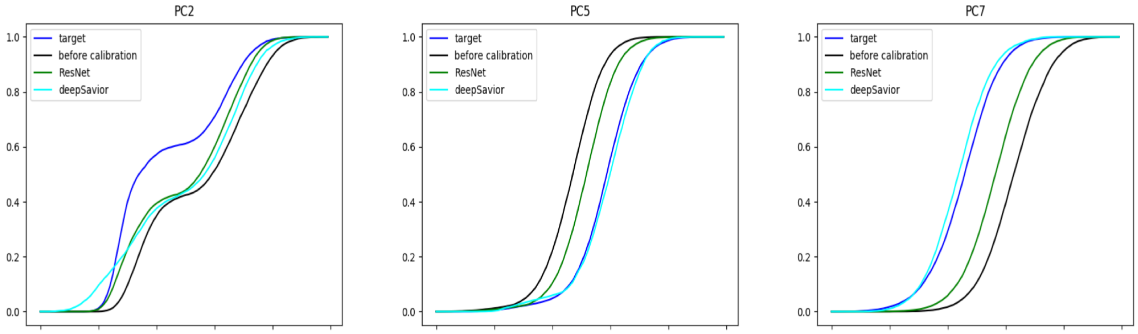


Figure 5.10 Cumulative distribution functions for principal component 2, 5 and 7. The features are selected based on having more 0.2 Ks statistics differences with ideal scenario.

In Sum, deepSavior performs well enough to compete with ResNet.

5.3 Methods

5.3.1 deepSavior Method

The activation function of each neuron in the deepSavior is a tanh function. The output range of tanh function is (-1, 1). If we keep the last layer as nodes with tanh activation, regardless of the range of input, the output will be truncated at -1 and 1. Hence, the reconstruction layer is followed by a linear transform layer to enforce the output range of tail of the distribution beyond -1 and 1.

The loss function of input and output is defined by the following.

L = L1 + L2 + L3 + L4 + L5 – C * corr(left[i], right[i]), where C is a constant and

L1 = square loss of input and output given CL and CR together

L2 (L3) = square loss of input and output given only CL (CR)

L4 (L5) = square of left (right) input and left (right) output given only UL (UR)

The above loss function has four components, minimizing self-reconstruction error, represented by L1, minimizing cross-reconstruction error from common data, represented by both L2 and L3, minimizing cross-reconstruction error from unique data, represented by both L4 and L5, and maximizing correlation encoding of left and right expression. We used Pearson correlation in the loss function and based on our experience the range between 0.1 to 5 works well as the value of C. However, the correlation between test data and predicted data does not vary much if the correlation is not imposed in the loss function (Figure 7.17).

5.3.2 Cell-matching algorithm

For single cell data, we, apriori, do not know the corresponding left and right cell expression. If we would have known the cell-type we can choose a cell either arbitrarily or based on correlation or Euclidean distance. In our algorithm of “with cell-type” version we choose two such expression arbitrarily from same cell-type.

For “without cell-type” version the assumption is we do know have the cell-type information and hence we match two such expression empirically. To this end, we measured the RBF kernel of each pair of expression from left and right. Next, we clustered the combined left-right expression data using Self-Organizing map in a grid of 7X7. The rational for choosing 7X7 grid is to make the clusters granular enough so that no two same cell types fall in the same cluster and we would not worry about if two same cell-types fall in different cluster. Given such clustering, we match two expression value which falls in same cluster and selected as one of the closest neighbors in the kernel space. If the candidat expression data from one side (e.g. right) is already taken by another expression from the other side (e.g. left), the next closest neighbor in the kernel space. The neighborhood is chosen arbitrarily as 300, i.e. beyond 300 data points, we keep the expression data as one of the unique data points (UL or UR).

5.4 Discussion

In this study, we have presented a novel deep learning based architecture to learn translation of gene expression across batches instead of correcting for batches. To the best of our knowledge, this is the first study for learning expression translation. The utility of such method can be tremendous, e.g. if we have data from n batches, and each (n-1) dataset can be translated into 1st batch then, together they can form a huge dataset. The most fundamental difference of proposed method and previous methods is we are bypassing the necessity of correction while incorporating the desired properties of multi-variate non-linear interactions. Even though there have been two other deep learning methods, namely ResNet and ADAGE, which offers such desirable properties, both require the same number of genes/features across batches. On the other hand, due to multi-modal input interface, *deepSavior* does not require having equal and same set of features between two sets of data. Therefore, when batches are due to platform differences, we do not need to throw away the information from unmapped probed between two platforms. However, the above results, so far, are expression translation across non-platform batches. For each platform of microarray or any other sequencing technology, the number of samples are very small to train neural network. As future work, we are developing an additional pipeline to generate large number of simulated data by looking at the covariance structure, correlation structure of the joint distribution and the marginal distribution of the original data. The simulated data is going to be used for training and the original data is going to be used for testing.

In addition, data simulation is going to be useful for cases where there are unequal number of left- and right-samples. Notably, is one side has much higher number of samples than the other side, the network parameters of the former side will get chance to be updated more. Therefore, this can lead to imbalance of parameter learning. To overcome this, we can throw out additional samples of larger side. However, throwing out samples will limit the learning. Hence, number of samples can be enlarged by simulating additional samples from the existing samples and keep equal number of samples on each side.

Given the success of deepSavior for translating expression across batches, it can be further utilized to translate expression in completely different setting. For example, learning translation of one tissue to another tissue and thus increasing the sample size at the population level.

6 Conclusion

In biology, variation is prevalent and it happens at multiple scale: across species, within species, and across cell types of same individual. This dissertation addressed part of the variations happening across cell-types of same individual. The two main sources for causing such variations are transcription factors (TF) and epi-genomics (e.g. methylation). Both factors cause transcriptomic variations generated from the same DNA across cell types and thus leading to cell-type specificity, or even disease. Collectively, this outcome is referred as phenotype and phenotype is function of genotype via the activity of transcription factors and epi-genetics. The presence of large scale genomics and epi-genomics data have enabled to understand such genotype-phenotype functional relationship using the art of Machine Learning (ML) and Artificial Intelligence (AI). Some representative examples of ML tasks in genomics can be prediction of epi-genetic state, prediction of TF binding, prediction of disease condition etc.

Design and interpretation of ML models can pose various challenges for effective understanding of mechanistic questions in genomics. For example, for proper mechanistic understanding of a biological process, the model built for that process should reflect the corresponding cell-type specificity. Often, even in same tissue, there exists heterogeneous groups of cells depending on their transcriptional properties. Having such heterogeneous groups of cells are very common phenomena in cancer. For effective interpretation, a good modelling should consider the presence of such heterogeneity. Finally, the dataset to build the models can be heterogeneous in nature in terms of their sources (e.g. lab) and technology (e.g. sequencing technology).

To keep the above challenges in mind, this dissertation has tried to address three big questions from genomics. The 1st question has asked for the possible determinants for loss of methylation in cancer. In cancer, loss of methylation or hypo-methylation happens in large blocks and it causes aberrant gene activity. Even though hypo-methylation is a common phenomenon, the underlying mechanism has not been investigated yet. In the 1st chapter, I designed a classification model for boundary vs. non-boundary of hypo-methylation blocks from colon cancer to reveal the associated genetic and epi-genetic markers. From our models and post-hoc analysis, we identified TF markers in the boundary which are involved in chromatin modification and the boundaries of methylation blocks behave as promoter although they are not promoter.

The 2nd chapter of the dissertation has studied the models of TF binding rules across cell-type. For decades, researchers are studying the models of TF binding to understand the functional consequence of TF binding rule. However, the relationship between TF binding rules and their functional consequences has not been not properly understood yet. This dissertation has shown that TF binding rule can exhibit significant amount of functional heterogeneity across cell-types which was previously unappreciated. In addition, in this chapter it has been discussed that such functional heterogeneity is exerted by the combinatorial effect of surrounding interaction partners which are responsible for both ubiquitous and cell-type specific regulatory functions.

The 3rd chapter of dissertation has studied the models of cancer metastasis for breast primary tissue. Metastasis is the spread of cancel cell from primary tissue to secondary tissue. Depending on the distant location, same primary tissue can end up showing multiple kinds of metastasis. This chapter has been focused on studying the heterogeneity of cancer metastasis using machine learning models.

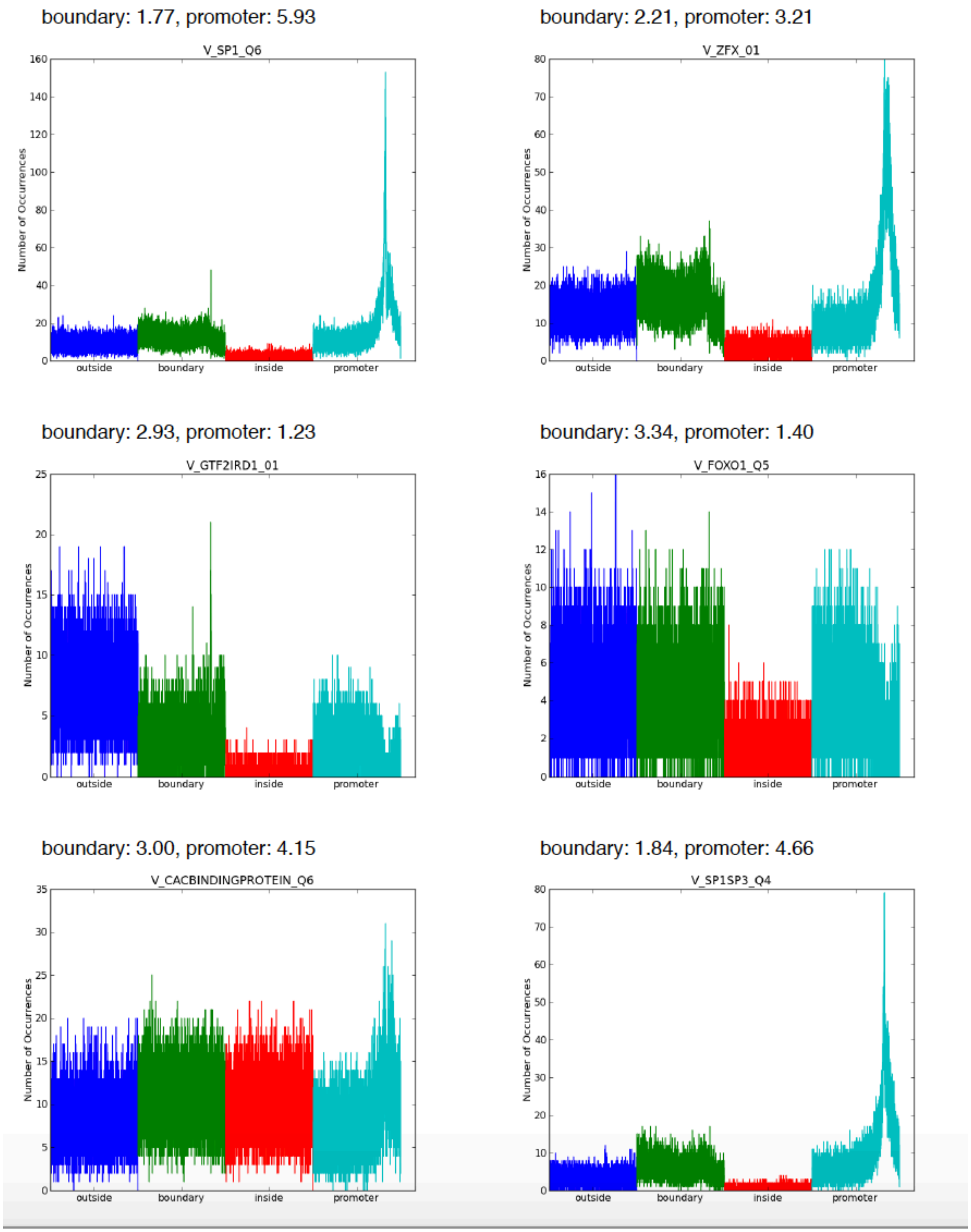
Building ML models often requires transcriptome from many patients. However, for disease, like metastasis, having patient samples from uniform sequencing technology is not possible. By necessity, patient data have been integrated across sources and technologies. However, data integration leads to noise which needs to be taken care of as it dilutes the real biological signal. This dissertation has studied the inefficacy of available noise correction methods and proposed a novel noise correction method for data integration.

To summarize, computation tools provided by Machine Learning and Artificial Intelligence offers both powerful and intelligent system design. On the other hand, nature remain to function as robust and resilient by its intelligent design. The theme followed by this dissertation can be stated as “Use the power of machine learning to demystify the wonder of biology, borrow the intelligence of machine learning to understand the intelligence of nature”.

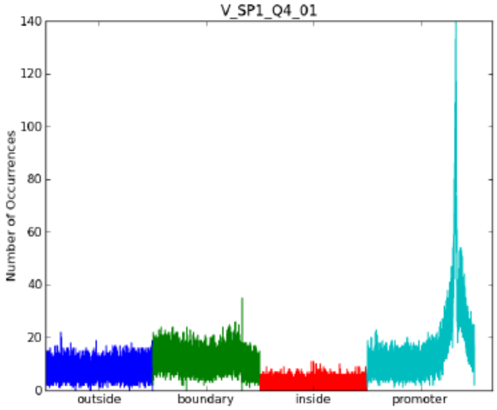
7 Supplementary Information

7.1 Supplementary for Chapter 2

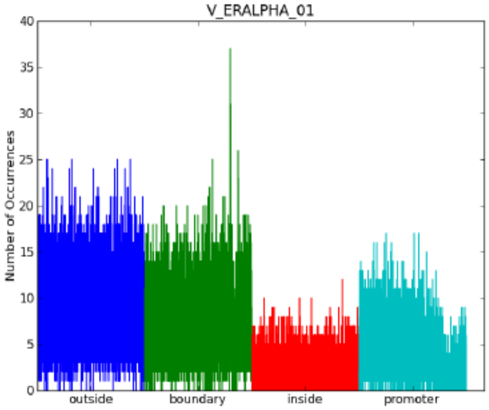
7.1.1 Supplementary Figures



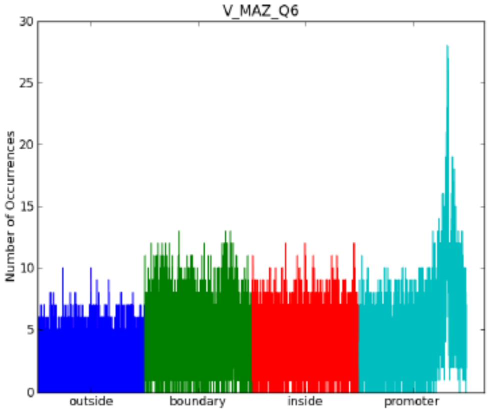
boundary: 1.75, promoter: 5.97



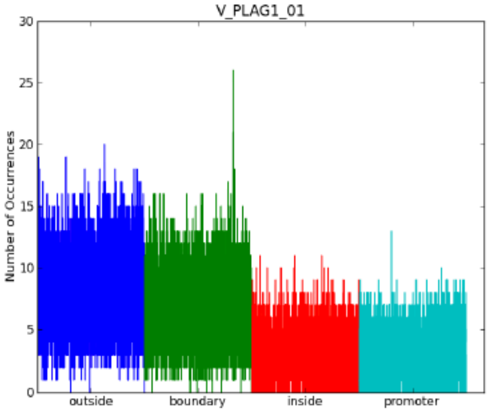
boundary: 4.35, promoter: 1.34



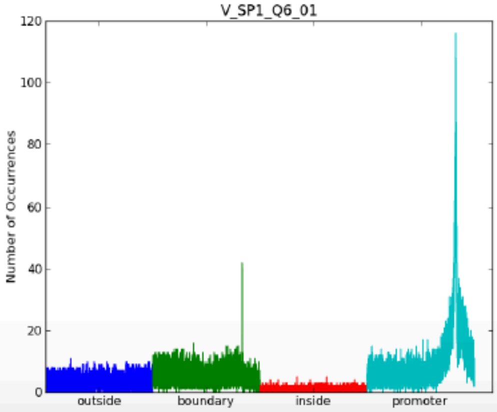
boundary: 2.45, promoter: 5.46



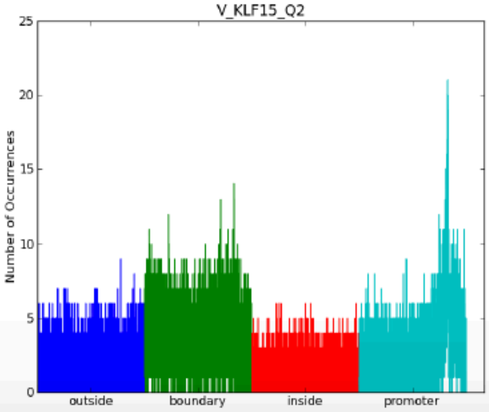
boundary: 3.55, promoter: 3.58



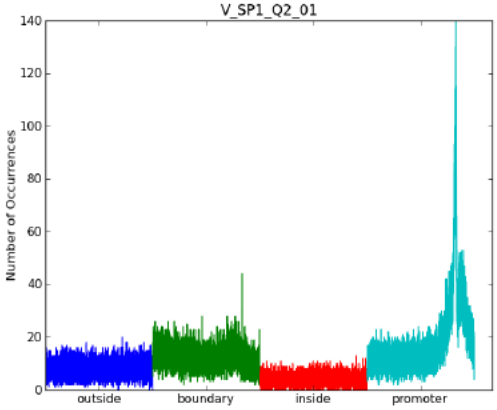
boundary: 2.09, promoter: 6.26



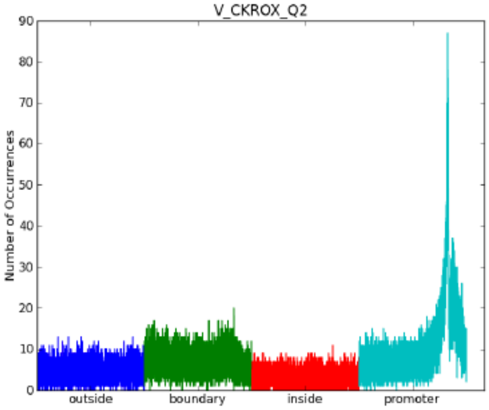
boundary: 2.48, promoter: 6.07



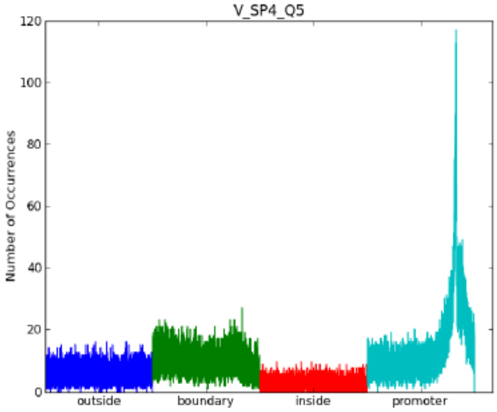
boundary: 1.75, promoter: 5.84



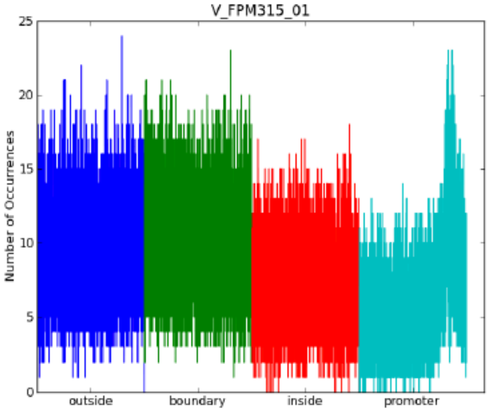
boundary: 2.20, promoter: 5.88



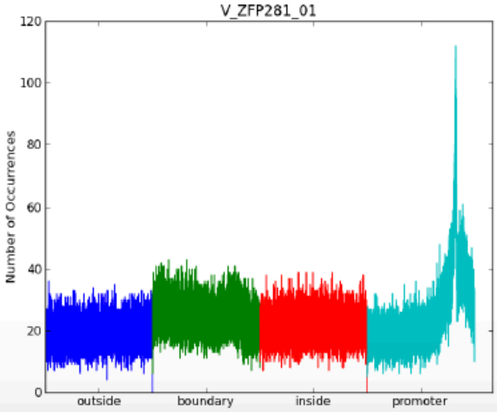
boundary: 1.63, promoter: 5.71



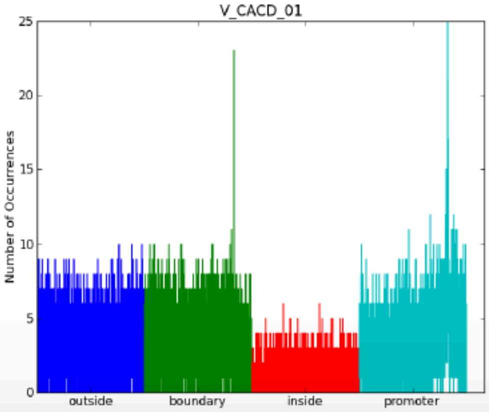
boundary: 2.76, promoter: 4.31



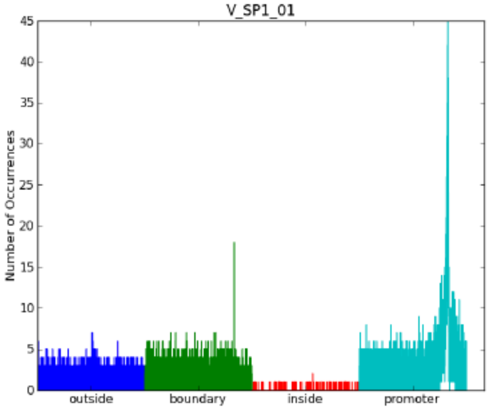
boundary: 2.44, promoter: 5.14



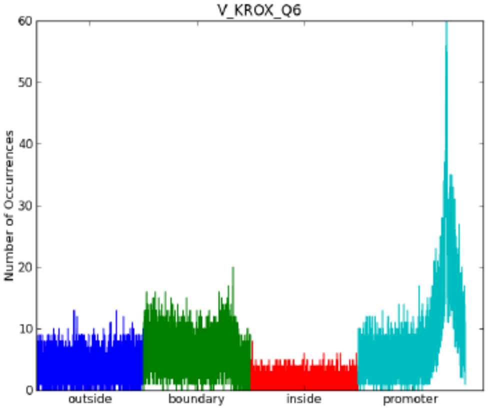
boundary: 2.27, promoter: 6.12



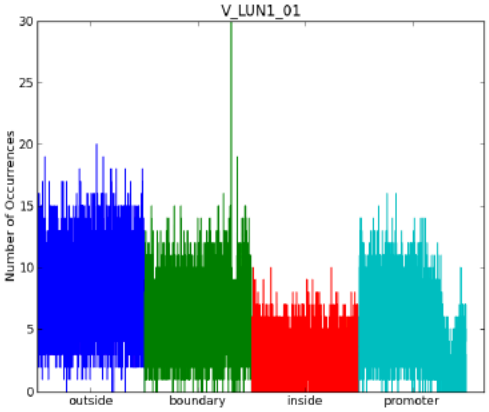
boundary: 2.56, promoter: 6.94



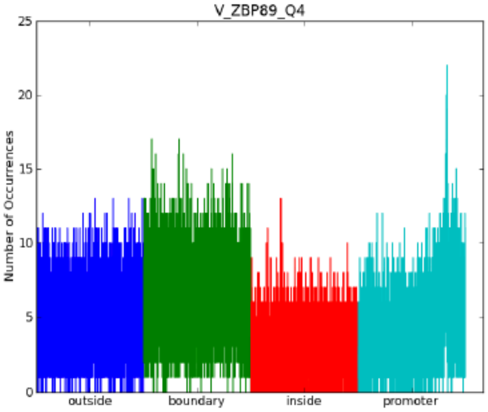
boundary: 1.96, promoter: 4.98



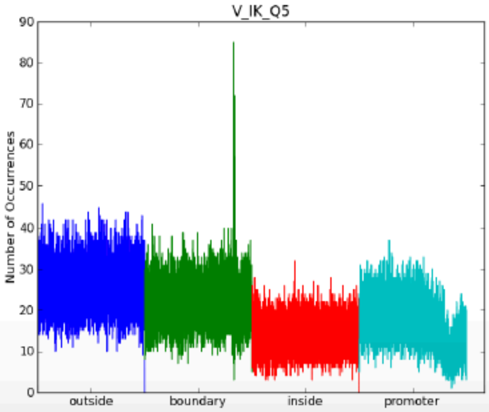
boundary: 5.04, promoter: 1.15



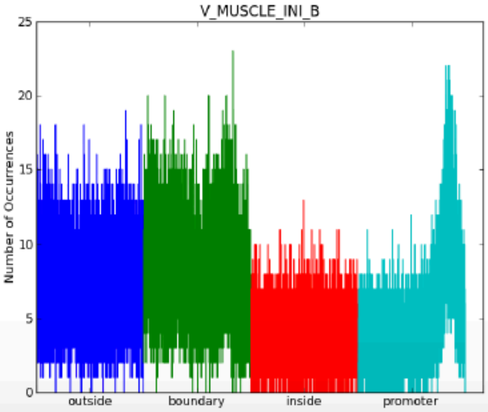
boundary: 2.12, promoter: 5.20



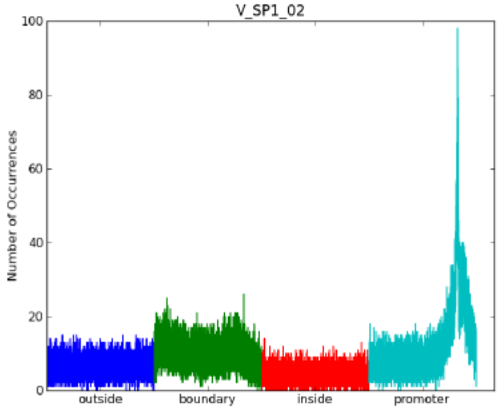
boundary: 3.14, promoter: 1.15



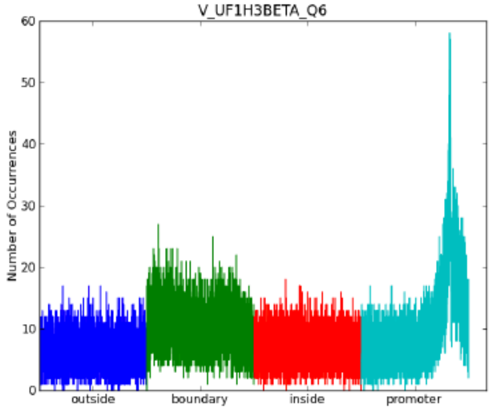
boundary: 2.03, promoter: 3.01



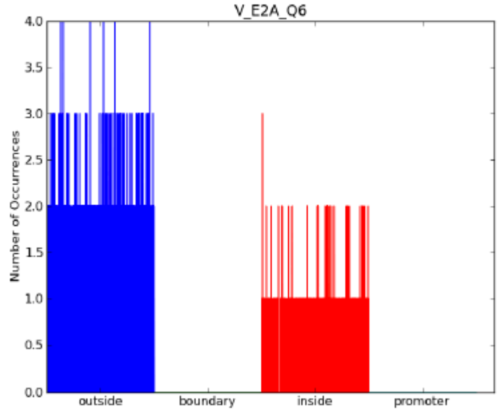
boundary: 1.87, promoter: 5.40



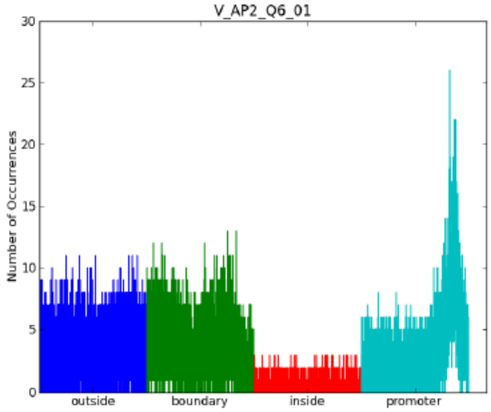
boundary: 2.01, promoter: 4.88



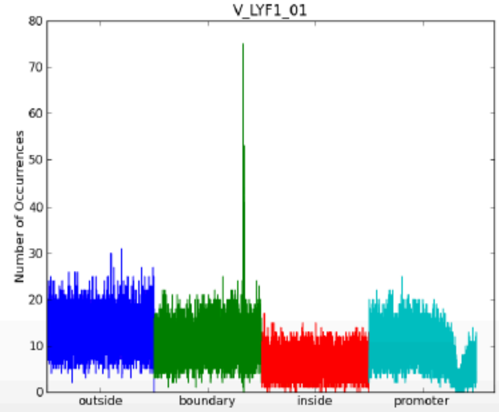
boundary: nan, promoter: nan



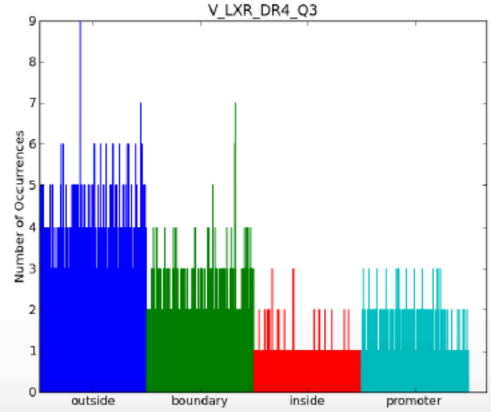
boundary: 2.34, promoter: 3.50



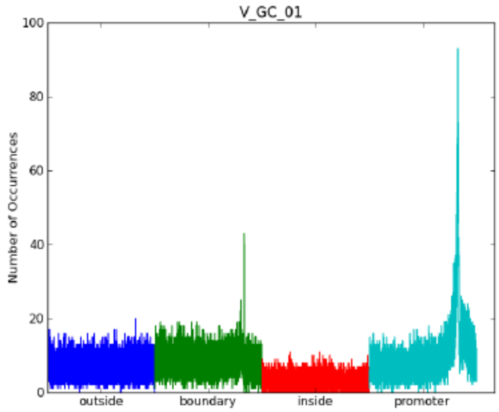
boundary: 4.84, promoter: 0.98



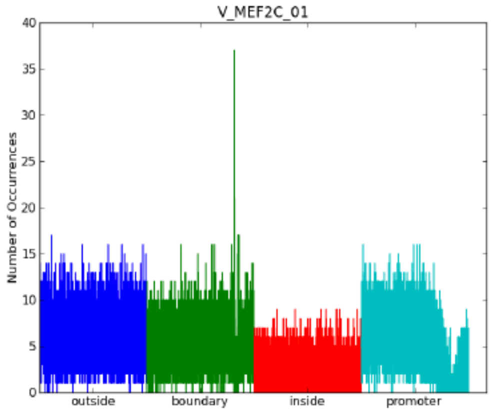
boundary: 3.23, promoter: 2.57



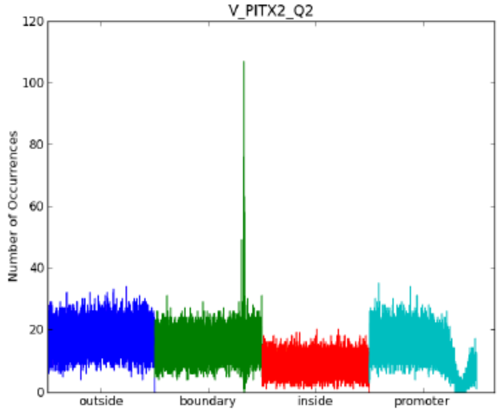
boundary: 3.12, promoter: 6.73



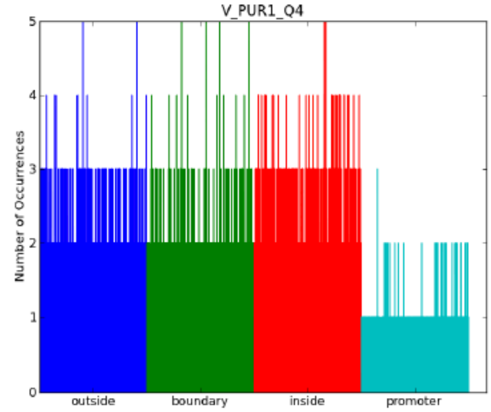
boundary: 3.38, promoter: 0.95



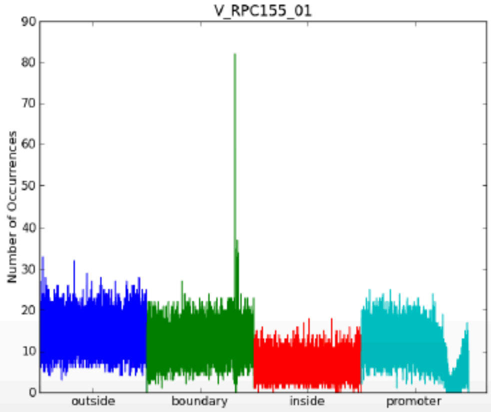
boundary: 3.35, promoter: 0.92



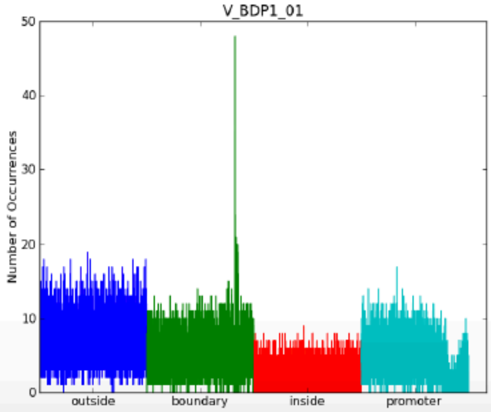
boundary: 4.35, promoter: 3.14

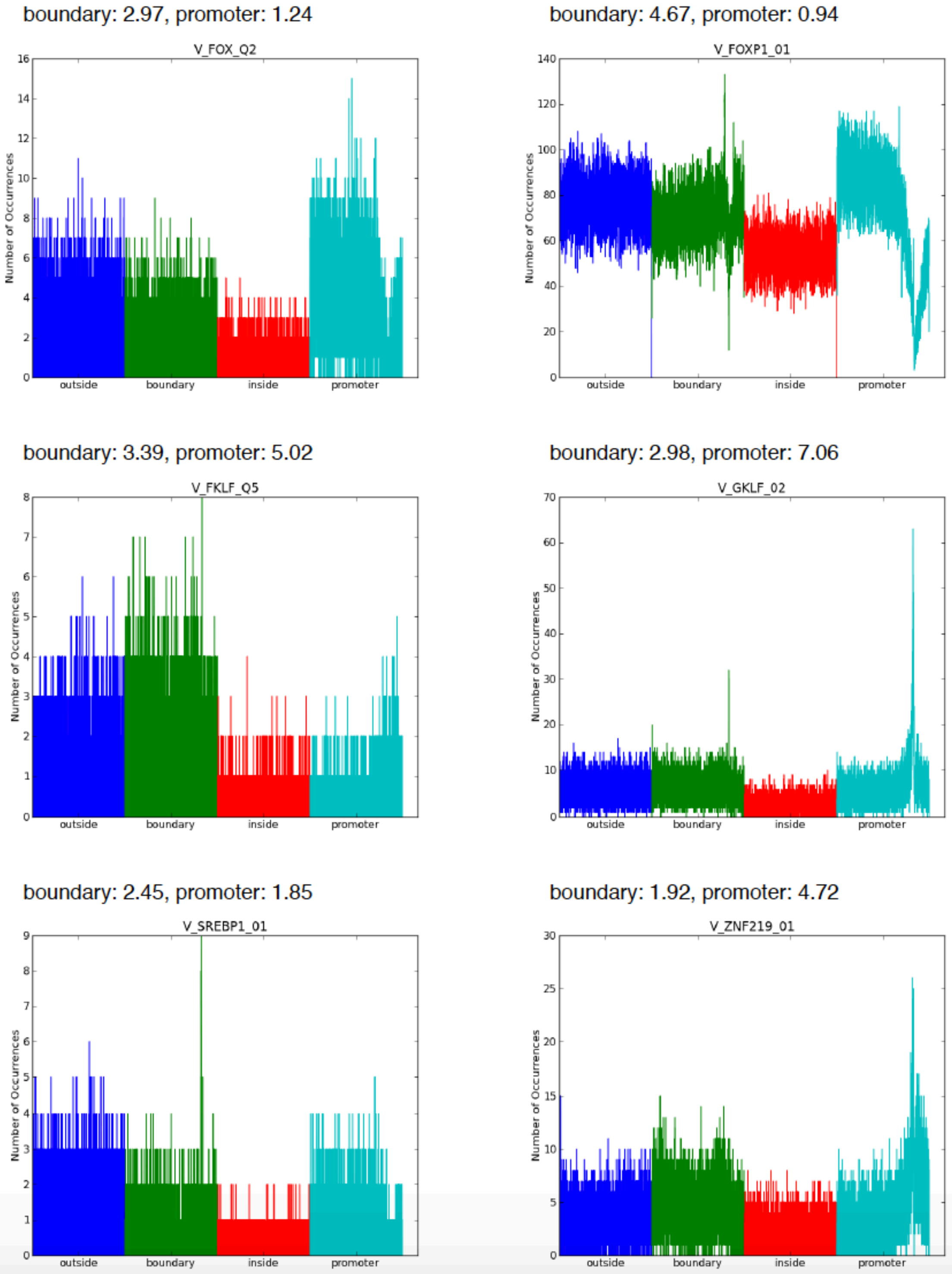


boundary: 3.48, promoter: 0.95



boundary: 4.06, promoter: 1.20




boundary: 2.97, promoter: 1.24

boundary: 4.67, promoter: 0.94

boundary: 3.39, promoter: 5.02

boundary: 2.98, promoter: 7.06

boundary: 2.45, promoter: 1.85

boundary: 1.92, promoter: 4.72

Figure 7.1 Positional profile of Frequency Plots for the TF motifs listed in Supplementary data.

7.1.2 Supplementary Data

Table 7.1 TF motifs from classification of boundary vs. promoter.

ENSG	motif	MeanDecreaseAccuracy
	V_NANOG_02	0.13648
ENSG00000129654	V_HFH4_01	0.09752
ENSG00000123405	V_MAF_Q6_01	0.06761
	V_DBX1_01	0.06727
ENSG00000168269	V_HFH3_01	0.05710
ENSG00000170608	V_HNF3_Q6_01	0.04865
ENSG00000170608	V_HNF3_Q6	0.03299
ENSG00000165556	V_CDX_Q5	0.01729
ENSG00000181690	V_PLAG1_01	0.01595
ENSG00000176678	V_FREAC7_01	0.01500
ENSG00000172845	V_SP1_Q6_01	0.01132
ENSG00000172845	V_SP1_Q4_01	0.00875
ENSG00000172845	V_SP1_Q2_01	0.00820
	V_SP1_01	0.00744
ENSG00000072310	V_SREBP_Q6	0.00668
ENSG00000006194	V_FPM315_01	0.00541
	V_TCF3_01	0.00535
	V_SP1_Q6	0.00525
	V_GTF2IRD1_01	0.00380
ENSG00000105866	V_SP4_Q5	0.00368
	V_CACD_01	0.00360
ENSG00000160685	V_CKROX_Q2	0.00327
	V_MUSCLE_INI_B	0.00320
ENSG00000072310	V_SREBP1_01	0.00307
	cpgoverlap	0.00306
	V_PITX2_Q2	0.00290
	V_CACBINDINGPROTEIN_Q6	0.00287
ENSG00000167182	V_SP2_01	0.00255
ENSG00000143190	V_OCT1_04	0.00250
ENSG00000163848	V_ZBP89_Q4	0.00250
ENSG00000148606	V_RPC155_01	0.00249
	V_ZFP281_01	0.00243
ENSG00000185811	V_IK_Q5	0.00227
ENSG00000103495	V_MAZ_Q6	0.00214
ENSG00000114861	V_FOXP1_01	0.00205
	V_SP1SP3_Q4	0.00202
	V_ZFX_01	0.00190

ENSG00000120738	V_KROX_Q6	0.00189
	V_MEF2C_01	0.00181
	V_FOXD3_01	0.00178
ENSG00000129514	V_HNF3ALPHA_Q6	0.00175
	V_SP1_02	0.00152
	V_GC_01	0.00148
	V_KLF15_Q2	0.00142
	V_UF1H3BETA_Q6	0.00135
ENSG00000091831	V_ERALPHA_01	0.00129
	V_BDP1_01	0.00123
ENSG00000071564	V_MYOD_Q6_01	0.00118
ENSG00000137203	V_AP2_Q6_01	0.00115
	V_FOXJ2_01	0.00112
ENSG00000136826	V_GKLF_02	0.00104
ENSG00000185811	V_LYF1_01	0.00104
ENSG00000111424	V_VDR_Q3	0.00101
ENSG00000150907	V_FOXO1_Q5	0.00098
ENSG00000172059	V_FKLF_Q5	0.00098
ENSG00000197579	V_LUN1_01	0.00095
ENSG00000185551	V_COUPTF_Q6	0.00088
ENSG00000126351	V_TERALPHA_Q6	0.00066
ENSG00000089225	V_TBX5_Q5	0.00062
ENSG00000165804	V_ZNF219_01	0.00056
ENSG00000130726	V_RNF96_01	0.00056
ENSG00000135363	V_LMO2COM_01	0.00054
ENSG00000124782	V_RREB1_01	0.00046
	V_MYF_01	0.00044
ENSG00000103241	V_FOX_Q2	0.00042
	V_LXR_DR4_Q3	0.00041
ENSG00000088038	V_CNOT3_01	0.00029
	V_PUR1_Q4	0.00026
ENSG00000100105	V_MAZR_01	0.00024
	V_MINI19_B	0.00023
ENSG00000185551	V_ARP1_01	0.00021
ENSG00000066336	V_PU1_Q4	0.00018
ENSG00000102974	V_CTCF_01	0.00018
	V_NCX_02	0.00017
ENSG00000184937	V_WT1_Q6	0.00017
ENSG00000099326	V_MZF1_02	0.00016
ENSG00000071564	V_E2A_Q2	0.00013
	V_LHX3_02	0.00012

ENSG00000077809	V_TFIII_Q6	0.00012
ENSG00000256683	V_ZBRK1_01	0.00011
ENSG00000162367	V_TAL1_Q6	0.00011
ENSG00000143190	V_OCT1_Q5_01	0.00009
ENSG00000137203	V_AP2_Q3	0.00009
	V_DBX2_01	0.00008
	V_P53_04	0.00008
ENSG00000111206	V_FOXM1_01	0.00007
ENSG00000143190	V_OCT1_08	0.00007
ENSG00000064835	V_PIT1_01	0.00007
ENSG00000171786	V_HEN1_02	0.00007
ENSG00000120738	V_EGR_Q6	0.00006
ENSG00000171786	V_HEN1_01	0.00006
ENSG00000156150	V_ALX3_01	0.00006
	V_MYOD_Q6	0.00005
ENSG00000162992	V_NEUROD_02	0.00004
	V_NFY_01	0.00004
ENSG00000106331	V_PAX4_04	0.00003
ENSG00000106331	V_PAX4_03	0.00003
	V_LHX5_01	0.00003
ENSG00000102974	V_CTCF_02	0.00003
	V_MINI20_B	0.00003
ENSG00000172216	V_CEBPB_02	0.00003
	V_ZFP206_01	0.00002
ENSG00000057657	V_BLIMP1_Q6	0.00002
ENSG00000084093	V_NRSF_01	0.00002
ENSG00000177374	V_HIC1_03	0.00002
ENSG00000188786	V_MTF1_Q4	0.00002
	V_REX1_03	0.00002
ENSG00000132170	V_PPARG_01	0.00002
ENSG00000196767	V_BRN4_01	0.00002
ENSG00000177374	V_HIC1_02	0.00001
ENSG00000071564	V_E47_01	0.00001
ENSG00000087510	V_AP2GAMMA_01	0.00001
ENSG00000167034	V_NKX3A_02	0.00001
ENSG00000137203	V_AP2ALPHA_01	0.00001
ENSG00000072310	V_SREBP1_Q5	0.00001
	V_IRX2_01	0.00001
ENSG00000185668	V_TST1_02	0.00001
	V_HOXC6_01	0.00001
ENSG00000007372	V_PAX6_02	0.00001

ENSG00000198914	V_OCTAMER_01	0.00001
ENSG00000084093	V_NRSE_B	0.00001
ENSG00000125347	V_IRF_Q6	0.00001
	V_DMRT3_01	0.00001
	V_FOXJ2_02	0.00001
ENSG00000116833	V_LRH1_Q5	0.00000
ENSG00000163848	V_CACCCBINDINGFACTOR_Q6	0.00000
ENSG00000162367	V_TAL1_01	0.00000
	V_POU2F3_01	0.00000
	V_HNF3B_01	0.00000
	V_OLF1_01	0.00000
ENSG00000082175	V_PR_02	0.00000
ENSG00000172845	V_SP3_Q3	0.00000
	V_TRF1_01	0.00000
ENSG00000184486	V_POU3F2_01	0.00000
ENSG00000100811	V_YY1_02	0.00000
ENSG00000068305	V_HMEF2_Q6	0.00000
	V_DMRT2_01	0.00000
	V_BARX1_01	0.00000
	V_HOXD8_01	0.00000
ENSG00000101076	V_HNF4_01_B	0.00000
ENSG00000166478	V_STAF_02	0.00000
ENSG00000185551	V_COUP_DR1_Q6	0.00000
ENSG00000128645	V_HOXD1_01	0.00000
ENSG00000245848	V_CEBP_C	0.00000
ENSG00000171634	V_FAC1_01	0.00000
	V_IRX5_01	0.00000
ENSG00000043039	V_BARX2_01	0.00000
ENSG00000159387	V_IRXB3_01	0.00000
	V_OBOX5_01	0.00000
ENSG00000113916	V_BCL6_01	0.00000
ENSG00000148200	V_GCNF_01	0.00000
	V_ISL2_01	0.00000
ENSG00000185122	V_HSF1_Q6	0.00000
ENSG00000068305	V_RSRFC4_Q2	0.00000
ENSG00000025156	V_HSF2_02	0.00000
ENSG00000135457	V_CP2_02	0.00000
ENSG00000101076	V_HNF4_Q6_01	0.00000
ENSG00000162772	V_ATF3_Q6	0.00000
ENSG00000160113	V_EAR2_Q2	0.00000
ENSG00000118513	V_CMYB_01	0.00000

ENSG00000160224	V_AIRE_01	0.00000
ENSG00000173153	V_ERR1_Q2	0.00000
	V_NKX29_01	0.00000
	V_ATF_01	0.00000
ENSG00000169083	V_AR_Q2	0.00000
ENSG00000090447	V_AP4_01	0.00000
ENSG00000157554	V_ETS_Q4	0.00000
ENSG00000147421	V_HMBOX1_01	0.00000
	V_PSX1_01	0.00000
	V_POLY_C	0.00000
ENSG00000156925	V_ZIC3_01	0.00000
	V_ETS1_B	0.00000
ENSG00000068305	V_MEF2_01	0.00000
ENSG00000064835	V_PIT1_Q6	0.00000
ENSG00000196092	V_PAX5_01	0.00000
ENSG00000169297	V_DAX1_01	0.00000
ENSG00000074047	V_GLI2_01	0.00000
ENSG00000182568	V_SATB1_Q3	0.00000
ENSG00000135100	V_HNF1_Q6	0.00000
	V_ELK1_01	0.00000
	V_SEF1_C	0.00000
ENSG00000251493	V_FREAC4_01	0.00000
ENSG00000136944	V_LMX1B_01	0.00000
ENSG00000141905	V_MYOGNF1_01	0.00000
	V_STRA13_01	0.00000
ENSG00000101076	V_HNF4_01	0.00000
	V_ZTA_Q2	0.00000
ENSG00000005102	V_MOX1_01	0.00000
	V_HOXD10_01	0.00000
	V_TBX15_01	0.00000
	V_PAX9_B	0.00000
	V_TAACC_B	0.00000
ENSG00000068305	V_RSRFC4_01	0.00000
ENSG00000149948	VS_HMGA2_01	0.00000
	V_AHRARNT_01	0.00000
	V_AHRARNT_02	0.00000
	V_AHR_01	0.00000
ENSG00000160224	V_AIRE_02	0.00000
ENSG00000068305	V_AMEF2_Q6	0.00000
ENSG00000159216	V_AML_Q6	0.00000
ENSG00000170345	V_AP1_01	0.00000

ENSG00000137203	V_AP2ALPHA_02	0.00000
ENSG00000137203	V_AP2ALPHA_03	0.00000
ENSG00000148516	V_AREB6_01	0.00000
	V_ARNT_01	0.00000
ENSG00000004848	V_ARX_01	0.00000
ENSG00000169083	V_AR_01	0.00000
ENSG00000169083	V_AR_02	0.00000
ENSG00000169083	V_AR_04	0.00000
ENSG00000169136	V_ATF5_01	0.00000
ENSG00000156273	V_BACH1_01	0.00000
	V_BACH2_01	0.00000
	V_BARBIE_01	0.00000
	V_BARHL1_01	0.00000
	V_BARHL2_01	0.00000
ENSG00000113916	V_BCL6_02	0.00000
	V_BEL1_B	0.00000
ENSG00000164458	V_BRACH_01	0.00000
ENSG00000184486	V_BRN2_01	0.00000
ENSG00000091010	V_BRN3C_01	0.00000
	V_BSX_01	0.00000
	V_CAAT_C	0.00000
	V_CART1_01	0.00000
	V_CART1_03	0.00000
	V_CBF_02	0.00000
	V_CDP_02	0.00000
	V_CDP_03	0.00000
ENSG00000113722	V_CDX1_01	0.00000
ENSG00000165556	V_CDX2_01	0.00000
ENSG00000165556	V_CDX2_Q5	0.00000
ENSG00000245848	V_CEBPA_01	0.00000
ENSG00000153879	V_CEBPGAMMA_Q6	0.00000
ENSG00000245848	V_CEBP_01	0.00000
ENSG00000245848	V_CEBP_Q2	0.00000
ENSG00000245848	V_CEBP_Q2_01	0.00000
	V_CETS1P54_03	0.00000
ENSG00000245848	V_CHOP_01	0.00000
	V_CHX10_01	0.00000
	V_CLOX_01	0.00000
ENSG00000178573	V_CMAF_01	0.00000
ENSG00000136997	V_CMYC_01	0.00000
ENSG00000136997	V_CMYC_02	0.00000

	V_COMP1_01	0.00000
ENSG00000175745	V_COUP_01	0.00000
ENSG00000115966	V_CREBP1_Q2	0.00000
ENSG00000118260	V_CREB_Q2_01	0.00000
ENSG00000141905	V_CTF1_01	0.00000
ENSG00000177030	V_DEAF1_01	0.00000
ENSG00000134107	V_DEC_Q1	0.00000
ENSG00000144355	V_DLX1_01	0.00000
ENSG00000115844	V_DLX2_01	0.00000
	V_DMRT1_01	0.00000
	V_DMRT4_01	0.00000
	V_DMRT7_01	0.00000
	V_DOBOX4_01	0.00000
	V_DOBOX5_01	0.00000
	V_DUXL_01	0.00000
ENSG00000101412	V_E2F1_Q3_01	0.00000
ENSG00000101412	V_E2F_01	0.00000
	V_E2_01	0.00000
	V_E2_Q6	0.00000
ENSG00000071564	V_E47_02	0.00000
	V_EBNA1_01	0.00000
ENSG00000132005	V_EFC_Q6	0.00000
ENSG00000122877	V_EGR2_01	0.00000
	V_ELF1_Q6	0.00000
ENSG00000135374	V_ELF5_01	0.00000
	V_ELK1_02	0.00000
ENSG00000163064	V_EN1_02	0.00000
ENSG00000164778	V_EN2_01	0.00000
ENSG00000173153	V_ERR1_Q3	0.00000
ENSG00000091831	V_ER_Q6	0.00000
	V_ESX1_01	0.00000
	V_ETS2_B	0.00000
ENSG00000085276	V_EVI1_04	0.00000
ENSG00000106038	V_EVX1_01	0.00000
	V_EVX2_01	0.00000
	V_FOXO3_01	0.00000
ENSG00000184481	V_FOXO4_02	0.00000
	V_FOXP3_Q4	0.00000
ENSG00000137273	V_FREAC2_01	0.00000
ENSG00000054598	V_FREAC3_01	0.00000
	V_FXR_IR1_Q6	0.00000

ENSG00000012504	V_FXR_Q3	0.00000
	V_GADP_01	0.00000
ENSG00000102145	V_GATA1_04	0.00000
	V_GBX1_01	0.00000
ENSG00000165702	V_GFI1B_01	0.00000
ENSG00000162676	V_GFI1_01	0.00000
ENSG00000111087	V_GLI1_01	0.00000
ENSG00000106571	V_GLI3_02	0.00000
ENSG00000111087	V_GLI_Q2	0.00000
ENSG00000113580	V_GRE_C	0.00000
ENSG00000113580	V_GR_01	0.00000
ENSG00000113580	V_GR_Q6	0.00000
ENSG00000180613	V_GSH2_01	0.00000
ENSG00000071564	V_HAND1E47_01	0.00000
	V_HB24_01	0.00000
ENSG00000130675	V_HB9_01	0.00000
	V_HDX_01	0.00000
	V_HES1_Q2	0.00000
	V_HFH1_01	0.00000
ENSG00000103241	V_HFH8_01	0.00000
ENSG00000100644	V_HIF1_Q3	0.00000
ENSG00000137309	V_HMGIY_Q3	0.00000
ENSG00000215612	V_HMX1_02	0.00000
ENSG00000108753	V_HNF1B_01	0.00000
ENSG00000135100	V_HNF1_01	0.00000
ENSG00000135100	V_HNF1_02	0.00000
ENSG00000135100	V_HNF1_C	0.00000
ENSG00000135100	V_HNF1_Q6_01	0.00000
ENSG00000101076	V_HNF4ALPHA_Q6	0.00000
ENSG00000101076	V_HNF4_DR1_Q3	0.00000
ENSG00000119547	V_HNF6_Q6	0.00000
	V_HOMEZ_01	0.00000
ENSG00000106004	V_HOX13_01	0.00000
ENSG00000253293	V_HOXA10_01	0.00000
ENSG00000105991	V_HOXA1_01	0.00000
ENSG00000197576	V_HOXA4_01	0.00000
ENSG00000106006	V_HOXA6_01	0.00000
	V_HOXA7_03	0.00000
ENSG00000078399	V_HOXA9_01	0.00000
ENSG00000159184	V_HOXB13_01	0.00000
ENSG00000120093	V_HOXB3_01	0.00000

ENSG00000182742	V_HOXB4_01	0.00000
ENSG00000120075	V_HOXB5_01	0.00000
	V_HOXB7_01	0.00000
ENSG00000120068	V_HOXB8_01	0.00000
ENSG00000170689	V_HOXB9_01	0.00000
ENSG00000123388	V_HOXC11_01	0.00000
ENSG00000198353	V_HOXC4_01	0.00000
ENSG00000172789	V_HOXC5_01	0.00000
ENSG00000037965	V_HOXC8_01	0.00000
ENSG00000180806	V_HOXC9_01	0.00000
ENSG00000100219	V_HTF_01	0.00000
ENSG00000140968	V_ICSBP_Q6	0.00000
ENSG00000185811	V_IK1_01	0.00000
ENSG00000185811	V_IK3_01	0.00000
	V_IPF1_Q4_01	0.00000
ENSG00000168310	V_IRF2_01	0.00000
	V_IRX3_02	0.00000
ENSG00000113430	V_IRX4_01	0.00000
ENSG00000116132	V_K2B_01	0.00000
ENSG00000115112	V_LBP9_01	0.00000
	V_LBX2_01	0.00000
	V_LDSPOLYA_B	0.00000
ENSG00000106689	V_LH2_01	0.00000
	V_LHX4_01	0.00000
	V_LHX8_01	0.00000
	V_LHX9_01	0.00000
	V_LIM1_01	0.00000
ENSG00000162761	V_LMX1_01	0.00000
ENSG00000025434	V_LXR_Q3	0.00000
	V_MAX_01	0.00000
	V_MAX_Q6	0.00000
ENSG00000169057	V_MECP2_01	0.00000
ENSG00000068305	V_MEF2_02	0.00000
ENSG00000068305	V_MEF2_03	0.00000
ENSG00000068305	V_MEF2_04	0.00000
ENSG00000068305	V_MEF2_05	0.00000
ENSG00000068305	V_MEF2_Q6_01	0.00000
	V_MEF3_B	0.00000
ENSG00000078399	V_MEIS1AHOXA9_01	0.00000
ENSG00000078399	V_MEIS1BHOXA9_02	0.00000
	V_MEIS1_02	0.00000

	V_MEIS2_01	0.00000
ENSG00000068305	V_MMEF2_Q6	0.00000
ENSG00000150347	V_MRF2_01	0.00000
	V_MRG2_01	0.00000
ENSG00000163132	V_MSX1_02	0.00000
ENSG00000120149	V_MSX2_01	0.00000
	V_MSX3_01	0.00000
	V_MTATA_B	0.00000
ENSG00000127989	V_MTERF_01	0.00000
ENSG00000188786	V_MTF1_01	0.00000
ENSG00000136997	V_MYCMAX_01	0.00000
ENSG00000136997	V_MYCMAX_03	0.00000
	V_NANOG_01	0.00000
ENSG00000141905	V_NF1_Q6	0.00000
ENSG00000141905	V_NF1_Q6_01	0.00000
ENSG00000109320	V_NFKB_Q6	0.00000
ENSG00000173039	V_NFKB_Q6_01	0.00000
ENSG00000120837	V_NFY_C	0.00000
ENSG00000120837	V_NFY_Q6_01	0.00000
	V_NKX11_01	0.00000
	V_NKX12_01	0.00000
	V_NKX21_01	0.00000
	V_NKX22_02	0.00000
	V_NKX23_01	0.00000
	V_NKX24_01	0.00000
ENSG00000183072	V_NKX25_03	0.00000
ENSG00000109705	V_NKX32_02	0.00000
	V_NKX52_01	0.00000
ENSG00000163623	V_NKX61_01	0.00000
ENSG00000163623	V_NKX61_02	0.00000
ENSG00000163623	V_NKX61_03	0.00000
	V_NKX63_01	0.00000
ENSG00000134323	V_NMYC_01	0.00000
ENSG00000116044	V_NRF2_Q4	0.00000
	V_OBOX1_01	0.00000
	V_OBOX2_01	0.00000
	V_OBOX5_02	0.00000
ENSG00000143190	V_OCT1_05	0.00000
ENSG00000143190	V_OCT1_06	0.00000
ENSG00000143190	V_OCT1_Q6	0.00000
ENSG00000204531	V_OCT4_01	0.00000

ENSG00000204531	V_OCT4_02	0.00000
ENSG00000184486	V_OCTAMER_02	0.00000
ENSG00000143190	V_OCT_C	0.00000
	V_OTP_01	0.00000
ENSG00000115507	V_OTX1_01	0.00000
ENSG00000165588	V_OTX2_01	0.00000
	V_OTX3_01	0.00000
ENSG00000100393	V_P300_01	0.00000
	V_P50P50_Q3	0.00000
	V_P53_05	0.00000
ENSG00000073282	V_P63_01	0.00000
ENSG00000075891	V_PAX2_01	0.00000
ENSG00000106331	V_PAX4_01	0.00000
ENSG00000106331	V_PAX4_05	0.00000
ENSG00000196092	V_PAX5_02	0.00000
ENSG00000125618	V_PAX8_01	0.00000
ENSG00000125618	V_PAX8_B	0.00000
	V_PBX1_02	0.00000
	V_PBX1_04	0.00000
ENSG00000159216	V_PEBP_Q6	0.00000
	V_PITX1_01	0.00000
	V_PITX2_01	0.00000
ENSG00000107859	V_PITX3_01	0.00000
ENSG00000165495	V_PKNOX2_01	0.00000
ENSG00000165462	V_PMX2A_01	0.00000
ENSG00000109132	V_PMX2B_01	0.00000
ENSG00000031544	V_PNR_01	0.00000
ENSG00000184271	V_POU6F1_03	0.00000
ENSG00000186951	V_PPARA_01	0.00000
ENSG00000132170	V_PPARG_02	0.00000
ENSG00000132170	V_PPARG_03	0.00000
ENSG00000160199	V_PREP1_01	0.00000
ENSG00000175325	V_PROP1_02	0.00000
ENSG00000082175	V_PR_01	0.00000
	V_PTF1BETA_Q6	0.00000
ENSG00000066336	V_PU1_01	0.00000
ENSG00000132005	V_RFX1_01	0.00000
	V_RHOX11_01	0.00000
	V_RHOX11_02	0.00000
ENSG00000102935	V_ROAZ_01	0.00000
	V_RXRLXRB_01	0.00000

	V_R_01	0.00000
	V_S8_01	0.00000
	V_S8_02	0.00000
	V_SIX1_01	0.00000
ENSG00000170577	V_SIX2_01	0.00000
	V_SIX3_01	0.00000
	V_SIX4_01	0.00000
	V_SIX6_01	0.00000
	V_SIX6_02	0.00000
ENSG00000125398	V_SOX9_B1	0.00000
ENSG00000184895	V_SOX_Q6	0.00000
ENSG00000142539	V_SPIB_01	0.00000
ENSG00000164299	V_SPZ1_01	0.00000
ENSG00000198911	V_SREBP2_Q6	0.00000
ENSG00000072310	V_SREBP_Q3	0.00000
ENSG00000112658	V_SRF_01	0.00000
ENSG00000112658	V_SRF_02	0.00000
ENSG00000112658	V_SRF_C	0.00000
ENSG00000112658	V_SRF_Q4	0.00000
ENSG00000112658	V_SRF_Q5_01	0.00000
ENSG00000112658	V_SRF_Q5_02	0.00000
ENSG00000112658	V_SRF_Q6	0.00000
ENSG00000115415	V_STAT1_01	0.00000
	V_STAT3STAT3_Q3	0.00000
ENSG00000168610	V_STAT3_01	0.00000
ENSG00000168610	V_STAT3_03	0.00000
ENSG00000138378	V_STAT4_Q4	0.00000
ENSG00000126561	V_STAT5A_01	0.00000
ENSG00000173757	V_STAT5B_01	0.00000
ENSG00000115415	V_STAT_Q6	0.00000
ENSG00000164048	V_SZF11_01	0.00000
	V_T3R_01	0.00000
ENSG00000071564	V_TAL1ALPHAE47_01	0.00000
ENSG00000071564	V_TAL1BETAE47_01	0.00000
ENSG00000162367	V_TAL1BETAITF2_01	0.00000
ENSG00000118260	V_TAXCREB_02	0.00000
ENSG00000112837	V_TBX18_01	0.00000
ENSG00000122145	V_TBX22_01	0.00000
ENSG00000082641	V_TCF11MAFG_01	0.00000
ENSG00000118707	V_TGIF2_01	0.00000
ENSG00000177426	V_TGIF_02	0.00000

	V_UNCX4.1_01	0.00000
ENSG00000158773	V_USF_01	0.00000
	V_VAX1_01	0.00000
	V_VAX2_01	0.00000
	V_VDRRXR_01	0.00000
	V_VJUN_01	0.00000
	V_VMAF_01	0.00000
ENSG00000100987	V_VSX1_01	0.00000
	V_XFD1_01	0.00000
	V_XFD2_01	0.00000
	V_XFD3_01	0.00000
ENSG00000100811	V_YY1_01	0.00000
ENSG00000100811	V_YY1_Q6_02	0.00000
ENSG00000198081	V_ZF5_B	0.00000
	V_ZID_01	0.00000
ENSG00000186350	V_PPARA_02	0.00000
	V_IRX3_01	0.00000
ENSG00000185551	V_DR1_Q3	0.00000
ENSG00000177030	V_DEAF1_02	0.00000
ENSG00000119715	V_ERR2_01	0.00000
ENSG00000186951	V_PPAR_DR1_Q2	0.00000
ENSG00000166478	V_STAF_01	0.00000
ENSG00000084093	V_NRSF_Q4	0.00000
ENSG00000143190	V_OCT1_02	0.00000
	V_P53_03	0.00000
ENSG00000169083	V_AR_03	0.00000
	V_NKX26_01	0.00000
ENSG00000188620	V_HMX3_02	0.00000
ENSG00000170365	V_SMAD1_01	0.00000
	V_P53_01	0.00000
	V_ISRE_01	0.00000
ENSG00000132005	V_RFX1_02	0.00000
ENSG00000185551	V_DR4_Q2	0.00000
ENSG00000085276	V_EVI1_01	0.00000
ENSG00000143190	V_OCT1_01	0.00000
	V_DMRT5_01	0.00000
ENSG00000185122	V_HSF_Q6	0.00000
ENSG00000111424	V_DR3_Q4	0.00000
ENSG00000028277	V_OCT2_01	0.00000
	V_TR4_03	0.00000
ENSG00000141646	V_SMAD4_Q6	0.00000

ENSG00000170370	V_EMX2_01	0.00000
	V_IPF1_06	0.00000
ENSG00000102145	V_GATA1_02	0.00000
	V_ARNT_02	0.00000
ENSG00000109381	V_NERF_Q2	0.00000
ENSG00000178573	V_MAF_Q6	0.00000
ENSG00000148516	V_AREB6_03	0.00000
	V_HOXD13_01	0.00000
	V_MIF1_01	0.00000
ENSG00000115415	V_STAT1_05	0.00000
ENSG00000137203	V_AP2_Q6	0.00000
	V_TBX15_02	0.00000
ENSG00000109906	V_PLZF_02	0.00000
ENSG00000185024	V_BRF1_01	0.00000
ENSG00000126561	V_STAT5A_02	0.00000
ENSG00000181449	V_SOX2_Q6	-0.00001
ENSG00000007372	V_PAX6_Q2	-0.00001
ENSG00000084093	V_REST_01	-0.00001

Table 7.2 TF motifs from classification of boundary vs. inside.

ENSG	motif	MeanDecreaseAccuracy
	V_SP1_Q6	0.08087
	V_ZFX_01	0.07906
ENSG00000150907	V_FOXO1_Q5	0.07825
ENSG00000091831	V_ERALPHA_01	0.05007
	V_SP1SP3_Q4	0.04075
	V_CACBINDINGPROTEIN_Q6	0.03286
ENSG00000172845	V_SP1_Q4_01	0.02955
ENSG00000172845	V_SP1_Q6_01	0.02199
	V_GTF2IRD1_01	0.02066
ENSG00000103495	V_MAZ_Q6	0.01563
	V_ZFP281_01	0.01261
ENSG00000172845	V_SP1_Q2_01	0.01147
ENSG00000006194	V_FPM315_01	0.01062
ENSG00000105866	V_SP4_Q5	0.01003
ENSG00000181690	V_PLAG1_01	0.00968
ENSG00000120738	V_KROX_Q6	0.00960
ENSG00000103241	V_FOX_Q2	0.00916
	V_PUR1_Q4	0.00833
	V_SP1_01	0.00823
	V_MUSCLE_INI_B	0.00806

	V_KLF15_Q2	0.00780
ENSG00000185811	V_LYF1_01	0.00721
	V_MEF2C_01	0.00713
ENSG00000197579	V_LUN1_01	0.00680
	V_GC_01	0.00636
ENSG00000137203	V_AP2_Q6_01	0.00608
ENSG00000160685	V_CKROX_Q2	0.00529
	V_PITX2_Q2	0.00510
ENSG00000185811	V_IK_Q5	0.00499
ENSG00000148606	V_RPC155_01	0.00468
	V_SP1_02	0.00435
	V_UF1H3BETA_Q6	0.00433
ENSG00000114861	V_FOXP1_01	0.00411
ENSG00000163848	V_ZBP89_Q4	0.00397
ENSG00000066336	V_PU1_Q4	0.00379
ENSG00000136826	V_GKLF_02	0.00366
	V_CACD_01	0.00344
ENSG00000135363	V_LMO2COM_01	0.00326
	V_BDP1_01	0.00319
ENSG00000185551	V_ARP1_01	0.00315
	cpgoverlap	0.00309
ENSG00000165804	V_ZNF219_01	0.00260
	V_FOXJ2_01	0.00253
ENSG00000162367	V_TAL1_01	0.00240
	V_FOXD3_01	0.00237
ENSG00000123405	V_MAF_Q6_01	0.00214
ENSG00000111424	V_VDR_Q3	0.00193
ENSG00000077809	V_TFIII_Q6	0.00185
ENSG00000185551	V_COUPTF_Q6	0.00165
ENSG00000126351	V_TERALPHA_Q6	0.00149
ENSG00000124782	V_RREB1_01	0.00145
ENSG00000162992	V_NEUROD_02	0.00135
ENSG00000129514	V_HNF3ALPHA_Q6	0.00128
ENSG00000137203	V_AP2_Q3	0.00114
ENSG00000071564	V_MYOD_Q6_01	0.00105
ENSG00000089225	V_TBX5_Q5	0.00099
ENSG00000177374	V_HIC1_02	0.00088
ENSG00000071564	V_E2A_Q2	0.00087
ENSG00000106571	V_GLI3_Q5_01	0.00074
	V_MINI19_B	0.00069
ENSG00000102974	V_CTCF_01	0.00066

ENSG00000143190	V_OCT1_Q5_01	0.00063
ENSG00000071564	V_E2A_Q6	0.00060
ENSG00000171786	V_HEN1_01	0.00055
	V_LXR_DR4_Q3	0.00054
ENSG00000072310	V_SREBP_Q6	0.00054
ENSG00000171786	V_HEN1_02	0.00046
ENSG00000099326	V_MZF1_02	0.00046
ENSG00000072310	V_SREBP1_01	0.00039
ENSG00000184937	V_WT1_Q6	0.00036
ENSG00000256683	V_ZBRK1_01	0.00035
ENSG00000106331	V_PAX4_03	0.00034
ENSG00000102974	V_CTCF_02	0.00030
ENSG00000143190	V_OCT_Q6	0.00023
	V_MYF_01	0.00022
ENSG00000172059	V_FKLF_Q5	0.00018
ENSG00000198911	V_SREBP2_Q6	0.00013
ENSG00000084093	V_REST_01	0.00010
ENSG00000100811	V_YY1_02	0.00009
ENSG00000172216	V_CEBPB_02	0.00009
ENSG00000106571	V_GLI3_01	0.00009
	V_P53_04	0.00009
ENSG00000100105	V_MAZR_01	0.00007
	V_NCX_02	0.00006
ENSG00000163848	V_CACCCBINDINGFACTOR_Q6	0.00006
ENSG00000119715	V_ERR2_01	0.00006
ENSG00000172845	V_SP3_Q3	0.00006
ENSG00000106331	V_PAX4_04	0.00006
ENSG00000072310	V_SREBP1_Q5	0.00006
	V_MINI20_B	0.00005
ENSG00000196767	V_BRN4_01	0.00005
ENSG00000188786	V_MTF1_Q4	0.00005
ENSG00000120738	V_EGR_Q6	0.00005
ENSG00000166478	V_STAF_02	0.00005
ENSG00000057657	V_BLIMP1_Q6	0.00004
ENSG00000185668	V_TST1_02	0.00004
	V_ETS1_B	0.00004
ENSG00000100393	V_P300_01	0.00003
ENSG00000084093	V_NRSF_01	0.00003
ENSG00000130726	V_RNF96_01	0.00003
	V_XFD3_01	0.00003
ENSG00000141646	V_SMAD4_Q6	0.00002

ENSG00000064835	V_PIT1_01	0.00002
	V_DBX1_01	0.00002
ENSG00000007372	V_PAX6_Q2	0.00002
ENSG00000084093	V_NRSF_Q4	0.00002
ENSG00000143190	V_OCT1_01	0.00001
ENSG00000071564	V_E47_01	0.00001
ENSG00000109381	V_NERF_Q2	0.00001
ENSG00000115415	V_STAT1_05	0.00001
ENSG00000185551	V_DR4_Q2	0.00001
ENSG00000185024	V_BRF1_01	0.00001
	V_IRX2_01	0.00001
ENSG00000170608	V_HNF3_Q6_01	0.00001
ENSG00000101076	V_HNF4_Q6_01	0.00001
ENSG00000101076	V_HNF4_01	0.00001
ENSG00000082641	V_TCF11MAFG_01	0.00001
ENSG00000148516	V_AREB6_03	0.00001
ENSG00000111206	V_FOXM1_01	0.00001
ENSG00000101076	V_HNF4_01_B	0.00001
	V_NANOG_02	0.00000
ENSG00000085276	V_EVI1_01	0.00000
	V_TCF3_01	0.00000
ENSG00000177374	V_HIC1_03	0.00000
	V_MYOD_Q6	0.00000
	V_HNF3B_01	0.00000
ENSG00000156925	V_ZIC3_01	0.00000
ENSG00000182568	V_SATB1_Q3	0.00000
ENSG00000183072	V_NKX25_03	0.00000
	V_TR4_03	0.00000
	V_ZTA_Q2	0.00000
ENSG00000084093	V_NRSE_B	0.00000
	V_NKX24_01	0.00000
ENSG00000186951	V_PPAR_DR1_Q2	0.00000
ENSG00000188620	V_HMX3_02	0.00000
	V_P50P50_Q3	0.00000
	V_DMRT3_01	0.00000
	V_OLF1_01	0.00000
ENSG00000136997	V_CMYC_02	0.00000
ENSG00000186350	V_PPARA_02	0.00000
	V_HOXD10_01	0.00000
ENSG00000074047	V_GLI2_01	0.00000
ENSG00000091831	V_ER_Q6_02	0.00000

	V_OBOX5_01	0.00000
ENSG00000101076	V_HNF4_DR1_Q3	0.00000
ENSG00000185551	V_COUP_DR1_Q6	0.00000
ENSG00000148200	V_GCNF_01	0.00000
ENSG00000068305	V_RSRFC4_01	0.00000
ENSG00000111087	V_GLI1_01	0.00000
ENSG00000120837	V_NFY_C	0.00000
ENSG00000071564	V_E12_Q6	0.00000
ENSG00000120075	V_HOXB5_01	0.00000
ENSG00000088038	V_CNOT3_01	0.00000
ENSG00000132170	V_PPARG_02	0.00000
ENSG00000167182	V_SP2_01	0.00000
ENSG00000116833	V_LRH1_Q5	0.00000
	V_BEL1_B	0.00000
	V_NKX21_01	0.00000
ENSG00000082175	V_PR_01	0.00000
ENSG00000107859	V_PITX3_01	0.00000
ENSG00000068305	V_MEF2_02	0.00000
	V_VDRRXR_01	0.00000
ENSG00000177426	V_TGIF_02	0.00000
ENSG00000125398	V_SOX9_B1	0.00000
ENSG00000066336	V_PU1_01	0.00000
	V_BARHL1_01	0.00000
	V_HB24_01	0.00000
	V_BARHL2_01	0.00000
	V_IRX3_01	0.00000
ENSG00000106004	V_HOX13_01	0.00000
ENSG00000176678	V_FREAC7_01	0.00000
ENSG00000073282	V_P63_01	0.00000
ENSG00000064835	V_PIT1_Q6	0.00000
ENSG00000164299	V_SPZ1_01	0.00000
ENSG00000054598	V_FREAC3_01	0.00000
	V_VJUN_01	0.00000
ENSG00000137203	V_AP2ALPHA_01	0.00000
	V_CAAT_C	0.00000
ENSG00000115507	V_OTX1_01	0.00000
	V_XFD1_01	0.00000
ENSG00000169083	V_AR_Q2	0.00000
	V_PBX1_04	0.00000
ENSG00000167034	V_NKX3A_02	0.00000
ENSG00000109906	V_PLZF_02	0.00000

ENSG00000143190	V_OCT1_05	0.00000
ENSG00000165556	V_CDX2_Q5	0.00000
ENSG00000171634	V_FAC1_01	0.00000
	V_PAX9_B	0.00000
ENSG00000115966	V_CREBP1_Q2	0.00000
ENSG00000100811	V_YY1_01	0.00000
ENSG00000100644	V_HIF1_Q3	0.00000
	V_TBX15_01	0.00000
	V_ARNT_01	0.00000
	V_HFH1_01	0.00000
ENSG00000169136	V_ATF5_01	0.00000
ENSG00000143190	V_OCT1_02	0.00000
ENSG00000184481	V_FOXO4_02	0.00000
	V_P53_03	0.00000
	V_CART1_01	0.00000
ENSG00000169083	V_AR_01	0.00000
	V_DMRT4_01	0.00000
ENSG00000068305	V_MEF2_03	0.00000
ENSG00000159387	V_IRXB3_01	0.00000
	V_LHX3_02	0.00000
	V_HDX_01	0.00000
ENSG00000126561	V_STAT5A_01	0.00000
ENSG00000004848	V_ARX_01	0.00000
	V_ARNT_02	0.00000
ENSG00000134107	V_DEC_Q1	0.00000
ENSG00000068305	V_MEF2_04	0.00000
	V_LHX4_01	0.00000
ENSG00000196092	V_PAX_Q6	0.00000
ENSG00000142539	V_SPIB_01	0.00000
ENSG00000087510	V_AP2GAMMA_01	0.00000
ENSG00000184486	V_OCTAMER_02	0.00000
ENSG00000111087	V_GLI_Q2	0.00000
ENSG00000132170	V_PPARG_01	0.00000
ENSG00000166478	V_STAF_01	0.00000
ENSG00000149948	VS_HMGA2_01	0.00000
	V_AHRARNT_01	0.00000
	V_AHR_01	0.00000
ENSG00000160224	V_AIRE_02	0.00000
ENSG00000156150	V_ALX3_01	0.00000
ENSG00000068305	V_AMEF2_Q6	0.00000
ENSG00000159216	V_AML_Q6	0.00000

ENSG00000170345	V_AP1_Q6_01	0.00000
ENSG00000137203	V_AP2ALPHA_02	0.00000
ENSG00000137203	V_AP2ALPHA_03	0.00000
ENSG00000137203	V_AP2_Q6	0.00000
	V_APOLYA_B	0.00000
ENSG00000148516	V_AREB6_01	0.00000
ENSG00000169083	V_AR_02	0.00000
ENSG00000169083	V_AR_03	0.00000
ENSG00000169083	V_AR_04	0.00000
	V_ATF_01	0.00000
ENSG00000156273	V_BACH1_01	0.00000
	V_BACH2_01	0.00000
	V_BARBIE_01	0.00000
	V_BARX1_01	0.00000
ENSG00000043039	V_BARX2_01	0.00000
ENSG00000164458	V_BRACH_01	0.00000
ENSG00000091010	V_BRN3C_01	0.00000
	V_BSX_01	0.00000
	V_CART1_03	0.00000
	V_CBF_02	0.00000
	V_CDP_03	0.00000
	V_CDP_04	0.00000
ENSG00000113722	V_CDX1_01	0.00000
ENSG00000165556	V_CDX2_01	0.00000
ENSG00000245848	V_CEBPA_01	0.00000
ENSG00000172216	V_CEBPB_01	0.00000
ENSG00000153879	V_CEBPGAMMA_Q6	0.00000
ENSG00000245848	V_CEBP_01	0.00000
ENSG00000245848	V_CEBP_C	0.00000
	V_CETS1P54_02	0.00000
	V_CETS1P54_03	0.00000
ENSG00000245848	V_CHOP_01	0.00000
	V_CIZ_01	0.00000
	V_COMP1_01	0.00000
ENSG00000175745	V_COUP_01	0.00000
ENSG00000118260	V_CREB_Q2_01	0.00000
ENSG00000118260	V_CREB_Q4_01	0.00000
ENSG00000162924	V_CREL_01	0.00000
ENSG00000141905	V_CTF1_01	0.00000
ENSG00000177030	V_DEAF1_01	0.00000
ENSG00000177030	V_DEAF1_02	0.00000

ENSG00000148516	V_DELTAEF1_01	0.00000
ENSG00000144355	V_DLX1_01	0.00000
ENSG00000115844	V_DLX2_01	0.00000
	V_DMRT1_01	0.00000
	V_DMRT2_01	0.00000
	V_DMRT7_01	0.00000
	V_DOBOX5_01	0.00000
ENSG00000111424	V_DR3_Q4	0.00000
	V_E2_01	0.00000
	V_E2_Q6	0.00000
ENSG00000071564	V_E47_02	0.00000
ENSG00000165030	V_E4BP4_01	0.00000
ENSG00000167967	V_E4F1_Q6	0.00000
	V_EBNA1_01	0.00000
	V_ELF1_Q6	0.00000
ENSG00000135374	V_ELF5_01	0.00000
	V_ELK1_01	0.00000
	V_ELK1_02	0.00000
ENSG00000170370	V_EMX2_01	0.00000
ENSG00000163064	V_EN1_02	0.00000
ENSG00000164778	V_EN2_01	0.00000
	V_ESX1_01	0.00000
	V_ETS2_B	0.00000
ENSG00000085276	V_EVI1_04	0.00000
	V_EVX2_01	0.00000
	V_FOXJ2_02	0.00000
	V_FOXO3A_Q1	0.00000
	V_FOXO3_01	0.00000
ENSG00000184481	V_FOXO4_01	0.00000
	V_FOXP3_Q4	0.00000
ENSG00000137273	V_FREAC2_01	0.00000
ENSG00000012504	V_FXR_Q3	0.00000
	V_GADP_01	0.00000
ENSG00000102145	V_GATA1_06	0.00000
	V_GBX1_01	0.00000
ENSG00000165702	V_GFI1B_01	0.00000
ENSG00000162676	V_GFI1_01	0.00000
ENSG00000111087	V_GLI1_Q2	0.00000
ENSG00000106571	V_GLI3_02	0.00000
ENSG00000113580	V_GRE_C	0.00000
ENSG00000113580	V_GR_01	0.00000

ENSG00000113580	V_GR_Q6	0.00000
ENSG00000130675	V_HB9_01	0.00000
ENSG00000030419	V_HELIOSA_02	0.00000
	V_HES1_Q2	0.00000
ENSG00000168269	V_HFH3_01	0.00000
ENSG00000129654	V_HFH4_01	0.00000
	V_HLF_01	0.00000
ENSG00000147421	V_HMBOX1_01	0.00000
ENSG00000068305	V_HMEF2_Q6	0.00000
ENSG00000215612	V_HMX1_02	0.00000
ENSG00000108753	V_HNF1B_01	0.00000
ENSG00000135100	V_HNF1_C	0.00000
ENSG00000135100	V_HNF1_Q6	0.00000
ENSG00000135100	V_HNF1_Q6_01	0.00000
ENSG00000170608	V_HNF3_Q6	0.00000
ENSG00000101076	V_HNF4ALPHA_Q6	0.00000
	V_HOMEZ_01	0.00000
ENSG00000106004	V_HOX13_02	0.00000
ENSG00000105991	V_HOXA1_01	0.00000
ENSG00000105996	V_HOXA2_01	0.00000
ENSG00000197576	V_HOXA4_01	0.00000
	V_HOXA7_03	0.00000
ENSG00000078399	V_HOXA9_01	0.00000
ENSG00000159184	V_HOXB13_01	0.00000
ENSG00000120093	V_HOXB3_01	0.00000
ENSG00000182742	V_HOXB4_01	0.00000
	V_HOXB7_01	0.00000
ENSG00000170689	V_HOXB9_01	0.00000
ENSG00000123388	V_HOXC11_01	0.00000
ENSG00000172789	V_HOXC5_01	0.00000
	V_HOXC6_01	0.00000
ENSG00000037965	V_HOXC8_01	0.00000
ENSG00000180806	V_HOXC9_01	0.00000
ENSG00000128645	V_HOXD1_01	0.00000
ENSG00000128652	V_HOXD3_01	0.00000
	V_HOXD8_01	0.00000
ENSG00000185122	V_HSF_Q6	0.00000
ENSG00000100219	V_HTF_01	0.00000
ENSG00000140968	V_ICSBP_Q6	0.00000
ENSG00000185811	V_IK1_01	0.00000
ENSG00000185811	V_IK3_01	0.00000

	V_IPF1_Q4_01	0.00000
	V_IRX3_02	0.00000
	V_ISL2_01	0.00000
	V_ISX_01	0.00000
ENSG00000116132	V_K2B_01	0.00000
ENSG00000115112	V_LBP9_01	0.00000
	V_LBX2_01	0.00000
	V_LDSPOLYA_B	0.00000
	V_LHX5_01	0.00000
	V_LHX61_01	0.00000
	V_LHX8_01	0.00000
	V_LHX9_01	0.00000
	V_LIM1_01	0.00000
ENSG00000136944	V_LMX1B_01	0.00000
ENSG00000162761	V_LMX1_01	0.00000
	V_MAX_01	0.00000
	V_MAX_Q6	0.00000
	V_MEF3_B	0.00000
ENSG00000078399	V_MEIS1AHOXA9_01	0.00000
ENSG00000078399	V_MEIS1BHOXA9_02	0.00000
	V_MEIS1_01	0.00000
	V_MEIS1_02	0.00000
ENSG00000068305	V_MMEF2_Q6	0.00000
ENSG00000005102	V_MOX1_01	0.00000
ENSG00000150347	V_MRF2_01	0.00000
	V_MRG2_01	0.00000
ENSG00000120149	V_MSX2_01	0.00000
ENSG00000127989	V_MTERF_01	0.00000
ENSG00000188786	V_MTF1_01	0.00000
ENSG00000136997	V_MYCMAX_01	0.00000
ENSG00000136997	V_MYCMAX_03	0.00000
	V_MYOD_01	0.00000
	V_NANOG_01	0.00000
ENSG00000141905	V_NF1_Q6_01	0.00000
ENSG00000123405	V_NFE2_01	0.00000
	V_NFKAPPAB50_01	0.00000
ENSG00000173039	V_NFKAPPAB65_01	0.00000
ENSG00000173039	V_NFKAPPAB_01	0.00000
ENSG00000109320	V_NFKB_Q6	0.00000
ENSG00000173039	V_NFKB_Q6_01	0.00000
ENSG00000120837	V_NFY_Q6_01	0.00000

	V_NKX12_01	0.00000
	V_NKX22_02	0.00000
	V_NKX23_01	0.00000
	V_NKX26_01	0.00000
ENSG00000109705	V_NKX32_02	0.00000
ENSG00000163623	V_NKX61_01	0.00000
ENSG00000163623	V_NKX61_02	0.00000
ENSG00000163623	V_NKX61_03	0.00000
	V_NKX63_01	0.00000
	V_OBOX1_01	0.00000
	V_OBOX2_01	0.00000
	V_OBOX5_02	0.00000
ENSG00000143190	V_OCT1_06	0.00000
ENSG00000143190	V_OCT1_08	0.00000
ENSG00000143190	V_OCT1_Q6	0.00000
ENSG00000028277	V_OCT2_01	0.00000
ENSG00000204531	V_OCT4_01	0.00000
ENSG00000204531	V_OCT4_02	0.00000
ENSG00000198914	V_OCTAMER_01	0.00000
ENSG00000143190	V_OCT_C	0.00000
	V_OTP_01	0.00000
ENSG00000165588	V_OTX2_01	0.00000
ENSG00000165588	V_OTX2_Q3	0.00000
	V_P53_01	0.00000
ENSG00000075891	V_PAX2_01	0.00000
ENSG00000106331	V_PAX4_01	0.00000
ENSG00000106331	V_PAX4_05	0.00000
ENSG00000196092	V_PAX5_02	0.00000
ENSG00000125618	V_PAX8_01	0.00000
ENSG00000125618	V_PAX8_B	0.00000
	V_PBX1_02	0.00000
	V_PITX1_01	0.00000
	V_PITX2_01	0.00000
ENSG00000165495	V_PKNOX2_01	0.00000
ENSG00000165462	V_PMX2A_01	0.00000
ENSG00000109132	V_PMX2B_01	0.00000
ENSG00000031544	V_PNR_01	0.00000
	V_POU2F3_01	0.00000
ENSG00000204531	V_POU5F1_01	0.00000
ENSG00000184271	V_POU6F1_01	0.00000
ENSG00000184271	V_POU6F1_02	0.00000

ENSG00000184271	V_POU6F1_03	0.00000
ENSG00000160199	V_PREP1_01	0.00000
ENSG00000175325	V_PROP1_02	0.00000
ENSG00000082175	V_PR_02	0.00000
	V_PSX1_01	0.00000
	V_RELBP52_01	0.00000
	V_REX1_03	0.00000
ENSG00000132005	V_RFX1_01	0.00000
ENSG00000087903	V_RFX_Q6	0.00000
	V_RHOX11_01	0.00000
	V_RHOX11_02	0.00000
ENSG00000102935	V_ROAZ_01	0.00000
ENSG00000069667	V_RORA1_01	0.00000
	V_R_01	0.00000
	V_S8_01	0.00000
	V_S8_02	0.00000
	V_SEF1_C	0.00000
ENSG00000168779	V_SHOX2_01	0.00000
	V_SIX1_01	0.00000
ENSG00000170577	V_SIX2_01	0.00000
	V_SIX4_01	0.00000
	V_SIX6_02	0.00000
ENSG00000170365	V_SMAD1_01	0.00000
ENSG00000072310	V_SREBP_Q3	0.00000
ENSG00000112658	V_SRF_01	0.00000
ENSG00000112658	V_SRF_C	0.00000
ENSG00000112658	V_SRF_Q4	0.00000
ENSG00000112658	V_SRF_Q5_01	0.00000
ENSG00000112658	V_SRF_Q5_02	0.00000
ENSG00000112658	V_SRF_Q6	0.00000
ENSG00000115415	V_STAT1_01	0.00000
	V_STAT3STAT3_Q3	0.00000
ENSG00000168610	V_STAT3_01	0.00000
ENSG00000126561	V_STAT5A_02	0.00000
	V_STRA13_01	0.00000
ENSG00000164048	V_SZF11_01	0.00000
	V_T3R_01	0.00000
ENSG00000077092	V_T3R_Q6	0.00000
	V_TAACC_B	0.00000
ENSG00000071564	V_TAL1ALPHAE47_01	0.00000
ENSG00000071564	V_TAL1BETAE47_01	0.00000

ENSG00000162367	V_TAL1BETAITF2_01	0.00000
ENSG00000112592	V_TATA_01	0.00000
ENSG00000118260	V_TAXCREB_02	0.00000
ENSG00000112837	V_TBX18_01	0.00000
ENSG00000122145	V_TBX22_01	0.00000
ENSG00000118707	V_TGIF2_01	0.00000
	V_TITF1_Q3	0.00000
ENSG00000177463	V_TR4_Q2	0.00000
ENSG00000185668	V_TST1_01	0.00000
	V_UNCX4.1_01	0.00000
ENSG00000158773	V_USF_01	0.00000
ENSG00000158773	V_USF_02	0.00000
ENSG00000158773	V_USF_Q6_01	0.00000
	V_VAX1_01	0.00000
	V_VAX2_01	0.00000
ENSG00000167074	V_VBP_01	0.00000
	V_VMAF_01	0.00000
	V_VMYB_01	0.00000
ENSG00000100987	V_VSX1_01	0.00000
ENSG00000100219	V_XBP1_01	0.00000
	V_ZBED6_01	0.00000
ENSG00000198081	V_ZF5_B	0.00000
	V_ZFP206_01	0.00000
ENSG00000152977	V_ZIC1_01	0.00000
ENSG00000071564	V_HAND1E47_01	0.00000
	V_P53_05	0.00000
ENSG00000159216	V_PEBP_Q6	0.00000
ENSG00000132005	V_RFX1_02	0.00000
ENSG00000181449	V_SOX2_Q6	0.00000
ENSG00000068305	V_RSRFC4_Q2	0.00000
ENSG00000157554	V_ETS_Q4	0.00000
ENSG00000113430	V_IRX4_01	0.00000
ENSG00000132170	V_PPARG_03	0.00000
ENSG00000170345	V_AP1_01	0.00000
	V_DOBOX4_01	0.00000
	V_MIF1_01	0.00000
ENSG00000141905	V_NF1_Q6	0.00000
ENSG00000113916	V_BCL6_01	0.00000
ENSG00000068305	V_MEF2_01	0.00000
ENSG00000169297	V_DAX1_01	0.00000
ENSG00000186951	V_PPARA_01	0.00000

	V_MSX3_01	0.00000
ENSG00000185551	V_DR1_Q3	0.00000
	V_HOXD13_01	0.00000
ENSG00000025434	V_LXR_Q3	0.00000
ENSG00000196092	V_PAX5_01	0.00000
ENSG00000160224	V_AIRE_01	0.00000
	V_POLY_C	0.00000
ENSG00000103241	V_HFH8_01	0.00000
	V_NKX29_01	0.00000
	V_RXRLXRB_01	0.00000
ENSG00000168610	V_STAT3_03	0.00000
	V_NFY_01	0.00000
ENSG00000091831	V_ER_Q6	0.00000
	V_PTF1BETA_Q6	0.00000
ENSG00000118513	V_CMYB_01	0.00000
ENSG00000116044	V_NRF2_Q4	0.00000
ENSG00000112658	V_SRF_02	0.00000
ENSG00000135457	V_CP2_02	0.00000
ENSG00000090447	V_AP4_01	0.00000
ENSG00000173153	V_ERR1_Q2	0.00000
ENSG00000185122	V_HSF1_Q6	0.00000
	V_TBX15_02	0.00000
ENSG00000162772	V_ATF3_Q6	0.00000
ENSG00000178573	V_MAF_Q6	0.00000
	V_SIX6_01	0.00000
ENSG00000173153	V_ERR1_Q3	0.00000
ENSG00000165556	V_CDX_Q5	0.00000
	V_DMRT5_01	0.00000
ENSG00000141905	V_MYOGNF1_01	0.00000
	V_NKX52_01	0.00000
ENSG00000160113	V_EAR2_Q2	0.00000
ENSG00000138378	V_STAT4_Q4	0.00000
ENSG00000253293	V_HOXA10_01	0.00000
	V_DBX2_01	0.00000
	V_IPF1_06	0.00000
ENSG00000115415	V_STAT_Q6	0.00000
	V_IRX5_01	0.00000
ENSG00000102145	V_GATA1_02	0.00000
	V_TRF1_01	0.00000
ENSG00000251493	V_FREAC4_01	0.00000
ENSG00000137309	V_HMGIY_Q3	0.00000

ENSG00000113916	V_BCL6_02	0.00000
	V_ISRE_01	0.00000
ENSG00000184486	V_POU3F2_01	0.00000
ENSG00000007372	V_PAX6_02	0.00000
	V_SIX3_01	0.00000
ENSG00000125347	V_IRF_Q6	0.00000
ENSG00000173757	V_STAT5B_01	0.00000
ENSG00000143190	V_OCT1_04	0.00000

Table 7.3 TF motifs from classification of boundary vs. outside.

ENSG	motif	MeanDecreaseAccuracy
	V_GTF2IRD1_01	0.07980
	V_CACBINDINGPROTEIN_Q6	0.04906
ENSG00000181690	V_PLAG1_01	0.02797
ENSG00000103495	V_MAZ_Q6	0.02714
	V_ZFX_01	0.02393
	V_SP1_Q6	0.02371
	V_SP1SP3_Q4	0.02178
ENSG00000172845	V_SP1_Q4_01	0.01994
ENSG00000160685	V_CKROX_Q2	0.01935
	V_KLF15_Q2	0.01916
	V_CACD_01	0.01295
ENSG00000071564	V_E2A_Q6	0.01140
	V_LXR_DR4_Q3	0.01122
ENSG00000163848	V_ZBP89_Q4	0.01083
ENSG00000105866	V_SP4_Q5	0.00981
ENSG00000006194	V_FPM315_01	0.00945
ENSG00000185811	V_IK_Q5	0.00940
ENSG00000172845	V_SP1_Q2_01	0.00937
	V_UF1H3BETA_Q6	0.00924
	V_SP1_02	0.00896
ENSG00000172845	V_SP1_Q6_01	0.00858
ENSG00000197579	V_LUN1_01	0.00854
	V_BDP1_01	0.00817
	V_SP1_01	0.00716
ENSG00000172059	V_FKLF_Q5	0.00701
	V_MUSCLE_INI_B	0.00671
	V_ZFP281_01	0.00641
ENSG00000137203	V_AP2_Q6_01	0.00584
	V_PITX2_Q2	0.00569
ENSG00000148606	V_RPC155_01	0.00558

ENSG00000185811	V_LYF1_01	0.00521
ENSG00000072310	V_SREBP1_01	0.00516
ENSG00000077092	V_T3R_Q6	0.00472
ENSG00000120738	V_KROX_Q6	0.00455
ENSG00000114861	V_FOXP1_01	0.00424
ENSG00000072310	V_SREBP_Q6	0.00402
ENSG00000256683	V_ZBRK1_01	0.00374
	V_GC_01	0.00357
ENSG00000111424	V_VDR_Q3	0.00344
ENSG00000136826	V_GKLF_02	0.00286
	V_MEF2C_01	0.00285
ENSG00000091831	V_ERALPHA_01	0.00245
ENSG00000123405	V_MAF_Q6_01	0.00236
ENSG00000071564	V_MYOD_Q6_01	0.00231
ENSG00000165804	V_ZNF219_01	0.00226
ENSG00000130726	V_RNF96_01	0.00169
ENSG00000185551	V_COUPTF_Q6	0.00165
ENSG00000126351	V_TERALPHA_Q6	0.00164
ENSG00000103241	V_FOX_Q2	0.00157
ENSG00000185551	V_ARP1_01	0.00153
	V_FOXJ2_01	0.00150
	cpgoverlap	0.00123
	V_PUR1_Q4	0.00121
ENSG00000124782	V_RREB1_01	0.00118
ENSG00000077809	V_TFIII_Q6	0.00112
ENSG00000184937	V_WT1_Q6	0.00111
ENSG00000172216	V_CEBPB_02	0.00102
ENSG00000137203	V_AP2_Q3	0.00086
ENSG00000089225	V_TBX5_Q5	0.00072
ENSG00000148516	V_AREB6_03	0.00066
ENSG00000099326	V_MZF1_02	0.00060
	V_MYF_01	0.00060
	V_FOXD3_01	0.00057
ENSG00000102974	V_CTCF_01	0.00055
ENSG00000150907	V_FOXO1_Q5	0.00054
ENSG00000066336	V_PU1_Q4	0.00052
ENSG00000100105	V_MAZR_01	0.00050
ENSG00000071564	V_E2A_Q2	0.00044
ENSG00000135363	V_LMO2COM_01	0.00040
ENSG00000072310	V_SREBP1_Q5	0.00039
	V_MINI19_B	0.00037

ENSG00000111206	V_FOXM1_01	0.00035
ENSG00000129514	V_HNF3ALPHA_Q6	0.00035
ENSG00000102974	V_CTCF_02	0.00032
	V_P53_04	0.00028
ENSG00000120738	V_EGR_Q6	0.00028
ENSG00000106331	V_PAX4_04	0.00025
ENSG00000171786	V_HEN1_02	0.00023
ENSG00000162992	V_NEUROD_02	0.00021
ENSG00000106331	V_PAX4_03	0.00021
ENSG00000171786	V_HEN1_01	0.00017
	V_MYOD_Q6	0.00014
ENSG00000177374	V_HIC1_02	0.00011
ENSG00000084093	V_REST_01	0.00010
ENSG00000188786	V_MTF1_Q4	0.00007
	V_MINI20_B	0.00007
ENSG00000057657	V_BLIMP1_Q6	0.00006
ENSG00000163848	V_CACCCBINDINGFACTOR_Q6	0.00006
ENSG00000084093	V_NRSF_Q4	0.00006
ENSG00000137203	V_AP2ALPHA_01	0.00006
ENSG00000087510	V_AP2GAMMA_01	0.00005
ENSG00000162367	V_TAL1_01	0.00005
ENSG00000088038	V_CNOT3_01	0.00005
ENSG00000196767	V_BRN4_01	0.00004
	V_NCX_02	0.00003
ENSG00000143190	V_OCT1_Q5_01	0.00002
ENSG00000177374	V_HIC1_03	0.00002
ENSG00000185668	V_TST1_02	0.00002
ENSG00000185024	V_BRF1_01	0.00001
ENSG00000090447	V_AP4_01	0.00001
	V_ARNT_01	0.00001
	V_NANOG_02	0.00001
ENSG00000071564	V_E47_01	0.00001
ENSG00000172845	V_SP3_Q3	0.00001
ENSG00000102145	V_GATA1_02	0.00001
	V_DMRT3_01	0.00001
	V_DBX2_01	0.00001
ENSG00000185551	V_COUP_DR1_Q6	0.00001
ENSG00000170608	V_HNF3_Q6_01	0.00001
ENSG00000251493	V_FREAC4_01	0.00001
ENSG00000101076	V_HNF4_Q6_01	0.00000
ENSG00000125398	V_SOX9_B1	0.00000

ENSG00000084093	V_NRSF_01	0.00000
ENSG00000166478	V_STAF_02	0.00000
	V_DMRT5_01	0.00000
ENSG00000196092	V_PAX5_01	0.00000
ENSG00000101076	V_HNF4_DR1_Q3	0.00000
ENSG00000186951	V_PPAR_DR1_Q2	0.00000
ENSG00000176678	V_FREAC7_01	0.00000
ENSG00000182568	V_SATB1_Q3	0.00000
	V_ISRE_01	0.00000
ENSG00000129654	V_HFH4_01	0.00000
ENSG00000113430	V_IRX4_01	0.00000
ENSG00000166478	V_STAF_01	0.00000
ENSG00000123405	V_NFE2_01	0.00000
ENSG00000178573	V_MAF_Q6	0.00000
ENSG00000078399	V_MEIS1BHOXA9_02	0.00000
ENSG00000118513	V_CMYB_01	0.00000
ENSG00000109381	V_NERF_Q2	0.00000
ENSG00000100393	V_P300_01	0.00000
	V_LDSPOLYA_B	0.00000
ENSG00000162676	V_GFI1_01	0.00000
	V_TR4_03	0.00000
ENSG00000160113	V_EAR2_Q2	0.00000
	V_SIX6_01	0.00000
ENSG00000064835	V_PIT1_01	0.00000
	V_XFD3_01	0.00000
ENSG00000137309	V_HMGIY_Q3	0.00000
ENSG00000173153	V_ERR1_Q3	0.00000
ENSG00000186350	V_PPARA_02	0.00000
	V_BEL1_B	0.00000
ENSG00000165556	V_CDX2_Q5	0.00000
ENSG00000068305	V_AMEF2_Q6	0.00000
ENSG00000073282	V_P63_01	0.00000
	V_MIF1_01	0.00000
ENSG00000091831	V_ER_Q6	0.00000
ENSG00000074047	V_GLI2_01	0.00000
ENSG00000116044	V_NRF2_Q4	0.00000
ENSG00000204531	V_OCT4_01	0.00000
	V_MSX3_01	0.00000
ENSG00000132005	V_RFX1_02	0.00000
ENSG00000173039	V_NFKB_Q6_01	0.00000
ENSG00000186951	V_PPARA_01	0.00000

ENSG00000028277	V_OCT2_01	0.00000
	V_CAAT_C	0.00000
ENSG00000165556	V_CDX_Q5	0.00000
ENSG00000068305	V_RSRFC4_01	0.00000
ENSG00000084093	V_NRSE_B	0.00000
ENSG00000068305	V_HMEF2_Q6	0.00000
ENSG00000101076	V_HNF4_01	0.00000
ENSG00000177463	V_TR4_Q2	0.00000
	V_P53_03	0.00000
ENSG00000141905	V_NF1_Q6	0.00000
ENSG00000112658	V_SRF_Q5_02	0.00000
ENSG00000113916	V_BCL6_01	0.00000
	V_SIX3_01	0.00000
ENSG00000031544	V_PNR_01	0.00000
ENSG00000106331	V_PAX4_01	0.00000
ENSG00000132170	V_PPARG_03	0.00000
ENSG00000164458	V_BRACH_01	0.00000
	V_ETS2_B	0.00000
	V_OLF1_01	0.00000
	V_IPF1_06	0.00000
ENSG00000111087	V_GLI1_01	0.00000
ENSG00000164778	V_EN2_01	0.00000
	V_LHX3_02	0.00000
	V_PBX1_04	0.00000
ENSG00000156925	V_ZIC3_01	0.00000
ENSG00000198914	V_OCTAMER_01	0.00000
	V_PTF1BETA_Q6	0.00000
	V_HOXA7_03	0.00000
	V_PSX1_01	0.00000
	V_HNF3B_01	0.00000
ENSG00000162761	V_LMX1_01	0.00000
	V_BARBIE_01	0.00000
ENSG00000136997	V_EBOX_Q6_01	0.00000
ENSG00000102935	V_ROAZ_01	0.00000
	V_ARNT_02	0.00000
ENSG00000147421	V_HMBOX1_01	0.00000
ENSG00000164299	V_SPZ1_01	0.00000
	V_P50P50_Q3	0.00000
ENSG00000100811	V_YY1_01	0.00000
	V_OBOX5_02	0.00000
ENSG00000143190	V_OCT1_05	0.00000

	V_SEF1_C	0.00000
	V_NKX52_01	0.00000
ENSG00000120093	V_HOXB3_01	0.00000
ENSG00000127989	V_MTERF_01	0.00000
	V_ZID_01	0.00000
ENSG00000120068	V_HOXB8_01	0.00000
ENSG00000112658	V_SRF_C	0.00000
	V_NKX22_02	0.00000
ENSG00000136944	V_LMX1B_01	0.00000
ENSG00000116833	V_LRH1_Q5	0.00000
ENSG00000173757	V_STAT5B_01	0.00000
ENSG00000163623	V_NKX61_01	0.00000
ENSG00000185122	V_HSF1_Q6	0.00000
	V_MEIS1_02	0.00000
	V_AHRARNT_01	0.00000
	V_AHRARNT_02	0.00000
	V_AHR_01	0.00000
ENSG00000160224	V_AIRE_02	0.00000
ENSG00000156150	V_ALX3_01	0.00000
	V_ALX4_01	0.00000
ENSG00000170345	V_AP1_01	0.00000
ENSG00000137203	V_AP2ALPHA_02	0.00000
ENSG00000137203	V_AP2ALPHA_03	0.00000
	V_APOLYA_B	0.00000
ENSG00000148516	V_AREB6_01	0.00000
ENSG00000169083	V_AR_02	0.00000
ENSG00000169083	V_AR_Q2	0.00000
ENSG00000169136	V_ATF5_01	0.00000
	V_ATF_01	0.00000
ENSG00000156273	V_BACH1_01	0.00000
	V_BACH2_01	0.00000
	V_BARHL1_01	0.00000
ENSG00000043039	V_BARX2_01	0.00000
ENSG00000113916	V_BCL6_02	0.00000
ENSG00000091010	V_BRN3C_01	0.00000
	V_BSX_01	0.00000
	V_CAAT_01	0.00000
	V_CART1_01	0.00000
	V_CART1_03	0.00000
	V_CBF_02	0.00000
	V_CDP_03	0.00000

	V_CDP_04	0.00000
ENSG00000113722	V_CDX1_01	0.00000
ENSG00000153879	V_CEBPGAMMA_Q6	0.00000
ENSG00000245848	V_CEBP_01	0.00000
ENSG00000245848	V_CEBP_Q2_01	0.00000
	V_CETS1P54_03	0.00000
ENSG00000245848	V_CHOP_01	0.00000
	V_CHX10_01	0.00000
	V_COMP1_01	0.00000
ENSG00000115966	V_CREBP1_Q2	0.00000
ENSG00000118260	V_CREB_Q2_01	0.00000
ENSG00000105392	V_CRX_02	0.00000
ENSG00000141905	V_CTF1_01	0.00000
ENSG00000177030	V_DEAF1_01	0.00000
ENSG00000177030	V_DEAF1_02	0.00000
ENSG00000134107	V_DEC_Q1	0.00000
ENSG00000144355	V_DLX1_01	0.00000
ENSG00000105880	V_DLX5_01	0.00000
	V_DMRT1_01	0.00000
	V_DMRT4_01	0.00000
	V_DOBOX4_01	0.00000
	V_DOBOX5_01	0.00000
ENSG00000111424	V_DR3_Q4	0.00000
	V_E2_01	0.00000
	V_E2_Q6	0.00000
ENSG00000071564	V_E47_02	0.00000
	V_EBNA1_01	0.00000
ENSG00000132005	V_EFC_Q6	0.00000
	V_ELF1_Q6	0.00000
ENSG00000135374	V_ELF5_01	0.00000
	V_ELK1_01	0.00000
	V_ELK1_02	0.00000
ENSG00000170370	V_EMX2_01	0.00000
ENSG00000163064	V_EN1_02	0.00000
ENSG00000119715	V_ERR2_01	0.00000
	V_ESX1_01	0.00000
ENSG00000157554	V_ETS_Q4	0.00000
ENSG00000085276	V_EVI1_03	0.00000
ENSG00000085276	V_EVI1_04	0.00000
	V_EVX2_01	0.00000
ENSG00000171634	V_FAC1_01	0.00000

	V_FOXJ2_02	0.00000
ENSG00000150907	V_FOXO1_02	0.00000
ENSG00000184481	V_FOXO4_02	0.00000
	V_FOXP3_Q4	0.00000
ENSG00000137273	V_FREAC2_01	0.00000
	V_FXR_IR1_Q6	0.00000
	V_GADP_01	0.00000
ENSG00000102145	V_GATA1_04	0.00000
	V_GBX1_01	0.00000
ENSG00000165702	V_GFI1B_01	0.00000
ENSG00000106571	V_GLI3_02	0.00000
ENSG00000111087	V_GLI_Q2	0.00000
ENSG00000113580	V_GRE_C	0.00000
ENSG00000113580	V_GR_01	0.00000
ENSG00000113580	V_GR_Q6	0.00000
	V_GSC_01	0.00000
ENSG00000180613	V_GSH2_01	0.00000
ENSG00000071564	V_HAND1E47_01	0.00000
ENSG00000130675	V_HB9_01	0.00000
	V_HDX_01	0.00000
ENSG00000030419	V_HELIOSA_02	0.00000
	V_HES1_Q2	0.00000
	V_HFH1_01	0.00000
ENSG00000103241	V_HFH8_01	0.00000
ENSG00000100644	V_HIF1_Q3	0.00000
ENSG00000215612	V_HMX1_02	0.00000
ENSG00000108753	V_HNF1B_01	0.00000
ENSG00000135100	V_HNF1_C	0.00000
ENSG00000135100	V_HNF1_Q6_01	0.00000
ENSG00000170608	V_HNF3_Q6	0.00000
ENSG00000101076	V_HNF4ALPHA_Q6	0.00000
ENSG00000119547	V_HNF6_Q6	0.00000
	V_HOMEZ_01	0.00000
ENSG00000106004	V_HOX13_01	0.00000
ENSG00000106004	V_HOX13_02	0.00000
ENSG00000253293	V_HOXA10_01	0.00000
ENSG00000105991	V_HOXA1_01	0.00000
ENSG00000105996	V_HOXA2_01	0.00000
ENSG00000197576	V_HOXA4_01	0.00000
ENSG00000106006	V_HOXA6_01	0.00000
	V_HOXA7_02	0.00000

ENSG00000078399	V_HOXA9_01	0.00000
ENSG00000159184	V_HOXB13_01	0.00000
ENSG00000182742	V_HOXB4_01	0.00000
ENSG00000108511	V_HOXB6_01	0.00000
	V_HOXB7_01	0.00000
ENSG00000170689	V_HOXB9_01	0.00000
ENSG00000123388	V_HOXC11_01	0.00000
ENSG00000172789	V_HOXC5_01	0.00000
	V_HOXC6_01	0.00000
ENSG00000037965	V_HOXC8_01	0.00000
ENSG00000180806	V_HOXC9_01	0.00000
	V_HOXD10_01	0.00000
	V_HOXD13_01	0.00000
ENSG00000128645	V_HOXD1_01	0.00000
ENSG00000128652	V_HOXD3_01	0.00000
	V_HOXD8_01	0.00000
	V_HP1SITEFACTOR_Q6	0.00000
ENSG00000025156	V_HSF2_02	0.00000
ENSG00000185122	V_HSF_Q6	0.00000
ENSG00000100219	V_HTF_01	0.00000
ENSG00000140968	V_ICSBP_Q6	0.00000
ENSG00000185811	V_IK1_01	0.00000
ENSG00000185811	V_IK3_01	0.00000
	V_IPF1_05	0.00000
	V_IPF1_Q4_01	0.00000
ENSG00000125347	V_IRF1_01	0.00000
ENSG00000168310	V_IRF2_01	0.00000
ENSG00000185507	V_IRF7_01	0.00000
	V_IRX3_02	0.00000
ENSG00000116132	V_K2B_01	0.00000
ENSG00000115112	V_LBP9_01	0.00000
	V_LBX2_01	0.00000
	V_LHX4_01	0.00000
	V_LHX5_01	0.00000
	V_LHX8_01	0.00000
	V_LHX9_01	0.00000
	V_LIM1_01	0.00000
	V_MAX_01	0.00000
	V_MAX_Q6	0.00000
ENSG00000068305	V_MEF2_02	0.00000
ENSG00000068305	V_MEF2_03	0.00000

	V_MEF3_B	0.00000
ENSG00000078399	V_MEIS1AHOXA9_01	0.00000
	V_MEIS2_01	0.00000
ENSG00000005102	V_MOX1_01	0.00000
ENSG00000150347	V_MRF2_01	0.00000
	V_MRG2_01	0.00000
ENSG00000188786	V_MTF1_01	0.00000
ENSG00000136997	V_MYCMAX_01	0.00000
ENSG00000136997	V_MYCMAX_03	0.00000
	V_MYOD_01	0.00000
ENSG00000141905	V_NF1_Q6_01	0.00000
ENSG00000109320	V_NFKB_Q6	0.00000
ENSG00000120837	V_NFY_Q6_01	0.00000
	V_NKX12_01	0.00000
	V_NKX23_01	0.00000
	V_NKX24_01	0.00000
ENSG00000183072	V_NKX25_03	0.00000
	V_NKX29_01	0.00000
ENSG00000109705	V_NKX32_02	0.00000
ENSG00000167034	V_NKX3A_02	0.00000
ENSG00000163623	V_NKX61_02	0.00000
ENSG00000163623	V_NKX61_03	0.00000
	V_NKX63_01	0.00000
	V_OBOX1_01	0.00000
	V_OBOX2_01	0.00000
	V_OBOX5_01	0.00000
ENSG00000143190	V_OCT1_06	0.00000
ENSG00000143190	V_OCT1_08	0.00000
ENSG00000143190	V_OCT1_Q6	0.00000
ENSG00000204531	V_OCT4_02	0.00000
ENSG00000184486	V_OCTAMER_02	0.00000
ENSG00000143190	V_OCT_C	0.00000
ENSG00000143190	V_OCT_Q6	0.00000
	V_OG2_02	0.00000
ENSG00000115507	V_OTX1_01	0.00000
ENSG00000165588	V_OTX2_01	0.00000
	V_OTX3_01	0.00000
ENSG00000073282	V_P53_DECAMER_Q2	0.00000
ENSG00000075891	V_PAX2_01	0.00000
ENSG00000106331	V_PAX4_05	0.00000
ENSG00000196092	V_PAX5_02	0.00000

ENSG00000007372	V_PAX6_01	0.00000
ENSG00000007372	V_PAX6_02	0.00000
ENSG00000125618	V_PAX8_01	0.00000
ENSG00000125618	V_PAX8_B	0.00000
	V_PBX1_02	0.00000
ENSG00000159216	V_PEBP_Q6	0.00000
	V_PITX1_01	0.00000
	V_PITX2_01	0.00000
ENSG00000107859	V_PITX3_01	0.00000
ENSG00000165495	V_PKNOX2_01	0.00000
ENSG00000109132	V_PMX2B_01	0.00000
	V_POU2F3_01	0.00000
ENSG00000184271	V_POU6F1_02	0.00000
ENSG00000184271	V_POU6F1_03	0.00000
ENSG00000132170	V_PPARG_02	0.00000
ENSG00000160199	V_PREP1_01	0.00000
	V_PROP1_01	0.00000
ENSG00000175325	V_PROP1_02	0.00000
	V_REX1_03	0.00000
ENSG00000132005	V_RFX1_01	0.00000
	V_RHOX11_01	0.00000
	V_RHOX11_02	0.00000
ENSG00000179456	V_RP58_01	0.00000
ENSG00000068305	V_RSRFC4_Q2	0.00000
	V_RXRLXRB_01	0.00000
	V_R_01	0.00000
	V_S8_01	0.00000
	V_S8_02	0.00000
	V_SIX1_01	0.00000
ENSG00000170577	V_SIX2_01	0.00000
	V_SIX4_01	0.00000
	V_SIX6_02	0.00000
ENSG00000170365	V_SMAD1_01	0.00000
ENSG00000170365	V_SMAD_Q6_01	0.00000
ENSG00000072310	V_SREBP_Q3	0.00000
ENSG00000112658	V_SRF_01	0.00000
ENSG00000112658	V_SRF_Q4	0.00000
ENSG00000112658	V_SRF_Q5_01	0.00000
ENSG00000112658	V_SRF_Q6	0.00000
ENSG00000184895	V_SRY_02	0.00000
ENSG00000115415	V_STAT1_01	0.00000

	V_STAT3STAT3_Q3	0.00000
ENSG00000168610	V_STAT3_01	0.00000
ENSG00000138378	V_STAT4_Q4	0.00000
ENSG00000126561	V_STAT5A_01	0.00000
	V_STRA13_01	0.00000
ENSG00000164048	V_SZF11_01	0.00000
	V_T3R_01	0.00000
	V_TAACC_B	0.00000
ENSG00000071564	V_TAL1ALPHAE47_01	0.00000
ENSG00000071564	V_TAL1BETAE47_01	0.00000
ENSG00000162367	V_TAL1BETAITF2_01	0.00000
ENSG00000112592	V_TATA_01	0.00000
ENSG00000118260	V_TAXCREB_02	0.00000
	V_TBX15_01	0.00000
	V_TBX15_02	0.00000
ENSG00000112837	V_TBX18_01	0.00000
ENSG00000122145	V_TBX22_01	0.00000
ENSG00000089225	V_TBX5_01	0.00000
ENSG00000118707	V_TGIF2_01	0.00000
ENSG00000177426	V_TGIF_02	0.00000
ENSG00000185668	V_TST1_01	0.00000
	V_UNCX4.1_01	0.00000
ENSG00000158773	V_USF_01	0.00000
	V_VAX1_01	0.00000
	V_VAX2_01	0.00000
	V_VDRRXR_01	0.00000
	V_VJUN_01	0.00000
	V_VMAF_01	0.00000
ENSG00000100987	V_VSX1_01	0.00000
ENSG00000100219	V_XBP1_01	0.00000
	V_XFD2_01	0.00000
ENSG00000198081	V_ZF5_B	0.00000
	V_ZTA_Q2	0.00000
ENSG00000245848	V_CEBP_C	0.00000
ENSG00000120837	V_NFY_C	0.00000
ENSG00000004848	V_ARX_01	0.00000
ENSG00000135100	V_HNF1_Q6	0.00000
	V_P53_01	0.00000
	V_BARHL2_01	0.00000
ENSG00000181449	V_SOX2_Q6	0.00000
	V_PAX9_B	0.00000

	V_TRF1_01	0.00000
	V_TCF3_01	0.00000
ENSG00000112658	V_SRF_02	0.00000
ENSG00000142539	V_SPIB_01	0.00000
ENSG00000162772	V_ATF3_Q6	0.00000
ENSG00000054598	V_FREAC3_01	0.00000
ENSG00000165462	V_PMX2A_01	0.00000
ENSG00000179388	V_EGR3_01	0.00000
ENSG00000126561	V_STAT5A_02	0.00000
	V_DMRT2_01	0.00000
ENSG00000169083	V_AR_03	0.00000
ENSG00000064835	V_PIT1_Q6	0.00000
	V_XFD1_01	0.00000
ENSG00000159216	V_AML_Q6	0.00000
ENSG00000109906	V_PLZF_02	0.00000
ENSG00000165556	V_CDX2_01	0.00000
	V_NANOG_01	0.00000
ENSG00000143190	V_OCT1_02	0.00000
ENSG00000066336	V_PU1_01	0.00000
ENSG00000175745	V_COUP_01	0.00000
	V_DMRT7_01	0.00000
	V_POLY_C	0.00000
	V_P53_05	0.00000
ENSG00000068305	V_MMEF2_Q6	0.00000
ENSG00000120075	V_HOXB5_01	0.00000
	V_BARX1_01	0.00000
	V_ISL2_01	0.00000
	V_IRX2_01	0.00000
	V_NKX26_01	0.00000
ENSG00000137203	V_AP2_Q6	0.00000
ENSG00000136997	V_CMYC_02	0.00000
	V_NKX21_01	0.00000
ENSG00000132170	V_PPARG_01	0.00000
	V_HB24_01	0.00000
ENSG00000085276	V_EVI1_01	0.00000
ENSG00000169083	V_AR_04	0.00000
ENSG00000188620	V_HMX3_02	0.00000
ENSG00000198911	V_SREBP2_Q6	0.00000
ENSG00000168610	V_STAT3_03	0.00000
ENSG00000185551	V_DR1_Q3	0.00000
ENSG00000082175	V_PR_01	0.00000

	V_IRX3_01	0.00000
ENSG00000169083	V_AR_01	0.00000
ENSG00000068305	V_MEF2_01	0.00000
ENSG00000101076	V_HNF4_01_B	0.00000
ENSG00000141905	V_MYOGNF1_01	0.00000
ENSG00000125347	V_IRF_Q6	0.00000
ENSG00000135457	V_CP2_02	0.00000
	V_IRX5_01	0.00000
ENSG00000082641	V_TCF11MAFG_01	0.00000
ENSG00000148200	V_GCNF_01	0.00000
ENSG00000115415	V_STAT_Q6	0.00000
ENSG00000159387	V_IRXB3_01	0.00000
ENSG00000025434	V_LXR_Q3	0.00000
ENSG00000100811	V_YY1_02	0.00000
ENSG00000007372	V_PAX6_Q2	0.00000
ENSG00000167182	V_SP2_01	0.00000
ENSG00000141646	V_SMAD4_Q6	0.00000
ENSG00000160224	V_AIRE_01	0.00000
ENSG00000143190	V_OCT1_01	0.00000
ENSG00000068305	V_MEF2_04	0.00000
ENSG00000082175	V_PR_02	0.00000
ENSG00000115415	V_STAT1_05	0.00000
ENSG00000169297	V_DAX1_01	0.00000
ENSG00000143190	V_OCT1_04	0.00000
	V_ETS1_B	0.00000
ENSG00000184486	V_POU3F2_01	0.00000
ENSG00000173153	V_ERR1_Q2	0.00000
	V_ZFP206_01	0.00000
	V_NFY_01	-0.00001
	V_DBX1_01	-0.00001
ENSG00000185551	V_DR4_Q2	-0.00001

Table 7.4 Union of top 20 motifs from three classifications.

V_SP1_Q6
V_ZFX_01
V_FOXO1_Q5
V_ERALPHA_01
V_SP1SP3_Q4
V_CACBINDINGPROTEIN_Q6
V_SP1_Q4_01
V_SP1_Q6_01

V_GTF2IRD1_01
V_MAZ_Q6
V_ZFP281_01
V_SP1_Q2_01
V_FPM315_01
V_SP4_Q5
V_PLAG1_01
V_KROX_Q6
V_FOX_Q2
V_PUR1_Q4
V_SP1_01
V_MUSCLE_INI_B
V_KLF15_Q2
V_LYF1_01
V_MEF2C_01
V_LUN1_01
V_GC_01
V_AP2_Q6_01
V_CKROX_Q2
V_PITX2_Q2
V_IK_Q5
V_RPC155_01
V_SP1_02
V_UF1H3BETA_Q6
V_FOXP1_01
V_ZBP89_Q4
V_PU1_Q4
V_GKLF_02
V_CACD_01
V_LXR_DR4_Q3
V_BDP1_01
V_FKLF_Q5
V_SREBP1_01
V_SREBP_Q6
V_MAF_Q6_01

Table 7.5 List of chromatin modifying enzymes interacting with top 20 available Ensemble Gene Id.

Ensembl Protein ID	Gene Name	Description
ENSP00000080059	HDAC7	histone deacetylase 7
ENSP00000200691	MT3	metallothionein 3
ENSP00000206249	ESR1	estrogen receptor 1

ENSP00000212015	SIRT1	sirtuin 1
ENSP00000221413	RUVBL2	RuvB-like AAA ATPase 2
ENSP00000225916	KAT2A	K(lysine) acetyltransferase 2A
ENSP00000225983	HDAC5	histone deacetylase 5
ENSP00000231487	SKP1	S-phase kinase-associated protein 1
ENSP00000231509	NR3C1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)
ENSP00000243914	CTCFL	CCCTC-binding factor (zinc finger protein)-like
ENSP00000245479	SOX9	SRY (sex determining region Y)-box 9
ENSP00000250003	MYOD1	myogenic differentiation 1
ENSP00000250448	FOXA1	forkhead box A1
ENSP00000257555	HNF1A	HNF1 homeobox A
ENSP00000257745	KMT2E	lysine (K)-specific methyltransferase 2E
ENSP00000260926	SATB2	SATB homeobox 2
ENSP00000262188	SMARCD3	SWI/SNF related, matrix associated, actin dependent chromatin regulator
ENSP00000262367	CREBBP	CREB binding protein
ENSP00000262376	UBN1	ubinuclein 1
ENSP00000262965	TCF3	transcription factor 3
ENSP00000263121	SMARCB1	SWI/SNF related, matrix associated, actin dependent chromatin regulator
ENSP00000263253	EP300	E1A binding protein p300
ENSP00000263360	EED	embryonic ectoderm development
ENSP00000263754	KAT2B	K(lysine) acetyltransferase 2B
ENSP00000264110	ATF2	activating transcription factor 2
ENSP00000264183	ARID4B	AT rich interactive domain 4B (RBP1-like)
ENSP00000264515	RBBP5	retinoblastoma binding protein 5
ENSP00000264606	HDAC4	histone deacetylase 4
ENSP00000264709	DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha
ENSP00000264834	KLF1	Kruppel-like factor 1 (erythroid)
ENSP00000265165	LEF1	lymphoid enhancer-binding factor 1
ENSP00000265773	SMARCA2	SWI/SNF related, matrix associated, actin dependent chromatin regulator
ENSP00000266970	CDK2	cyclin-dependent kinase 2
ENSP00000267163	RB1	retinoblastoma 1
ENSP00000268712	NCOR1	nuclear receptor corepressor 1
ENSP00000274764	HIST1H2BA	histone cluster 1, H2ba
ENSP00000275780	TLK2	tousled-like kinase 2
ENSP00000278616	ATM	ataxia telangiectasia mutated
ENSP00000278916	CHEK1	checkpoint kinase 1

ENSP00000283131	SMARCA5	SWI/SNF related, matrix associated, actin dependent chromatin regulator
ENSP00000284898	L3MBTL4	l(3)mbt-like 4 (Drosophila)
ENSP00000289352	HIST1H4H	histone cluster 1, H4h
ENSP00000296930	NPM1	nucleophosmin (nucleolar phosphoprotein B23, numatrin)
ENSP00000299402	APBB1	amyloid beta (A4) precursor protein-binding, family B, member 1 (Fe65)
ENSP00000299440	RAG1	recombination activating gene 1
ENSP00000301067	KMT2D	lysine (K)-specific methyltransferase 2D
ENSP00000302967	HDAC3	histone deacetylase 3
ENSP00000304004	FOXA3	forkhead box A3
ENSP00000305355	PRKCB	protein kinase C, beta
ENSP00000305899	SUV420H1	suppressor of variegation 4-20 homolog 1 (Drosophila)
ENSP00000307208	BPTF	bromodomain PHD finger transcription factor
ENSP00000307684	TADA3	transcriptional adaptor 3
ENSP00000307803	TET3	tet methylcytosine dioxygenase 3
ENSP00000308227	HMGA1	high mobility group AT-hook 1
ENSP00000308620	RAG2	recombination activating gene 2
ENSP00000309555	HCFC1	host cell factor C1 (VP16-accessory protein)
ENSP00000309992	ZMYND11	zinc finger, MYND-type containing 11
ENSP00000311513	RSF1	remodeling and spacing factor 1
ENSP00000311816	REST	RE1-silencing transcription factor
ENSP00000316578	SUZ12	SUZ12 polycomb repressive complex 2 subunit
ENSP00000318094	SCMH1	sex comb on midleg homolog 1 (Drosophila)
ENSP00000318297	RUVBL1	RuvB-like AAA ATPase 1
ENSP00000320147	EZH2	enhancer of zeste homolog 2 (Drosophila)
ENSP00000320940	NCOA1	nuclear receptor coactivator 1
ENSP00000323967	SMARCE1	SWI/SNF related, matrix associated, actin dependent chromatin regulator
ENSP00000324444	MBIP	MAP3K12 binding inhibitory protein 1
ENSP00000328547	DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta
ENSP00000331614	IKZF1	IKAROS family zinc finger 1 (Ikaros)
ENSP00000333640	EYA2	eyes absent homolog 2 (Drosophila)
ENSP00000337088	MEN1	multiple endocrine neoplasia I
ENSP00000338868	PHF8	PHD finger protein 8
ENSP00000339250	DPPA3	developmental pluripotency associated 3
ENSP00000339992	MYB	v-myb avian myeloblastosis viral oncogene homolog
ENSP00000340896	ASH2L	ash2 (absent, small, or homeotic)-like (Drosophila)
ENSP00000342434	BAZ1B	bromodomain adjacent to zinc finger domain, 1B

ENSP00000342626	EYA1	eyes absent homolog 1 (Drosophila)
ENSP00000343282	HIST1H4D	histone cluster 1, H4d
ENSP00000343325	PKN1	protein kinase N1
ENSP00000346148	PRKAA1	protein kinase, AMP-activated, alpha 1 catalytic subunit
ENSP00000346316	HIST1H4I	histone cluster 1, H4i
ENSP00000346986	WAC	WW domain containing adaptor with coiled-coil
ENSP00000347168	HIST1H4J	histone cluster 1, H4j
ENSP00000347733	TRRAP	transformation/transcription domain-associated protein
ENSP00000348258	HIST1H4L	histone cluster 1, H4l
ENSP00000348610	MED24	mediator complex subunit 24
ENSP00000349213	PBRM1	polybromo 1
ENSP00000349508	CHD4	chromodomain helicase DNA binding protein 4
ENSP00000350681	ELK4	ELK4, ETS-domain protein (SRF accessory protein 1)
ENSP00000350720	SMARCA4	SWI/SNF related, matrix associated, actin dependent chromatin regulator
ENSP00000352516	DNMT1	DNA (cytosine-5-)-methyltransferase 1
ENSP00000354522	TOP1	topoisomerase (DNA) I
ENSP00000354850	MGEA5	meningioma expressed antigen 5 (hyaluronidase)
ENSP00000355153	CDKN2A	cyclin-dependent kinase inhibitor 2A
ENSP00000357311	CENPW	centromere protein W
ENSP00000357965	SETDB1	SET domain, bifurcated 1
ENSP00000358335	MAP3K7	mitogen-activated protein kinase kinase kinase 7
ENSP00000359290	DR1	down-regulator of transcription 1, TBP-binding (negative cofactor 2)
ENSP00000359321	MTF2	metal response element binding transcription factor 2
ENSP00000360163	SMARCA1	SWI/SNF related, matrix associated, actin dependent chromatin regulator
ENSP00000360290	PRKAA2	protein kinase, AMP-activated, alpha 2 catalytic subunit
ENSP00000361066	NCOA3	nuclear receptor coactivator 3
ENSP00000361219	GTF3C4	general transcription factor IIIC, polypeptide 4, 90kDa
ENSP00000362592	RBBP4	retinoblastoma binding protein 4
ENSP00000362649	HDAC1	histone deacetylase 1
ENSP00000362674	HDAC8	histone deacetylase 8
ENSP00000362748	TET1	tet methylcytosine dioxygenase 1
ENSP00000362824	OGT	O-linked N-acetylglucosamine (GlcNAc) transferase
ENSP00000363958	BRD2	bromodomain containing 2
ENSP00000364524	PAX7	paired box 7
ENSP00000364597	PADI4	peptidyl arginine deiminase, type IV
ENSP00000364839	ASXL1	additional sex combs like 1 (Drosophila)

ENSP00000365380	FOXP3	forkhead box P3
ENSP00000365877	SUV39H1	suppressor of variegation 3-9 homolog 1 (Drosophila)
ENSP00000367207	MYC	v-myc avian myelocytomatosis viral oncogene homolog
ENSP00000369351	TET2	tet methylcytosine dioxygenase 2
ENSP00000369681	USP3	ubiquitin specific peptidase 3
ENSP00000369716	CHD3	chromodomain helicase DNA binding protein 3
ENSP00000370343	IRF4	interferon regulatory factor 4
ENSP00000371067	JAK2	Janus kinase 2
ENSP00000376611	TDG	thymine-DNA glycosylase
ENSP00000378414	SMARCD1	SWI/SNF related, matrix associated, actin dependent chromatin regulator
ENSP00000380414	DEK	DEK oncogene
ENSP00000380695	SUDS3	suppressor of defective silencing 3 homolog (S. cerevisiae
ENSP00000381331	HDAC2	histone deacetylase 2
ENSP00000381522	CHD9	chromodomain helicase DNA binding protein 9
ENSP00000382204	JMJD1C	jumonji domain containing 1C
ENSP00000382688	KDM5A	lysine (K)-specific demethylase 5A
ENSP00000384026	HMGA2	high mobility group AT-hook 2
ENSP00000384708	FSHR	follicle stimulating hormone receptor
ENSP00000392028	CHD7	chromodomain helicase DNA binding protein 7
ENSP00000395535	MECP2	methyl CpG binding protein 2 (Rett syndrome)
ENSP00000405574	TBL1XR1	transducin (beta)-like 1 X-linked receptor 1
ENSP00000408617	HDAC9	histone deacetylase 9
ENSP00000418379	TAF1L	TAF1 RNA polymerase II, TBP-associated factor
ENSP00000419494	RYBP	RING1 and YY1 binding protein

7.1.3

7.2 Supplementary for Chapter 3

7.2.1 Supplementary Text

Supplemental Note 1. A model is built using Adaboost method where multiple classifiers are combined to represent the final output of the composite boosted classifier. In this approach, for each weighted bootstrap sample, a new sub-model (in our case a decision tree) is built and added to the model until no further improvement can be made.

In the *Interaction* model, the composite boosted model includes multiple decision trees, each of which captures a set of binding rules based on co-occurring motifs (potential interaction partners or co-factors) in the weighted training sequences bound by the reference TF. Each path from root to leaf in an estimated decision tree sub-model captures one such binding rule, asserting how a combination of motifs and along with their binding affinities relative to thresholds defined by the sub-model contribute to the target TF's binding. Each sub-model is built allowing interaction (tree) depth of 15. We found that 79% of all sub-models include the reference motif. However, this percentage increases up to 85% with increasing interaction depth (Figure 7.3F of Supplemental Data) and no performance loss (Figure 7.3D of Supplemental Data). The sub-models without a reference motif may be explained by the possible absence of the reference motif sequence from the training set due a sequence-length restriction, PWM match threshold, indirect binding, or by other unknown confounders. Notably, by virtue of physical space, the number of non-overlapping features fit in the sequence of restricted length should be limited, e.g. with average size (8bp) of PWM the 15 features need at least 120bp. However, we have ascertained that only 0.8% of all possible single paths (encompassing only ~13% of all the sub-models) have more than 12 features. Thus, in almost all cases, the features fit in a 100bp physical space (12*8=96).

We clustered the sub-models based on feature importance, meaning the contribution of each co-factor in the set of binding rules specified by each decision tree. Therefore, by design, sub-models, common across cell types, will have increased similarity in the set of co-factors and the weight of their contribution whereas cell type-specific sub-models will either have different sets of co-factors, or similar sets of co-factors but with different contribution. For example, none of the CEBPB sub-models have the rule: "presence of IRF8 *and* presence of NFATC4 leads to the binding of CEBPB" (highlighted in Figure 3.1B) except in Gm12878. On the other hand, the following rule exists in multiple sub-models of all cell types except Gm12878 generated sub-models: "presence of CEPBE and one of the reference PWM of CEBPB increases the binding probability of CEBPB". In Figure 7.3A-B of Supplemental Data the corresponding rules are highlighted. Overall, when we looked at the ubiquitous sub-models and cell-specific sub-models, we found that ubiquitous sub-models contribute more co-factors than cell type-specific sub-models (Figure 7.3L of Supplemental Data).

This leads to cell-specific sub-models having a more skewed feature importance than ubiquitous sub-models (Figure 7.3M of Supplemental Data). Furthermore, the co-factors contributed by cell-specific sub-models exhibit a slightly more skewed gene expression across cell types than those contributed by ubiquitous sub-models (Figure 7.3N of Supplemental Data, see Methods for details).

Supplemental Note 2. Each cluster of sub-models can itself serve as a composite, or ensemble, classifier. We determined a cluster-specific score for each TF-bound sequence based on these new cluster-based ensemble classifiers and assigned each sequence to one or more sub-model clusters based on this score (see Methods). Independently, for each TF, we partitioned all bound sequences into those that are bound uniquely in a cell type and those that are bound in multiple cell types. In general, if the clustering of sub-models in different cell types is simply due to sequence sharing then we expect to see a large fraction of overlapping sequence pairs, and not the cell type-specific sequence pairs, assigned to same cluster. We trained EMT using 75% of the sequences in each cell type dataset followed by clustering, and assessed the aforementioned fractions for the remaining 25% of the sequences to avoid training bias. As shown in Figure 7.5 of Supplemental Data, we expect pairs of overlapping sequences to be assigned to the same cluster and hence the size of dark orange box (same cluster) is greater than dark purple box (different cluster). However, many pairs of non-overlapping sequences are also assigned to the same cluster (light orange). We conducted a chi-squared test to assess whether the proportion of non-overlapped sequence pairs assigned to the same cluster is smaller than expected, indicating that co-clustering is driven by sequence overlap. For each cluster, we obtain the proportion of overlapping and non-overlapping sequence pairs and computed the expected proportion from the overall proportion of overlapping vs. non-overlapping sequence pairs. We conducted one chi-squared test per TF using data pooled from all clusters and all cell types, there was no evidence for depletion of non-overlapped sequence pairs assigned to the same cluster (all P-values > 0.05). These results suggest that co-clustering of sub-models across cell types are not simply due to sequence overlap, but rather, represent shared binding rules.

Supplemental Note 3. In clustering the sub-models, our goal was not to find the precise number of distinct binding rules, but rather to assess the modularity and sharing of binding rules across cell types. That's why we decided to choose k in such a way that the coherence among the sub-models in the same clusters is still detectable (i.e. k, not too high) while still revealing the cross-cell type sharing (i.e., k not too low). We checked the value of within-cluster sum of squares (normalized by the cluster-size) for different cluster sizes (Figure 7.7A of Supplemental Data) [153]. For some TFs the suggested clusters seem ~15 (e.g. CTCF), for others ~20 (e.g. FOS), and in extreme cases the desired number of clusters seems to be more than 30 (e.g. ATF3, MYC). Based on these results a cluster size ranging from 15 to 25 seemed a reasonable choice. As a compromise across TFs and to make the analyses comparable we selected k=16 for all TFs.

Supplemental Note 4. We collected 22 position frequency matrices for the zinger motifs reported in [109]. We identified the corresponding TRANSFAC id by matching the PWM-similarity by TFBSTools R package [154]. Allowing 90% (85%) PWM-similarity gave us 16 (42) TRANSFAC ids as zinger motifs; Figure 7.65 of Supplemental Data lists all the zinger TRANSFAC ids. We found that only 5.5% (14%) of the identified the co-factors are zinger motifs suggesting that these motifs have little impact on the models. Moreover, we checked the clustering pattern of the sub-models after removing the zinger motifs and found that the *sparsity* of the cluster-membership matrix is highly correlated with the original clustering pattern (spearman correlation = 0.96, p.value = 2.4×10^{-13}). This suggests that our overall findings are not affected by the zinger motifs.

Supplemental Note 5. It is possible that *EMT*s can falsely yield multiple sub-models, even in absence of heterogeneity, and those sub-models can be falsely clustered. We ascertained heterogeneity across sub-models for a TF from multiple cell types using a *Duda-Hart test* [155] and assessed the clustering tendency of the sub-models in the *d*-dimensional feature space using *Hopkins statistics* [156]. The *Duda-Hart* test verifies whether or not a set of data points should be split into two clusters from the estimate of within-cluster sum of squares for all pairs of clusters versus overall sum of squares; the ratio of the two sum of squares is quantified as the *dh-ratio*; the smaller the value, the greater the clustering. On the other hand, the *Hopkins statistic (H)* compares the nearest neighbor distribution for a random set of points to the same distribution for the clustered sub-models (see Methods). A value close to 0.5 indicates the sub-models are random sets of points with no clustering, a value close to 1 indicates that they form cohesive clusters. Figure 7.7B-C of Supplemental Data summarize the *dh-ratio* and *Hopkins statistic* respectively for 135 TF-cell pairs based on sub-models of TF-cell type pair, and for each TF after gathering all sub-models under a TF. We found that in all cases the *dh-ratio* is less than 1, and the *Hopkins statistic* > 0.5, consistent with heterogeneity; all tests rejected homogeneity (p.value <0.001). Together, the *Duda-Hart test* and *Hopkins statistic* strongly suggest that the sub-models are distinct and cluster-able, i.e., TF binding rules are heterogeneous and partly shared across cell types.

Supplemental Note 6. Figure 7.76H of Supplemental Data shows the distribution of enrichment scores for the co-factors identified per TF. Except CTCF, the minimum of median enrichment score is ~1.2. If we choose a cutoff greater than 1.2, we might lose true positive co-factors for TFs like NRF1, REST, TBP etc. On the other hand, a lower threshold will likely yield many false positive co-factors for the other TFs.

Supplemental Note 7. The enriched GO terms (only biological processes at <=10% false discovery rate) for the cell type-specific co-factors are listed in Table 7.14 from Supplemental Data. Here we discuss the literature evidence supporting the TF functionality in different tissues related to the enriched terms for some of the TFs studied. For other TF's, based on our limited literature survey, we did not find a compelling support for tissue-specific functions of the TF. The following should be considered a selected sampling and an absence of support below should not necessarily be considered as a contradiction.

1. **BHLHE40** – BHLHE40 is known to be associated with many biological processes including circadian rhythm [157], [158], chondrogenesis [159], cell growth, cell differentiation [160], immune response, and apoptosis [161]. Our enrichment analysis of Hepg2 co-factors is consistent with the link between Bhlhe40 expression and hepatic clock and metabolic functions of the liver [162], [163]. Gm12878 co-factors are enriched for cell differentiation, and signaling pathway which are related to inhibition of cell growth and immune response. Enrichment of with BMP response in leukemia cell line is consistent with stimulation of BMP response in certain kinds of leukemia [164].
2. **CEBPB** - The enrichment analysis of Gm12878 co-factors supports the known roles of CEBPB in the "regulation of genes involved in immune and inflammatory responses" [165], "binding to the IL-1 response element in the IL-6 gene, as well as to regulatory regions of several acute-phase and cytokine genes" [166], high induction of CEBPB in blood leukocytes to strengthen muscle [167] etc. Association of CEBPB in AML (Acute myeloid leukemia) [168] is known, where encouragingly coagulation is enriched among the co-factor functions. Studies have found metastasis in Helas3 via ER stress of unfolded protein response [169], [170] and GO analysis shows that the Helas3 co-factors are enriched for ER and unfolded protein response, strongly supporting CEBPB's role. The function of liver and lung depend of the circadian cycle [171], [172].
3. **EP300** - EP300 is acetyl-transferase gene involved in tumor suppression [173], [174], cell proliferation specially in myeloproliferative disorders [175], enhance beta-catenin activity [176], chromatin modelling [173], alu-expression [177], induction of epithelial and mesenchymal proteins and cell-adhesion [178] etc. These are broadly consistent with the enrichment analysis. Enrichment of cell signaling, cell communication in epithelial cancer, limb bud formation in H1hesc, different type of immune and cellular response in normal blood and liver cancer etc. Co-factors identified in Sknsh (brain cancer) are enriched for cortex related hormone-secretion and stimulus, drug response etc. In literature also, there are many evidence about involvement of EP3oo with neuronal disease and its potential as drug for neuronal disorders [179]–[182]. Not surprisingly, EP300 co-factors in liver are involved in response to alcohol and several other metabolic processes. Interestingly, enrichment of several hormone-mediated processes is consistent with the role of EP300 in hepatic encephalopathy [183].
4. **FOS** - FOS processes many extracellular signals via NOTCH signaling [184], or stimulating transcription of AP-1 responsive genes [185]. Therefore, it is not surprising to see enrichment of various type cell-signaling terms among FOS co-factors. FOS is also involved in other cellular events like

differentiation and survival, hypoxia and EMT (epithelial-mesenchymal-transition) [186], metastatis growth in mammary epithelial cells [187], [188]. Furthermore, FOS is a predictor for decreased survival rate in breast cancer [189] and is induced by VEGF which plays an important role in the neovascularization in primary breast cancer [190]. We found that breast-specific co-factors are enriched for organ regeneration.

5. **GABPA** - GABPA is known for maintaining homeostasis [191], mitochondrial respiration [192], and cellular oxidative stress [193]–[195]. The enrichment analysis revealed homeostasis in Gm12878, oxygen-containing compound in Hepg2, and DNA replication in H1hesc, which are consistent with literature. In addition, ETS TF family pays role in the development of vasculature in endothelial cell and its progenitor [196] and we find similar evidence of the role of GABPA, an Ets-family member, in K562.
6. **JUN** - AP-1 (JUN/FOS) complex modulates apoptosis in blood cells [125], controls cell proliferation, cell cycle progression [197], [198], and is involved in angiogenesis [199], [200]. This gene is the putative transforming gene of avian sarcoma virus. We find enrichment of defense mechanism, immune response, homeostasis, estrogen response etc. in blood cells (Gm12878, Huvec, K562). There is also some evidence of involvement of JUN in the development of liver tumor [201], and cervical cancer [202] via co-factors.
7. **MAFK** - MAFK regulates globin genes and plays significant role in coagulation system during embryonic growth and placental development [203]. In addition to that, perturbation in MAFK function is highly associated with carcinogenesis, especially leukemia [204]–[208]. Consistently, we found enrichment of mitotic cell cycle in stem cells, meiosis in liver cancer cell, and response to various metal ion in K562 cell.
8. **MAZ** - MAZ regulates MMP genes, gamma fibrinozen, and serum amyliod A [209], [210] which are consistent with the enriched terms among K562 co-factors, blood coagulation, and hemostasis. Immune and viral response functions among Gm12878 co-factors are supported by the study that MAZ plays functional role in CD4 expression [211].
9. **MYC** - MYC is an oncogene, and in Huvec and Gm12878, we found enrichment of cell cycle check points, DNA damage consistent with its role as oncogene. We found that MYC co-factors of H1hesc are enriched for spinal cord development, glial cell fat regulation, limb bud formation, consistent with its role in determining growth size [212], controlling glial cell in stem cells [213], [214], developing limb link with skeletal size [215]. Gonadotropin up-regulates myeloid protein leukemia-1 [216] and is induced by MYC expression [217]. In addition to that, MYC regulates intestinal intraepithelial lymphocytes and is involved in the homeostasis of adult intestinal epithelium [218], [219]. Consistently, K562 co-factors show enrichment of gonadotropin protein, intestinal epithelial cell differentiation along with cell cycle and cell-cell signaling. MCF co-factors were found to be enriched for viral transcription. It has been shown that knockdown of MYC inhibits breast tumor

growth by RNA interference which treats cancer by viral infection [220], [221] which is consistent with our findings that breats tumor (Mcf7) cell co-factors are enriched with viral transcription and high RNA production by carbon catabolite.

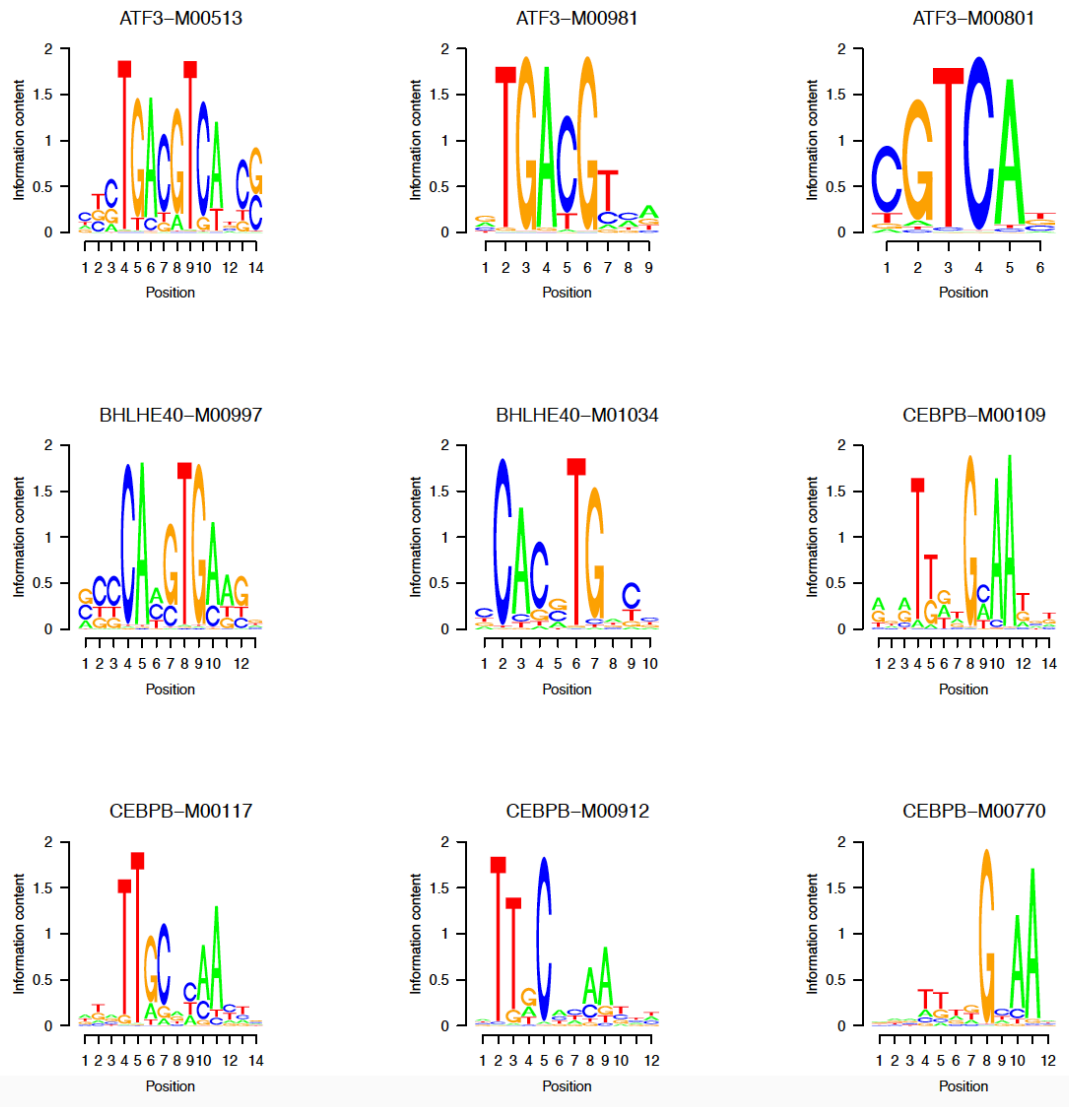
10. **NRF1** - NRF1 activates the expression of key metabolic genes regulating cellular growth and nuclear genes required for mitochondrial respiration [222]–[225]. We found an enrichment of terms related to mitochondrial respiration and biosynthetic process in most cell lines. Interestingly, the enrichment is evident even though cell specific co-factors are used. Interestingly, in K562 NRF1 shows enrichment of a diverse set of terms, several of which are consistent with literature, e.g. association with neurite outgrowth in rodent [226], oxidative stress response [227], mitochondrial biogenesis [228] etc.
11. **REST** - REST acts as a repressor neuronal genes in non-neuronal cell types, and has activation role in neuronal functions [229]–[231]. In the enrichment analysis of REST co-factors, gliblastoma cell line shows cognition, memory, pattern recognition etc. Furthermore, Pancreas cell line shows enrichment for cell differentiation which is consistent with the role of NRSF/REST in pancreas via induction of Pax4 gene [232]. We see extremely high enrichment (>93) of intestinal epithelial cell differentiation among K562 co-factors. We did not find any direct support for this, however, there is evidence that lung and colon epithelial cells show abnormal expression of NRSF in respective cancers [233].
12. **RFX5** - A lack of MHC-II expression results in a severe immunodeficiency syndrome called MHC-II deficiency, or the bare lymphocyte syndrome [234]. Helas3 co-factors are enriched for immune response related terms. RFX5 regulates collagen gene expression [235], which in turn modulate angiogenesis [236]. Consistently, K562 co-factors are enriched with positive regulation of angiogenesis. RFX5 is found to be up-regulated in primary lung budding and mesenchymal cells of branchial arches and stomach in sub-epithelial layer of mouse [237]. Consistently, Hepg2 co-factors are enriched with epithelial tube branching involved in lung morphogenesis. RFX5 complex interacts with the collagen in human fibroblasts [238] and consistently, regulation of fibroblast proliferation is enriched in Gm12878.
13. **USF1** – Upstream regulatory factor 1 is known for regulating multiple genes of glucose and lipid metabolism [239], [240]. In almost all cell lines, the USF1 co-factors are enriched with hormone mediated signaling pathway; especially epithelium cancer cell line (A549) shows enrichment of lipid metabolism related terms. In addition, ovulation, reproductive process, female pregnancy are significantly enriched in A549 co-factors which is consistent with suppression of Follicle-stimulating hormone receptor activity by USF1 [241].
14. **YY1** - YY1 is known as ubiquitous TF. Still the co-factor enrichment analysis shows some of its cell-specific roles. For example, enrichment of epithelial cell maturation in prostate gland development in Gm12878 co-factors [242],

various biosynthetic process in K562 and Hct116 [243], [244] are consistent with literature. Other ontology associated with YY1 co-factors are cell cycle/DNA damage [245], [246]. Nt2d1 and Hepg2 co-factors are enriched with cell cycles and DNA metabolic process respectively.

15. **ZNF143** – ZNF143 plays role as one of the key components of three-dimensional chromatin structure [247], [248], regulates dna-replication and cell-cycle-associated genes [249], [250]. Among the co-factors, we found enrichment of cell-adhesion and cell-proliferation in 3 out of 4 cells.

Supplemental Note 8. We have detailed lists of cell type-specific co-factors for all TFs that we can provide as supplementary online material, and which can serve as a resource for others. Here we discuss a few cases which demonstrate that the co-factors revealed by *TRISECT* are supported by previous experimental research. Recall that P300 is not a typical TF with a DNA binding motif. It nevertheless is expected to interact with other DNA-binding co-factors to achieve specificity. *TRISECT* revealed TEAD as one of the most influential co-factors of P300 in multiple cell types. Indeed, TEAD is known to form a complex with P300 providing locus-specificity to P300 [251]. Likewise, CEBP is known to recruit P300 [252] and ATF interacts with P300's HAT domain [253], and both were detected as P300's ubiquitous co-factors. NR2F2 (also known as COUP-TF2) has a liver-specific function [254] and is known to interact with P300, although this was shown in a different context [255]. It is this interesting that our method detects NR2F2 as Hepg2-specific co-factor of P300. Likewise, members of GATA family are core regulators in liver. We found that in many (but not all) cell lines, and notably in HepG2, members of GATA family are co-factors of P300, consistent with [256]. Serum response factor (SRF) is a ubiquitous protein and with a specific function in liver [257]. FOXA TFs are critical for liver development and function [258]. Our analysis reveals FOXA TFs as HepG2-specific co-factors of SRF. On the other hand, we found ELK4 to be broadly used co-factor of SRF, consistent with their broad expression and known physical interaction with SRF [259]. As yet another example, PAX1 and SOX4 are key TFs in embryogenesis, and both are revealed as co-factors of the core promoter factor *Tata Binding Protein* (TBP) specifically in hESC.

7.2.2 Supplemental Data


ATF3–M00513

ATF3–M00981

ATF3–M00801

BHLHE40–M00997

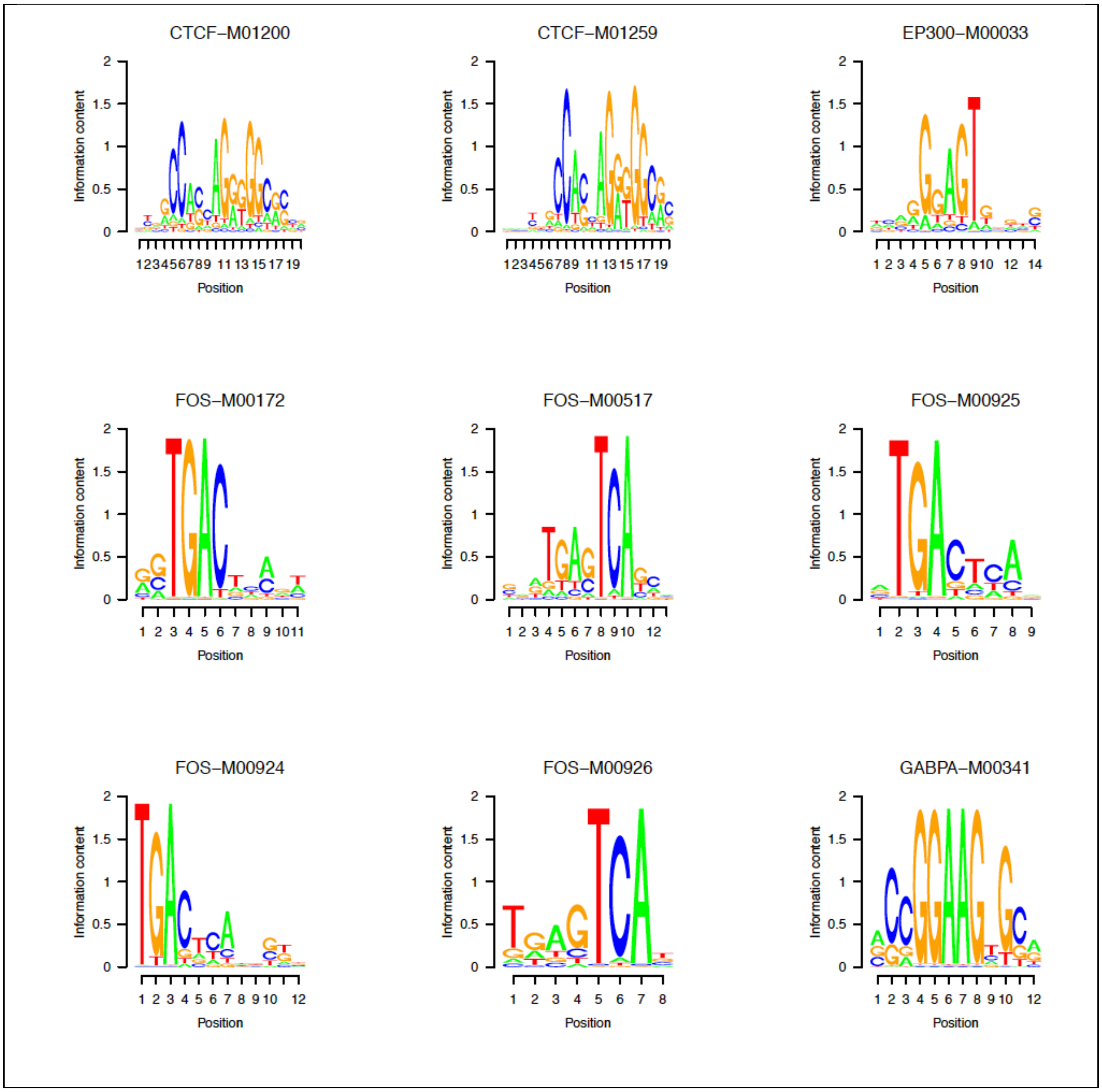
BHLHE40–M01034

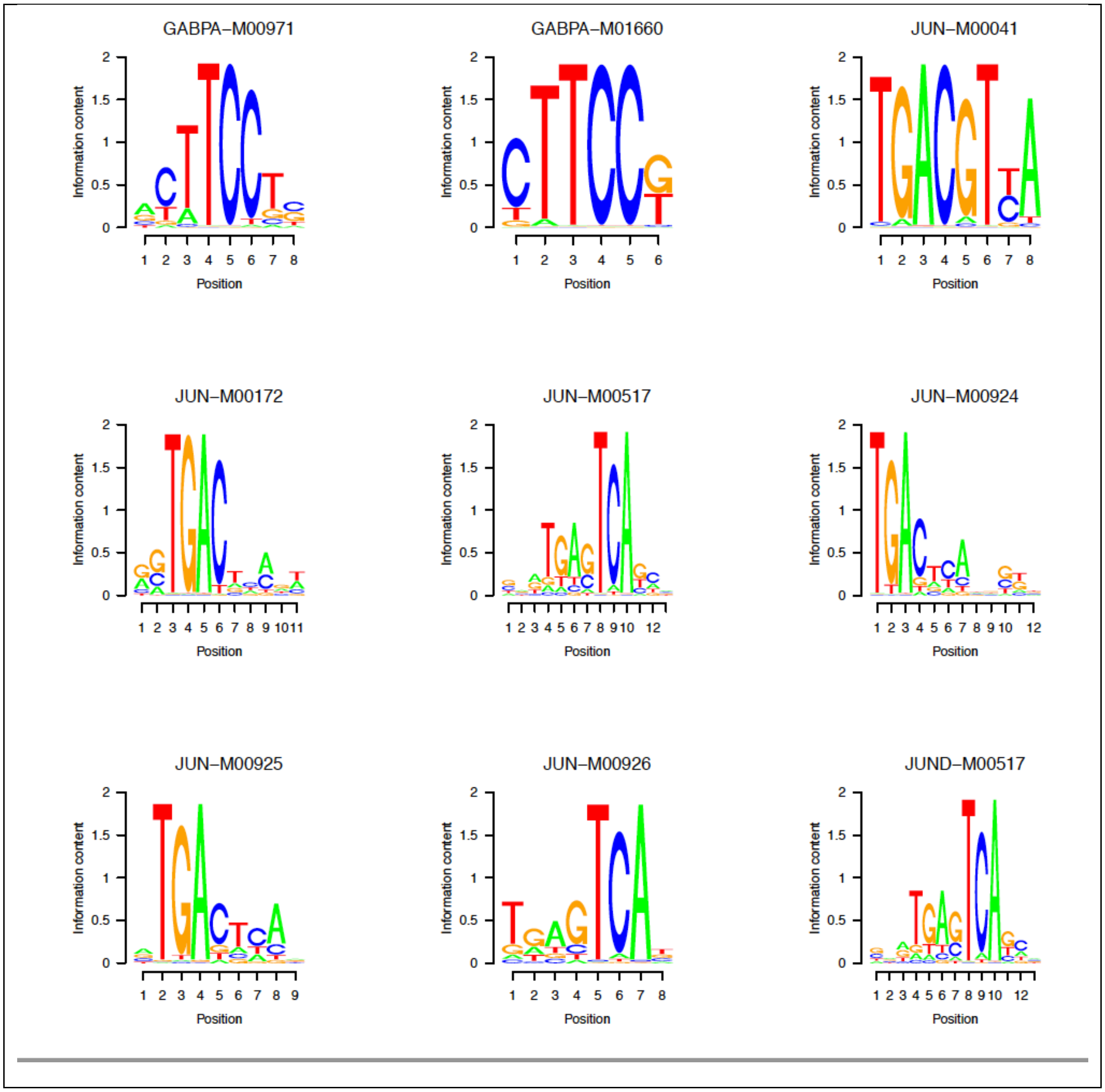
CEBPB–M00109

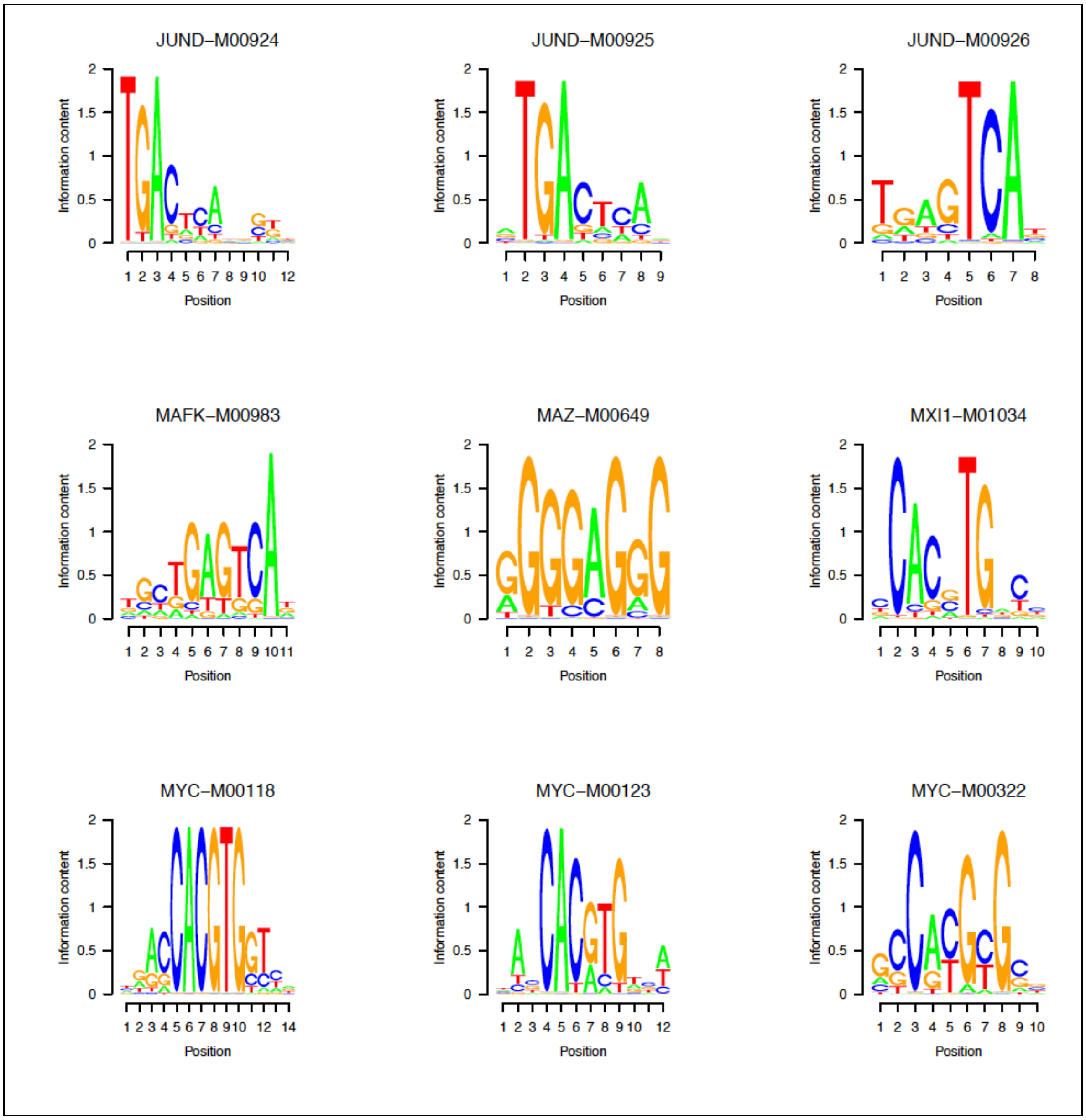
CEBPB–M00117

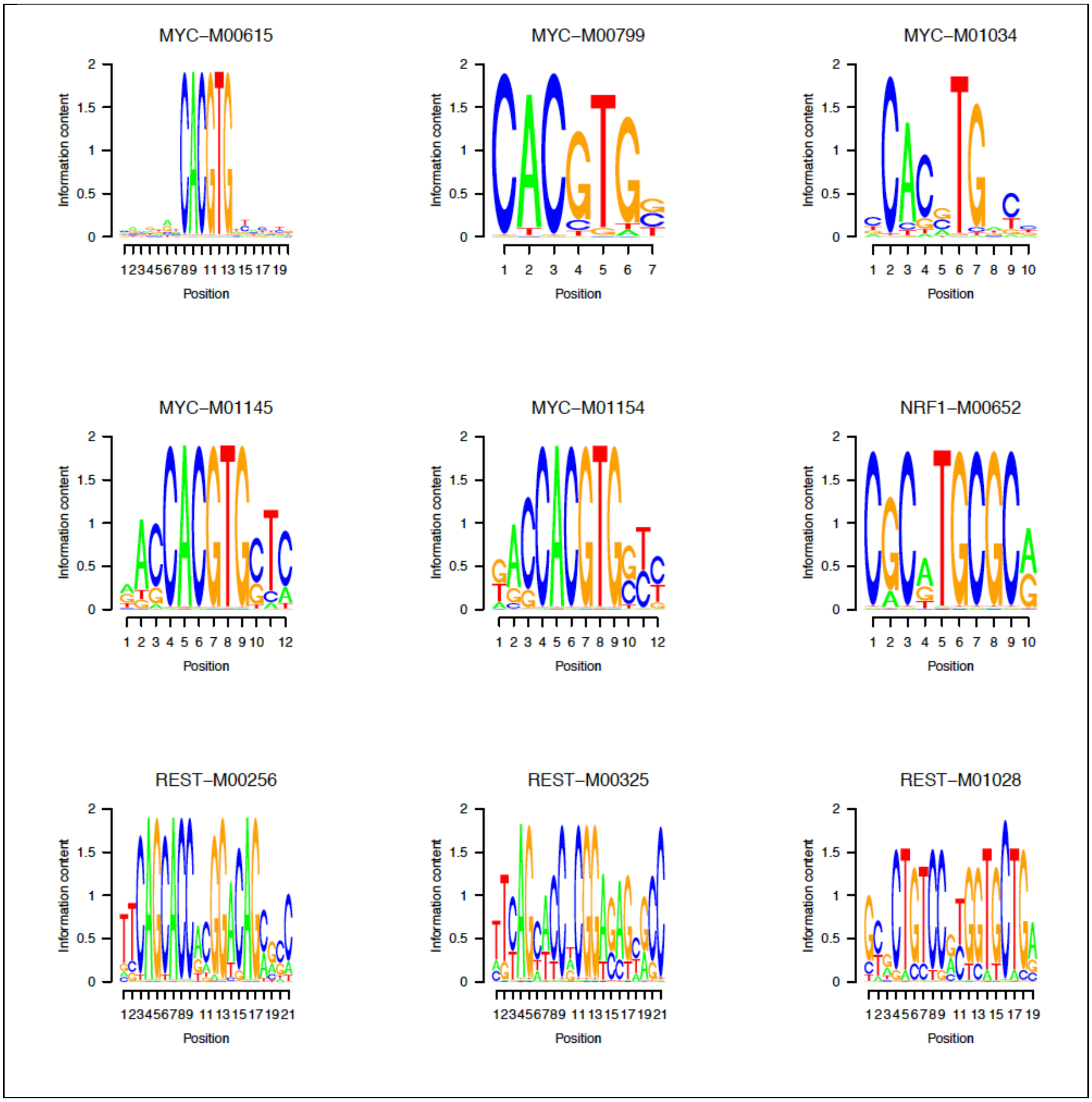
CEBPB–M00912

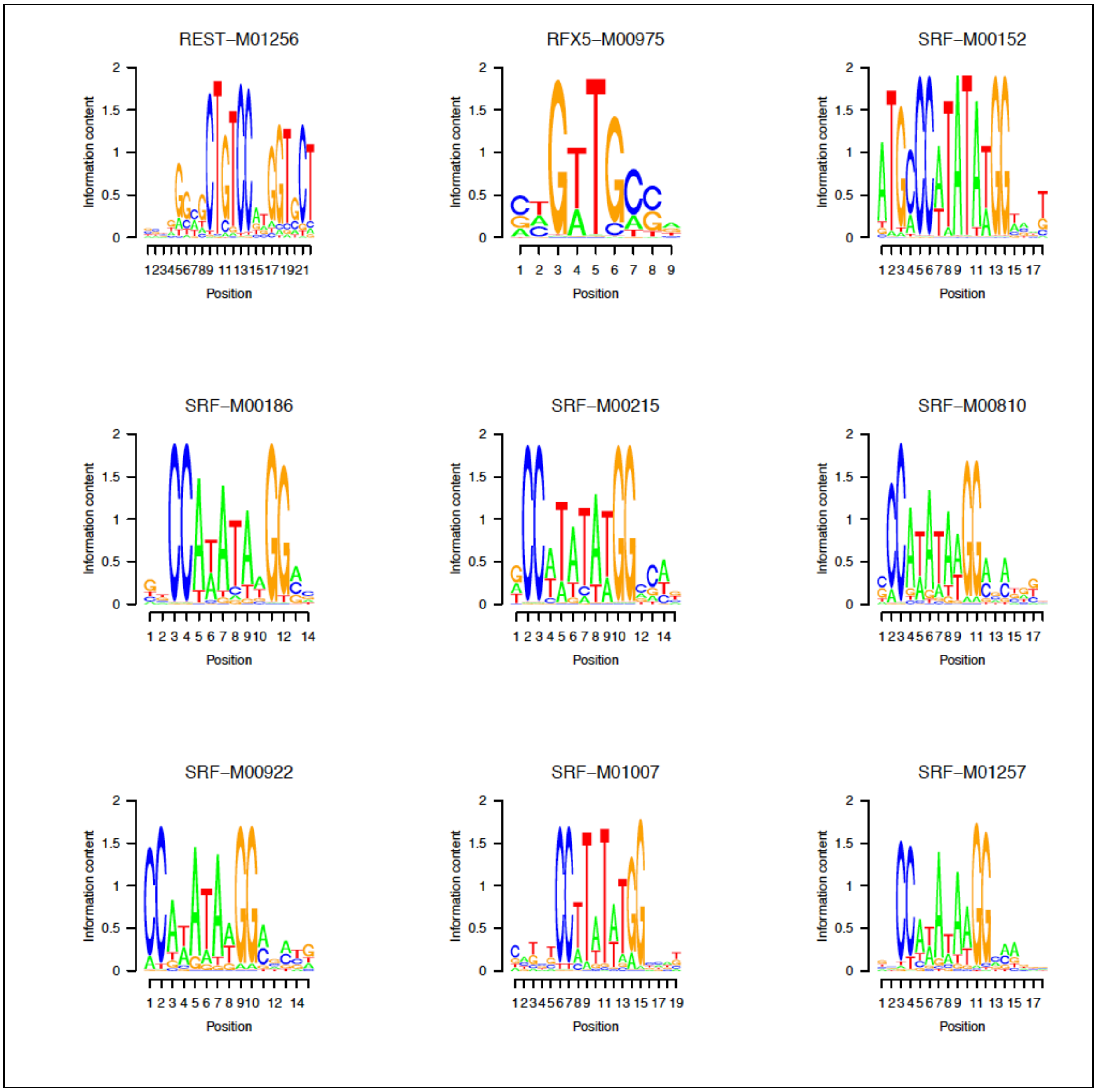
CEBPB–M00770

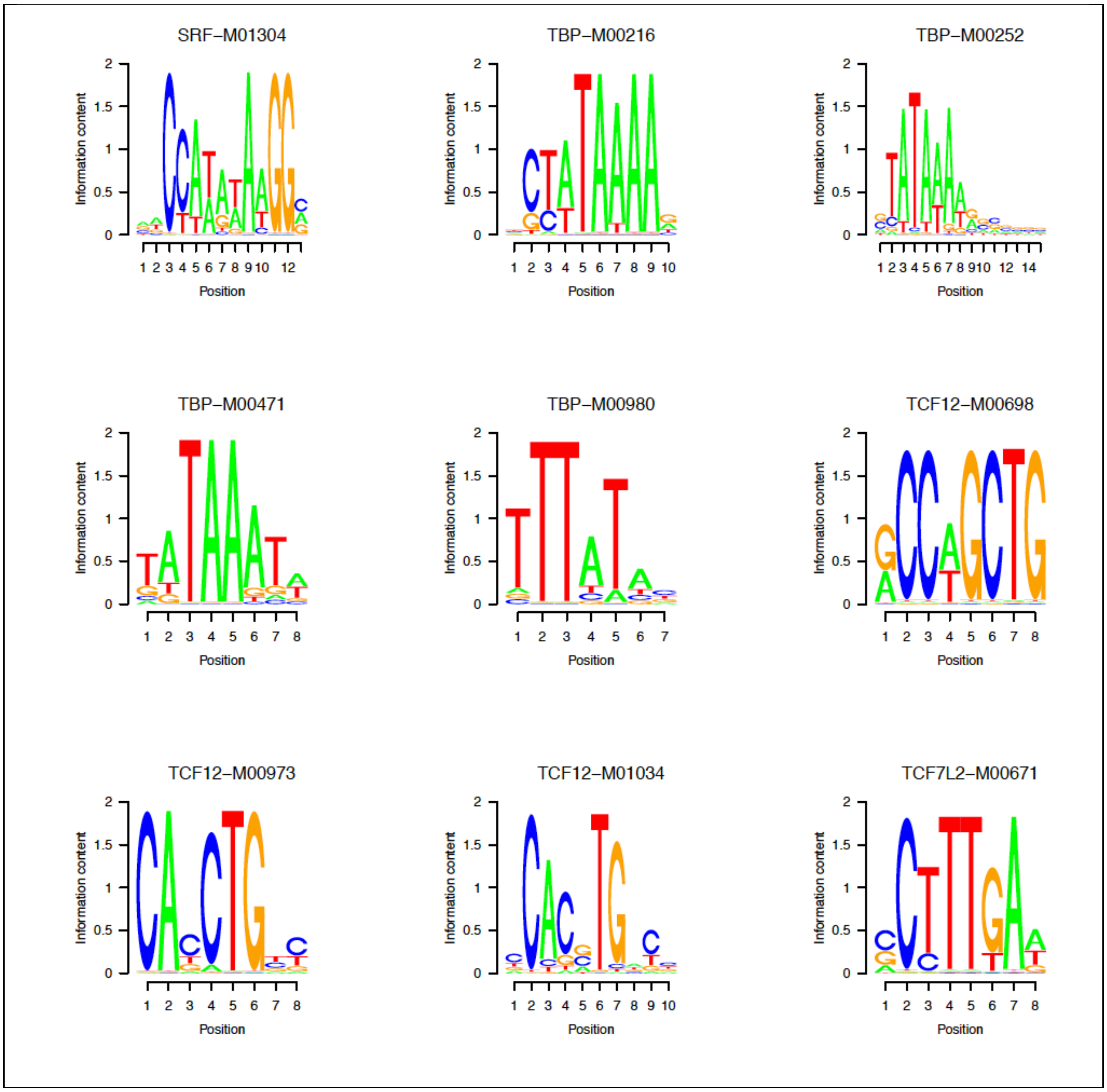

CTCF–M01200
CTCF–M01259
EP300–M00033
FOS–M00172
FOS–M00517
FOS–M00925
FOS–M00924
FOS–M00926
GABPA–M00341
Information content
0
0.5
1
1.5
2
Position

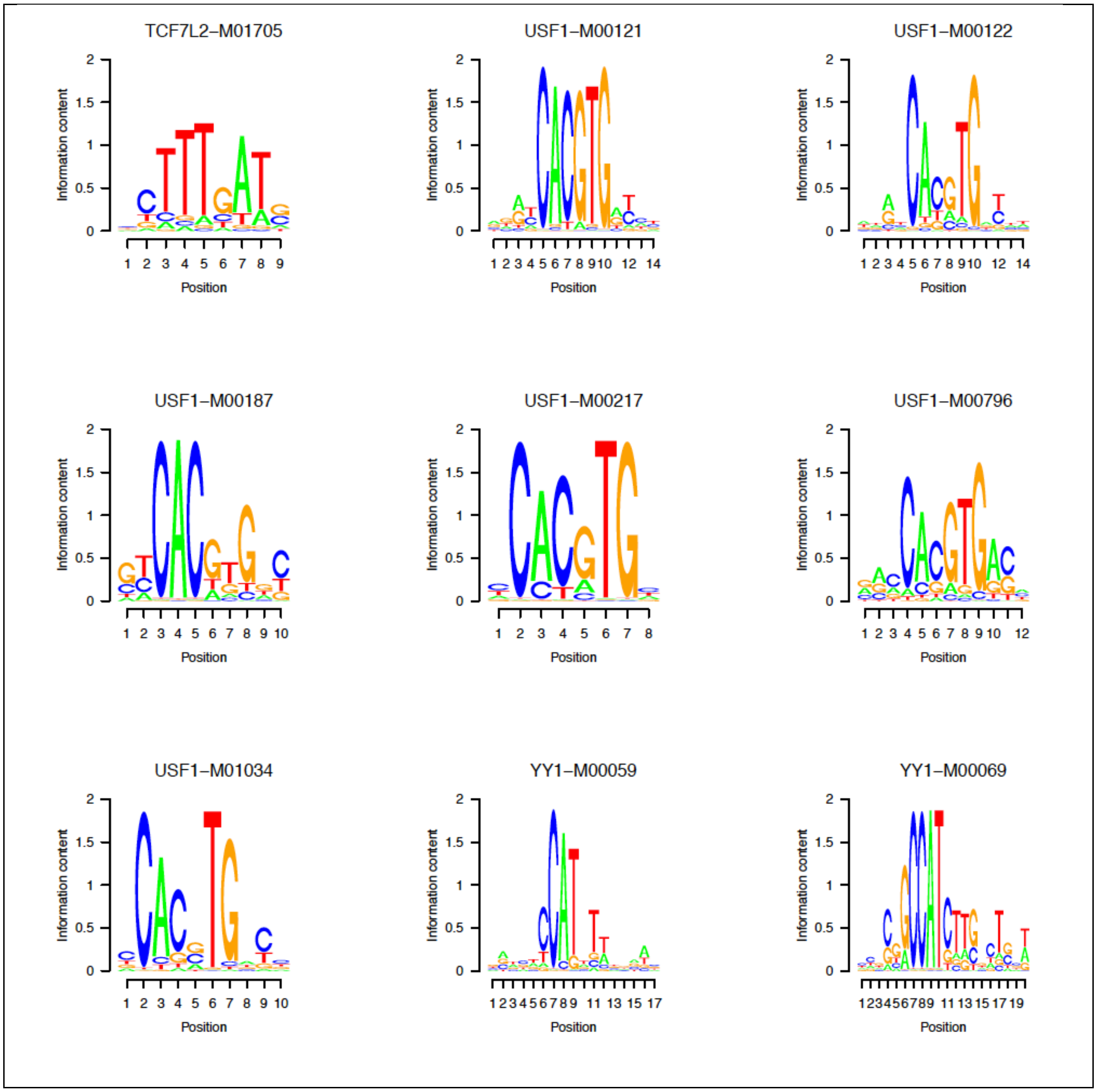

GABPA–M00971
GABPA–M01660
JUN–M00041
JUN–M00172
JUN–M00517
JUN–M00924
JUN–M00925
JUN–M00926
JUND–M00517
Information content
0
0.5
1
1.5
2
Position


JUND–M00924
JUND–M00925
JUND–M00926
MAFK–M00983
MAZ–M00649
MXI1–M01034
MYC–M00118
MYC–M00123
MYC–M00322
Information content
0
0.5
1
1.5
2
Position


MYC–M00615
MYC–M00799
MYC–M01034
MYC–M01145
MYC–M01154
NRF1–M00652
REST–M00256
REST–M00325
REST–M01028
Information content
0
0.5
1
1.5
2
Position


REST–M01256
RFX5–M00975
SRF–M00152
SRF–M00186
SRF–M00215
SRF–M00810
SRF–M00922
SRF–M01007
SRF–M01257
Information content
0
0.5
1
1.5
2
Position


SRF–M01304
TBP–M00216
TBP–M00252
TBP–M00471
TBP–M00980
TCF12–M00698
TCF12–M00973
TCF12–M01034
TCF7L2–M00671
Information content
0
0.5
1
1.5
2
Position
1 2 3 4 5 6 7 8 9 10 12
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10 12 14
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7



TCF7L2–M01705
USF1–M00121
USF1–M00122
USF1–M00187
USF1–M00217
USF1–M00796
USF1–M01034
YY1–M00059
YY1–M00069
Information content
0
0.5
1
1.5
2
Position

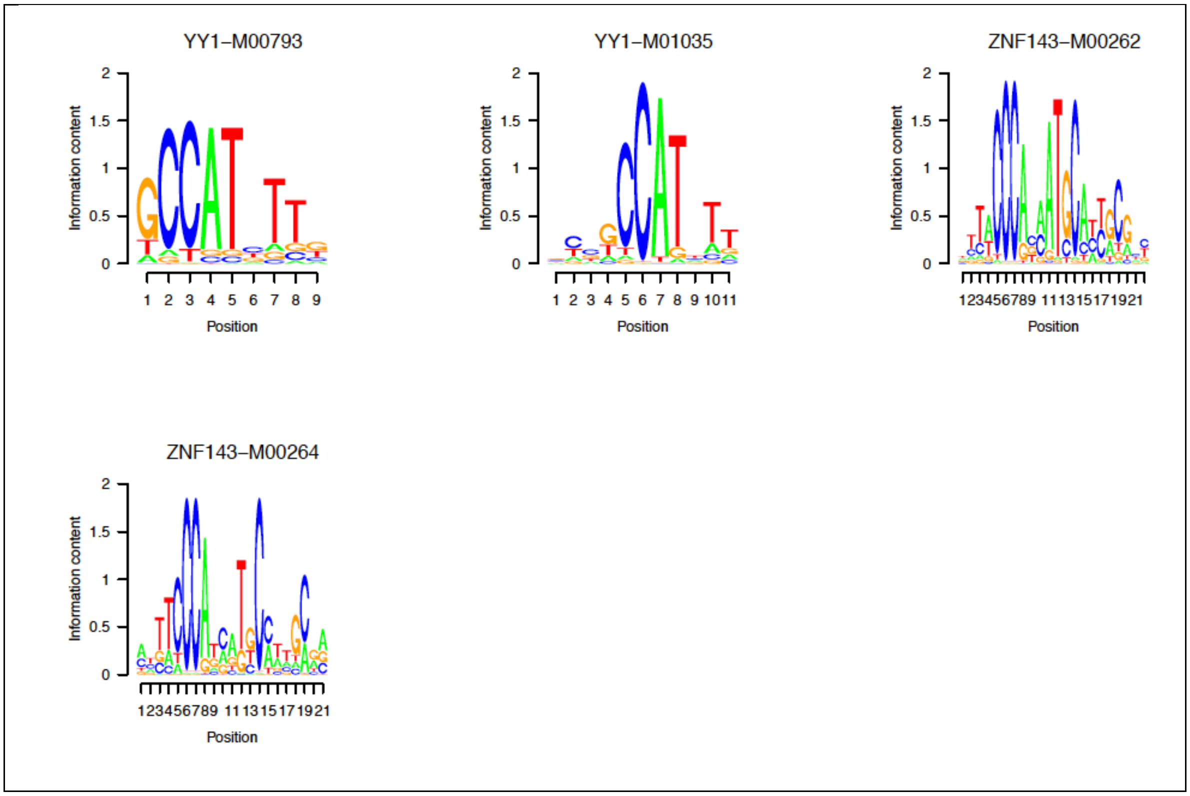
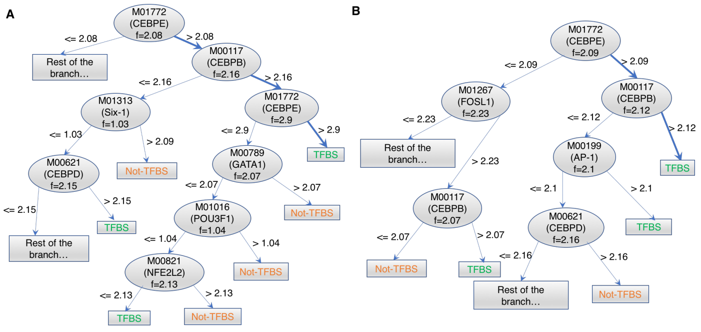
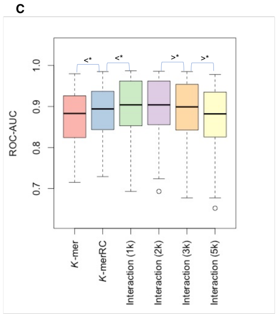


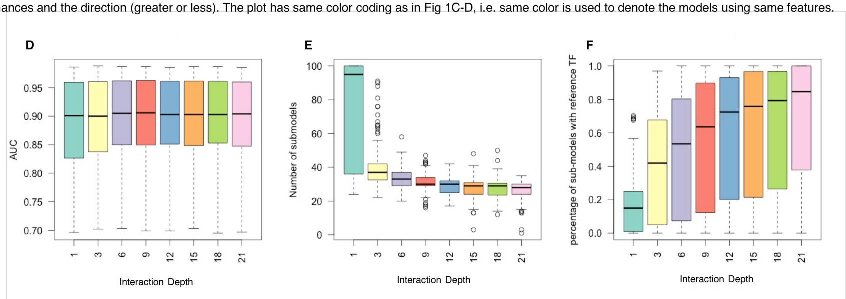
Figure 7.2 Web-logos of position weight matrices for 23 TFs investigated. Each weblogo is labeled with the TF name and TRANSFAC id.



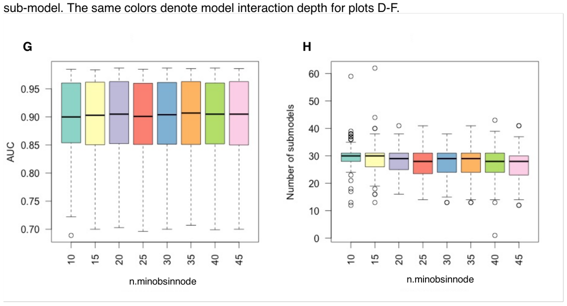


(A-B). Part of a sub-model taken from the *Interaction* model for CEBPB in H1hesc and Helas3 respectively. Each node in the tree is labeled with the TRANSFAC id, corresponding gene name and the threshold at which the feature is split. In each tree, one binding rule is highlighted: they are identical with respect to leading or increasing the probability of leading the binding probability of CEBPB.

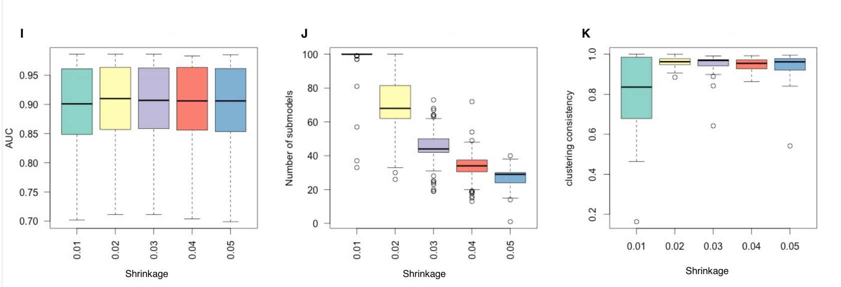
(C). Accuracy (ROC-AUC) distribution for 6 choices of EMT feature sets. "<*" (">*") denotes significant difference (one sided Wilcoxon p-value < 0.05) between the two sets of performances and the direction (greater or less). The plot has same color coding as in Fig 1C-D, i.e. same color is used to denote the models using same features.



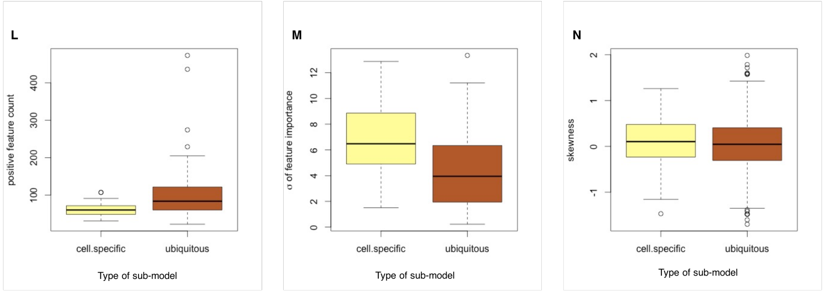
(D-F). ROC-AUC (D), number of sub-models (E), and fraction of sub-models (F) that included reference TFs, for different interaction depth of the models. Interaction depth defines maximum allowable features in a sub-model. The same colors denote model interaction depth for plots D-F.



(G-H). ROC-AUC (G), and number of sub-models (H) for different n.minobsinnode of the models. 'n.minobsinnode' denotes minimum number of observations made at each node while building the decision tree. Plots G and H use same the colors to indicate models using 'n.minobsinnode'.

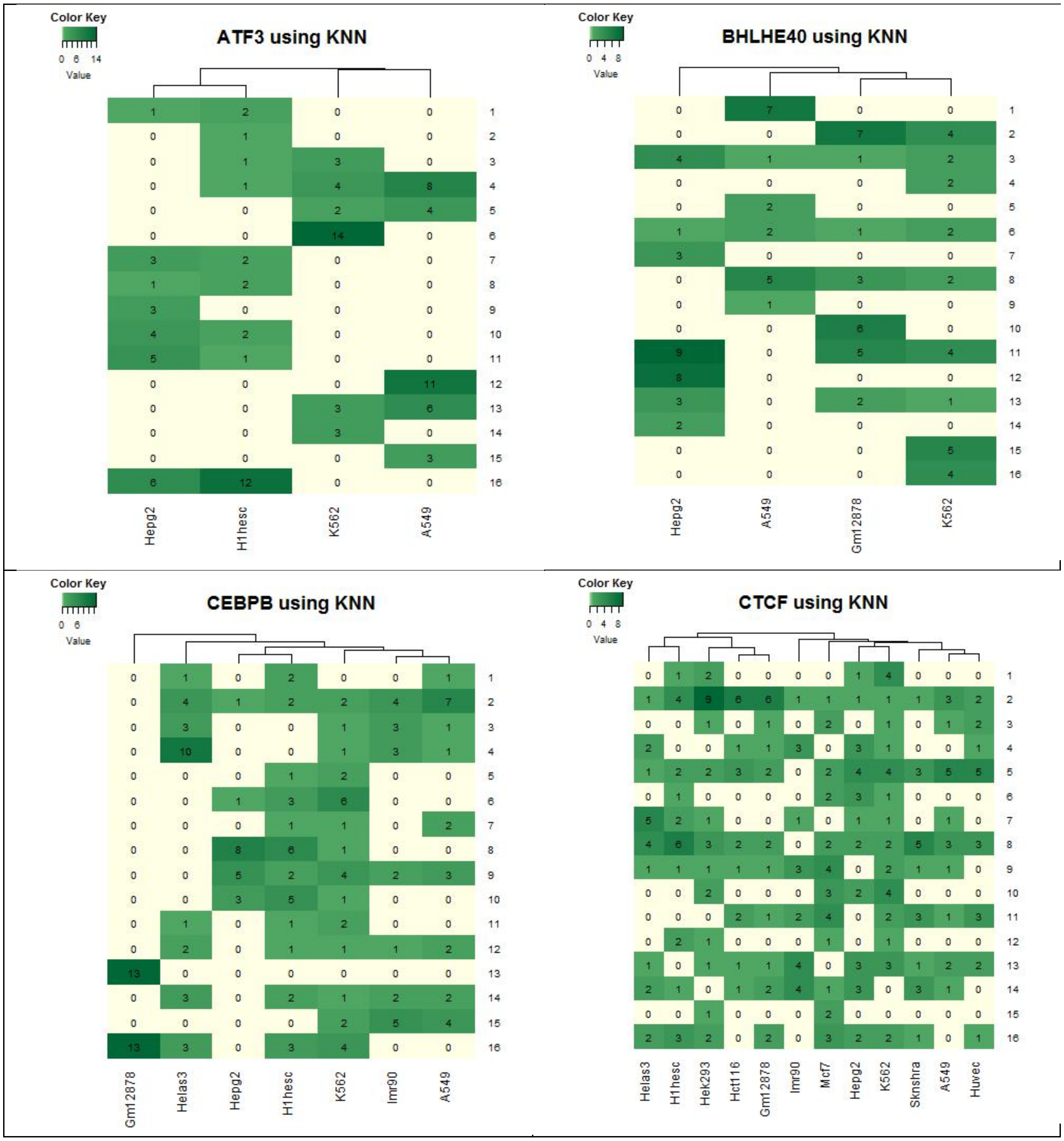


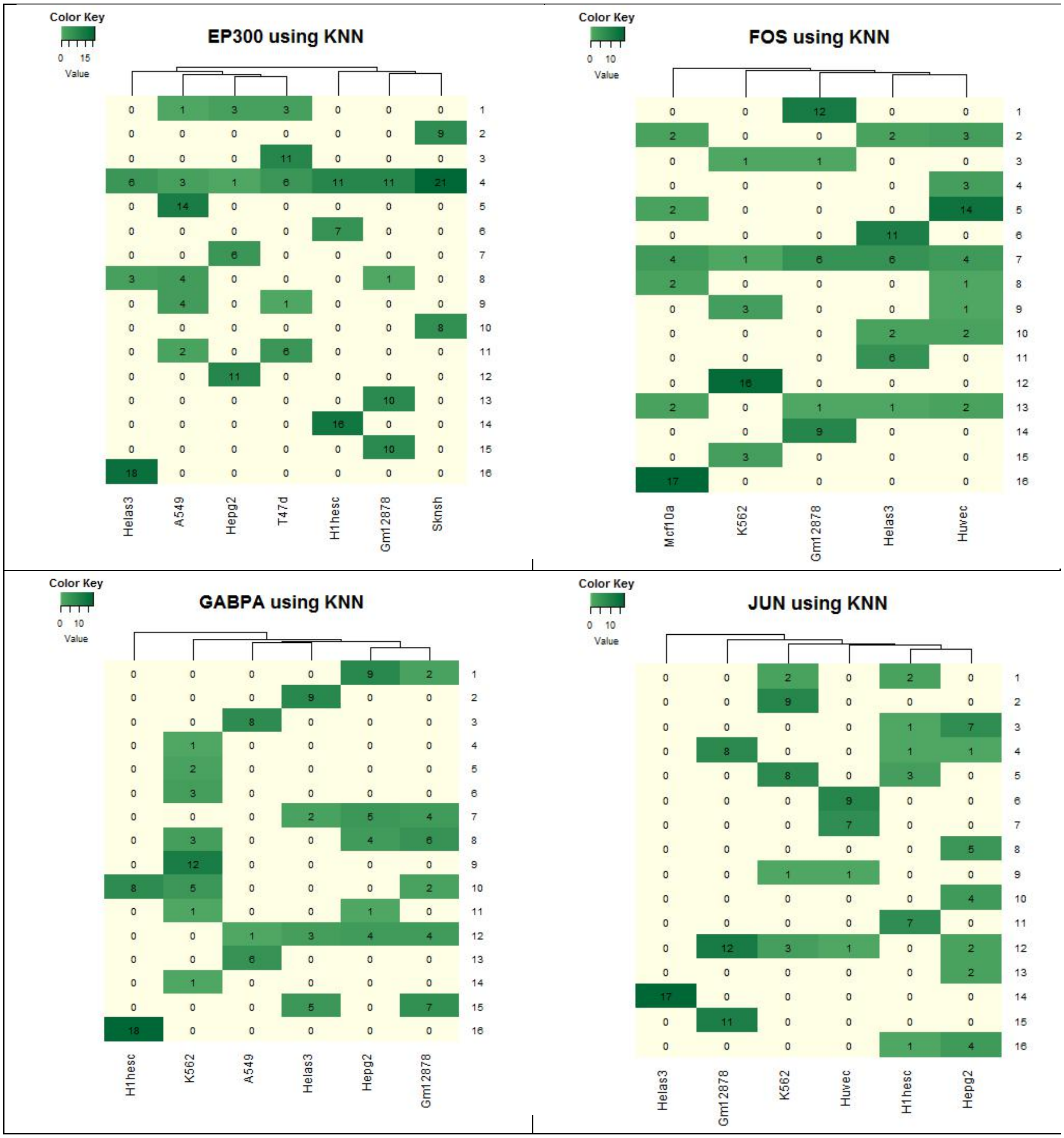
(I-K). ROC-AUC (I), number of sub-models (J), and clustering consistency (K) of the sequences (percentage of sequence-pairs that fall in same clusters) for different values of the shrinkage parameter. Shrinkage indicates the learning rate of the model. Plots I-K the use same colors to denote models using same learning rate.



(L-M). Comparison of ubiquitous and cell type-specific sub-models: (L) Number of relevant features (i.e. features with non-zero importance in any cell type-specific model), (M) Standard deviation of feature importance for each sub-model, (N) Skewness of gene expression in each cells for the co-factors. Yellow and brown colors denote ubiquitous and cell type-specific sub-models respectively.

Figure 7.3 Robustness of TRISECT.



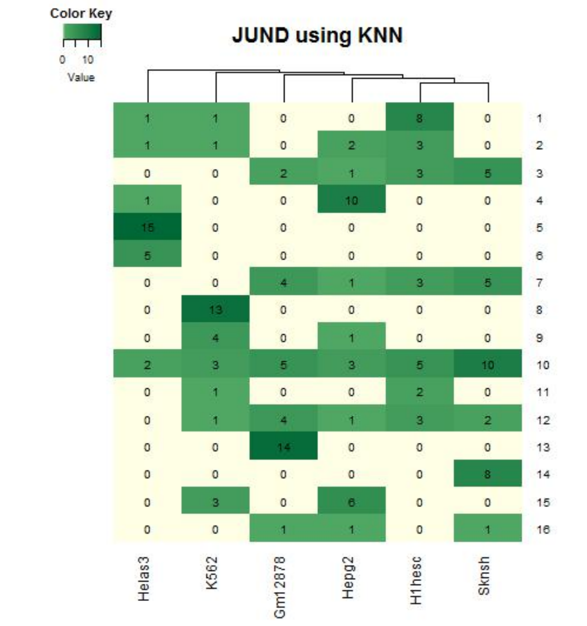


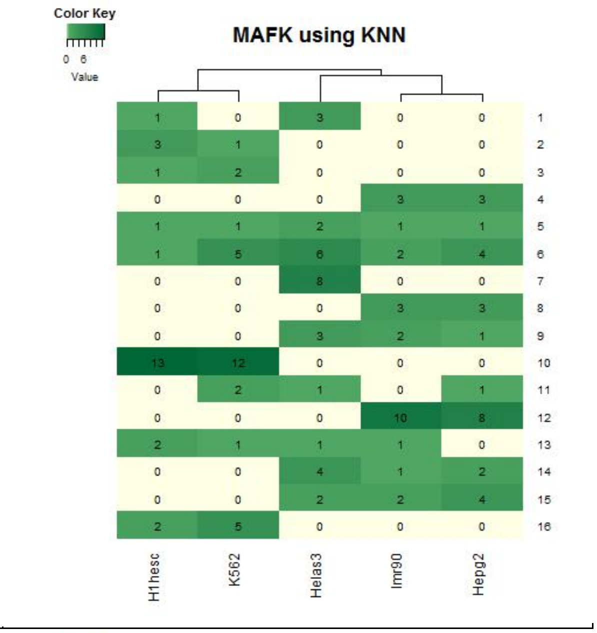
	Helas3	A549	Hepg2	T47d	H1hesc	Gm12878	Sknsh
1	0	1	3	3	0	0	0
2	0	0	0	0	0	0	9
3	0	0	0	11	0	0	0
4	6	3	1	6	11	11	21
5	0	14	0	0	0	0	0
6	0	0	0	0	7	0	0
7	0	0	6	0	0	0	0
8	3	4	0	0	0	1	0
9	0	4	0	1	0	0	0
10	0	0	0	0	0	0	8
11	0	2	0	6	0	0	0
12	0	0	11	0	0	0	0
13	0	0	0	0	0	10	0
14	0	0	0	0	16	0	0
15	0	0	0	0	0	10	0
16	18	0	0	0	0	0	0

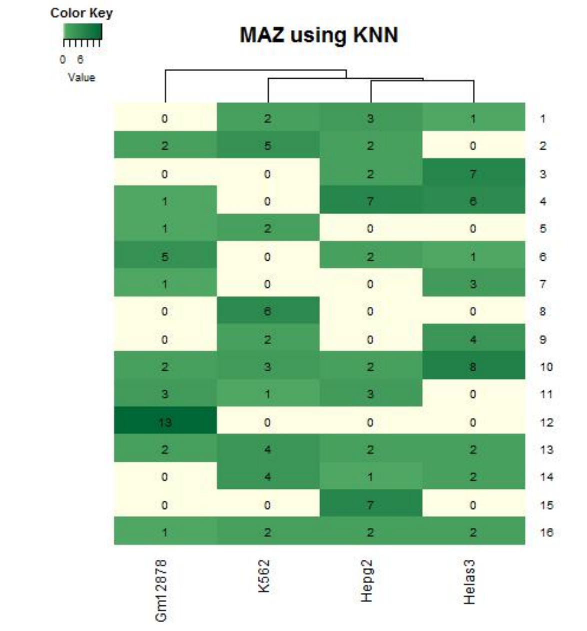
	Mcf10a	K562	Gm12878	Helas3	Huvec
1	0	0	12	0	0
2	2	0	0	2	3
3	0	1	1	0	0
4	0	0	0	0	3
5	2	0	0	0	14
6	0	0	0	11	0
7	4	1	6	6	4
8	2	0	0	0	1
9	0	3	0	0	1
10	0	0	0	2	2
11	0	0	0	6	0
12	0	16	0	0	0
13	2	0	1	1	2
14	0	0	9	0	0
15	0	3	0	0	0
16	17	0	0	0	0

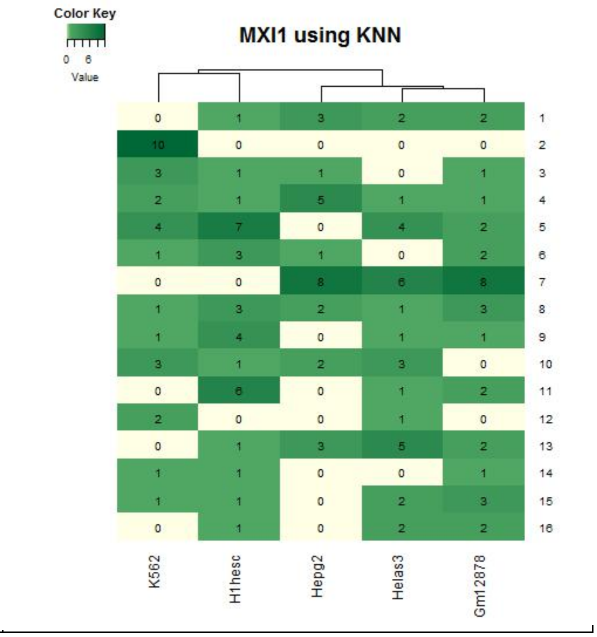
	H1hesc	K562	A549	Helas3	Hepg2	Gm12878
1	0	0	0	0	9	2
2	0	0	0	9	0	0
3	0	0	8	0	0	0
4	0	1	0	0	0	0
5	0	2	0	0	0	0
6	0	3	0	0	0	0
7	0	0	0	2	5	4
8	0	3	0	0	4	6
9	0	12	0	0	0	0
10	8	5	0	0	0	2
11	0	1	0	0	1	0
12	0	0	1	3	4	4
13	0	0	6	0	0	0
14	0	1	0	0	0	0
15	0	0	0	5	0	7
16	18	0	0	0	0	0

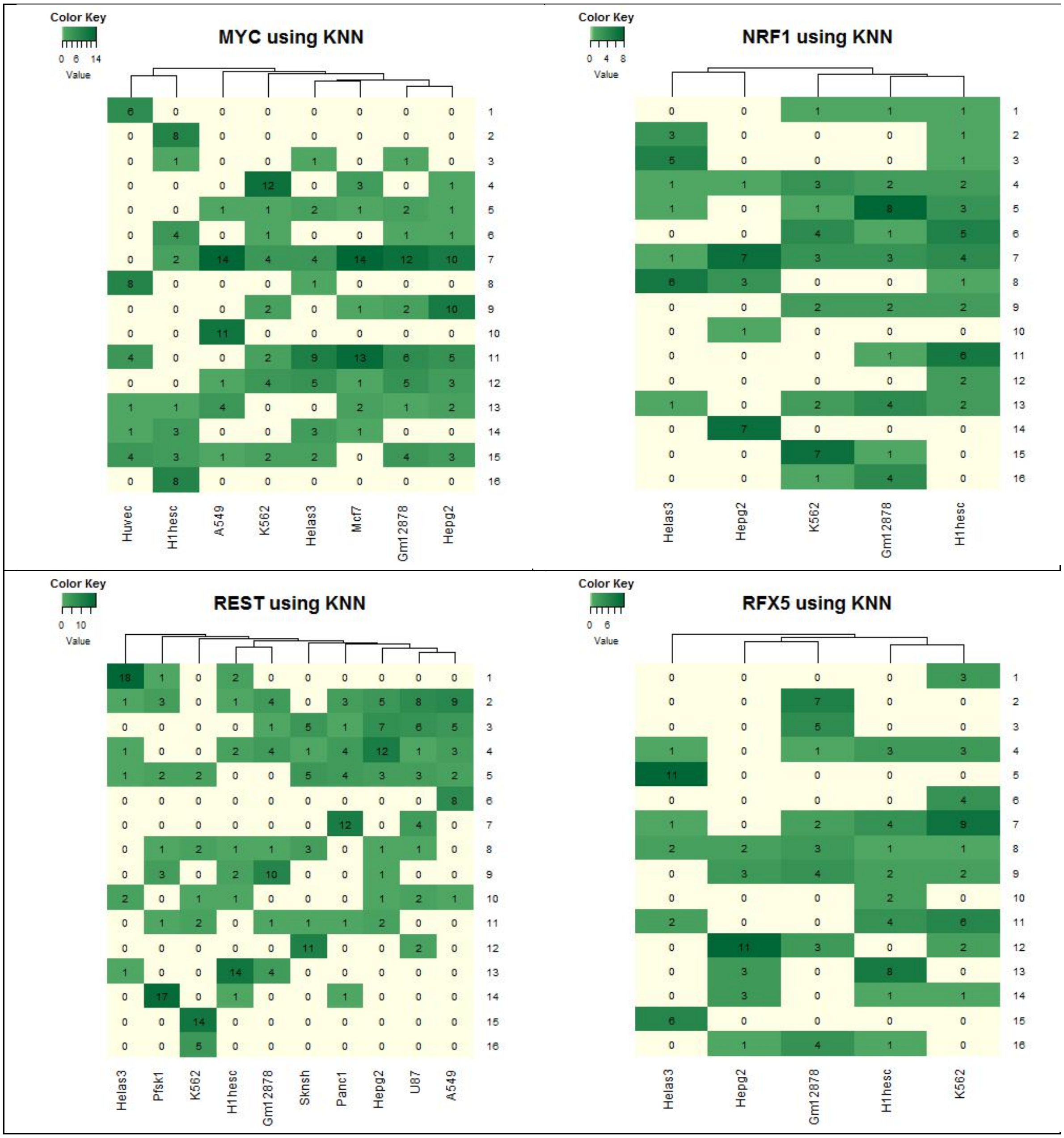
	Helas3	Gm12878	K562	Huvec	H1hesc	Hepg2
1	0	0	2	0	2	0
2	0	0	9	0	0	0
3	0	0	0	0	1	7
4	0	8	0	0	1	1
5	0	0	8	0	3	0
6	0	0	0	9	0	0
7	0	0	0	7	0	0
8	0	0	0	0	0	5
9	0	0	1	1	0	0
10	0	0	0	0	0	4
11	0	0	0	0	7	0
12	0	12	3	1	0	2
13	0	0	0	0	0	2
14	17	0	0	0	0	0
15	0	11	0	0	0	0
16	0	0	0	0	1	4


Color Key
0 10
Value
JUND using KNN
1 1 0 0 8 0 1
1 1 0 2 3 0 2
0 0 2 1 3 5 3
1 0 0 10 0 0 4
15 0 0 0 0 0 5
5 0 0 0 0 0 6
0 0 4 1 3 5 7
0 13 0 0 0 0 8
0 4 0 1 0 0 9
2 3 5 3 5 10 10
0 1 0 0 2 0 11
0 1 4 1 3 2 12
0 0 14 0 0 0 13
0 0 0 0 0 8 14
0 3 0 6 0 0 15
0 0 1 1 0 1 16
Helas3
K562
Gm12878
Hepg2
H1hesc
Sknsh


Color Key
0 6
Value
MAFK using KNN
1 0 3 0 0 1
3 1 0 0 0 2
1 2 0 0 0 3
0 0 0 3 3 4
1 1 2 1 1 5
1 5 6 2 4 6
0 0 8 0 0 7
0 0 0 3 3 8
0 0 3 2 1 9
13 12 0 0 0 10
0 2 1 0 1 11
0 0 0 10 8 12
2 1 1 1 0 13
0 0 4 1 2 14
0 0 2 2 4 15
2 5 0 0 0 16
H1hesc
K562
Helas3
Imr90
Hepg2


Color Key
0 6
Value
MAZ using KNN
0 2 3 1 1
2 5 2 0 2
0 0 2 7 3
1 0 7 6 4
1 2 0 0 5
5 0 2 1 6
1 0 0 3 7
0 6 0 0 8
0 2 0 4 9
2 3 2 8 10
3 1 3 0 11
13 0 0 0 12
2 4 2 2 13
0 4 1 2 14
0 0 7 0 15
1 2 2 2 16
Gm12878
K562
Hepg2
Helas3


Color Key
0 6
Value
MXI1 using KNN
0 1 3 2 2 1
10 0 0 0 0 2
3 1 1 0 1 3
2 1 5 1 1 4
4 7 0 4 2 5
1 3 1 0 2 6
0 0 8 6 8 7
1 3 2 1 3 8
1 4 0 1 1 9
3 1 2 3 0 10
0 6 0 1 2 11
2 0 0 1 0 12
0 1 3 5 2 13
1 1 0 0 1 14
1 1 0 2 3 15
0 1 0 2 2 16
K562
H1hesc
Hepg2
Helas3
Gm12878

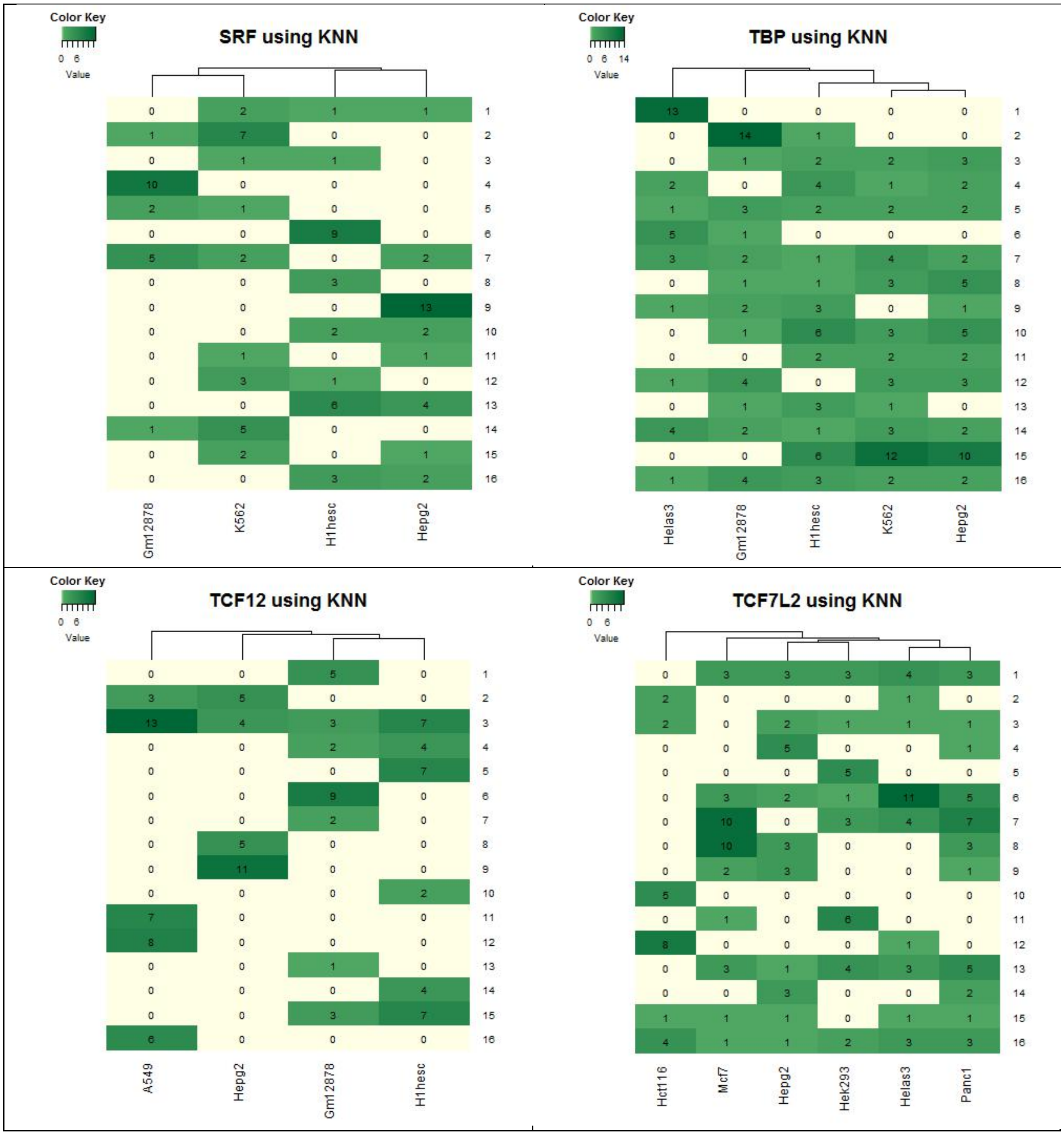


Huvec	H1hesc	A549	K562	Helas3	Mcf7	Gm12878	Hepg2	
6	0	0	0	0	0	0	0	1
0	8	0	0	0	0	0	0	2
0	1	0	0	1	0	1	0	3
0	0	0	12	0	3	0	1	4
0	0	1	1	2	1	2	1	5
0	4	0	1	0	0	1	1	6
0	2	14	4	4	14	12	10	7
8	0	0	0	1	0	0	0	8
0	0	0	2	0	1	2	10	9
0	0	11	0	0	0	0	0	10
4	0	0	2	9	13	6	5	11
0	0	1	4	5	1	5	3	12
1	1	4	0	0	2	1	2	13
1	3	0	0	3	1	0	0	14
4	3	1	2	2	0	4	3	15
0	8	0	0	0	0	0	0	16

Helas3	Hepg2	K562	Gm12878	H1hesc	
0	0	1	1	1	1
3	0	0	0	1	2
5	0	0	0	1	3
1	1	3	2	2	4
1	0	1	8	3	5
0	0	4	1	5	6
1	7	3	3	4	7
6	3	0	0	1	8
0	0	2	2	2	9
0	1	0	0	0	10
0	0	0	1	6	11
0	0	0	0	2	12
1	0	2	4	2	13
0	7	0	0	0	14
0	0	7	1	0	15
0	0	1	4	0	16

Helas3	Pfsk1	K562	H1hesc	Gm12878	Sknsh	Panc1	Hepg2	U87	A549	
18	1	0	2	0	0	0	0	0	0	1
1	3	0	1	4	0	3	5	8	9	2
0	0	0	0	1	5	1	7	6	5	3
1	0	0	2	4	1	4	12	1	3	4
1	2	2	0	0	5	4	3	3	2	5
0	0	0	0	0	0	0	0	0	8	6
0	0	0	0	0	0	12	0	4	0	7
0	1	2	1	1	3	0	1	1	0	8
0	3	0	2	10	0	0	1	0	0	9
2	0	1	1	0	0	0	1	2	1	10
0	1	2	0	1	1	1	2	0	0	11
0	0	0	0	0	11	0	0	2	0	12
1	0	0	14	4	0	0	0	0	0	13
0	17	0	1	0	0	1	0	0	0	14
0	0	14	0	0	0	0	0	0	0	15
0	0	5	0	0	0	0	0	0	0	16

Helas3	Hepg2	Gm12878	H1hesc	K562	
0	0	0	0	3	1
0	0	7	0	0	2
0	0	5	0	0	3
1	0	1	3	3	4
11	0	0	0	0	5
0	0	0	0	4	6
1	0	2	4	9	7
2	2	3	1	1	8
0	3	4	2	2	9
0	0	0	2	0	10
2	0	0	4	6	11
0	11	3	0	2	12
0	3	0	8	0	13
0	3	0	1	1	14
6	0	0	0	0	15
0	1	4	1	0	16



SRF using KNN

Gm12878	K562	H1hesc	Hepg2	
0	2	1	1	1
1	7	0	0	2
0	1	1	0	3
10	0	0	0	4
2	1	0	0	5
0	0	9	0	6
5	2	0	2	7
0	0	3	0	8
0	0	0	13	9
0	0	2	2	10
0	1	0	1	11
0	3	1	0	12
0	0	6	4	13
1	5	0	0	14
0	2	0	1	15
0	0	3	2	16

TBP using KNN

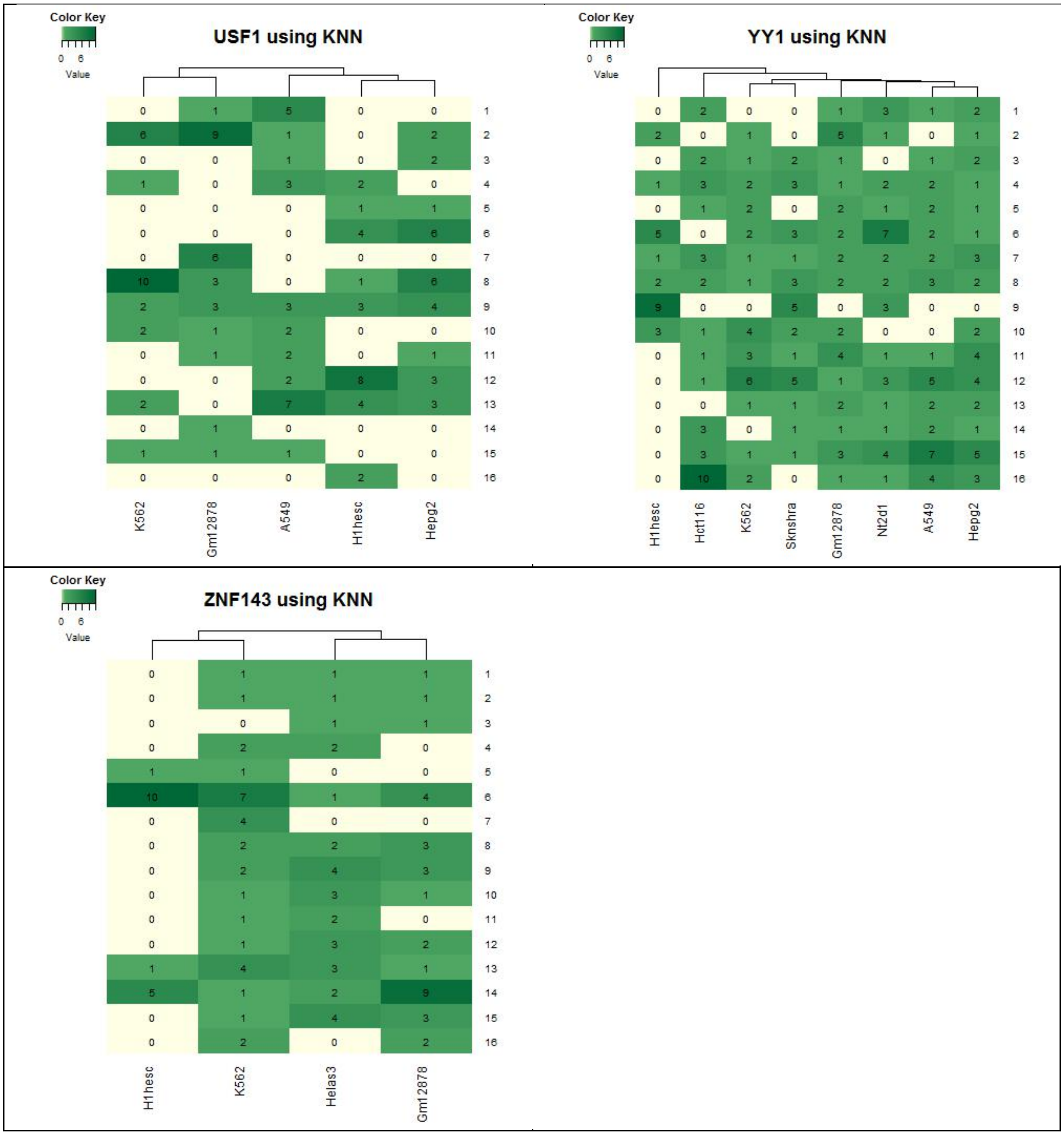
Helas3	Gm12878	H1hesc	K562	Hepg2	
13	0	0	0	0	1
0	14	1	0	0	2
0	1	2	2	3	3
2	0	4	1	2	4
1	3	2	2	2	5
5	1	0	0	0	6
3	2	1	4	2	7
0	1	1	3	5	8
1	2	3	0	1	9
0	1	6	3	5	10
0	0	2	2	2	11
1	4	0	3	3	12
0	1	3	1	0	13
4	2	1	3	2	14
0	0	6	12	10	15
1	4	3	2	2	16

TCF12 using KNN

A549	Hepg2	Gm12878	H1hesc	
0	0	5	0	1
3	5	0	0	2
13	4	3	7	3
0	0	2	4	4
0	0	0	7	5
0	0	9	0	6
0	0	2	0	7
0	5	0	0	8
0	11	0	0	9
0	0	0	2	10
7	0	0	0	11
8	0	0	0	12
0	0	1	0	13
0	0	0	4	14
0	0	3	7	15
6	0	0	0	16

TCF7L2 using KNN

Hct116	Mcf7	Hepg2	Hek293	Helas3	Panc1	
0	3	3	3	4	3	1
2	0	0	0	1	0	2
2	0	2	1	1	1	3
0	0	5	0	0	1	4
0	0	0	5	0	0	5
0	3	2	1	11	5	6
0	10	0	3	4	7	7
0	10	3	0	0	3	8
0	2	3	0	0	1	9
5	0	0	0	0	0	10
0	1	0	6	0	0	11
8	0	0	0	1	0	12
0	3	1	4	3	5	13
0	0	3	0	0	2	14
1	1	1	0	1	1	15
4	1	1	2	3	3	16

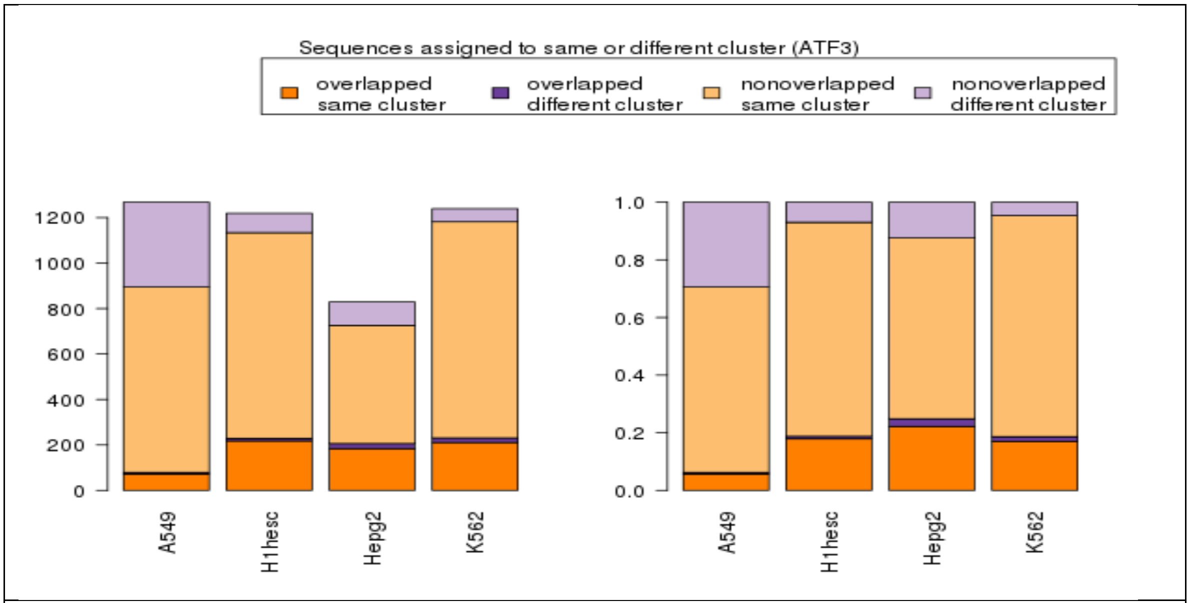


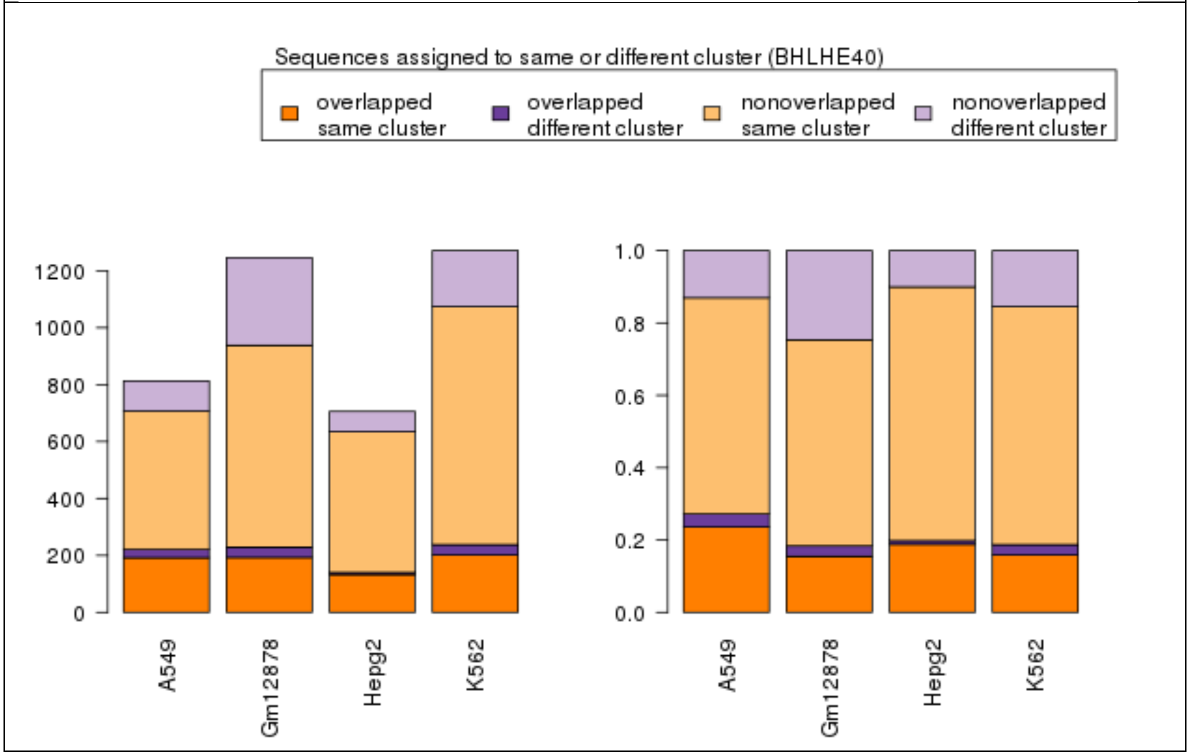
K562	Gm12878	A549	H1hesc	Hepg2	
0	1	5	0	0	1
6	9	1	0	2	2
0	0	1	0	2	3
1	0	3	2	0	4
0	0	0	1	1	5
0	0	0	4	6	6
0	6	0	0	0	7
10	3	0	1	6	8
2	3	3	3	4	9
2	1	2	0	0	10
0	1	2	0	1	11
0	0	2	8	3	12
2	0	7	4	3	13
0	1	0	0	0	14
1	1	1	0	0	15
0	0	0	2	0	16

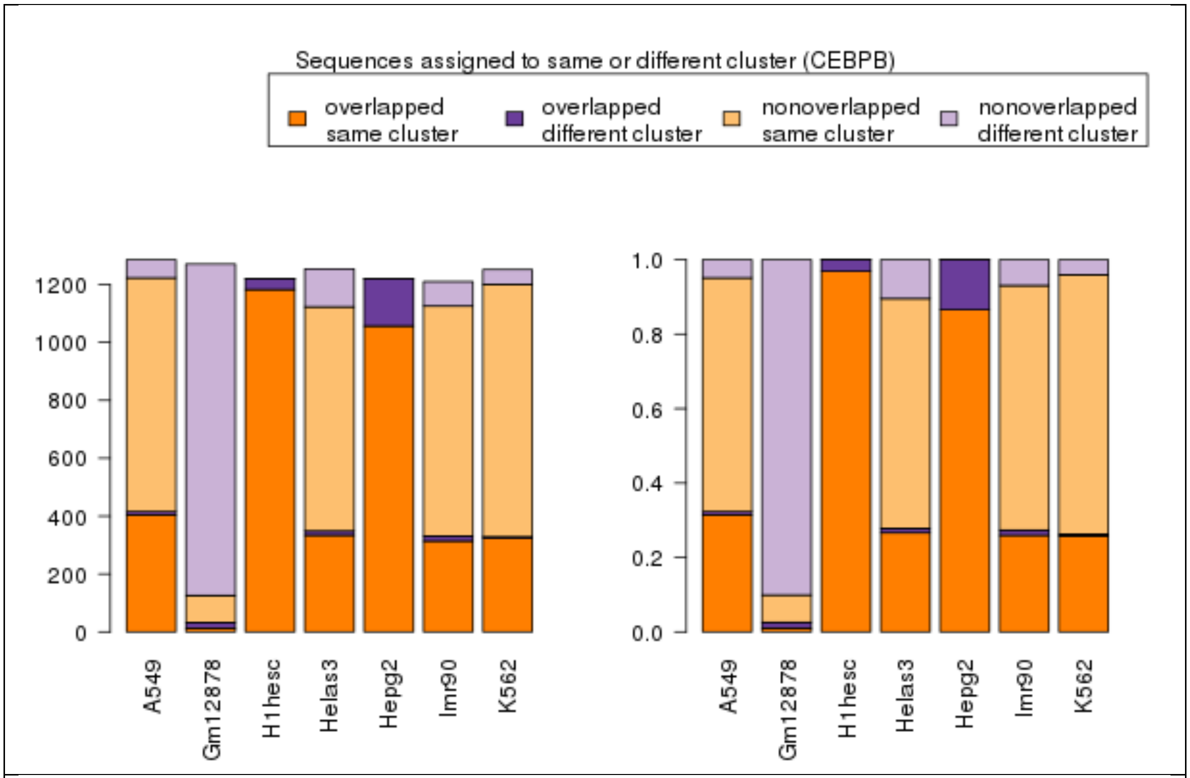
H1hesc	Hct116	K562	Sknshra	Gm12878	Nt2d1	A549	Hepg2	
0	2	0	0	1	3	1	2	1
2	0	1	0	5	1	0	1	2
0	2	1	2	1	0	1	2	3
1	3	2	3	1	2	2	1	4
0	1	2	0	2	1	2	1	5
5	0	2	3	2	7	2	1	6
1	3	1	1	2	2	2	3	7
2	2	1	3	2	2	3	2	8
9	0	0	5	0	3	0	0	9
3	1	4	2	2	0	0	2	10
0	1	3	1	4	1	1	4	11
0	1	6	5	1	3	5	4	12
0	0	1	1	2	1	2	2	13
0	3	0	1	1	1	2	1	14
0	3	1	1	3	4	7	5	15
0	10	2	0	1	1	4	3	16

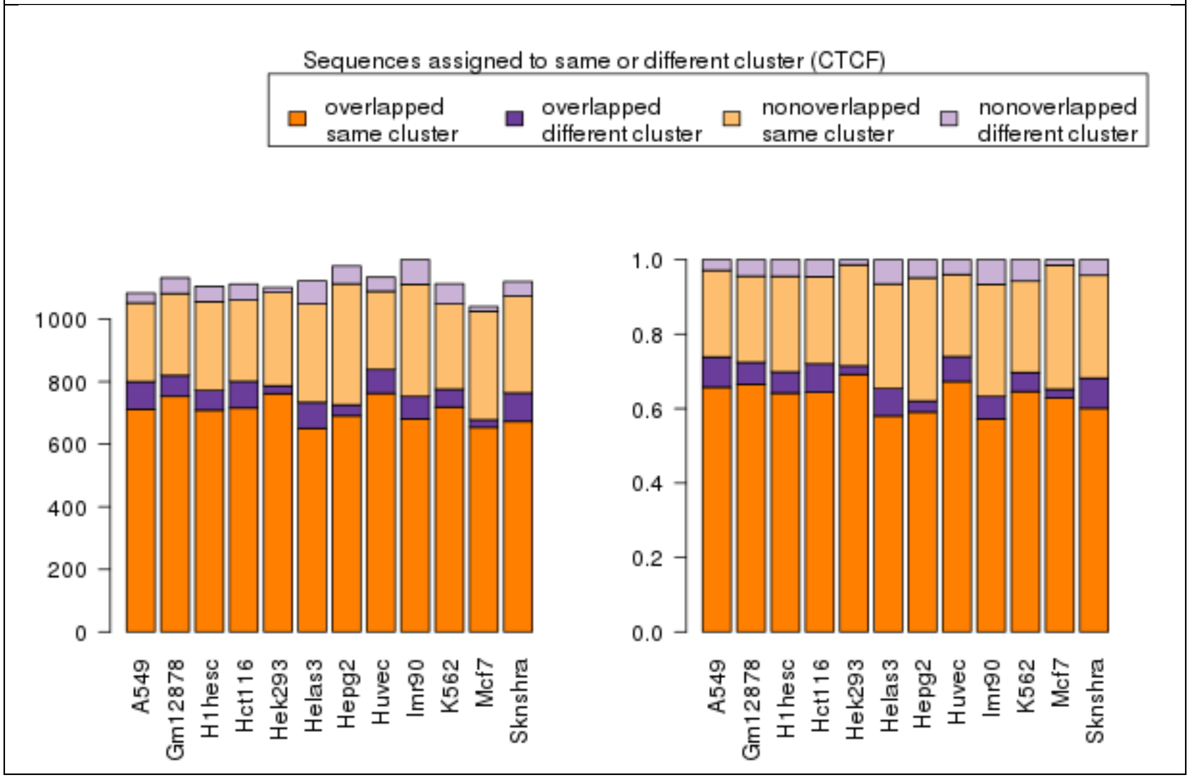
H1hesc	K562	Helas3	Gm12878	
0	1	1	1	1
0	1	1	1	2
0	0	1	1	3
0	2	2	0	4
1	1	0	0	5
10	7	1	4	6
0	4	0	0	7
0	2	2	3	8
0	2	4	3	9
0	1	3	1	10
0	1	2	0	11
0	1	3	2	12
1	4	3	1	13
5	1	2	9	14
0	1	4	3	15
0	2	0	2	16

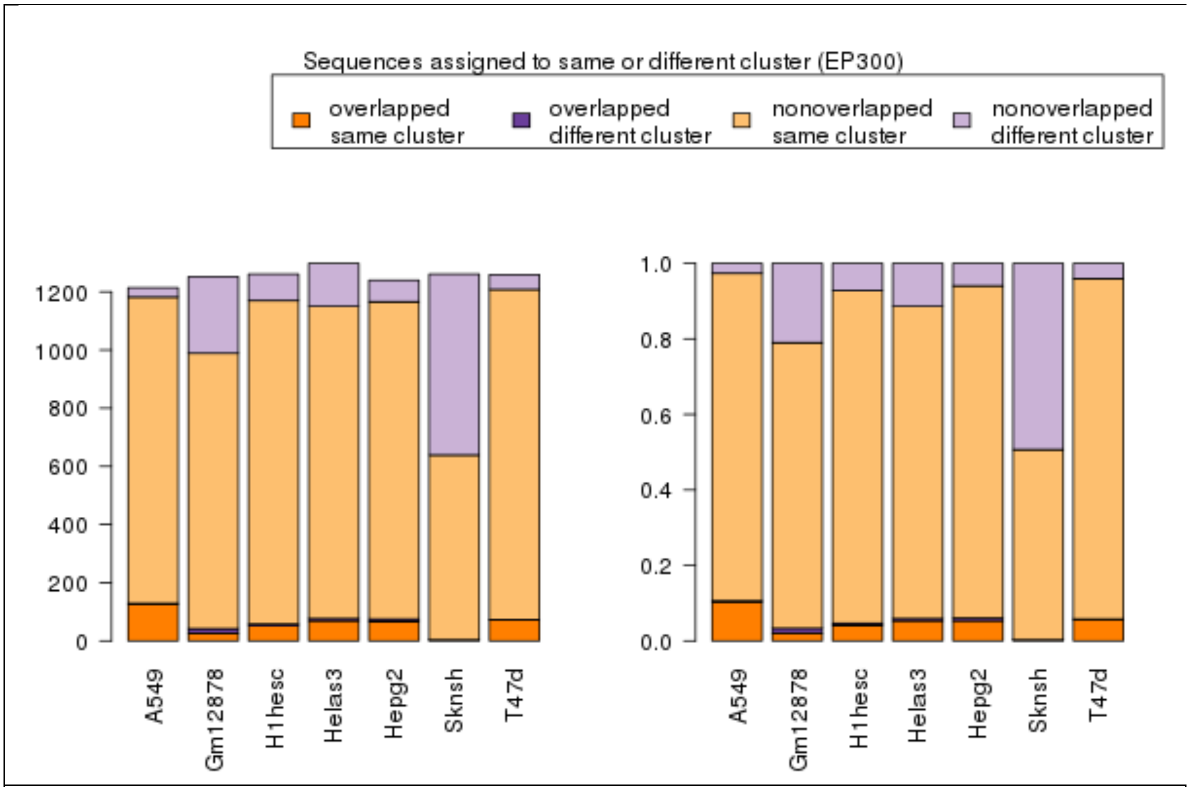
Figure 7.4 Cluster membership matrix for k-Nearest Neighbor (k-NN) algorithm for k = 16. In each matrix, a row represents a cluster and a column represents a cell type. Elements in the matrix denote the number of sub-models in the cluster belonging to a specific cell type.

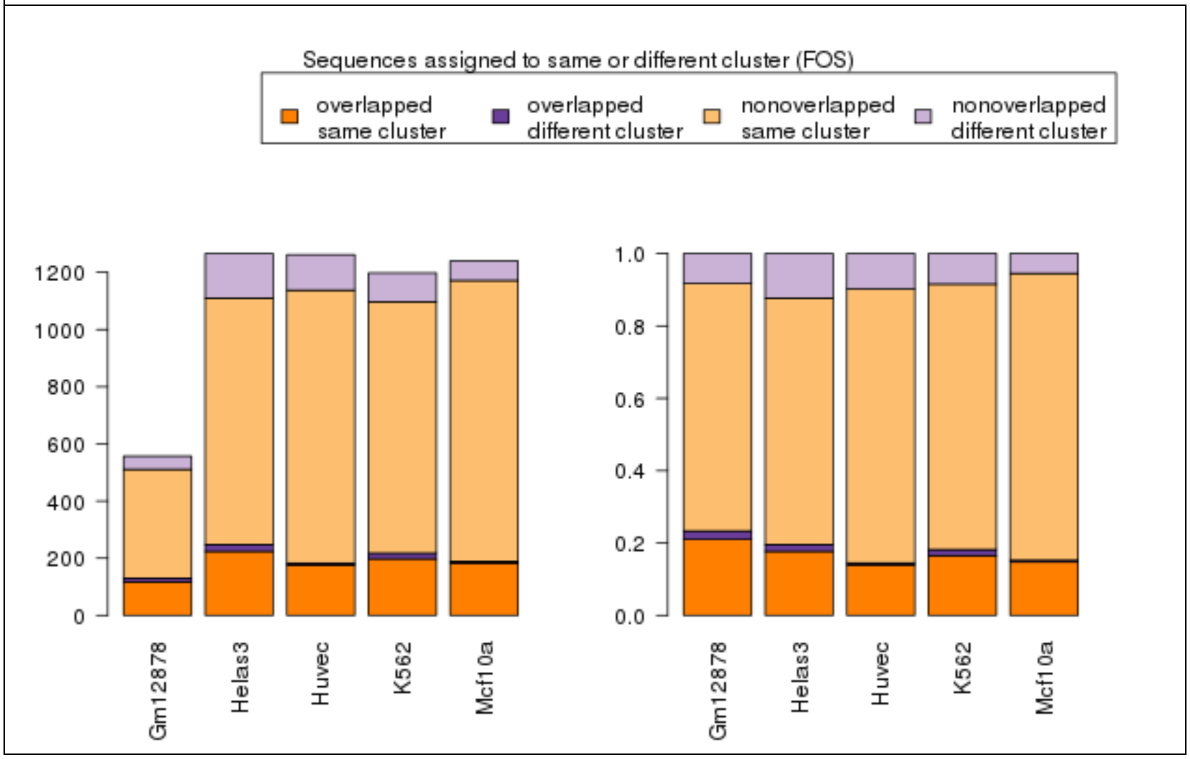

Sequences assigned to same or different cluster (ATF3)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
A549
H1hesc
Hepg2
K562
0.0
0.2
0.4
0.6
0.8
1.0
A549
H1hesc
Hepg2
K562

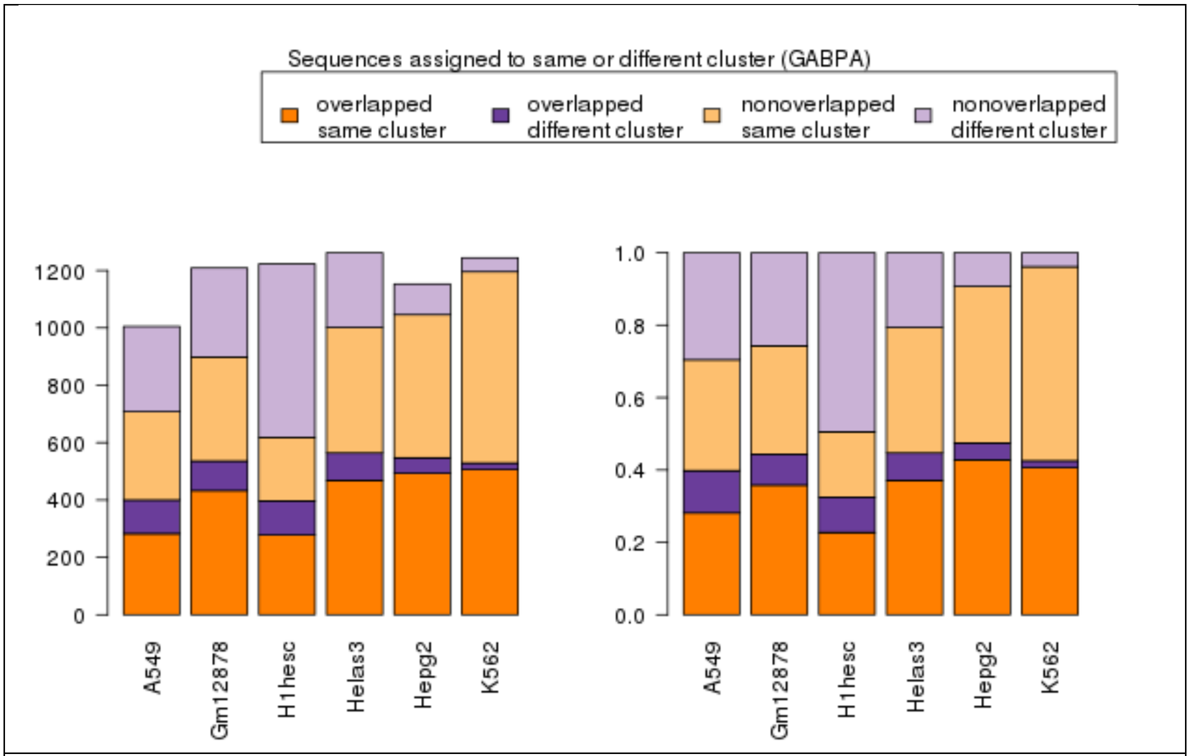

Sequences assigned to same or different cluster (BHLHE40)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
A549
Gm12878
Hepg2
K562
0.0
0.2
0.4
0.6
0.8
1.0
A549
Gm12878
Hepg2
K562

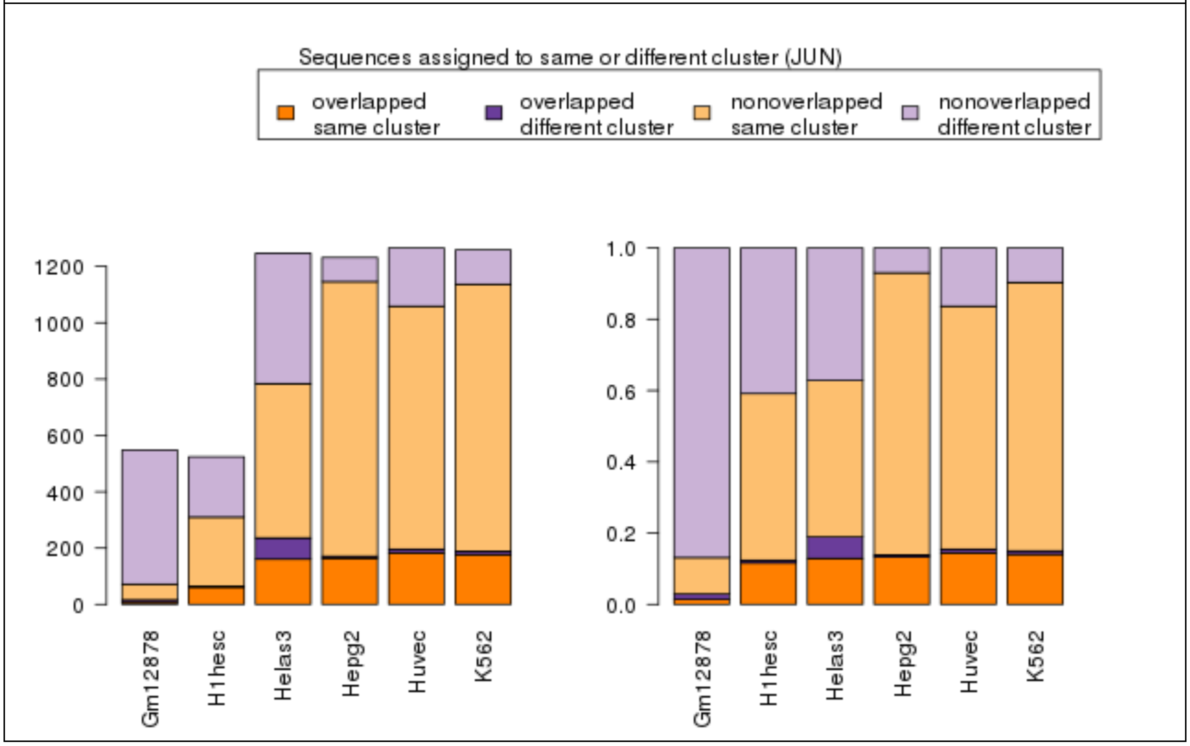

Sequences assigned to same or different cluster (CEBPB)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
A549
Gm12878
H1hesc
Helas3
Hepg2
Imr90
K562
A549
Gm12878
H1hesc
Helas3
Hepg2
Imr90
K562

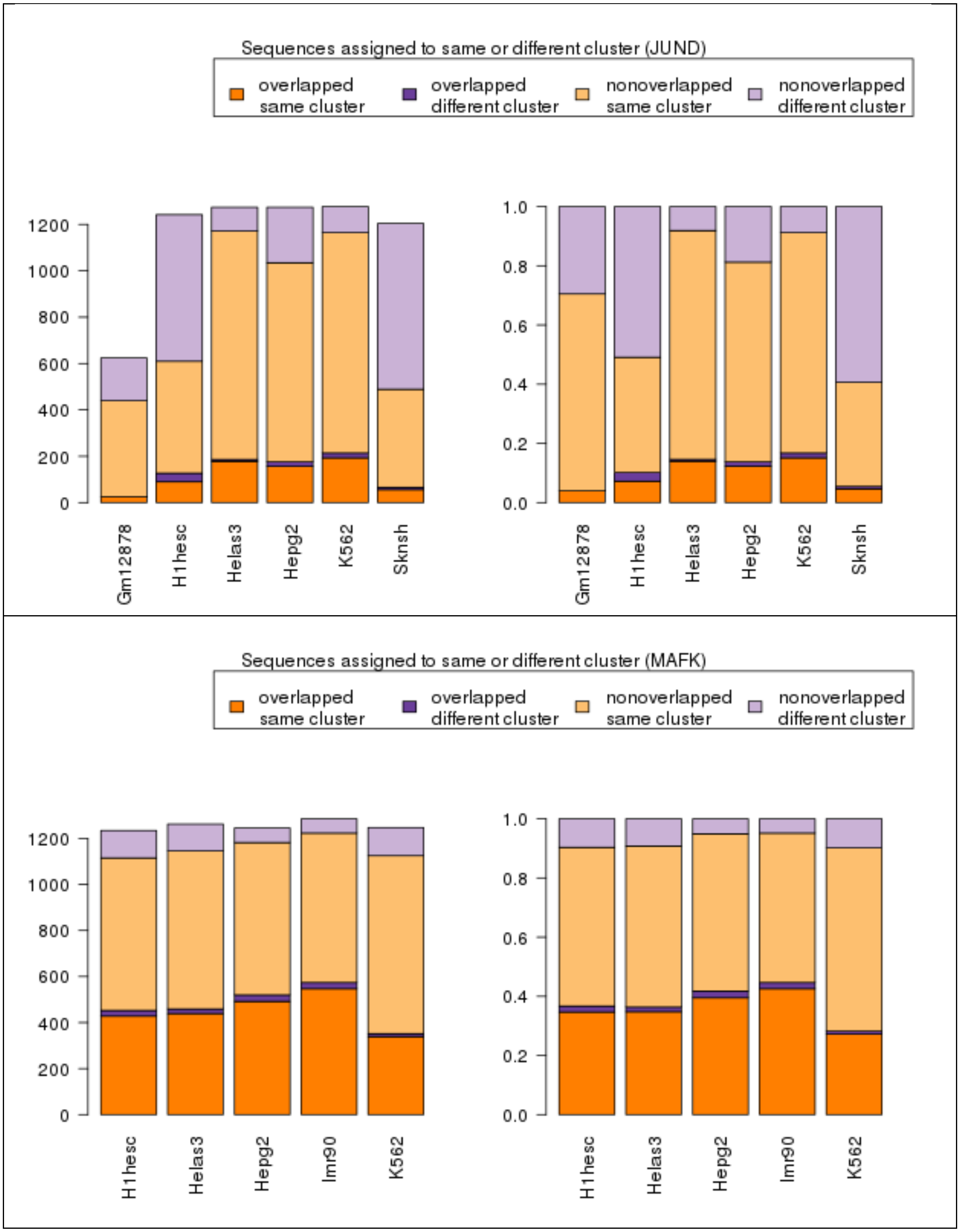

Sequences assigned to same or different cluster (CTCF)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
0.0
0.2
0.4
0.6
0.8
1.0
A549
Gm12878
H1hesc
Hct116
Hek293
Helas3
Hepg2
Huvec
Imr90
K562
Mcf7
Sknshra
A549
Gm12878
H1hesc
Hct116
Hek293
Helas3
Hepg2
Huvec
Imr90
K562
Mcf7
Sknshra

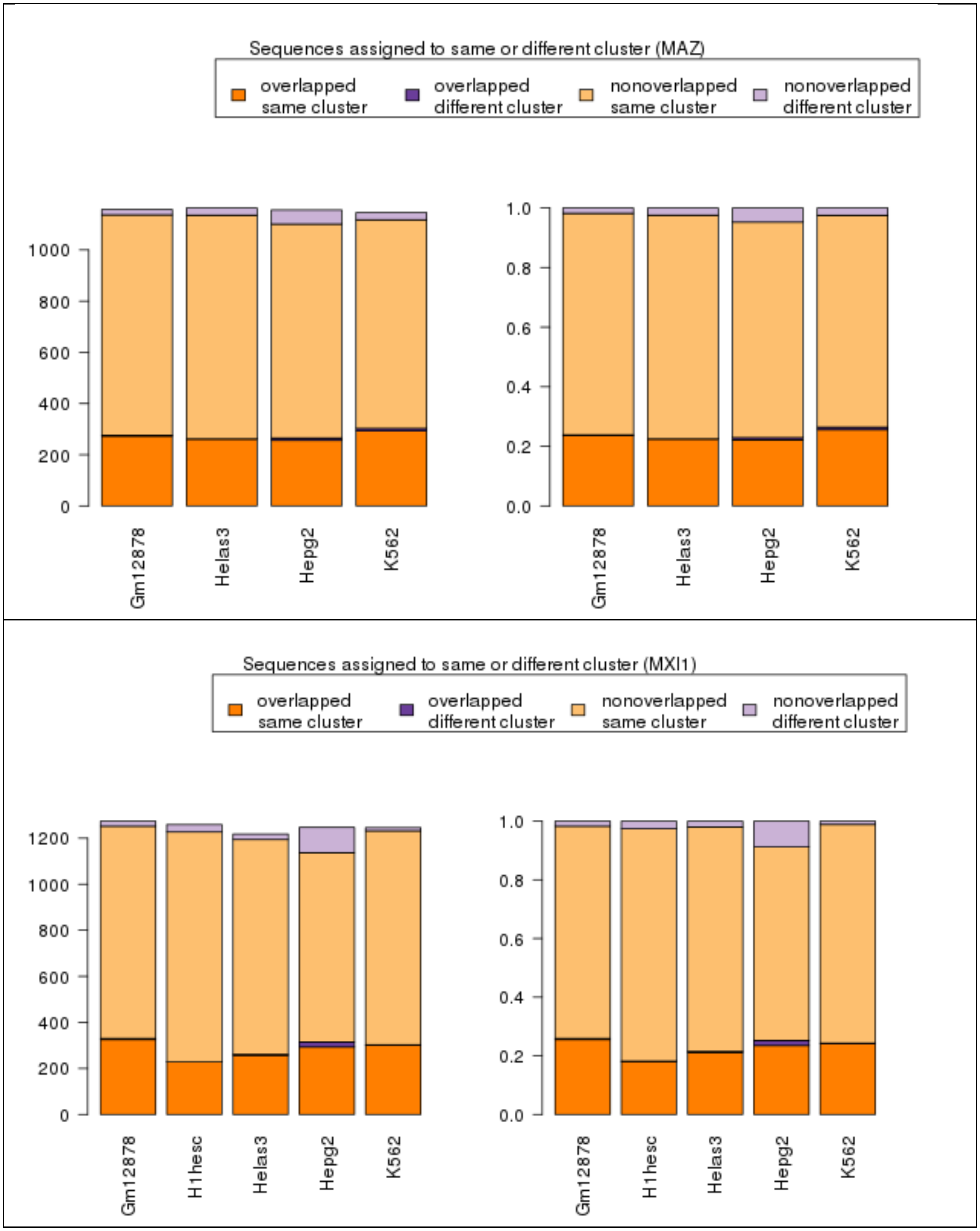

Sequences assigned to same or different cluster (EP300)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
A549
Gm12878
H1hesc
Helas3
Hepg2
Sknsh
T47d
0.0
0.2
0.4
0.6
0.8
1.0
A549
Gm12878
H1hesc
Helas3
Hepg2
Sknsh
T47d

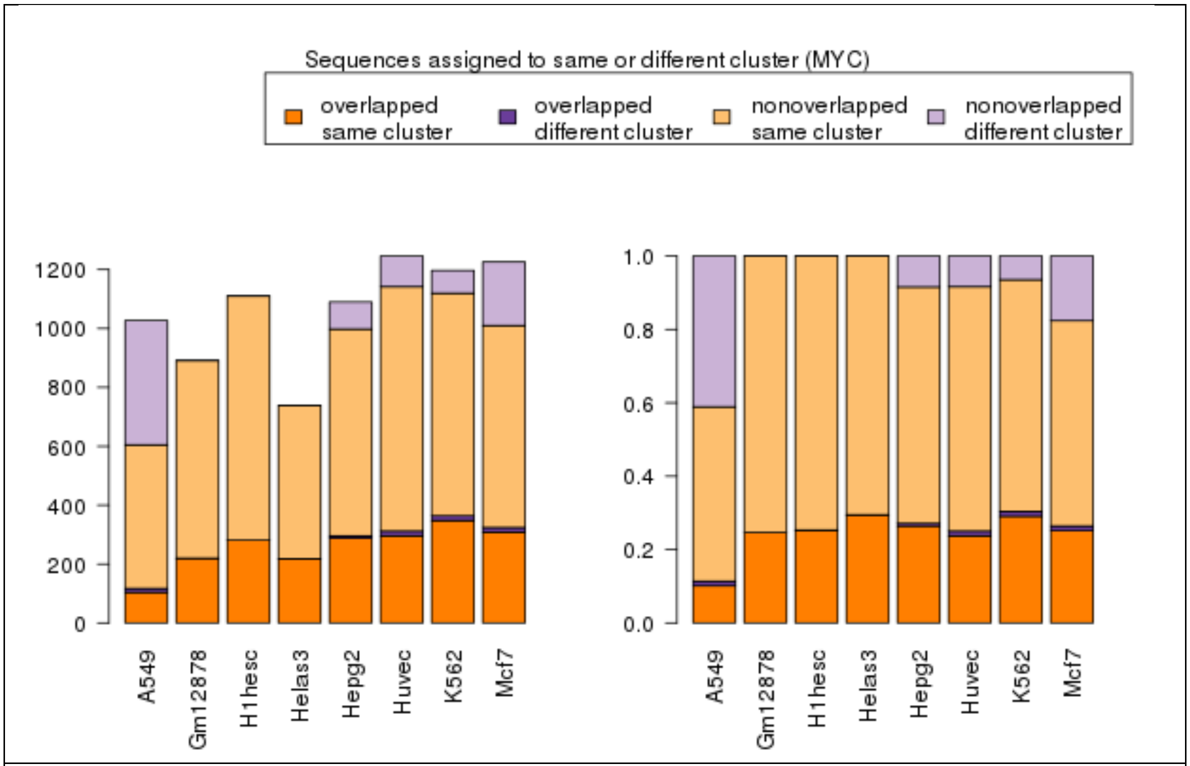

Sequences assigned to same or different cluster (FOS)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
Gm12878
Helas3
Huvec
K562
Mcf10a
0.0
0.2
0.4
0.6
0.8
1.0
Gm12878
Helas3
Huvec
K562
Mcf10a

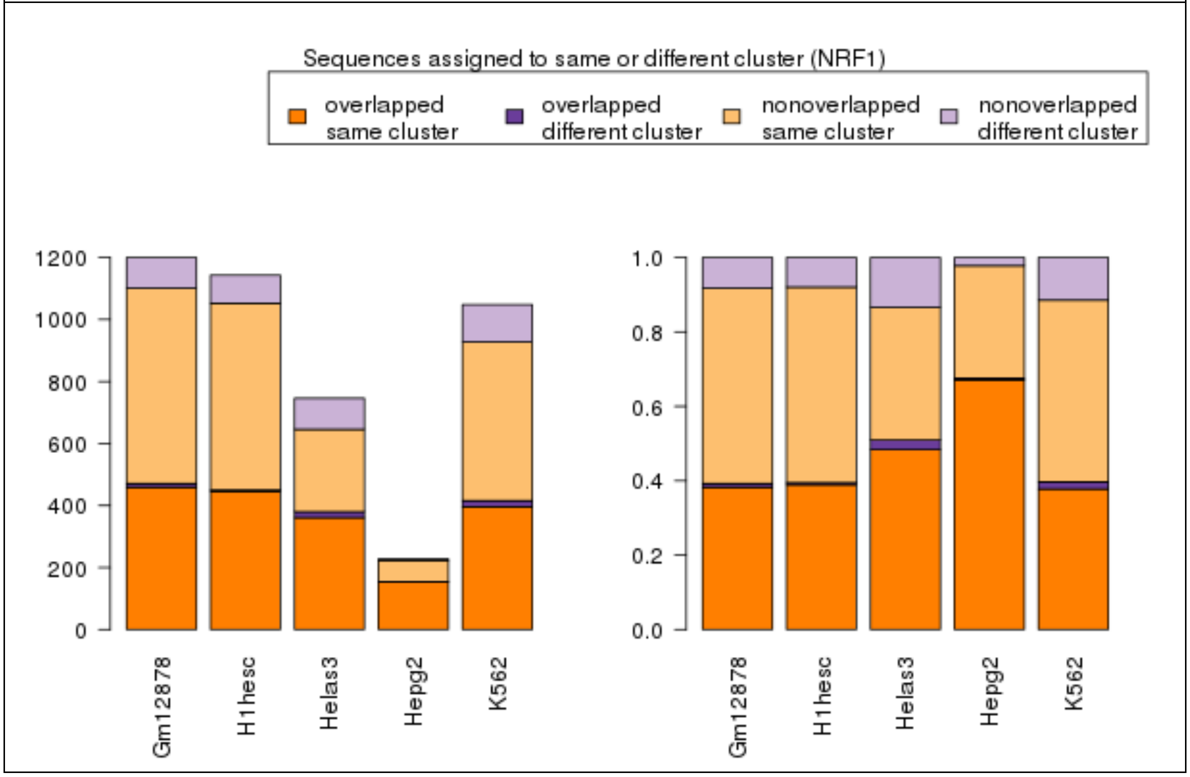

Sequences assigned to same or different cluster (GABPA)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
A549
Gm12878
H1hesc
Helas3
Hepg2
K562

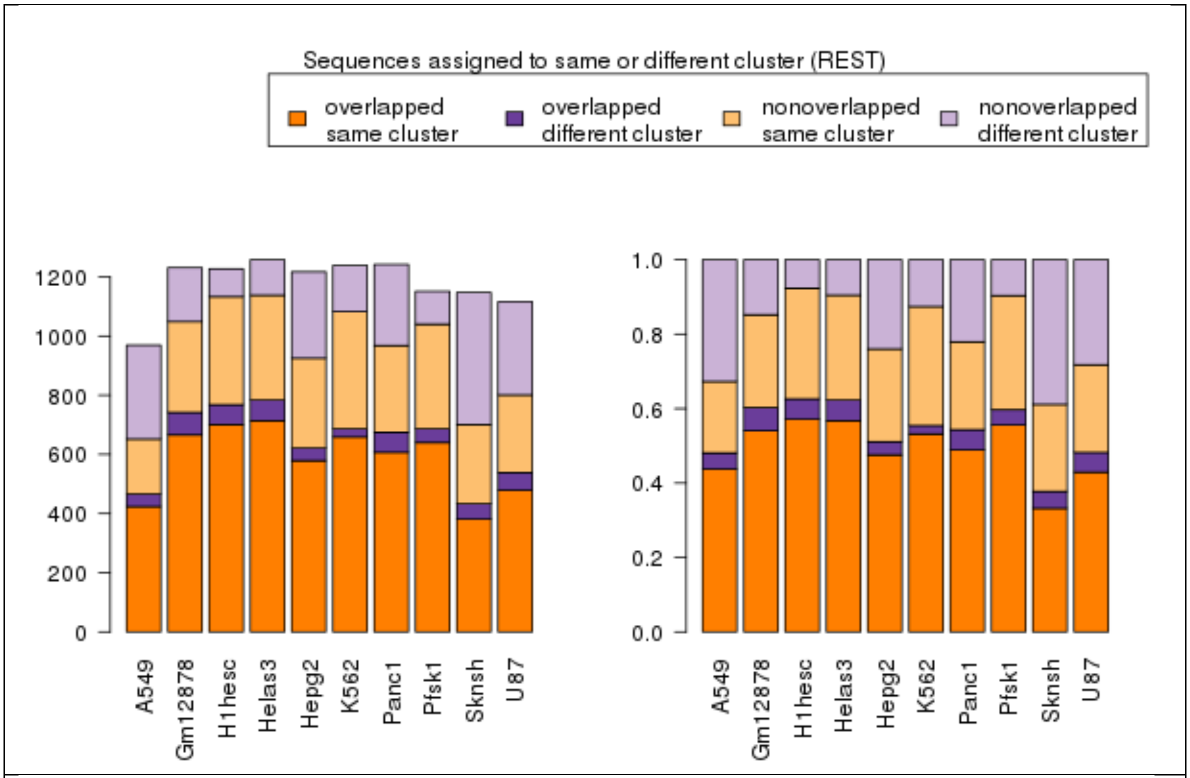

Sequences assigned to same or different cluster (JUN)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562

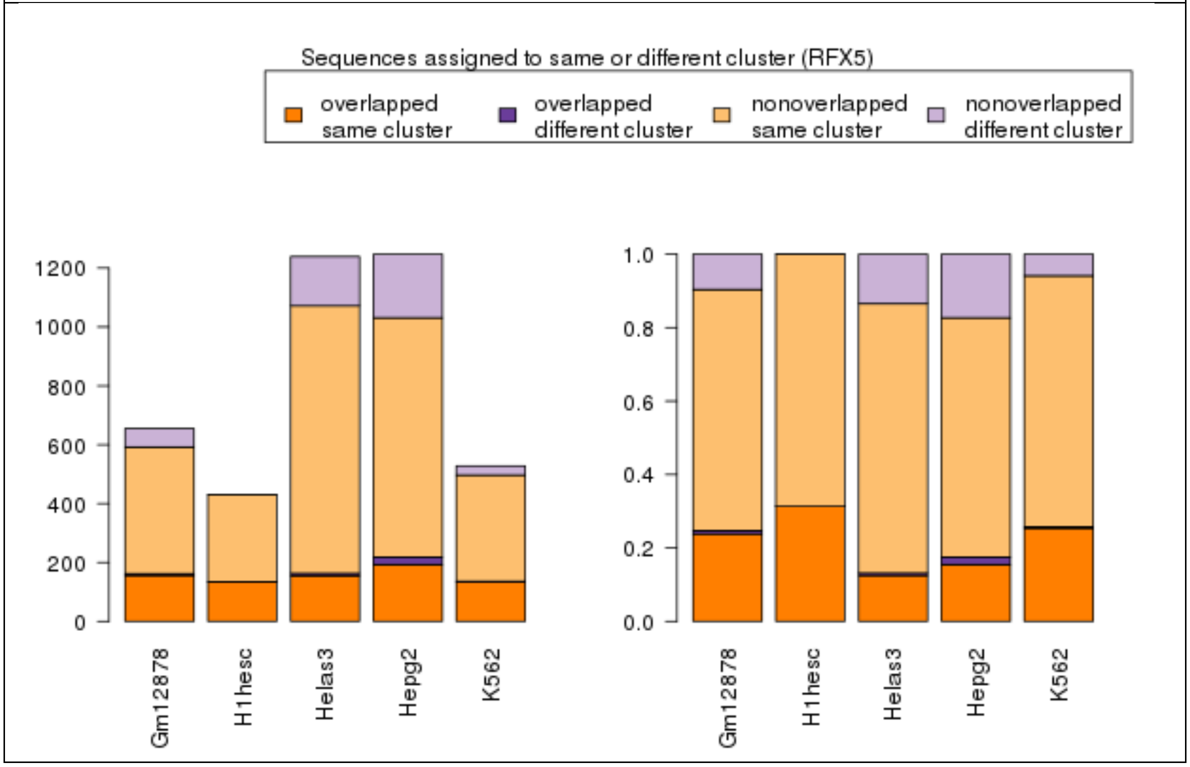

Sequences assigned to same or different cluster (JUND)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
Gm12878
H1hesc
Helas3
Hepg2
K562
Sknsh
Sequences assigned to same or different cluster (MAFK)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
H1hesc
Helas3
Hepg2
Imr90
K562

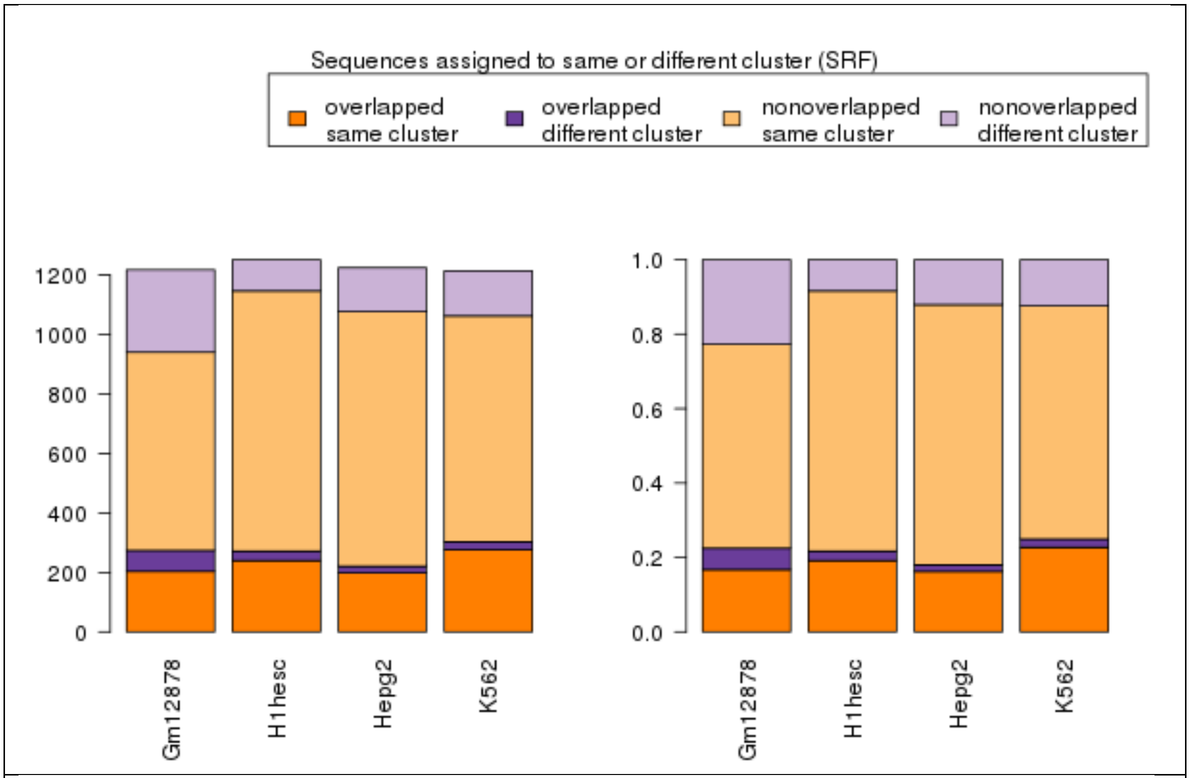

Sequences assigned to same or different cluster (MAZ)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
Gm12878
Helas3
Hepg2
K562
0.0
0.2
0.4
0.6
0.8
1.0
Gm12878
Helas3
Hepg2
K562
Sequences assigned to same or different cluster (MXI1)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
Gm12878
H1hesc
Helas3
Hepg2
K562
0.0
0.2
0.4
0.6
0.8
1.0
Gm12878
H1hesc
Helas3
Hepg2
K562

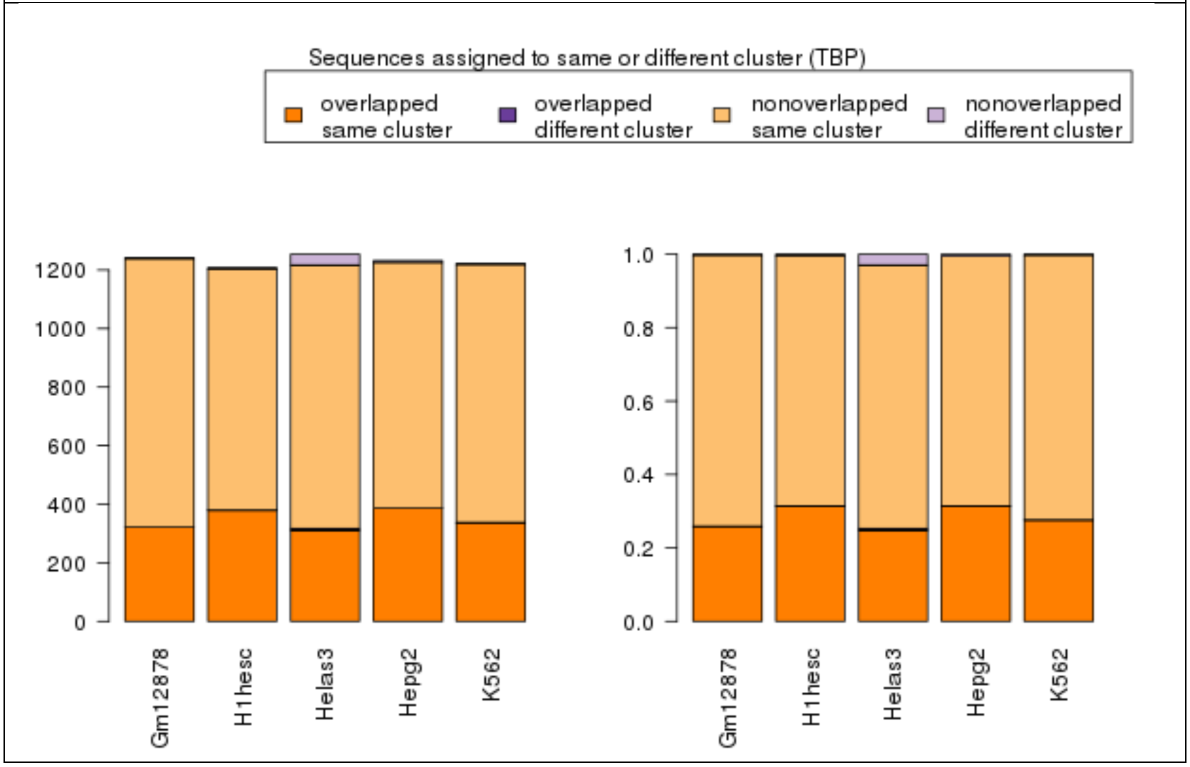

Sequences assigned to same or different cluster (MYC)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
A549
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562
Mcf7
0.0
0.2
0.4
0.6
0.8
1.0
A549
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562
Mcf7

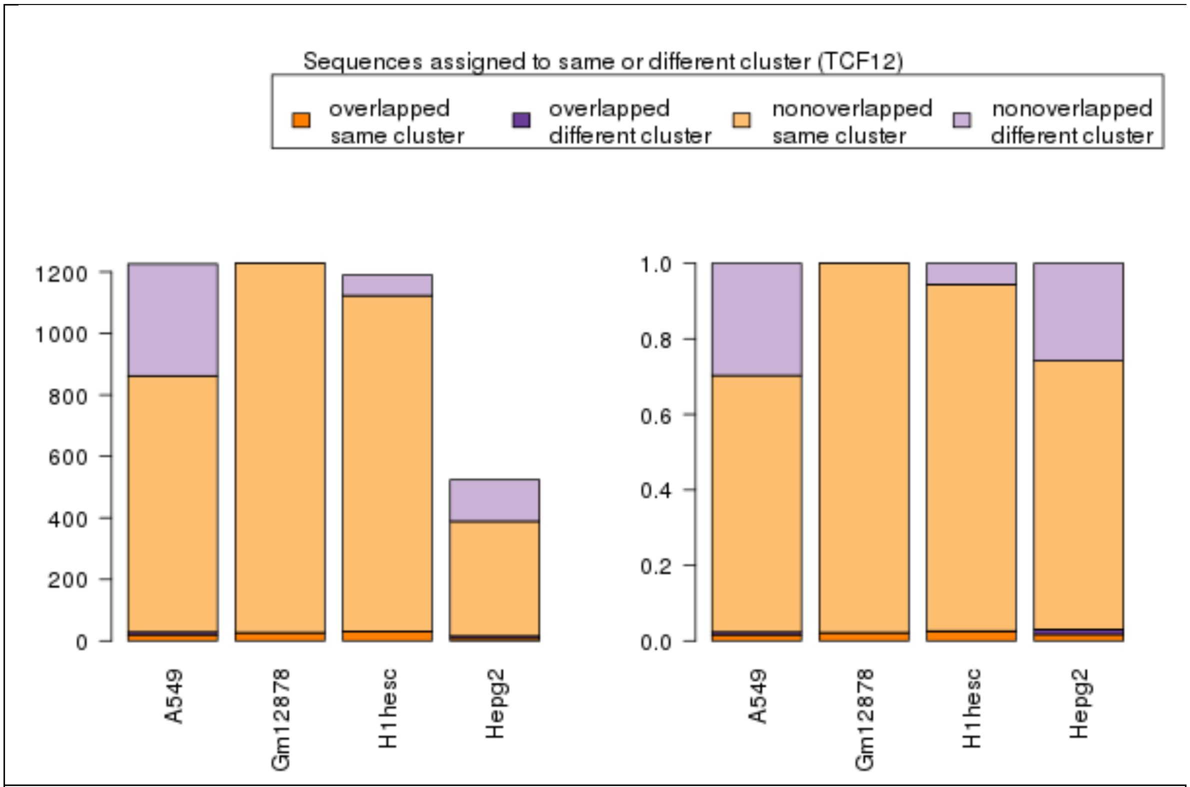

Sequences assigned to same or different cluster (NRF1)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
Gm12878
H1hesc
Helas3
Hepg2
K562
0.0
0.2
0.4
0.6
0.8
1.0
Gm12878
H1hesc
Helas3
Hepg2
K562

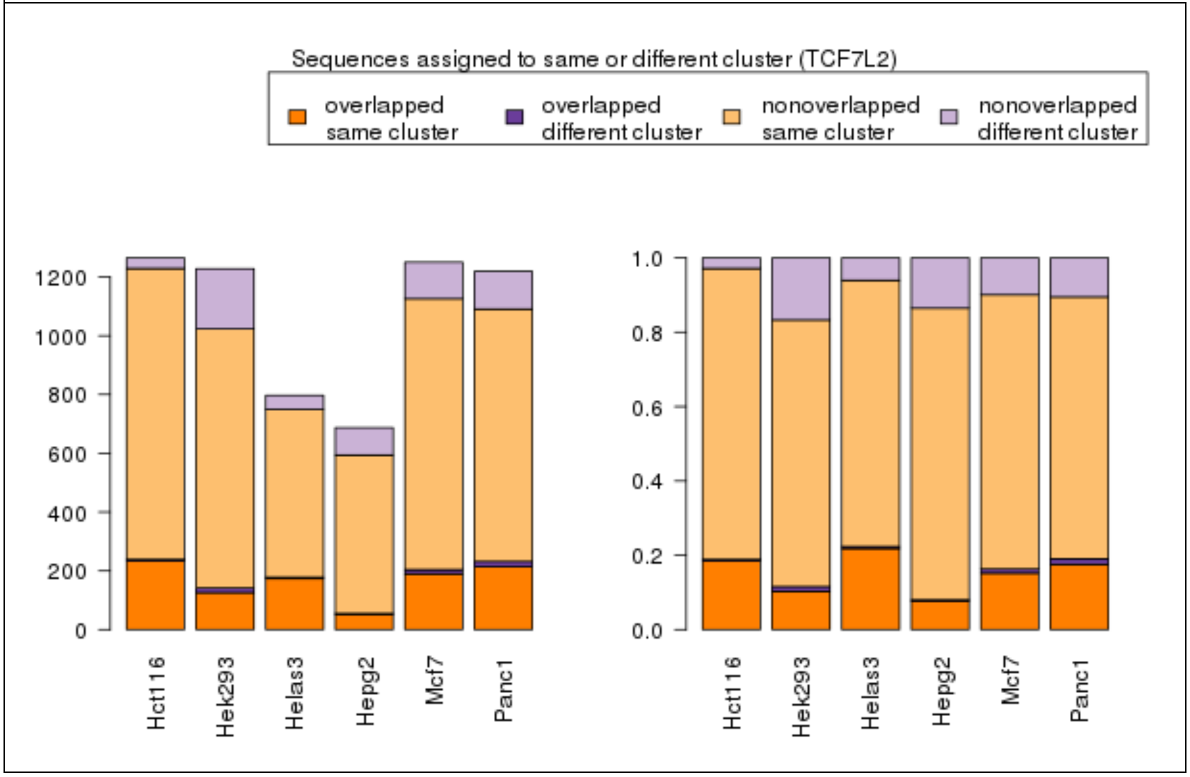

Sequences assigned to same or different cluster (REST)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
A549
Gm12878
H1hesc
Helas3
Hepg2
K562
Panc1
Pfsk1
Sknsh
U87

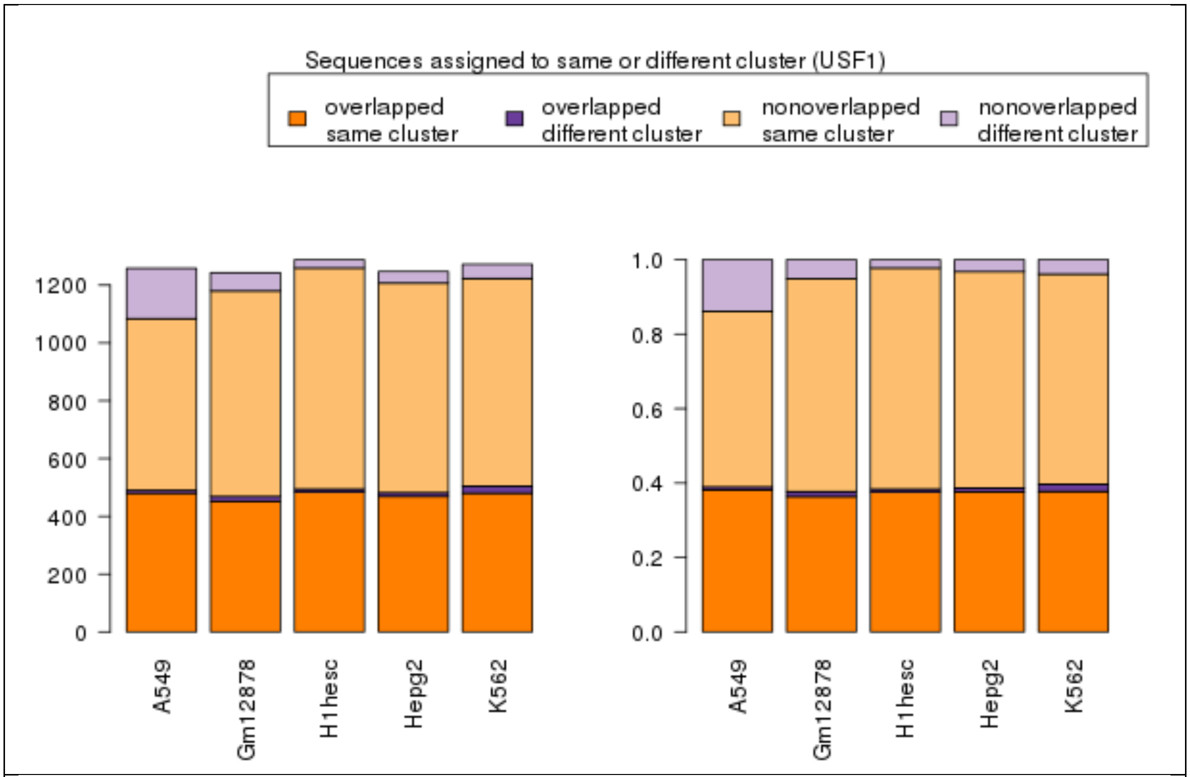

Sequences assigned to same or different cluster (RFX5)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
Gm12878
H1hesc
Helas3
Hepg2
K562

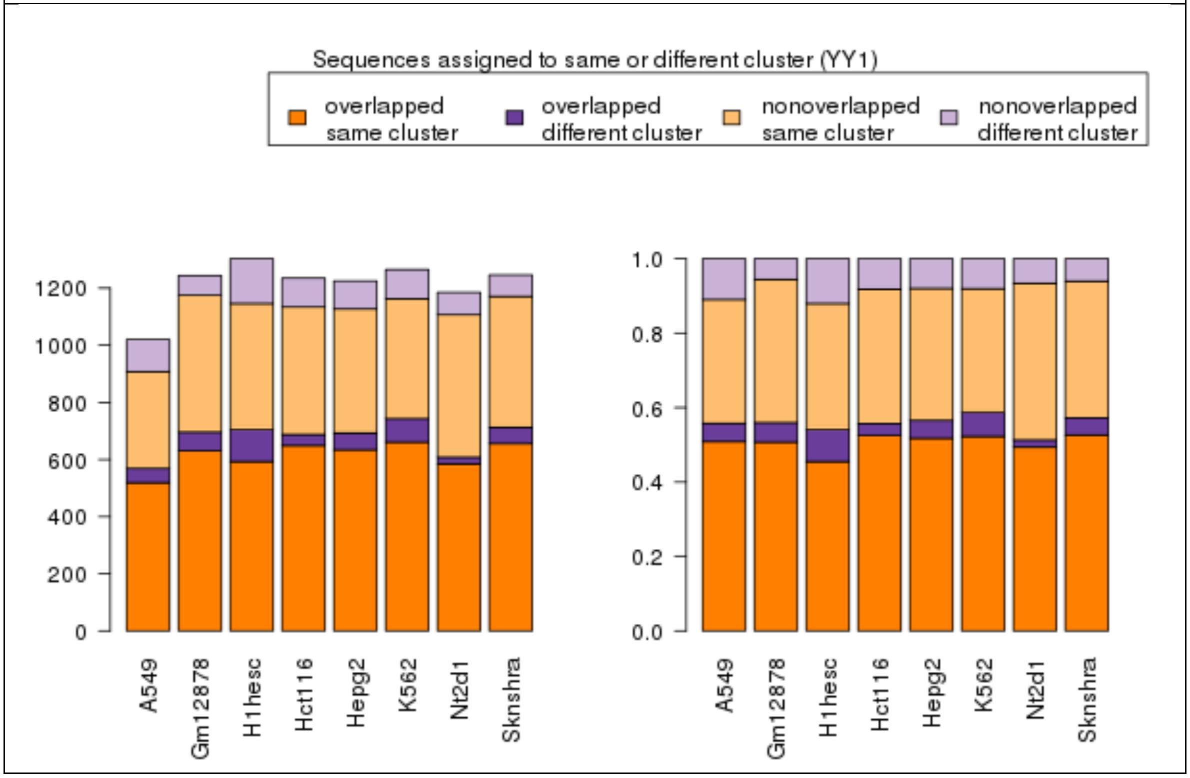

Sequences assigned to same or different cluster (SRF)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
Gm12878
H1hesc
Hepg2
K562
0.0
0.2
0.4
0.6
0.8
1.0
Gm12878
H1hesc
Hepg2
K562


Sequences assigned to same or different cluster (TBP)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
Gm12878
H1hesc
Helas3
Hepg2
K562
0.0
0.2
0.4
0.6
0.8
1.0
Gm12878
H1hesc
Helas3
Hepg2
K562


Sequences assigned to same or different cluster (TCF12)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
1200
1000
800
600
400
200
0
A549
Gm12878
H1hesc
Hepg2
1.0
0.8
0.6
0.4
0.2
0.0
A549
Gm12878
H1hesc
Hepg2


Sequences assigned to same or different cluster (TCF7L2)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
1200
1000
800
600
400
200
0
Hct116
Hek293
Helas3
Hepg2
Mcf7
Panc1
1.0
0.8
0.6
0.4
0.2
0.0
Hct116
Hek293
Helas3
Hepg2
Mcf7
Panc1


Sequences assigned to same or different cluster (USF1)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
A549
Gm12878
H1hesc
Hepg2
K562
0.0
0.2
0.4
0.6
0.8
1.0
A549
Gm12878
H1hesc
Hepg2
K562


Sequences assigned to same or different cluster (YY1)
overlapped same cluster
overlapped different cluster
nonoverlapped same cluster
nonoverlapped different cluster
0
200
400
600
800
1000
1200
A549
Gm12878
H1hesc
Hct116
Hepg2
K562
Nt2d1
Sknshra
0.0
0.2
0.4
0.6
0.8
1.0
A549
Gm12878
H1hesc
Hct116
Hepg2
K562
Nt2d1
Sknshra

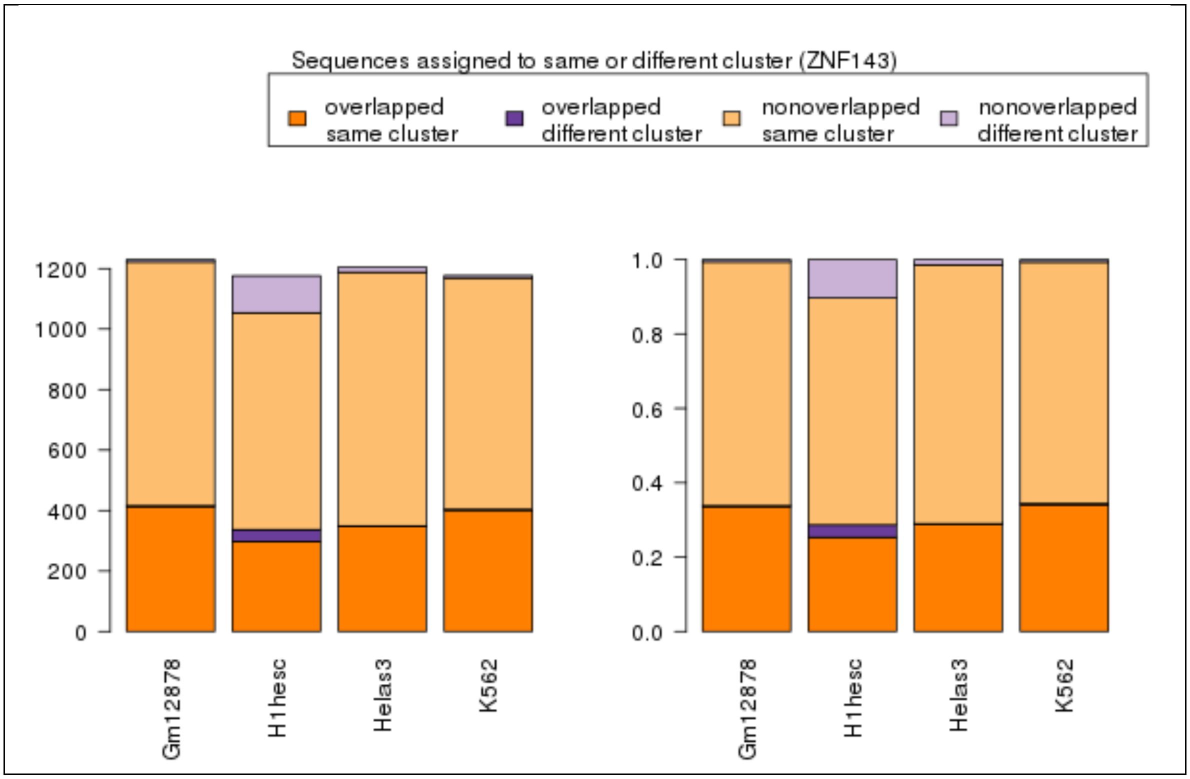
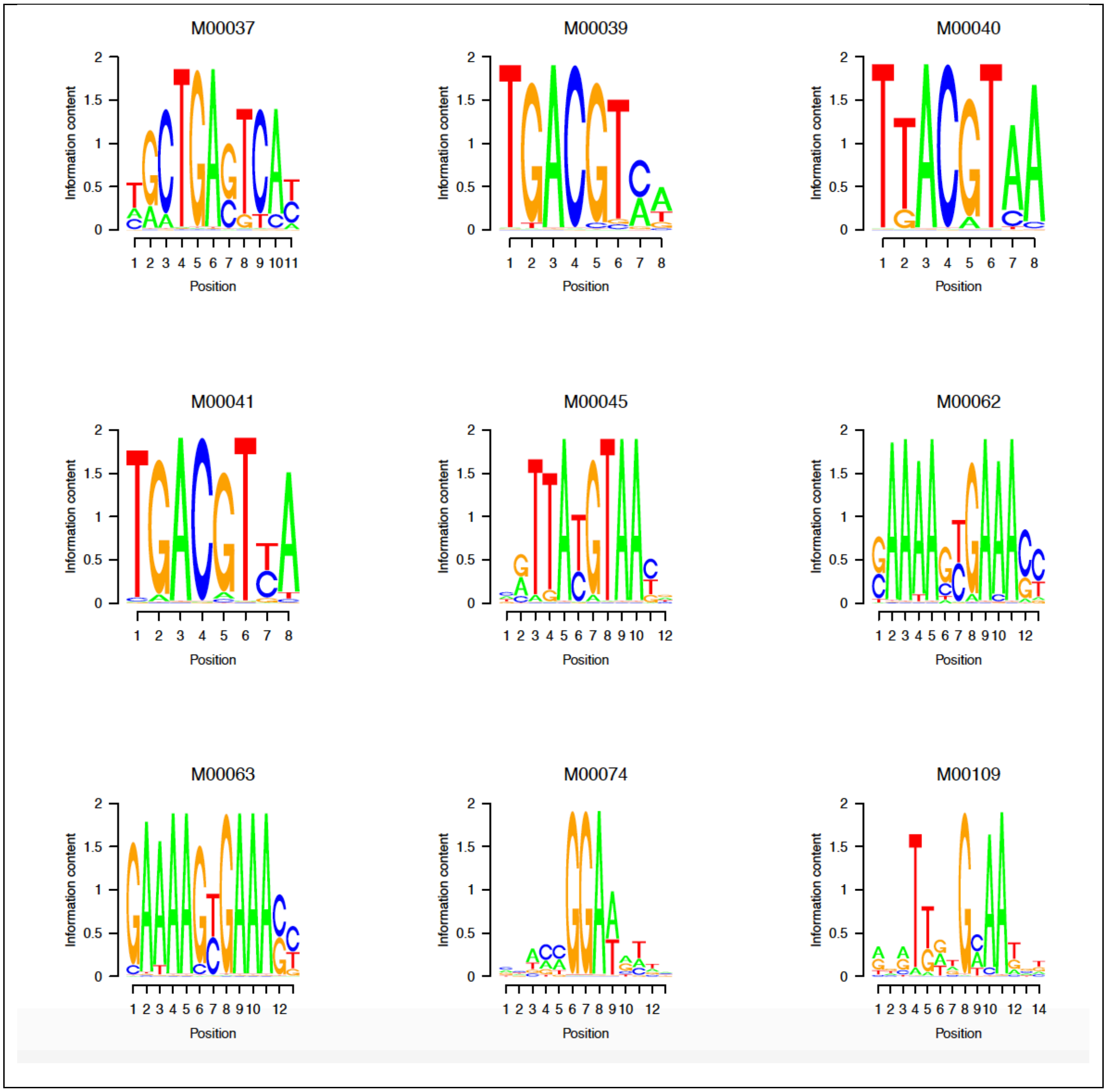
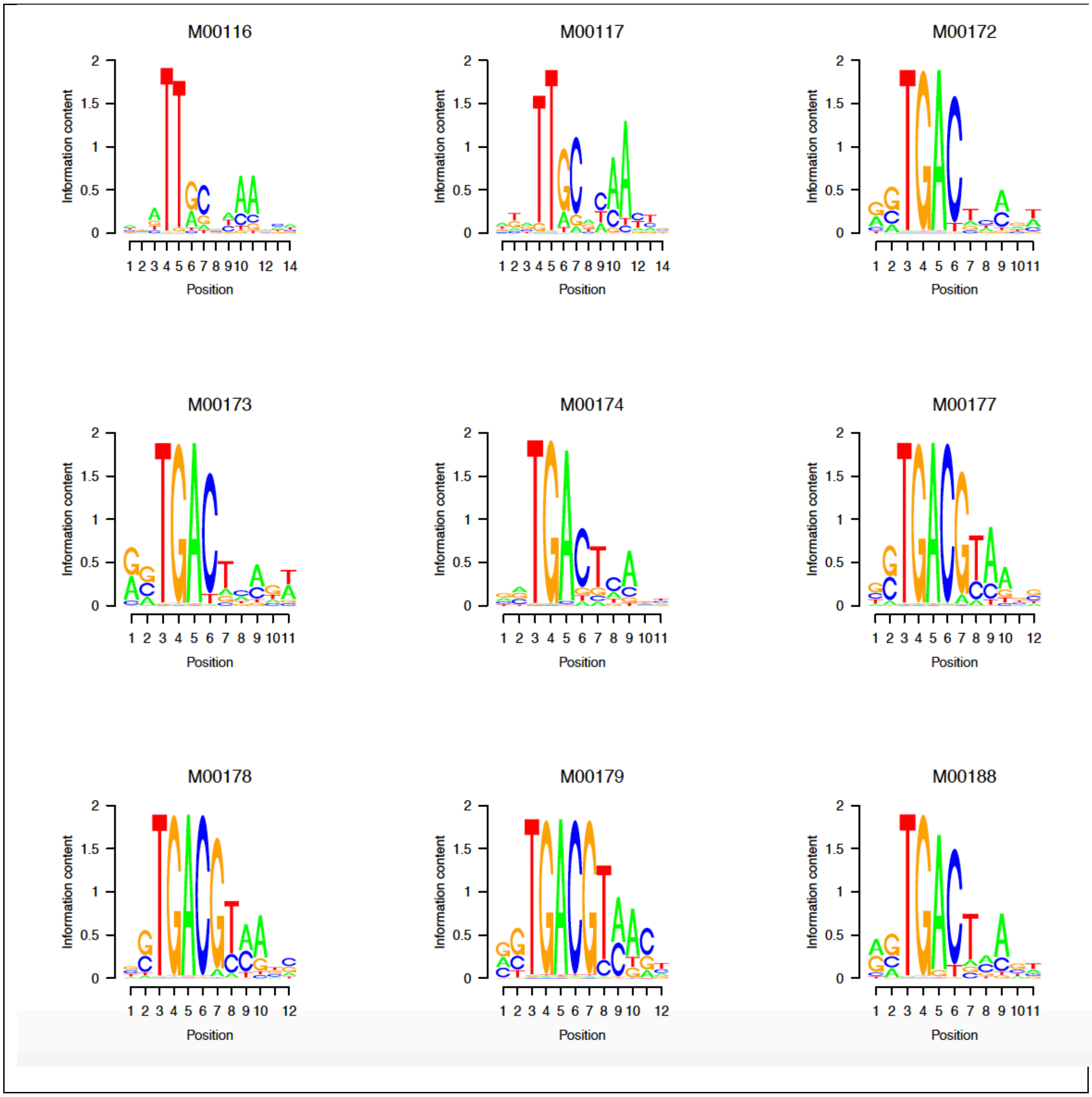
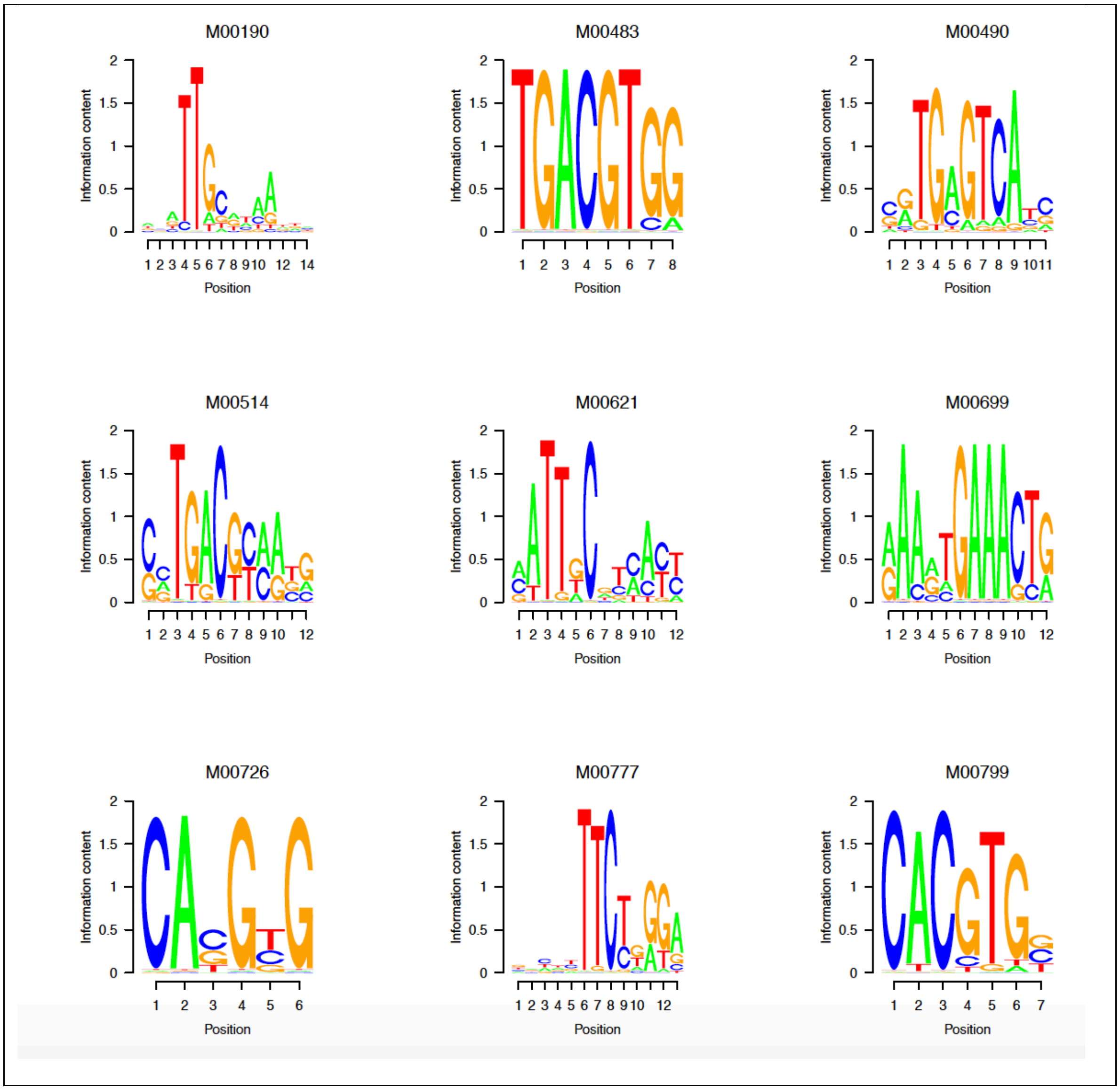
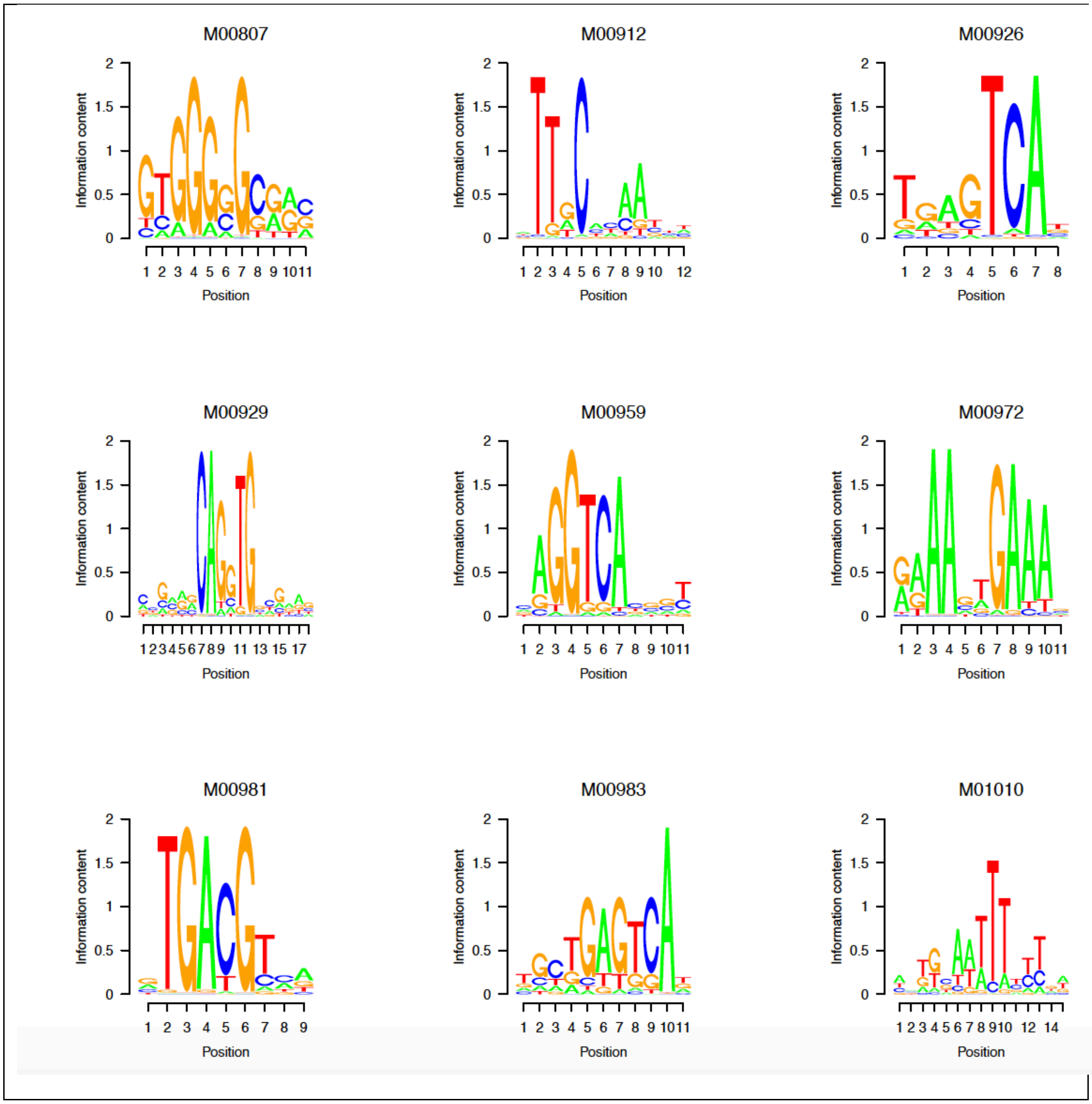


Figure 7.5 Fraction of overlapped and non-overlapped sequences which fall in the same or different clusters. Dark orange represents the fraction of overlapped sequences falling in the same cluster, whereas light orange represents non-overlapped sequences. Dark purple represents the fraction of overlapped sequences falling in different clusters, with light purple non-overlapped sequences.


M00037
M00039
M00040
M00041
M00045
M00062
M00063
M00074
M00109
Information content
0
0.5
1
1.5
2
Position
1 2 3 4 5 6 7 8 9 10 11 12 14


M00116
M00117
M00172
M00173
M00174
M00177
M00178
M00179
M00188
Information content
Position
0
0.5
1
1.5
2


M00190
M00483
M00490
M00514
M00621
M00699
M00726
M00777
M00799
Information content
0
0.5
1
1.5
2
Position


M00807
Information content
0
0.5
1
1.5
2
1 2 3 4 5 6 7 8 9 10 11
Position
M00912
Information content
0
0.5
1
1.5
2
1 2 3 4 5 6 7 8 9 10 12
Position
M00926
Information content
0
0.5
1
1.5
2
1 2 3 4 5 6 7 8
Position
M00929
Information content
0
0.5
1
1.5
2
1 2 3 4 5 6 7 8 9 11 13 15 17
Position
M00959
Information content
0
0.5
1
1.5
2
1 2 3 4 5 6 7 8 9 10 11
Position
M00972
Information content
0
0.5
1
1.5
2
1 2 3 4 5 6 7 8 9 10 11
Position
M00981
Information content
0
0.5
1
1.5
2
1 2 3 4 5 6 7 8 9
Position
M00983
Information content
0
0.5
1
1.5
2
1 2 3 4 5 6 7 8 9 10 11
Position
M01010
Information content
0
0.5
1
1.5
2
1 2 3 4 5 6 7 8 9 10 12 14
Position

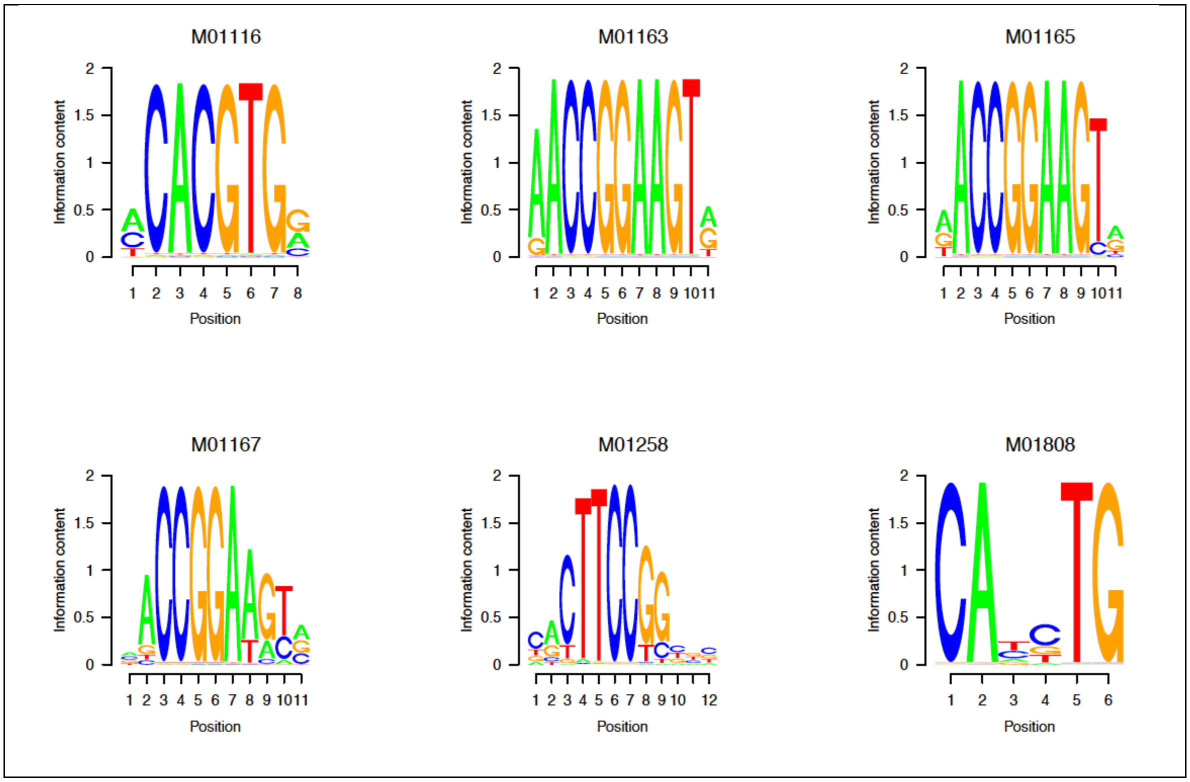
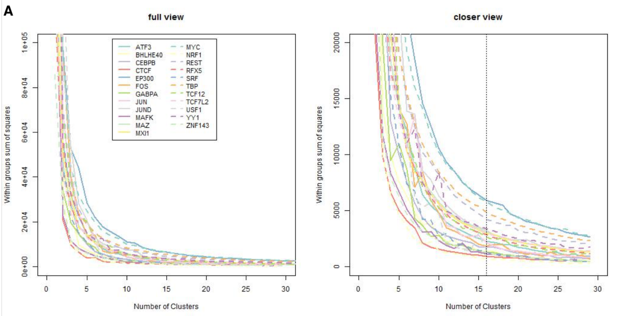
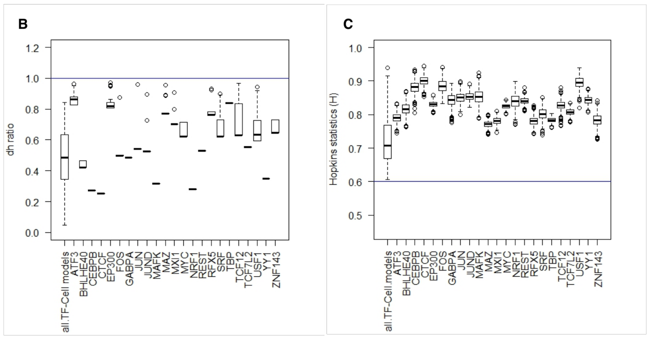


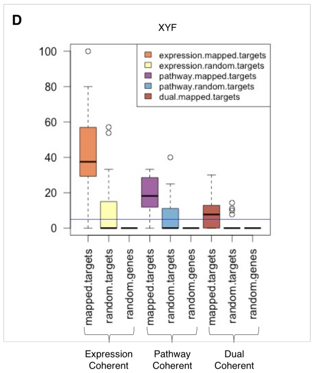
Figure 7.6 Weblogos of the TRANSFAC ids with 85% similar to any zinger motifs.

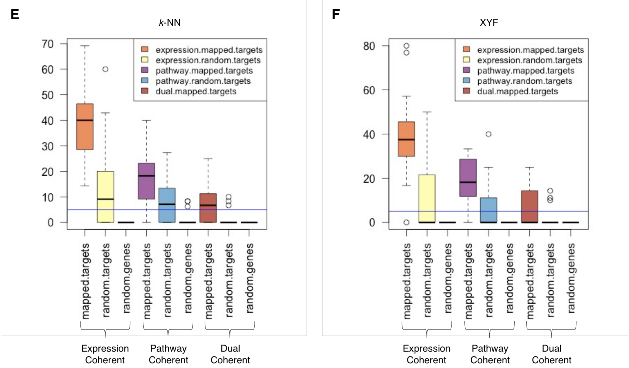


(A). Normalized within-cluster sum of squares stabilizes when the number of cluster is between 10 to 25; we choose k=16 (denoted by the vertical line in the closer view) as a representative for all TFs.



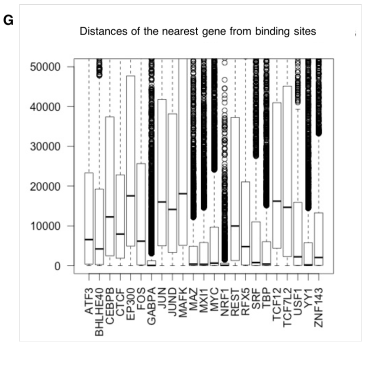
(B-C). Boxplot of *dh-ratio* (B), and *Hopkins statistic* (C) for 135 TF-cell pairs based on their sub-models, and pooled sub-models by TF. In (B), the horizontal line at Y=1 denotes the maximum limit of *dh-ratio*. In (C), the horizontal line at Y=0.6 denotes, current lowest value of *Hopkins statistic*.

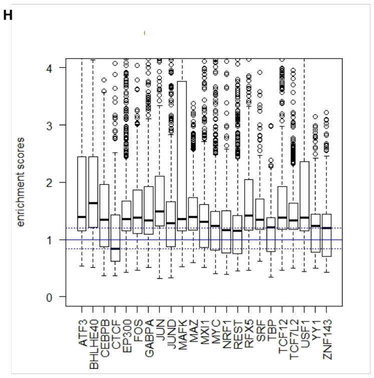




(D). Same plot as in Fig 2C, except the sub-models are clustered by XY-Fused (XYF) self-organizing map. In plot (D-F), the 'blue' horizontal line denotes the coherence in 5% of the total multi-clusters.

(E-F). Functional and Expression coherence of sub-model clusters with expression threshold of log2CPM>=5, i.e. a gene is considered as on when the log2CPM>=5. ~40% (~18%) multi-cell clusters show higher expression-coherence (pathway-coherence). Dual coherence denotes both expression and pathway coherence. (E) is drawn for *k*-NN (*k*-Nearest Neighbor) and (F) is drawn for XYF (XY fused network).

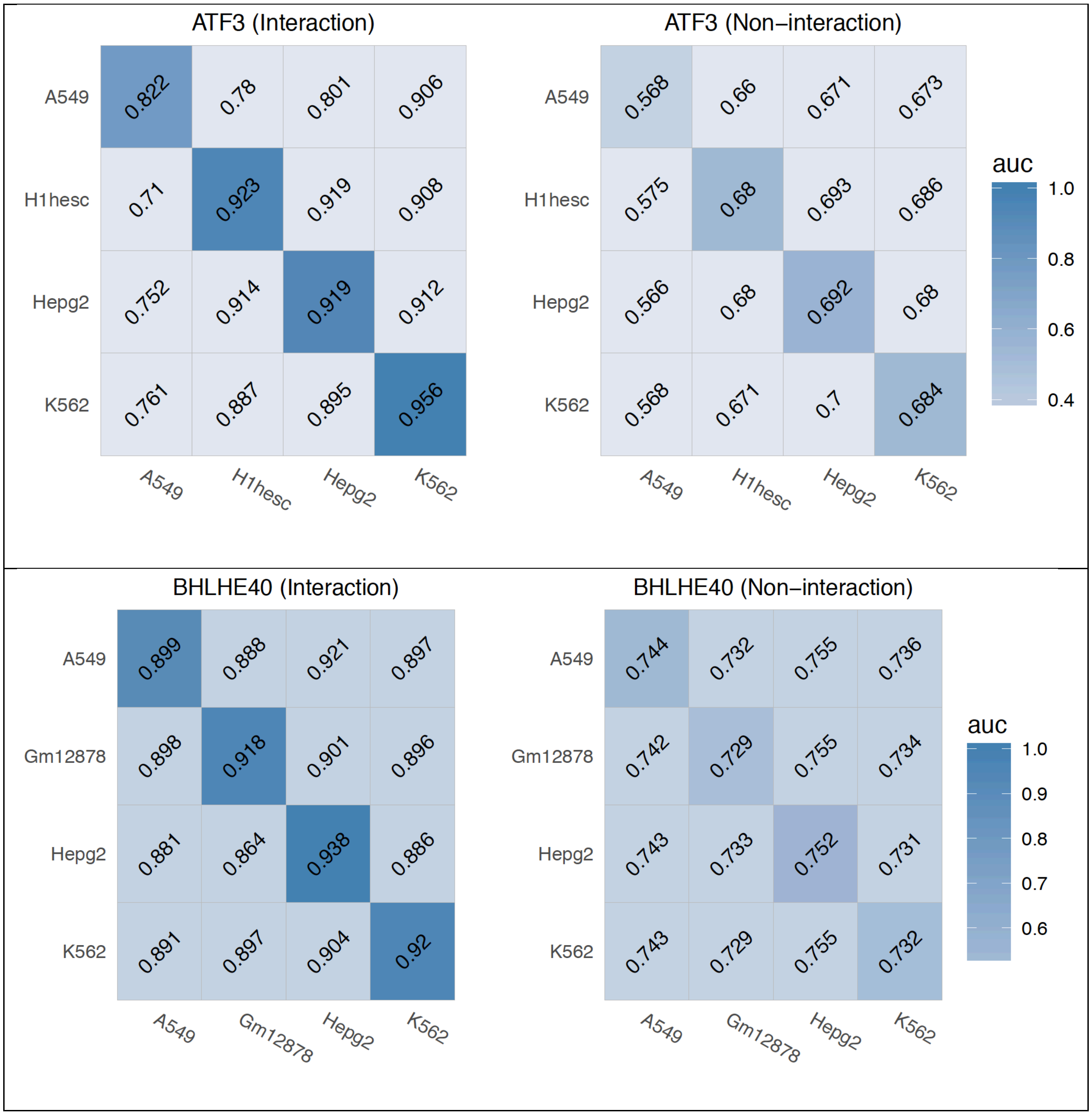


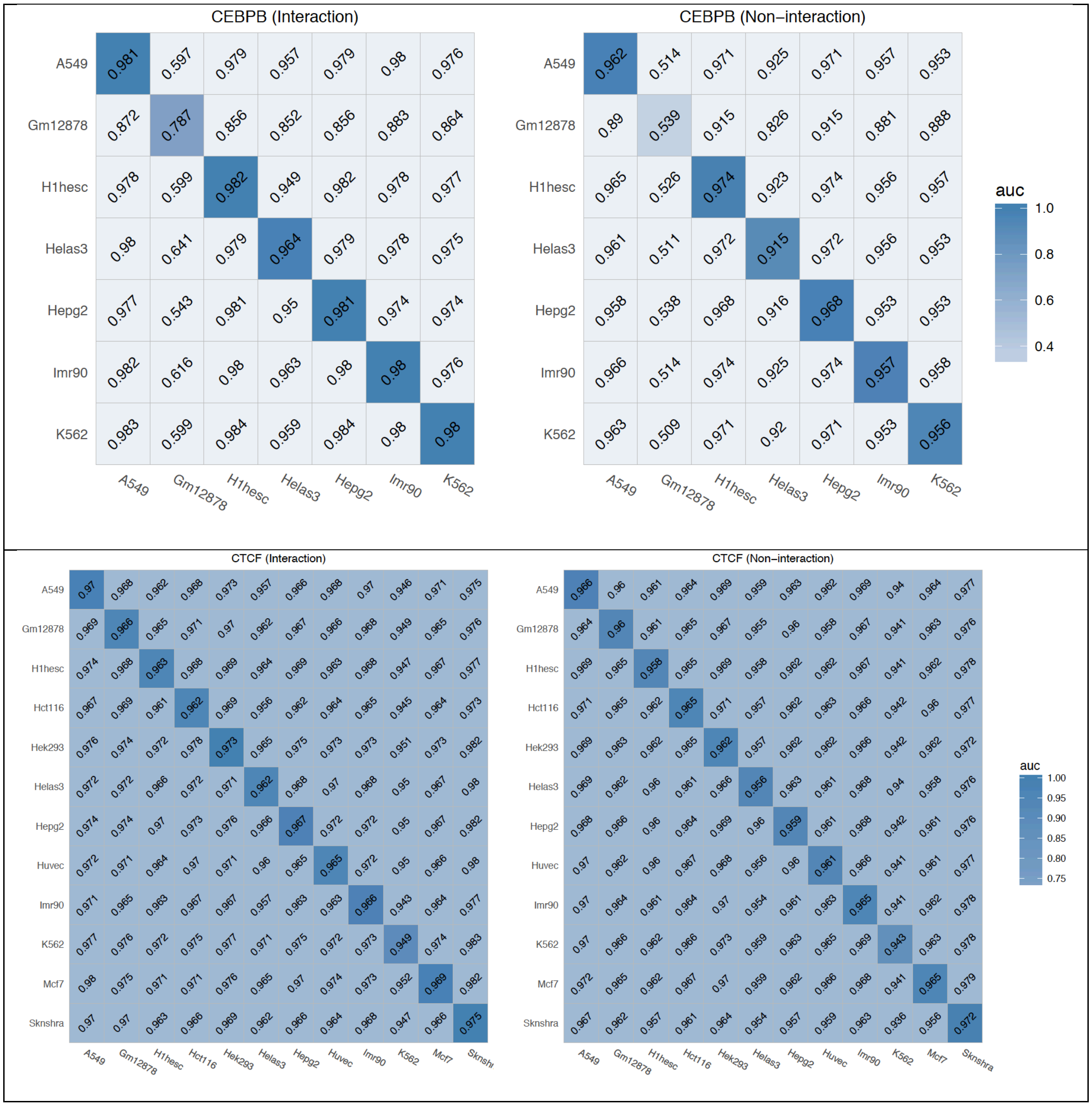


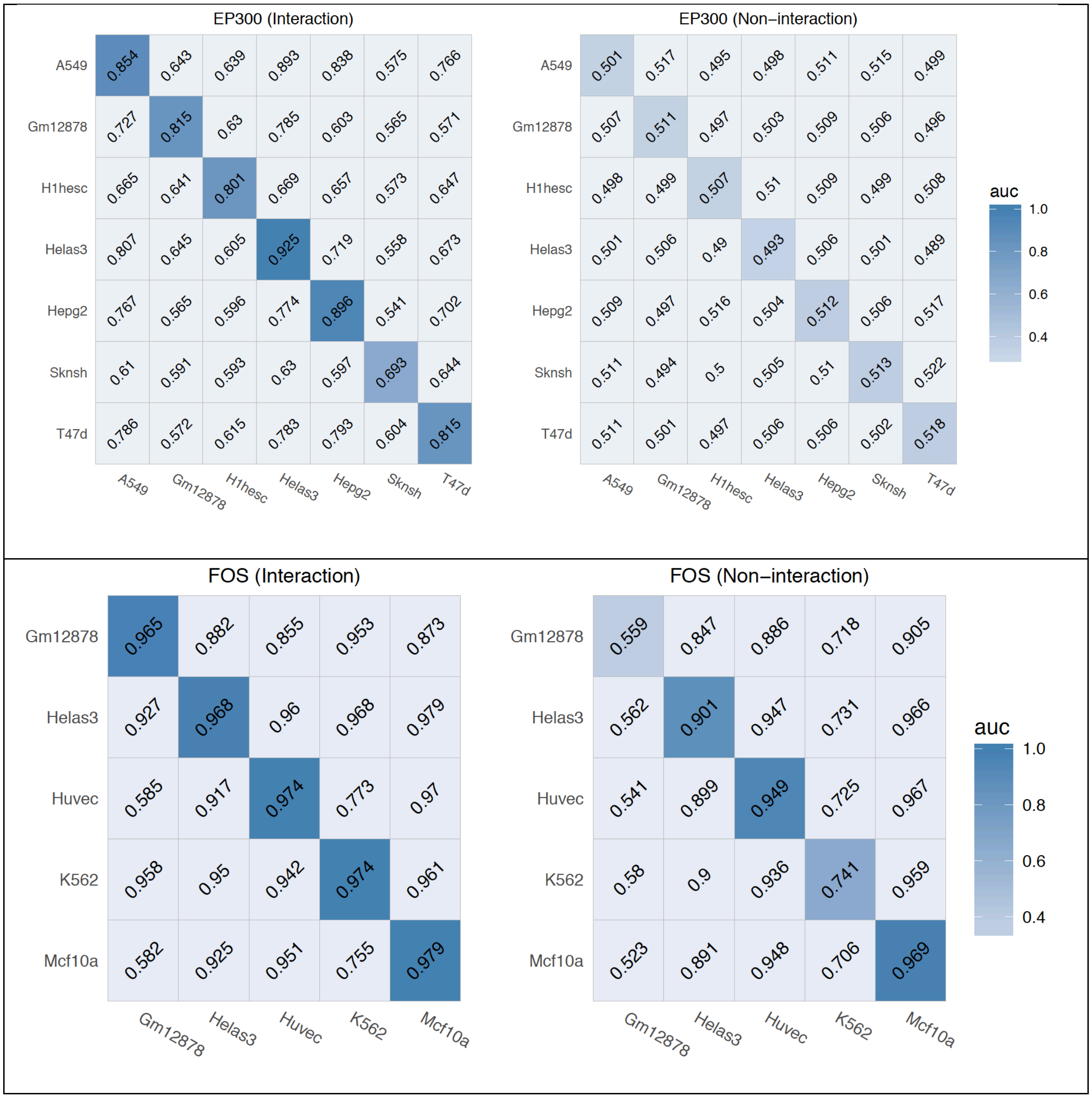
(G). Distance between binding sites and their nearest gene.

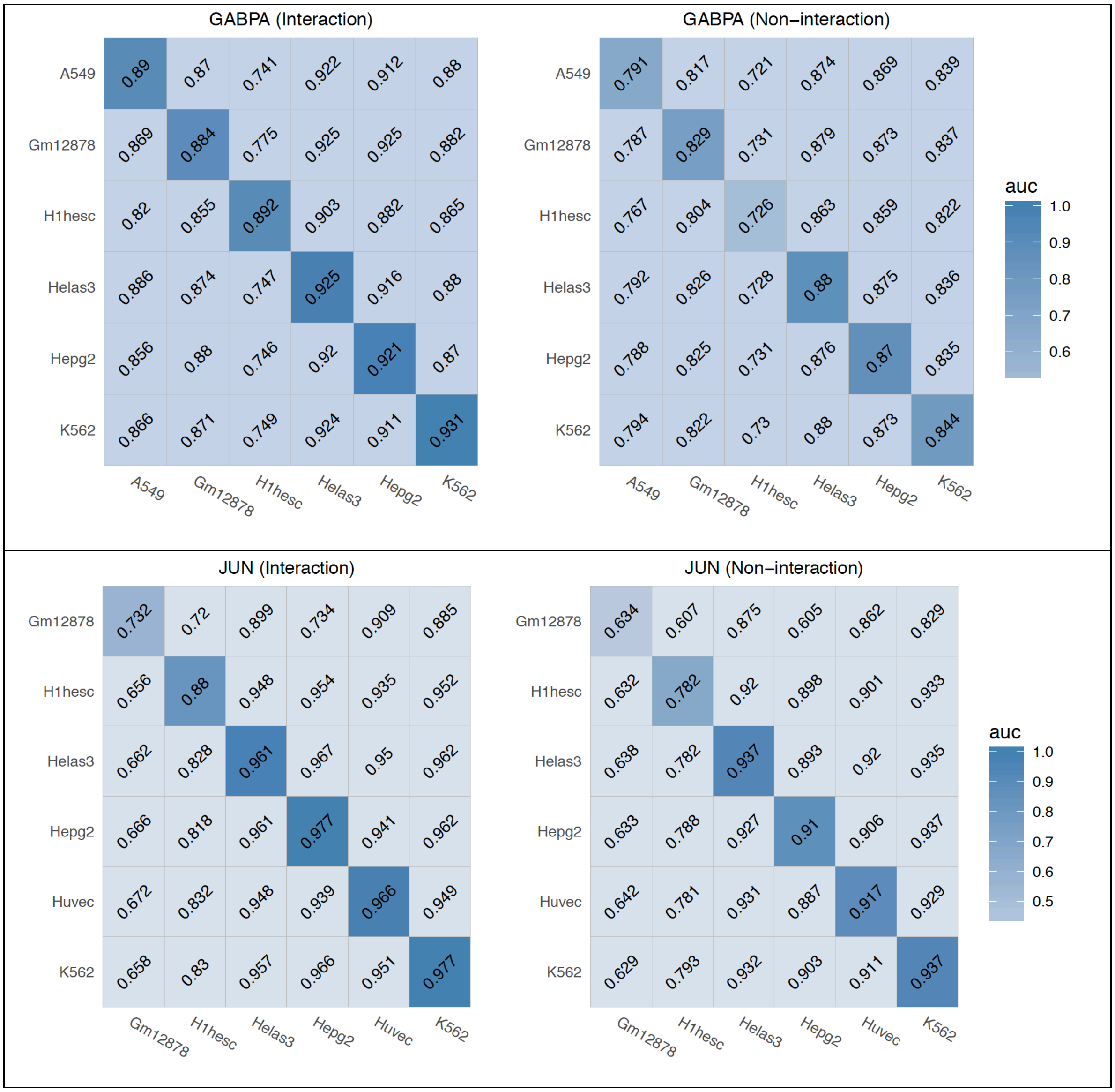
(H). Distribution of enrichment scores of all relevant co-factor motifs (with nonzero feature importance), for each TF separately. The horizontal line at Y=1 denotes no enrichment/depletion, and the upper and lower dotted horizontal line denotes enrichment and depletion of 1.2 respectively.

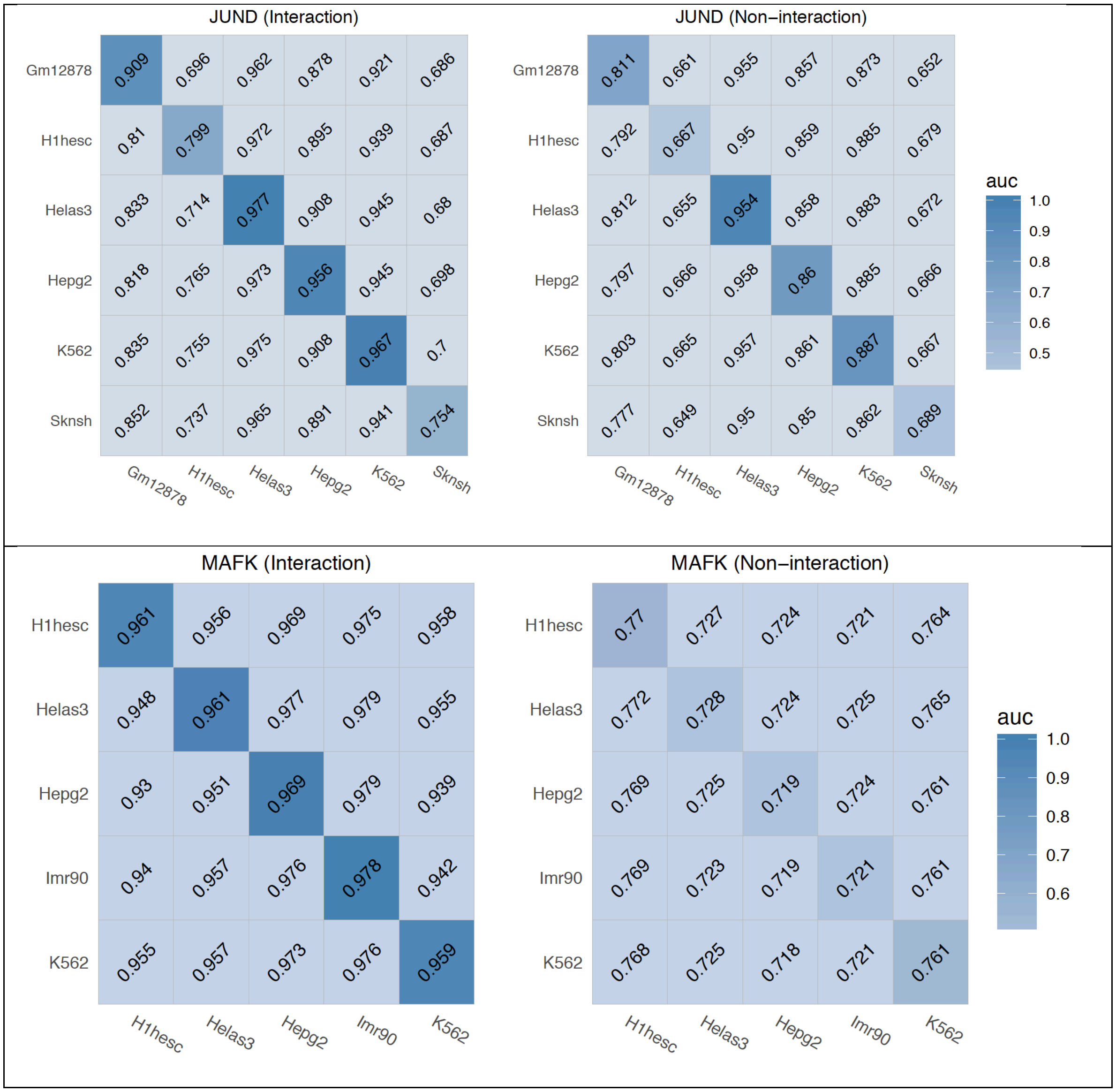
Figure 7.7 Assessment of clusters and associated genes.

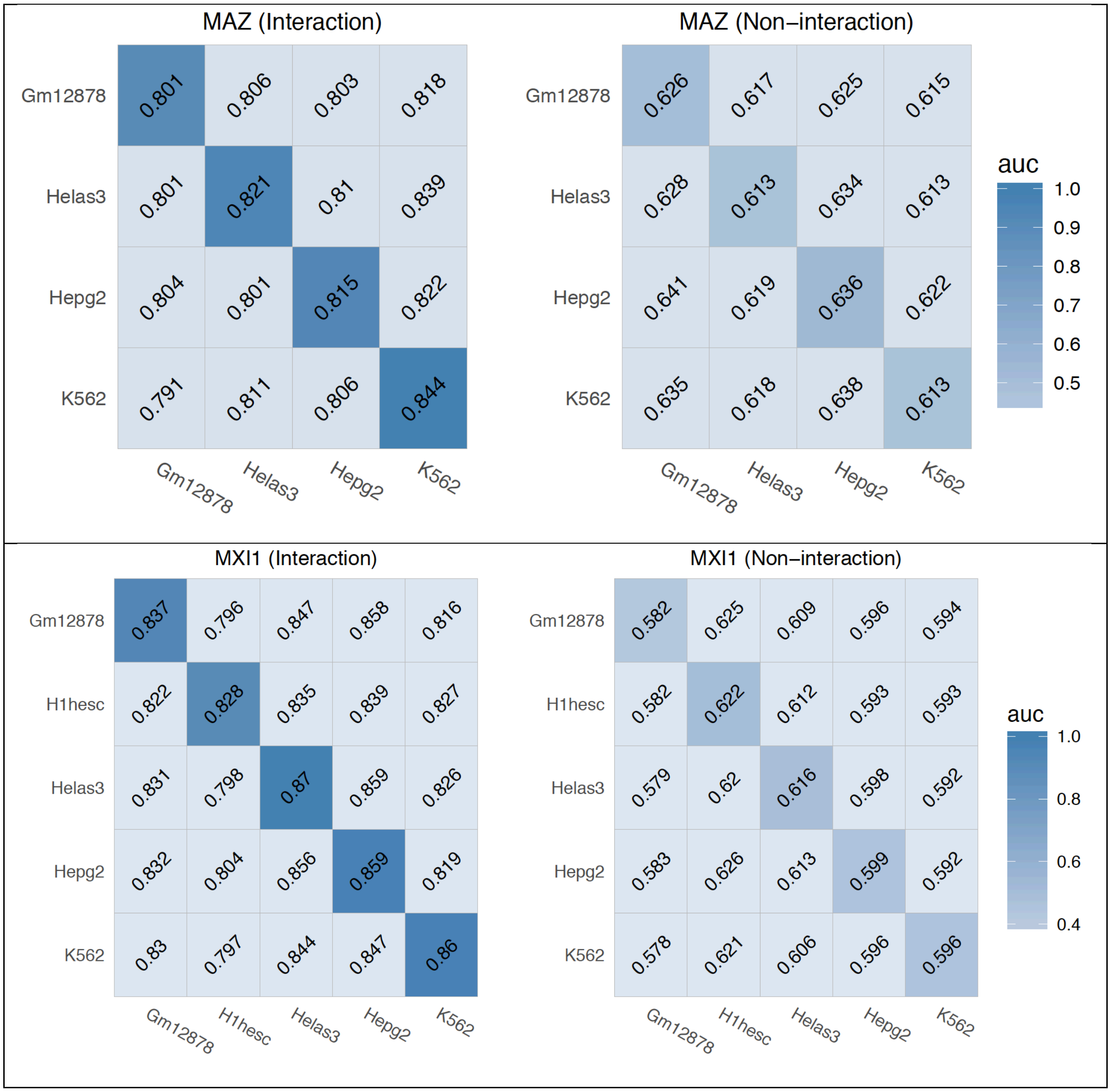


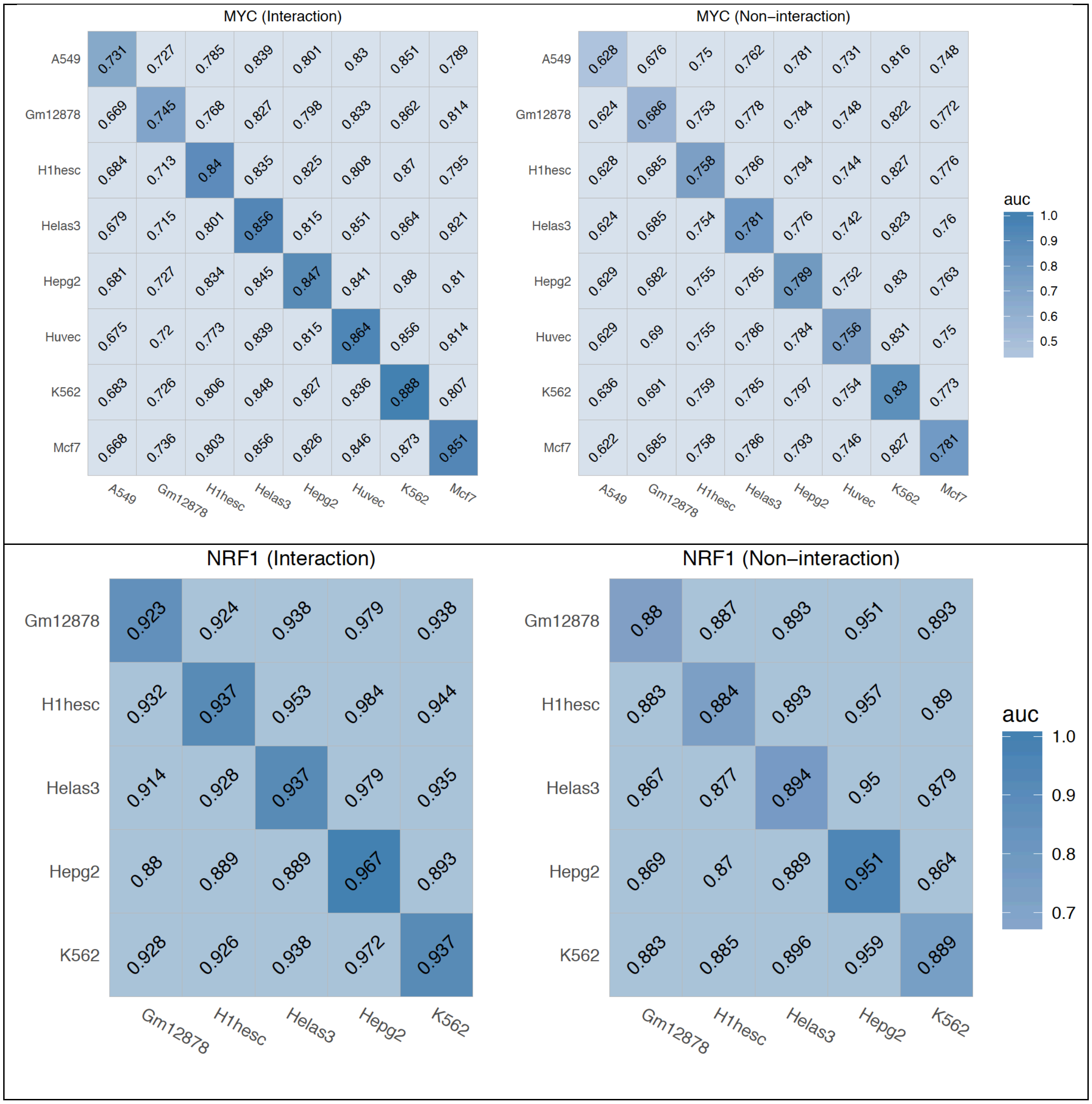

CEBPB (Interaction)
A549 0.981 0.597 0.979 0.957 0.979 0.98 0.976
Gm12878 0.872 0.787 0.856 0.852 0.856 0.883 0.864
H1hesc 0.978 0.599 0.982 0.949 0.982 0.978 0.977
Helas3 0.98 0.641 0.979 0.964 0.979 0.978 0.975
Hepg2 0.977 0.543 0.981 0.95 0.981 0.974 0.974
Imr90 0.982 0.616 0.98 0.963 0.98 0.98 0.976
K562 0.983 0.599 0.984 0.959 0.984 0.98 0.98
A549 Gm12878 H1hesc Helas3 Hepg2 Imr90 K562
CEBPB (Non−interaction)
A549 0.962 0.514 0.971 0.925 0.971 0.957 0.953
Gm12878 0.89 0.539 0.915 0.826 0.915 0.881 0.888
H1hesc 0.965 0.526 0.974 0.923 0.974 0.956 0.957
Helas3 0.961 0.511 0.972 0.915 0.972 0.956 0.953
Hepg2 0.958 0.538 0.968 0.916 0.968 0.953 0.953
Imr90 0.966 0.514 0.974 0.925 0.974 0.957 0.958
K562 0.963 0.509 0.971 0.92 0.971 0.953 0.956
A549 Gm12878 H1hesc Helas3 Hepg2 Imr90 K562
auc
1.0
0.8
0.6
0.4
CTCF (Interaction)
A549 0.97 0.968 0.962 0.968 0.973 0.957 0.966 0.968 0.97 0.946 0.971 0.975
Gm12878 0.969 0.966 0.965 0.971 0.97 0.962 0.967 0.966 0.968 0.949 0.965 0.976
H1hesc 0.974 0.968 0.963 0.968 0.969 0.964 0.969 0.963 0.968 0.947 0.967 0.977
Hct116 0.967 0.969 0.961 0.962 0.969 0.956 0.962 0.964 0.965 0.945 0.964 0.973
Hek293 0.976 0.974 0.972 0.978 0.973 0.965 0.975 0.973 0.973 0.951 0.973 0.982
Helas3 0.972 0.972 0.966 0.972 0.971 0.962 0.968 0.97 0.968 0.95 0.967 0.98
Hepg2 0.974 0.974 0.97 0.973 0.976 0.966 0.967 0.972 0.972 0.95 0.967 0.982
Huvec 0.972 0.971 0.964 0.97 0.971 0.96 0.965 0.965 0.972 0.95 0.966 0.98
Imr90 0.971 0.965 0.963 0.967 0.967 0.957 0.963 0.963 0.966 0.943 0.964 0.977
K562 0.977 0.976 0.972 0.975 0.977 0.971 0.975 0.972 0.973 0.949 0.974 0.983
Mcf7 0.98 0.975 0.971 0.971 0.976 0.965 0.97 0.974 0.973 0.952 0.969 0.982
Sknshra 0.97 0.97 0.963 0.966 0.969 0.962 0.966 0.964 0.968 0.947 0.966 0.975
A549 Gm12878 H1hesc Hct116 Hek293 Helas3 Hepg2 Huvec Imr90 K562 Mcf7 Sknsh
CTCF (Non−interaction)
A549 0.966 0.96 0.961 0.964 0.969 0.959 0.963 0.962 0.969 0.94 0.964 0.977
Gm12878 0.964 0.96 0.961 0.965 0.967 0.955 0.96 0.958 0.967 0.941 0.963 0.976
H1hesc 0.969 0.965 0.958 0.965 0.969 0.958 0.962 0.962 0.967 0.941 0.962 0.978
Hct116 0.971 0.965 0.962 0.965 0.971 0.957 0.962 0.963 0.966 0.942 0.96 0.977
Hek293 0.969 0.963 0.962 0.965 0.962 0.957 0.962 0.962 0.966 0.942 0.962 0.972
Helas3 0.969 0.962 0.96 0.961 0.966 0.956 0.963 0.961 0.968 0.94 0.958 0.976
Hepg2 0.968 0.966 0.96 0.964 0.969 0.96 0.959 0.961 0.968 0.942 0.961 0.976
Huvec 0.97 0.962 0.96 0.967 0.968 0.956 0.96 0.961 0.966 0.941 0.961 0.977
Imr90 0.97 0.964 0.961 0.964 0.97 0.954 0.961 0.963 0.965 0.941 0.962 0.978
K562 0.97 0.966 0.962 0.966 0.973 0.959 0.963 0.965 0.969 0.943 0.963 0.978
Mcf7 0.972 0.965 0.962 0.967 0.97 0.959 0.962 0.966 0.968 0.941 0.965 0.979
Sknshra 0.967 0.962 0.957 0.961 0.964 0.954 0.957 0.959 0.963 0.936 0.956 0.972
A549 Gm12878 H1hesc Hct116 Hek293 Helas3 Hepg2 Huvec Imr90 K562 Mcf7 Sknshra
auc
1.00
0.95
0.90
0.85
0.80
0.75

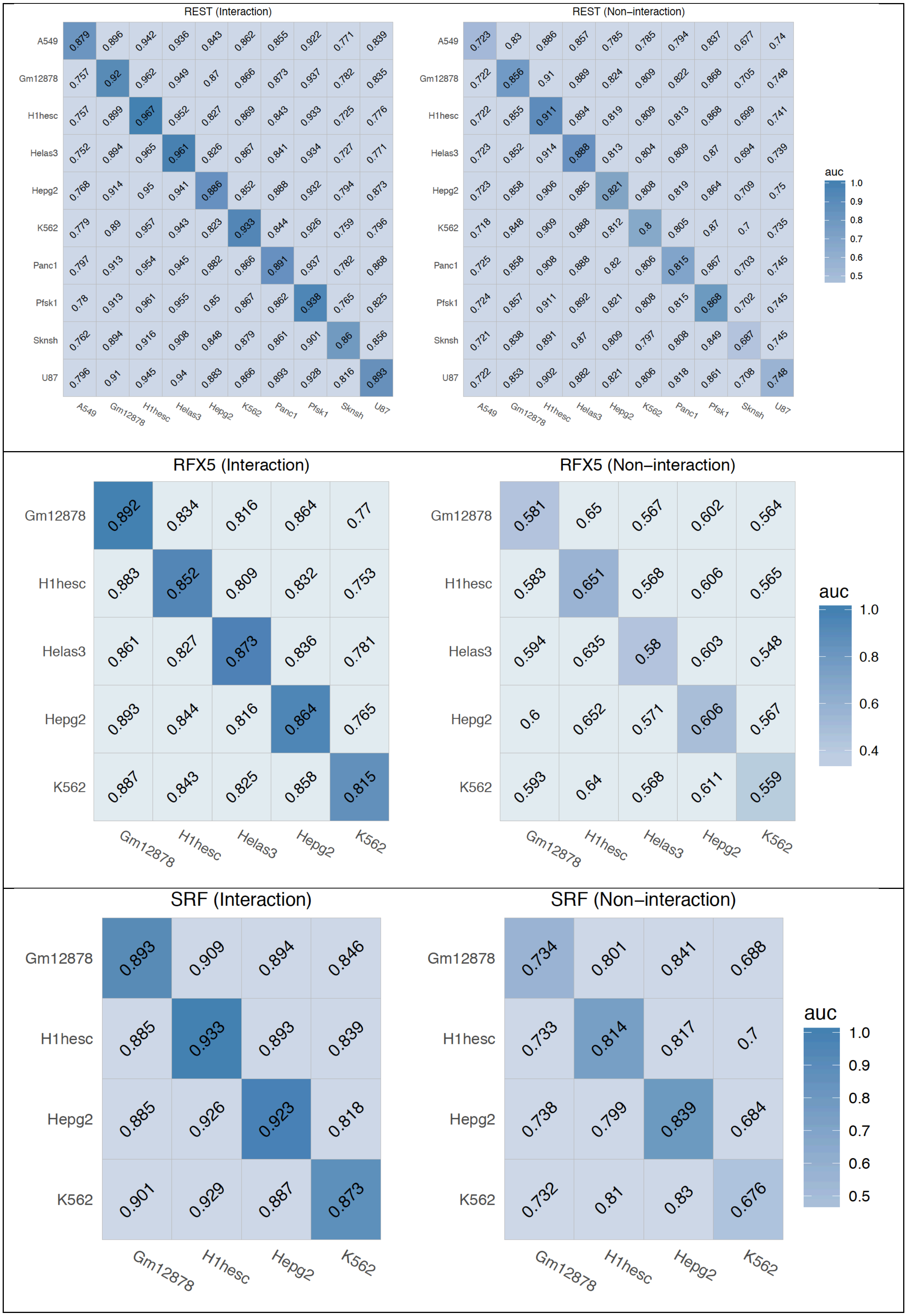

EP300 (Interaction)
A549 0.854 0.643 0.639 0.893 0.838 0.575 0.766
Gm12878 0.727 0.815 0.63 0.785 0.603 0.565 0.571
H1hesc 0.665 0.641 0.801 0.669 0.657 0.573 0.647
Helas3 0.807 0.645 0.605 0.925 0.719 0.558 0.673
Hepg2 0.767 0.565 0.596 0.774 0.896 0.541 0.702
Sknsh 0.61 0.591 0.593 0.63 0.597 0.693 0.644
T47d 0.786 0.572 0.615 0.783 0.793 0.604 0.815
A549 Gm12878 H1hesc Helas3 Hepg2 Sknsh T47d
EP300 (Non–interaction)
A549 0.501 0.517 0.495 0.498 0.511 0.515 0.499
Gm12878 0.507 0.511 0.497 0.503 0.509 0.506 0.496
H1hesc 0.498 0.499 0.507 0.51 0.509 0.499 0.508
Helas3 0.501 0.506 0.49 0.493 0.506 0.501 0.489
Hepg2 0.509 0.497 0.516 0.504 0.512 0.506 0.517
Sknsh 0.511 0.494 0.5 0.505 0.51 0.513 0.522
T47d 0.511 0.501 0.497 0.506 0.506 0.502 0.518
A549 Gm12878 H1hesc Helas3 Hepg2 Sknsh T47d
auc
1.0
0.8
0.6
0.4
FOS (Interaction)
Gm12878 0.965 0.882 0.855 0.953 0.873
Helas3 0.927 0.968 0.96 0.968 0.979
Huvec 0.585 0.917 0.974 0.773 0.97
K562 0.958 0.95 0.942 0.974 0.961
Mcf10a 0.582 0.925 0.951 0.755 0.979
Gm12878 Helas3 Huvec K562 Mcf10a
FOS (Non–interaction)
Gm12878 0.559 0.847 0.886 0.718 0.905
Helas3 0.562 0.901 0.947 0.731 0.966
Huvec 0.541 0.899 0.949 0.725 0.967
K562 0.58 0.9 0.936 0.741 0.959
Mcf10a 0.523 0.891 0.948 0.706 0.969
Gm12878 Helas3 Huvec K562 Mcf10a
auc
1.0
0.8
0.6
0.4

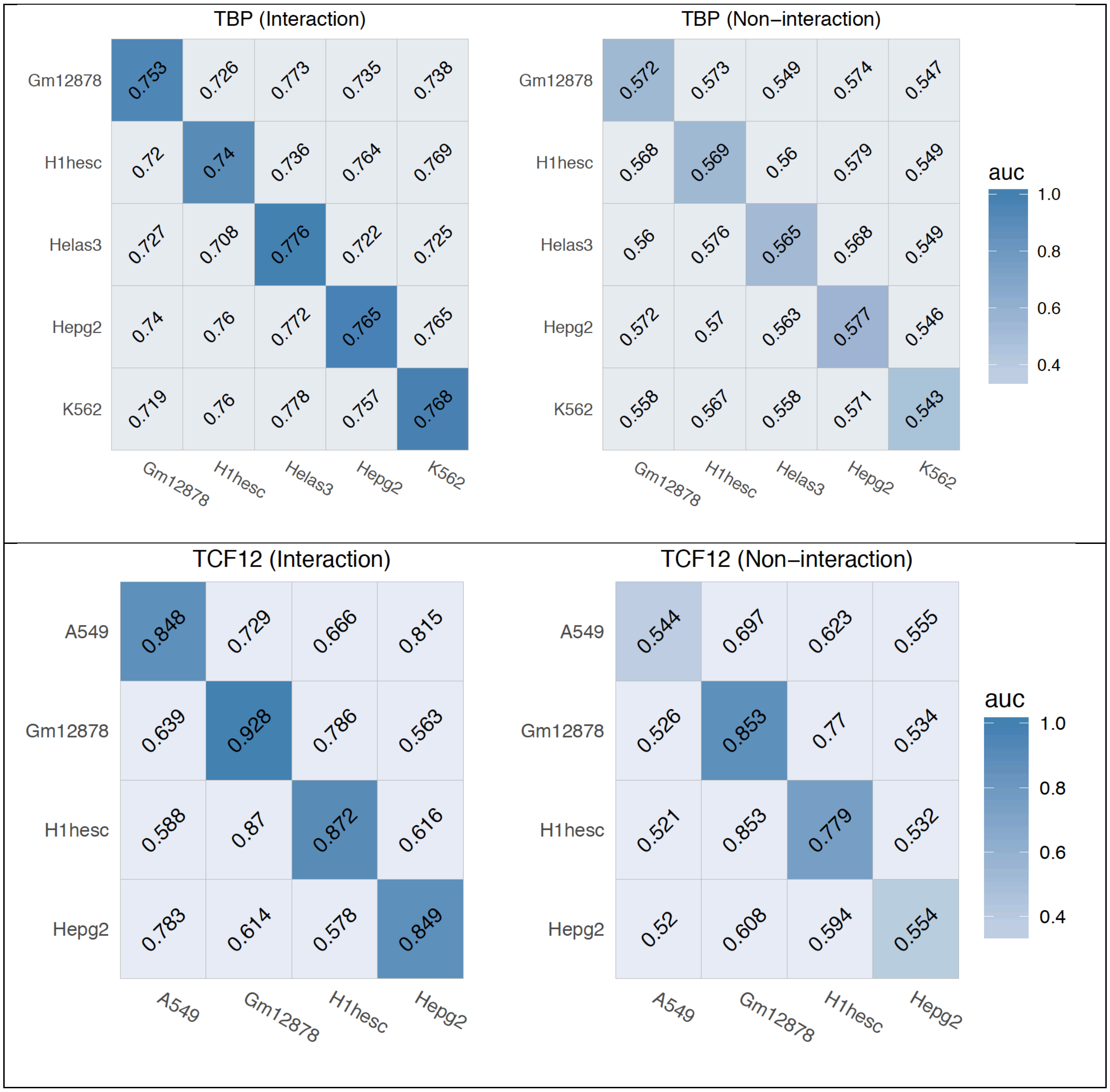

GABPA (Interaction)
A549
Gm12878
H1hesc
Helas3
Hepg2
K562
0.89 0.87 0.741 0.922 0.912 0.88
0.869 0.884 0.775 0.925 0.925 0.882
0.82 0.855 0.892 0.903 0.882 0.865
0.886 0.874 0.747 0.925 0.916 0.88
0.856 0.88 0.746 0.92 0.921 0.87
0.866 0.871 0.749 0.924 0.911 0.931
A549
Gm12878
H1hesc
Helas3
Hepg2
K562
GABPA (Non–interaction)
A549
Gm12878
H1hesc
Helas3
Hepg2
K562
0.791 0.817 0.721 0.874 0.869 0.839
0.787 0.829 0.731 0.879 0.873 0.837
0.767 0.804 0.726 0.863 0.859 0.822
0.792 0.826 0.728 0.88 0.875 0.836
0.788 0.825 0.731 0.876 0.87 0.835
0.794 0.822 0.73 0.88 0.873 0.844
A549
Gm12878
H1hesc
Helas3
Hepg2
K562
auc
1.0
0.9
0.8
0.7
0.6
JUN (Interaction)
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562
0.732 0.72 0.899 0.734 0.909 0.885
0.656 0.88 0.948 0.954 0.935 0.952
0.662 0.828 0.961 0.967 0.95 0.962
0.666 0.818 0.961 0.977 0.941 0.962
0.672 0.832 0.948 0.939 0.966 0.949
0.658 0.83 0.957 0.966 0.951 0.977
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562
JUN (Non–interaction)
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562
0.634 0.607 0.875 0.605 0.862 0.829
0.632 0.782 0.92 0.898 0.901 0.933
0.638 0.782 0.937 0.893 0.92 0.935
0.633 0.788 0.927 0.91 0.906 0.937
0.642 0.781 0.931 0.887 0.917 0.929
0.629 0.793 0.932 0.903 0.911 0.937
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562
auc
1.0
0.9
0.8
0.7
0.6
0.5

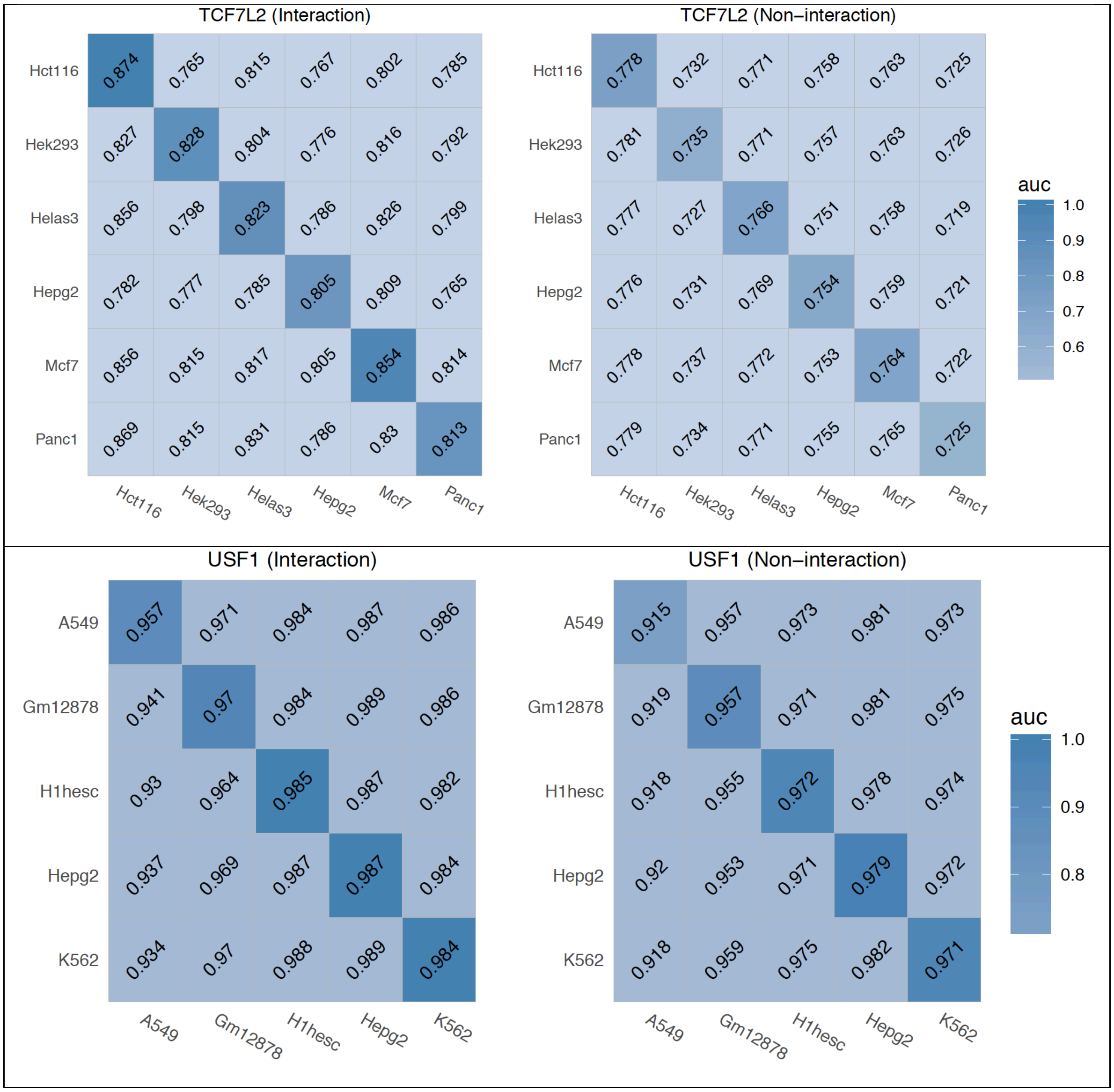

JUND (Interaction)
Gm12878 0.909 0.696 0.962 0.878 0.921 0.686
H1hesc 0.81 0.799 0.972 0.895 0.939 0.687
Helas3 0.833 0.714 0.977 0.908 0.945 0.68
Hepg2 0.818 0.765 0.973 0.956 0.945 0.698
K562 0.835 0.755 0.975 0.908 0.967 0.7
Sknsh 0.852 0.737 0.965 0.891 0.941 0.754
Gm12878 H1hesc Helas3 Hepg2 K562 Sknsh
JUND (Non–interaction)
Gm12878 0.811 0.661 0.955 0.857 0.873 0.652
H1hesc 0.792 0.667 0.95 0.859 0.885 0.679
Helas3 0.812 0.655 0.954 0.858 0.883 0.672
Hepg2 0.797 0.666 0.958 0.86 0.885 0.666
K562 0.803 0.665 0.957 0.861 0.887 0.667
Sknsh 0.777 0.649 0.95 0.85 0.862 0.689
Gm12878 H1hesc Helas3 Hepg2 K562 Sknsh
auc
1.0
0.9
0.8
0.7
0.6
0.5
MAFK (Interaction)
H1hesc 0.961 0.956 0.969 0.975 0.958
Helas3 0.948 0.961 0.977 0.979 0.955
Hepg2 0.93 0.951 0.969 0.979 0.939
Imr90 0.94 0.957 0.976 0.978 0.942
K562 0.955 0.957 0.973 0.976 0.959
H1hesc Helas3 Hepg2 Imr90 K562
MAFK (Non–interaction)
H1hesc 0.77 0.727 0.724 0.721 0.764
Helas3 0.772 0.728 0.724 0.725 0.765
Hepg2 0.769 0.725 0.719 0.724 0.761
Imr90 0.769 0.723 0.719 0.721 0.761
K562 0.768 0.725 0.718 0.721 0.761
H1hesc Helas3 Hepg2 Imr90 K562
auc
1.0
0.9
0.8
0.7
0.6

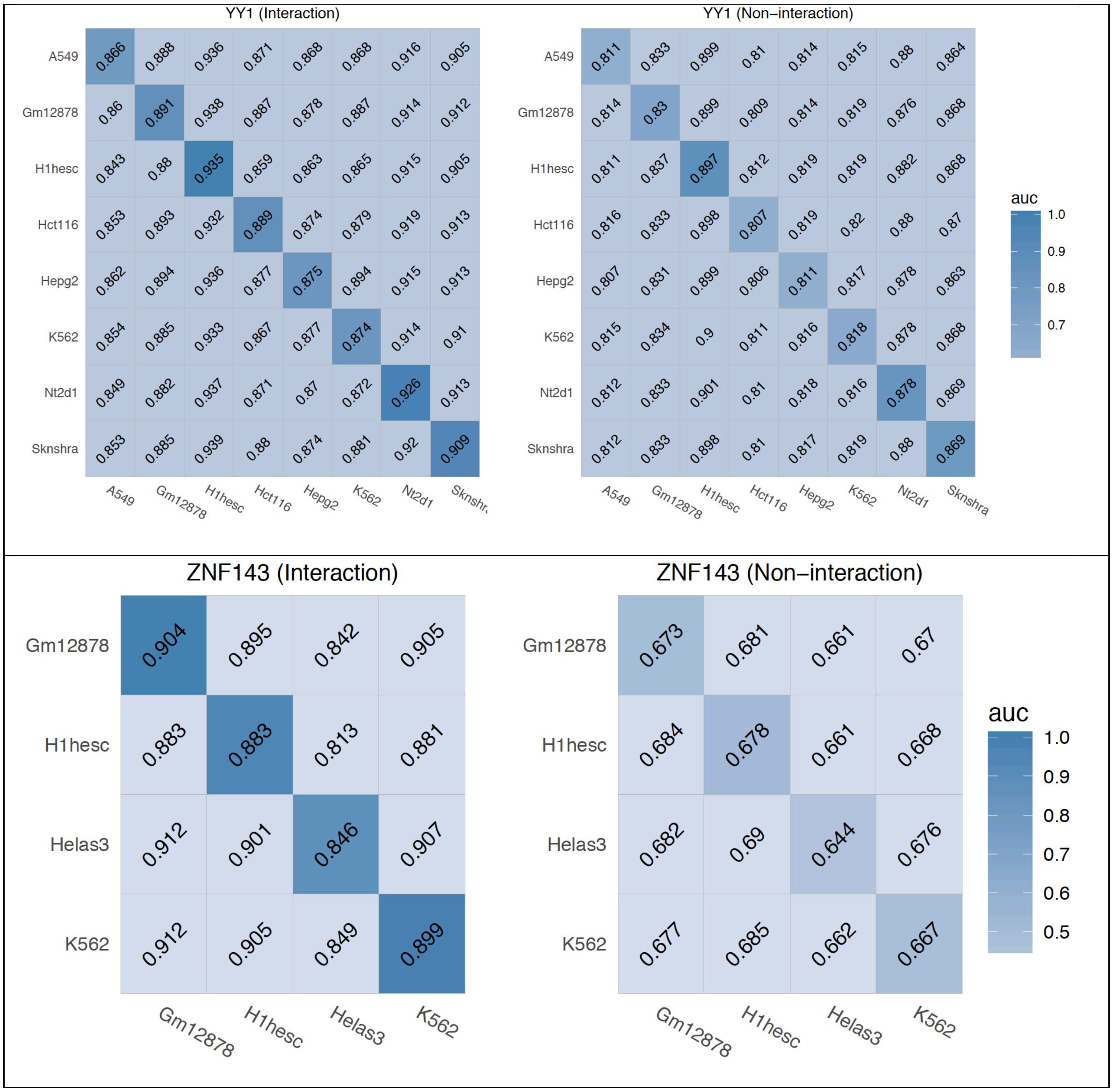

MAZ (Interaction)
Gm12878 0.801 0.806 0.803 0.818
Helas3 0.801 0.821 0.81 0.839
Hepg2 0.804 0.801 0.815 0.822
K562 0.791 0.811 0.806 0.844
Gm12878 Helas3 Hepg2 K562
MAZ (Non−interaction)
Gm12878 0.626 0.617 0.625 0.615
Helas3 0.628 0.613 0.634 0.613
Hepg2 0.641 0.619 0.636 0.622
K562 0.635 0.618 0.638 0.613
Gm12878 Helas3 Hepg2 K562
auc
1.0
0.9
0.8
0.7
0.6
0.5
MXI1 (Interaction)
Gm12878 0.837 0.796 0.847 0.858 0.816
H1hesc 0.822 0.828 0.835 0.839 0.827
Helas3 0.831 0.798 0.87 0.859 0.826
Hepg2 0.832 0.804 0.856 0.859 0.819
K562 0.83 0.797 0.844 0.847 0.86
Gm12878 H1hesc Helas3 Hepg2 K562
MXI1 (Non−interaction)
Gm12878 0.582 0.625 0.609 0.596 0.594
H1hesc 0.582 0.622 0.612 0.593 0.593
Helas3 0.579 0.62 0.616 0.598 0.592
Hepg2 0.583 0.626 0.613 0.599 0.592
K562 0.578 0.621 0.606 0.596 0.596
Gm12878 H1hesc Helas3 Hepg2 K562
auc
1.0
0.8
0.6
0.4


MYC (Interaction)
A549 0.731 0.727 0.785 0.839 0.801 0.83 0.851 0.789
Gm12878 0.669 0.745 0.768 0.827 0.798 0.833 0.862 0.814
H1hesc 0.684 0.713 0.84 0.835 0.825 0.808 0.87 0.795
Helas3 0.679 0.715 0.801 0.856 0.815 0.851 0.864 0.821
Hepg2 0.681 0.727 0.834 0.845 0.847 0.841 0.88 0.81
Huvec 0.675 0.72 0.773 0.839 0.815 0.864 0.856 0.814
K562 0.683 0.726 0.806 0.848 0.827 0.836 0.888 0.807
Mcf7 0.668 0.736 0.803 0.856 0.826 0.846 0.873 0.851
A549 Gm12878 H1hesc Helas3 Hepg2 Huvec K562 Mcf7
MYC (Non–interaction)
A549 0.628 0.676 0.75 0.762 0.781 0.731 0.816 0.748
Gm12878 0.624 0.686 0.753 0.778 0.784 0.748 0.822 0.772
H1hesc 0.628 0.685 0.758 0.786 0.794 0.744 0.827 0.776
Helas3 0.624 0.685 0.754 0.781 0.776 0.742 0.823 0.76
Hepg2 0.629 0.682 0.755 0.785 0.789 0.752 0.83 0.763
Huvec 0.629 0.69 0.755 0.786 0.784 0.756 0.831 0.75
K562 0.636 0.691 0.759 0.785 0.797 0.754 0.83 0.773
Mcf7 0.622 0.685 0.758 0.786 0.793 0.746 0.827 0.781
A549 Gm12878 H1hesc Helas3 Hepg2 Huvec K562 Mcf7
auc
1.0
0.9
0.8
0.7
0.6
0.5
NRF1 (Interaction)
Gm12878 0.923 0.924 0.938 0.979 0.938
H1hesc 0.932 0.937 0.953 0.984 0.944
Helas3 0.914 0.928 0.937 0.979 0.935
Hepg2 0.88 0.889 0.889 0.967 0.893
K562 0.928 0.926 0.938 0.972 0.937
Gm12878 H1hesc Helas3 Hepg2 K562
NRF1 (Non–interaction)
Gm12878 0.88 0.887 0.893 0.951 0.893
H1hesc 0.883 0.884 0.893 0.957 0.89
Helas3 0.867 0.877 0.894 0.95 0.879
Hepg2 0.869 0.87 0.889 0.951 0.864
K562 0.883 0.885 0.896 0.959 0.889
Gm12878 H1hesc Helas3 Hepg2 K562
auc
1.0
0.9
0.8
0.7


REST (Interaction)
A549 0.879 0.896 0.942 0.936 0.843 0.862 0.855 0.922 0.771 0.839
Gm12878 0.757 0.92 0.962 0.949 0.87 0.866 0.873 0.937 0.782 0.835
H1hesc 0.757 0.899 0.967 0.952 0.827 0.869 0.843 0.933 0.725 0.776
Helas3 0.752 0.894 0.965 0.961 0.826 0.867 0.841 0.934 0.727 0.771
Hepg2 0.768 0.914 0.95 0.941 0.886 0.852 0.888 0.932 0.794 0.873
K562 0.779 0.89 0.957 0.943 0.823 0.933 0.844 0.926 0.759 0.796
Panc1 0.797 0.913 0.954 0.945 0.882 0.866 0.891 0.937 0.782 0.868
Pfsk1 0.78 0.913 0.961 0.955 0.85 0.867 0.862 0.938 0.765 0.825
Sknsh 0.762 0.894 0.916 0.908 0.848 0.879 0.861 0.901 0.86 0.856
U87 0.796 0.91 0.945 0.94 0.883 0.866 0.893 0.928 0.816 0.893
A549 Gm12878 H1hesc Helas3 Hepg2 K562 Panc1 Pfsk1 Sknsh U87
REST (Non–interaction)
A549 0.723 0.83 0.886 0.857 0.785 0.785 0.794 0.837 0.677 0.74
Gm12878 0.722 0.856 0.91 0.889 0.824 0.809 0.822 0.868 0.705 0.748
H1hesc 0.722 0.855 0.911 0.894 0.819 0.809 0.813 0.868 0.699 0.741
Helas3 0.723 0.852 0.914 0.888 0.813 0.804 0.809 0.87 0.694 0.739
Hepg2 0.723 0.858 0.906 0.885 0.821 0.808 0.819 0.864 0.709 0.75
K562 0.718 0.848 0.909 0.888 0.812 0.8 0.805 0.87 0.7 0.735
Panc1 0.725 0.858 0.908 0.888 0.82 0.806 0.815 0.867 0.703 0.745
Pfsk1 0.724 0.857 0.911 0.892 0.821 0.808 0.815 0.868 0.702 0.745
Sknsh 0.721 0.838 0.891 0.87 0.809 0.797 0.808 0.849 0.687 0.745
U87 0.722 0.853 0.902 0.882 0.821 0.806 0.818 0.861 0.708 0.748
A549 Gm12878 H1hesc Helas3 Hepg2 K562 Panc1 Pfsk1 Sknsh U87
auc
1.0
0.9
0.8
0.7
0.6
0.5
RFX5 (Interaction)
Gm12878 0.892 0.834 0.816 0.864 0.77
H1hesc 0.883 0.852 0.809 0.832 0.753
Helas3 0.861 0.827 0.873 0.836 0.781
Hepg2 0.893 0.844 0.816 0.864 0.765
K562 0.887 0.843 0.825 0.858 0.815
Gm12878 H1hesc Helas3 Hepg2 K562
RFX5 (Non–interaction)
Gm12878 0.581 0.65 0.567 0.602 0.564
H1hesc 0.583 0.651 0.568 0.606 0.565
Helas3 0.594 0.635 0.58 0.603 0.548
Hepg2 0.6 0.652 0.571 0.606 0.567
K562 0.593 0.64 0.568 0.611 0.559
Gm12878 H1hesc Helas3 Hepg2 K562
auc
1.0
0.8
0.6
0.4
SRF (Interaction)
Gm12878 0.893 0.909 0.894 0.846
H1hesc 0.885 0.933 0.893 0.839
Hepg2 0.885 0.926 0.923 0.818
K562 0.901 0.929 0.887 0.873
Gm12878 H1hesc Hepg2 K562
SRF (Non–interaction)
Gm12878 0.734 0.801 0.841 0.688
H1hesc 0.733 0.814 0.817 0.7
Hepg2 0.738 0.799 0.839 0.684
K562 0.732 0.81 0.83 0.676
Gm12878 H1hesc Hepg2 K562
auc
1.0
0.9
0.8
0.7
0.6
0.5


TBP (Interaction)
Gm12878
H1hesc
Helas3
Hepg2
K562
0.753
0.726
0.773
0.735
0.738
0.72
0.74
0.736
0.764
0.769
0.727
0.708
0.776
0.722
0.725
0.74
0.76
0.772
0.765
0.765
0.719
0.76
0.778
0.757
0.768
Gm12878
H1hesc
Helas3
Hepg2
K562
TBP (Non−interaction)
Gm12878
H1hesc
Helas3
Hepg2
K562
0.572
0.573
0.549
0.574
0.547
0.568
0.569
0.56
0.579
0.549
0.56
0.576
0.565
0.568
0.549
0.572
0.57
0.563
0.577
0.546
0.558
0.567
0.558
0.571
0.543
Gm12878
H1hesc
Helas3
Hepg2
K562
auc
1.0
0.8
0.6
0.4
TCF12 (Interaction)
A549
Gm12878
H1hesc
Hepg2
0.848
0.729
0.666
0.815
0.639
0.928
0.786
0.563
0.588
0.87
0.872
0.616
0.783
0.614
0.578
0.849
A549
Gm12878
H1hesc
Hepg2
TCF12 (Non−interaction)
A549
Gm12878
H1hesc
Hepg2
0.544
0.697
0.623
0.555
0.526
0.853
0.77
0.534
0.521
0.853
0.779
0.532
0.52
0.608
0.594
0.554
A549
Gm12878
H1hesc
Hepg2
auc
1.0
0.8
0.6
0.4


TCF7L2 (Interaction)
Hct116
0.874
0.765
0.815
0.767
0.802
0.785
Hek293
0.827
0.828
0.804
0.776
0.816
0.792
Helas3
0.856
0.798
0.823
0.786
0.826
0.799
Hepg2
0.782
0.777
0.785
0.805
0.809
0.765
Mcf7
0.856
0.815
0.817
0.805
0.854
0.814
Panc1
0.869
0.815
0.831
0.786
0.83
0.813
Hct116
Hek293
Helas3
Hepg2
Mcf7
Panc1
TCF7L2 (Non–interaction)
Hct116
0.778
0.732
0.771
0.758
0.763
0.725
Hek293
0.781
0.735
0.771
0.757
0.763
0.726
Helas3
0.777
0.727
0.766
0.751
0.758
0.719
Hepg2
0.776
0.731
0.769
0.754
0.759
0.721
Mcf7
0.778
0.737
0.772
0.753
0.764
0.722
Panc1
0.779
0.734
0.771
0.755
0.765
0.725
Hct116
Hek293
Helas3
Hepg2
Mcf7
Panc1
auc
1.0
0.9
0.8
0.7
0.6
USF1 (Interaction)
A549
0.957
0.971
0.984
0.987
0.986
Gm12878
0.941
0.97
0.984
0.989
0.986
H1hesc
0.93
0.964
0.985
0.987
0.982
Hepg2
0.937
0.969
0.987
0.987
0.984
K562
0.934
0.97
0.988
0.989
0.984
A549
Gm12878
H1hesc
Hepg2
K562
USF1 (Non–interaction)
A549
0.915
0.957
0.973
0.981
0.973
Gm12878
0.919
0.957
0.971
0.981
0.975
H1hesc
0.918
0.955
0.972
0.978
0.974
Hepg2
0.92
0.953
0.971
0.979
0.972
K562
0.918
0.959
0.975
0.982
0.971
A549
Gm12878
H1hesc
Hepg2
K562
auc
1.0
0.9
0.8



YY1 (Interaction)

	A549	Gm12878	H1hesc	Hct116	Hepg2	K562	Nt2d1	Sknshr
A549	0.866	0.888	0.936	0.871	0.868	0.868	0.916	0.905
Gm12878	0.86	0.891	0.938	0.887	0.878	0.887	0.914	0.912
H1hesc	0.843	0.88	0.935	0.859	0.863	0.865	0.915	0.905
Hct116	0.853	0.893	0.932	0.889	0.874	0.879	0.919	0.913
Hepg2	0.862	0.894	0.936	0.877	0.875	0.894	0.915	0.913
K562	0.854	0.885	0.933	0.867	0.877	0.874	0.914	0.91
Nt2d1	0.849	0.882	0.937	0.871	0.87	0.872	0.926	0.913
Sknshra	0.853	0.885	0.939	0.88	0.874	0.881	0.92	0.909

YY1 (Non–interaction)

	A549	Gm12878	H1hesc	Hct116	Hepg2	K562	Nt2d1	Sknshra
A549	0.811	0.833	0.899	0.81	0.814	0.815	0.88	0.864
Gm12878	0.814	0.83	0.899	0.809	0.814	0.819	0.876	0.868
H1hesc	0.811	0.837	0.897	0.812	0.819	0.819	0.882	0.868
Hct116	0.816	0.833	0.898	0.807	0.819	0.82	0.88	0.87
Hepg2	0.807	0.831	0.899	0.806	0.811	0.817	0.878	0.863
K562	0.815	0.834	0.9	0.811	0.816	0.818	0.878	0.868
Nt2d1	0.812	0.833	0.901	0.81	0.818	0.816	0.878	0.869
Sknshra	0.812	0.833	0.898	0.81	0.817	0.819	0.88	0.869

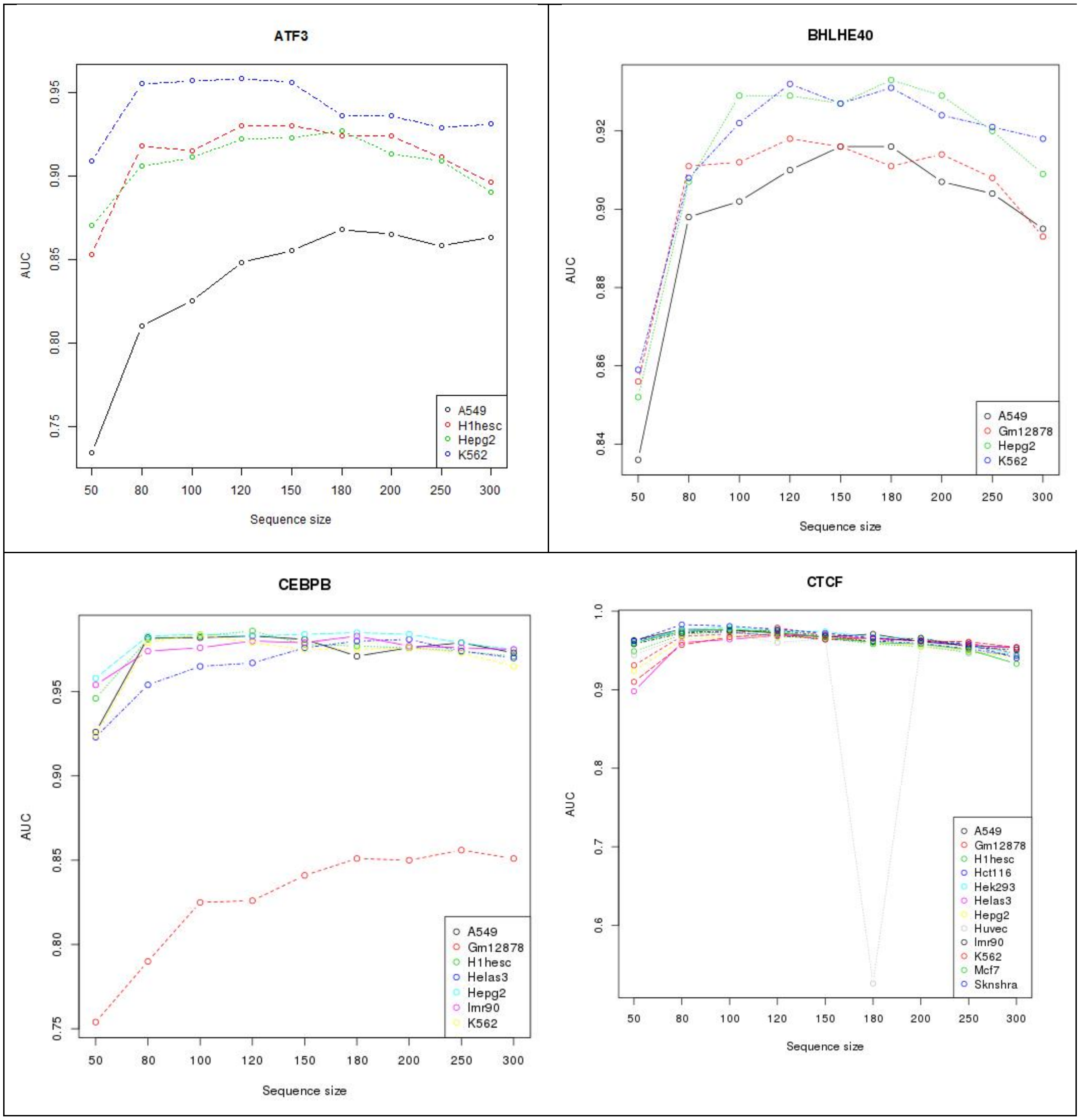
ZNF143 (Interaction)

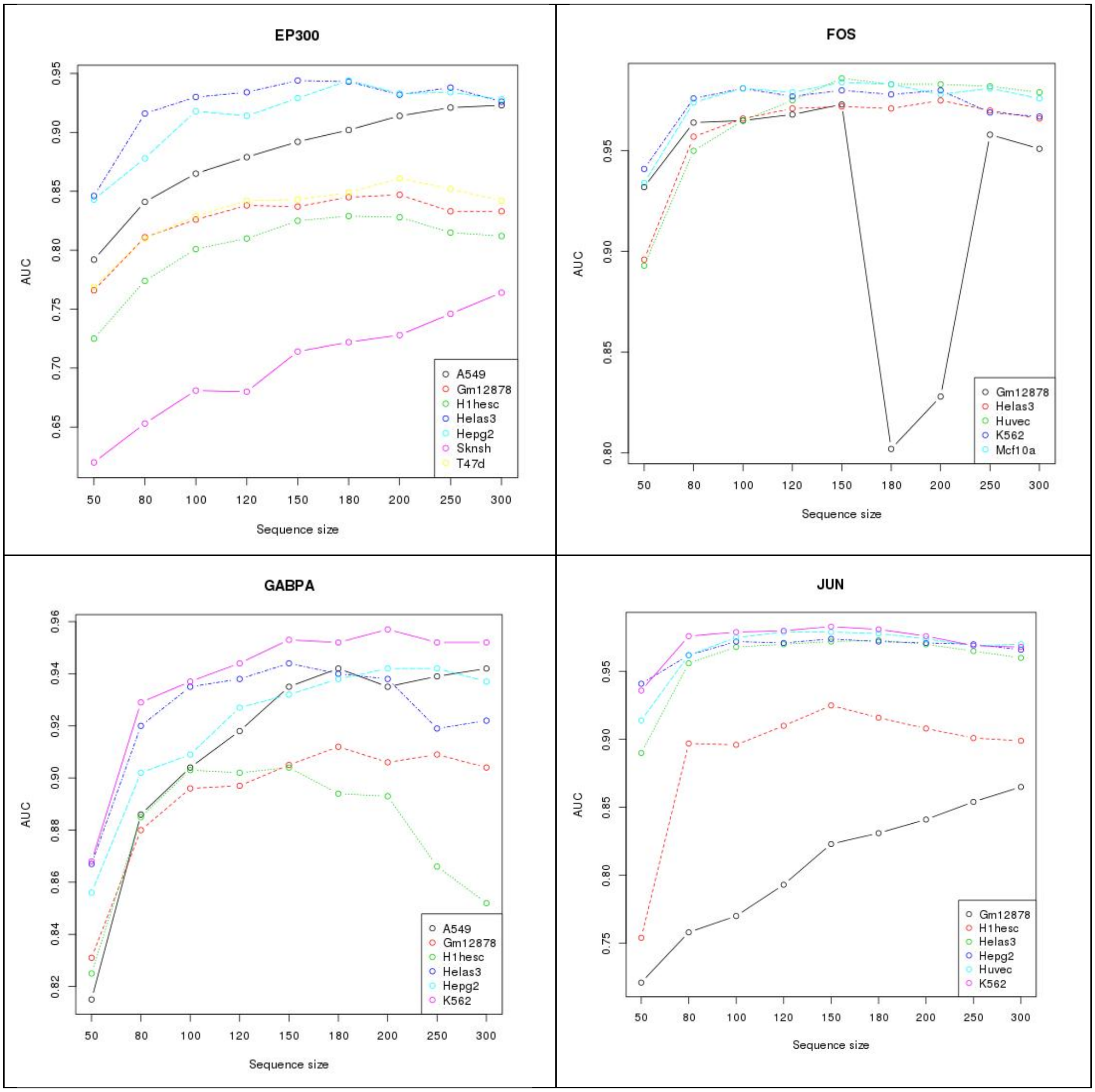
	Gm12878	H1hesc	Helas3	K562
Gm12878	0.904	0.895	0.842	0.905
H1hesc	0.883	0.883	0.813	0.881
Helas3	0.912	0.901	0.846	0.907
K562	0.912	0.905	0.849	0.899

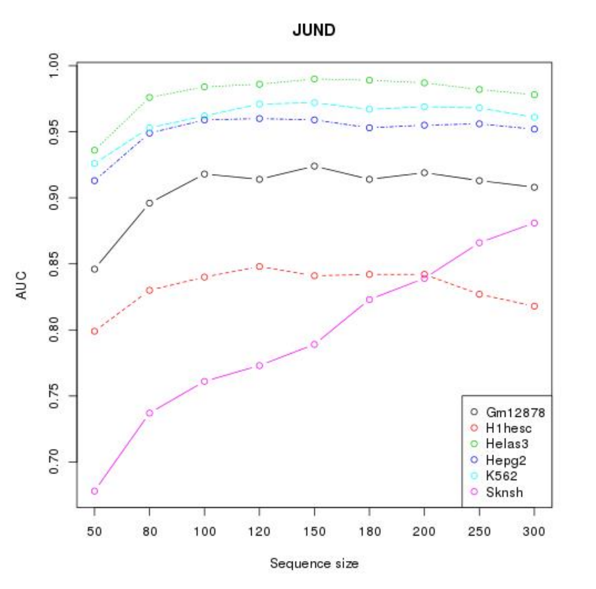
ZNF143 (Non–interaction)

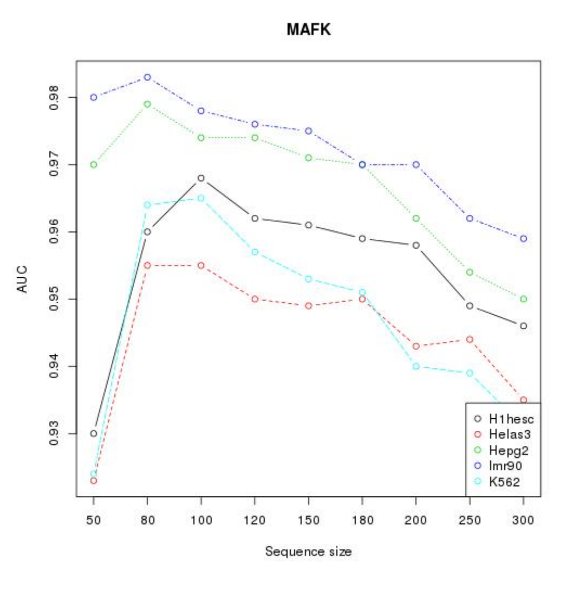
	Gm12878	H1hesc	Helas3	K562
Gm12878	0.673	0.681	0.661	0.67
H1hesc	0.684	0.678	0.661	0.668
Helas3	0.682	0.69	0.644	0.676
K562	0.677	0.685	0.662	0.667

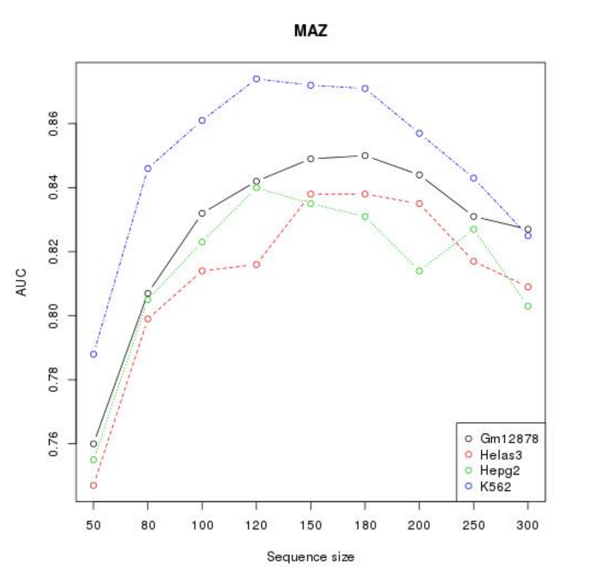
Figure 7.8 Cross-cell type performance matrix for Interaction and Noninteraction models. In each matrix, row represents the cell line used to build the model and column represents the cell line from which the test data is used. Diagonal elements are within cell type performance and only diagnal elements are colored according to the ROC-AUC to show the difference between Interaction and Noninteraction models.

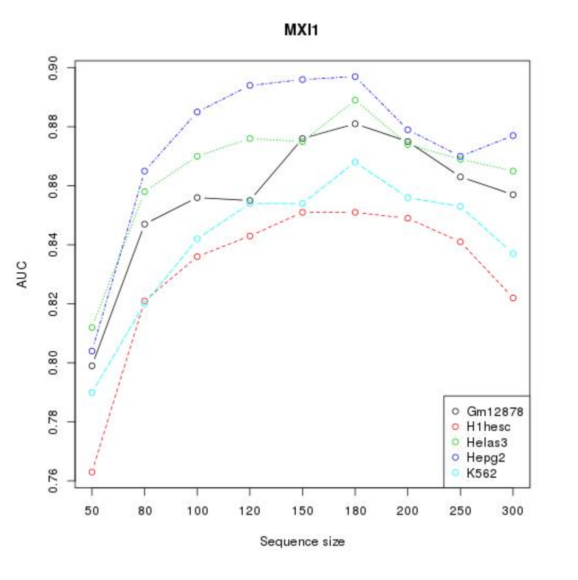

ATF3
AUC
0.75
0.80
0.85
0.90
0.95
50
80
100
120
150
180
200
250
300
Sequence size
A549
H1hesc
Hepg2
K562
BHLHE40
AUC
0.84
0.86
0.88
0.90
0.92
Sequence size
A549
Gm12878
Hepg2
K562
CEBPB
AUC
0.75
0.80
0.85
0.90
0.95
Sequence size
A549
Gm12878
H1hesc
Helas3
Hepg2
Imr90
K562
CTCF
AUC
0.6
0.7
0.8
0.9
1.0
Sequence size
A549
Gm12878
H1hesc
Hct116
Hek293
Helas3
Hepg2
Huvec
Imr90
K562
Mcf7
Sknshra

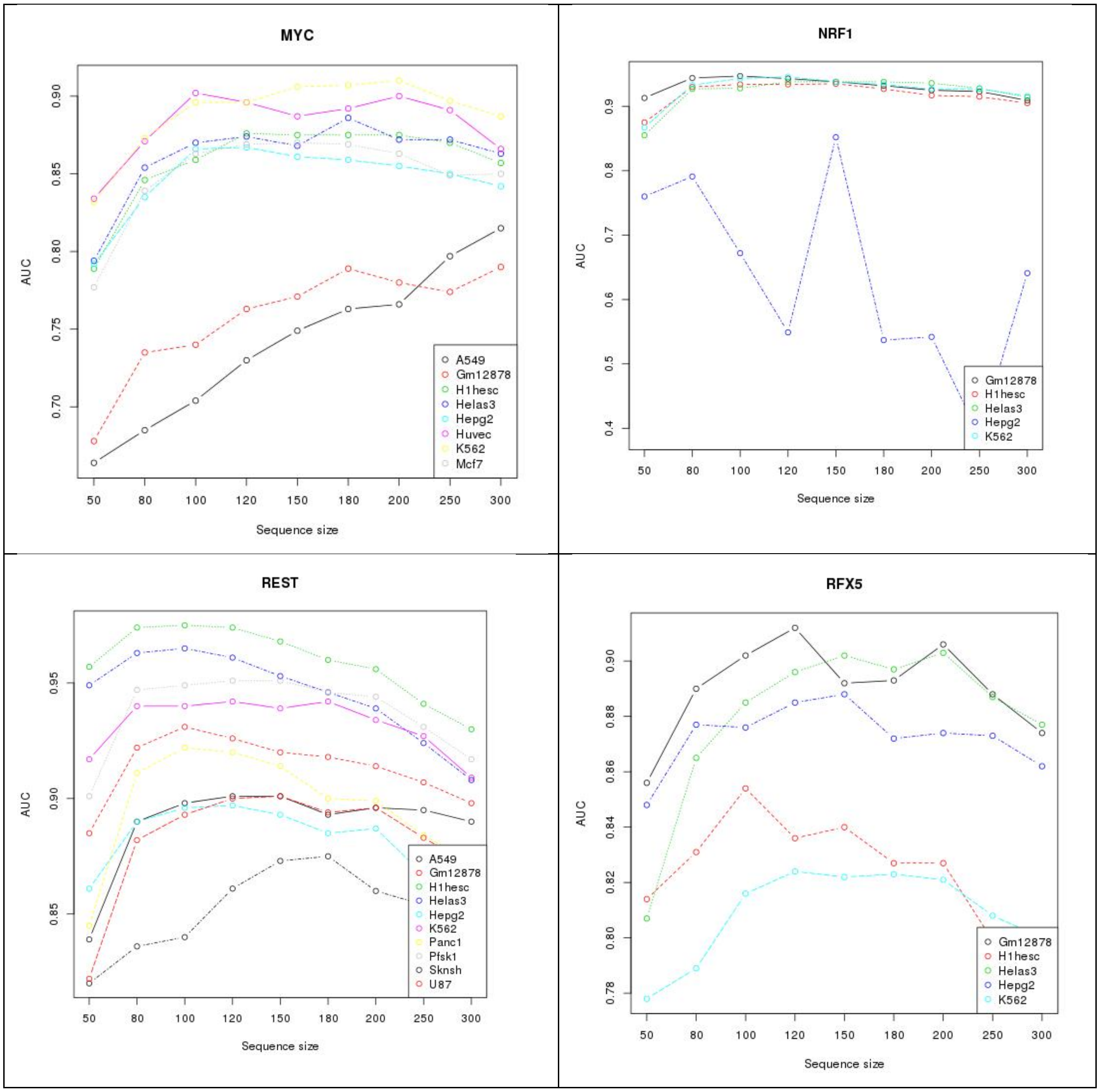

EP300
AUC
0.65
0.70
0.75
0.80
0.85
0.90
0.95
Sequence size
50
80
100
120
150
180
200
250
300
A549
Gm12878
H1hesc
Helas3
Hepg2
Sknsh
T47d
FOS
AUC
0.80
0.85
0.90
0.95
Sequence size
50
80
100
120
150
180
200
250
300
Gm12878
Helas3
Huvec
K562
Mcf10a
GABPA
AUC
0.82
0.84
0.86
0.88
0.90
0.92
0.94
0.96
Sequence size
50
80
100
120
150
180
200
250
300
A549
Gm12878
H1hesc
Helas3
Hepg2
K562
JUN
AUC
0.75
0.80
0.85
0.90
0.95
Sequence size
50
80
100
120
150
180
200
250
300
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562

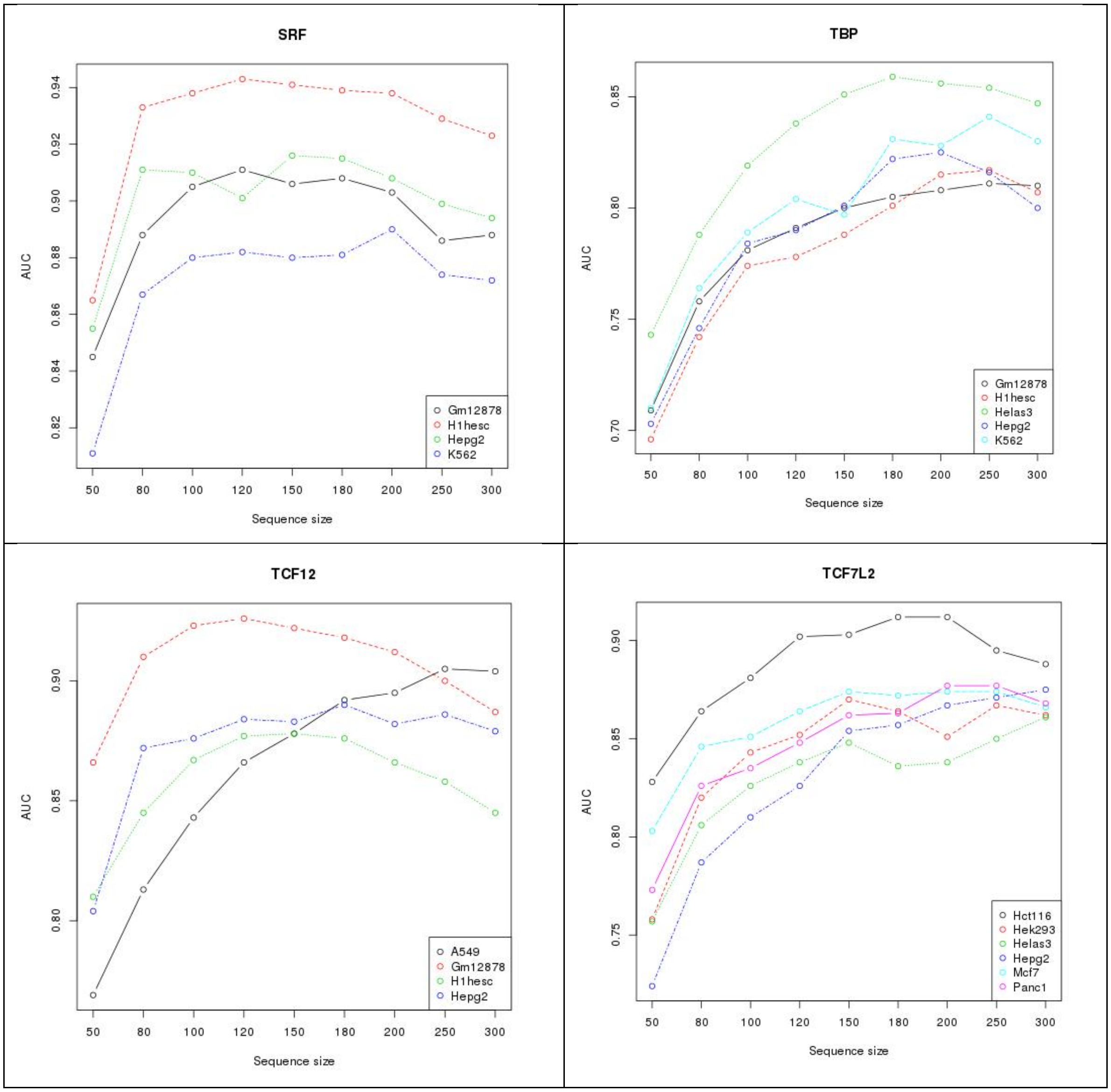

JUND
AUC
1.00
0.95
0.90
0.85
0.80
0.75
0.70
50
80
100
120
150
180
200
250
300
Sequence size
Gm12878
H1hesc
Helas3
Hepg2
K562
Sknsh


MAFK
AUC
0.98
0.97
0.96
0.95
0.94
0.93
50
80
100
120
150
180
200
250
300
Sequence size
H1hesc
Helas3
Hepg2
Imr90
K562


MAZ
AUC
0.86
0.84
0.82
0.80
0.78
0.76
50
80
100
120
150
180
200
250
300
Sequence size
Gm12878
Helas3
Hepg2
K562


MXI1
AUC
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
50
80
100
120
150
180
200
250
300
Sequence size
Gm12878
H1hesc
Helas3
Hepg2
K562


MYC
AUC
0.70
0.75
0.80
0.85
0.90
Sequence size
50
80
100
120
150
180
200
250
300
A549
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562
Mcf7
NRF1
AUC
0.4
0.5
0.6
0.7
0.8
0.9
Sequence size
50
80
100
120
150
180
200
250
300
Gm12878
H1hesc
Helas3
Hepg2
K562
REST
AUC
0.85
0.90
0.95
Sequence size
50
80
100
120
150
180
200
250
300
A549
Gm12878
H1hesc
Helas3
Hepg2
K562
Panc1
Pfsk1
Sknsh
U87
RFX5
AUC
0.78
0.80
0.82
0.84
0.86
0.88
0.90
Sequence size
50
80
100
120
150
180
200
250
300
Gm12878
H1hesc
Helas3
Hepg2
K562


SRF
AUC
0.82
0.84
0.86
0.88
0.90
0.92
0.94
Sequence size
50
80
100
120
150
180
200
250
300
Gm12878
H1hesc
Hepg2
K562
TBP
AUC
0.70
0.75
0.80
0.85
Sequence size
50
80
100
120
150
180
200
250
300
Gm12878
H1hesc
Helas3
Hepg2
K562
TCF12
AUC
0.80
0.85
0.90
Sequence size
50
80
100
120
150
180
200
250
300
A549
Gm12878
H1hesc
Hepg2
TCF7L2
AUC
0.75
0.80
0.85
0.90
Sequence size
50
80
100
120
150
180
200
250
300
Hct116
Hek293
Helas3
Hepg2
Mcf7
Panc1

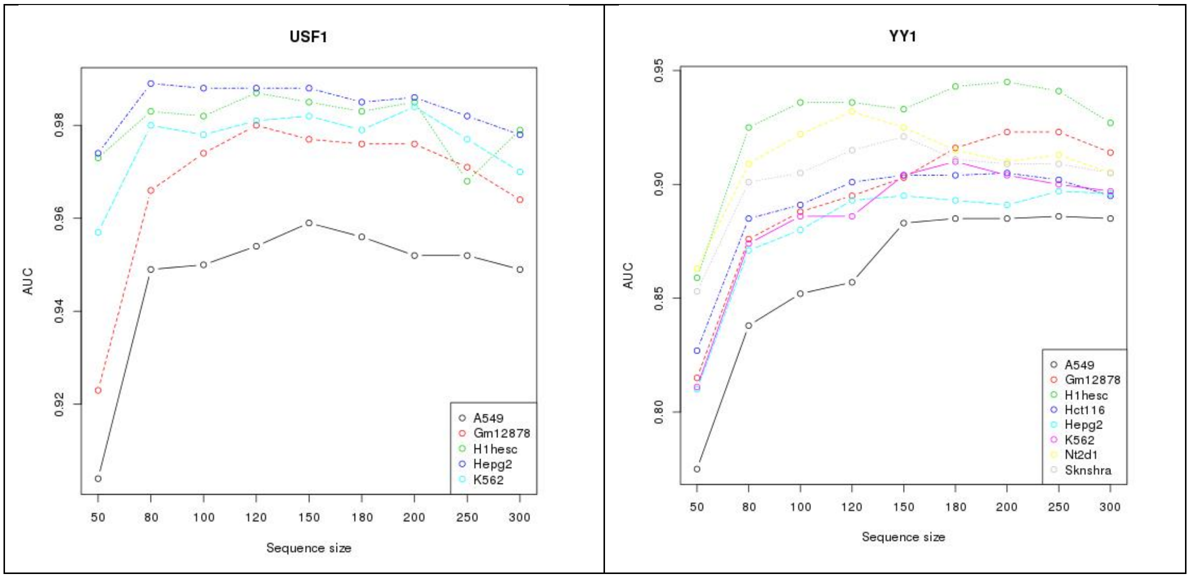
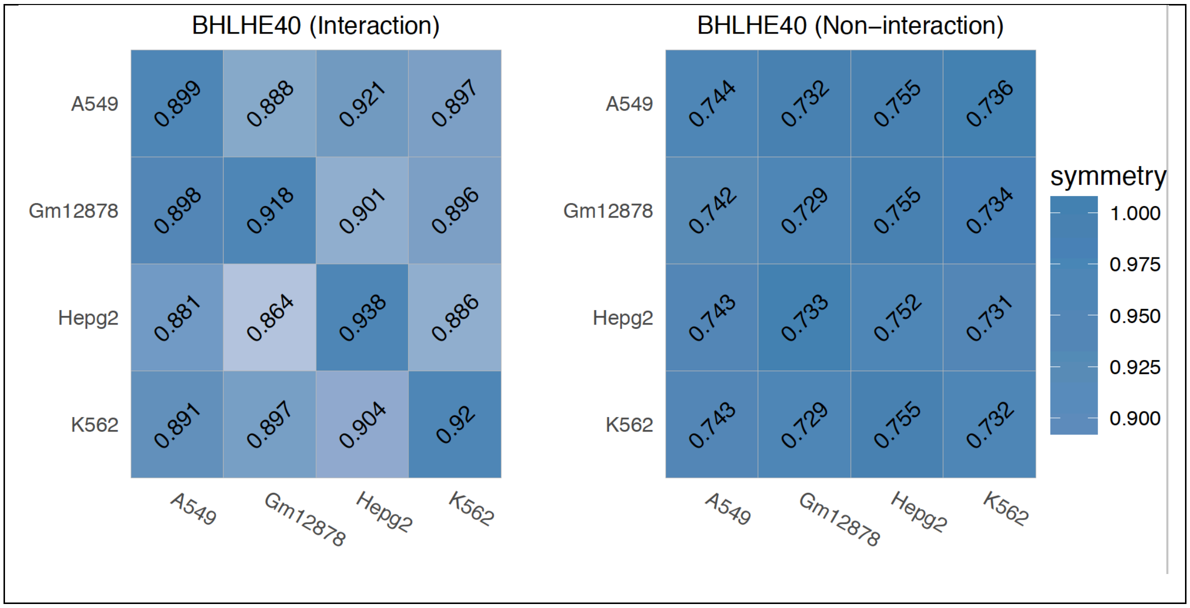


Figure 7.9 Relationship between model accuracy and sequence size. In each plot, color is used to indicate models from different cell lines.

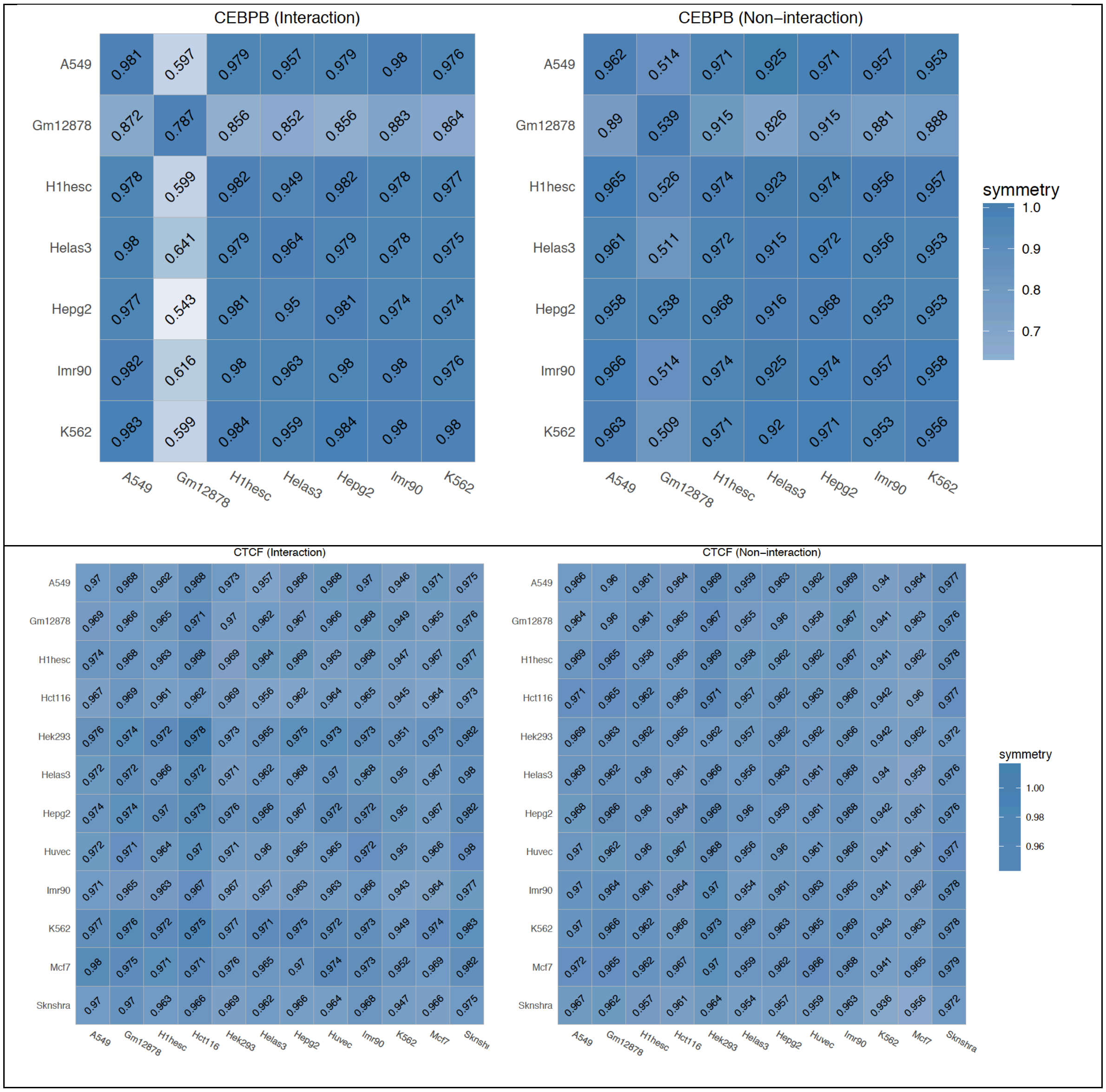


BHLHE40 (Interaction)

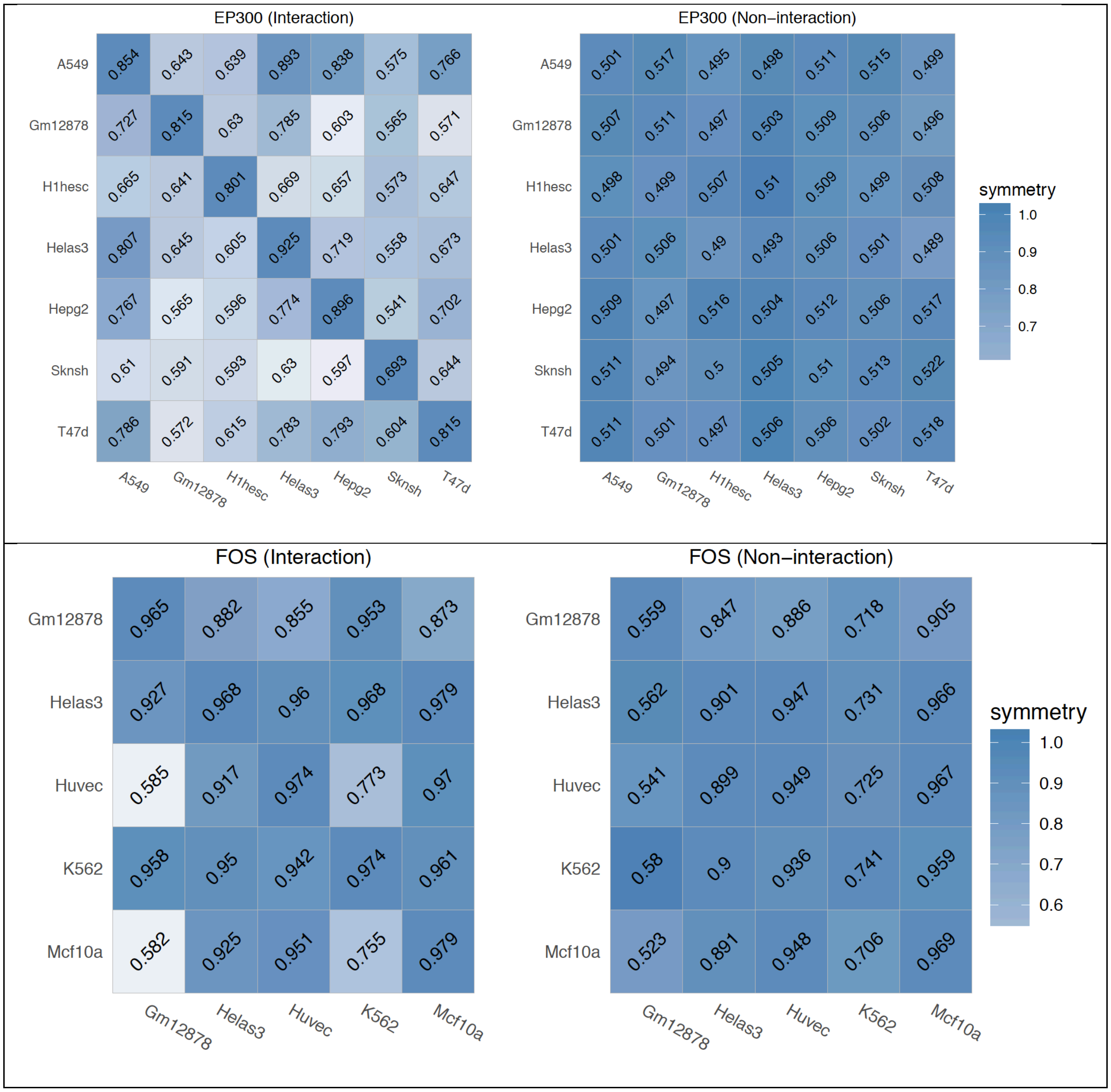
	A549	Gm12878	Hepg2	K562
A549	0.899	0.888	0.921	0.897
Gm12878	0.898	0.918	0.901	0.896
Hepg2	0.881	0.864	0.938	0.886
K562	0.891	0.897	0.904	0.92

BHLHE40 (Non−interaction)

	A549	Gm12878	Hepg2	K562
A549	0.744	0.732	0.755	0.736
Gm12878	0.742	0.729	0.755	0.734
Hepg2	0.743	0.733	0.752	0.731
K562	0.743	0.729	0.755	0.732


CEBPB (Interaction)
A549 0.981 0.597 0.979 0.957 0.979 0.98 0.976
Gm12878 0.872 0.787 0.856 0.852 0.856 0.883 0.864
H1hesc 0.978 0.599 0.982 0.949 0.982 0.978 0.977
Helas3 0.98 0.641 0.979 0.964 0.979 0.978 0.975
Hepg2 0.977 0.543 0.981 0.95 0.981 0.974 0.974
Imr90 0.982 0.616 0.98 0.963 0.98 0.98 0.976
K562 0.983 0.599 0.984 0.959 0.984 0.98 0.98
A549 Gm12878 H1hesc Helas3 Hepg2 Imr90 K562
CEBPB (Non−interaction)
A549 0.962 0.514 0.971 0.925 0.971 0.957 0.953
Gm12878 0.89 0.539 0.915 0.826 0.915 0.881 0.888
H1hesc 0.965 0.526 0.974 0.923 0.974 0.956 0.957
Helas3 0.961 0.511 0.972 0.915 0.972 0.956 0.953
Hepg2 0.958 0.538 0.968 0.916 0.968 0.953 0.953
Imr90 0.966 0.514 0.974 0.925 0.974 0.957 0.958
K562 0.963 0.509 0.971 0.92 0.971 0.953 0.956
A549 Gm12878 H1hesc Helas3 Hepg2 Imr90 K562
symmetry
1.0
0.9
0.8
0.7
CTCF (Interaction)
A549 0.97 0.968 0.962 0.968 0.973 0.957 0.966 0.968 0.97 0.946 0.971 0.975
Gm12878 0.969 0.966 0.965 0.971 0.97 0.962 0.967 0.966 0.968 0.949 0.965 0.976
H1hesc 0.974 0.968 0.963 0.968 0.969 0.964 0.969 0.963 0.968 0.947 0.967 0.977
Hct116 0.967 0.969 0.961 0.962 0.969 0.956 0.962 0.964 0.965 0.945 0.964 0.973
Hek293 0.976 0.974 0.972 0.978 0.973 0.965 0.975 0.973 0.973 0.951 0.973 0.982
Helas3 0.972 0.972 0.966 0.972 0.971 0.962 0.968 0.97 0.968 0.95 0.967 0.98
Hepg2 0.974 0.974 0.97 0.973 0.976 0.966 0.967 0.972 0.972 0.95 0.967 0.982
Huvec 0.972 0.971 0.964 0.97 0.971 0.96 0.965 0.965 0.972 0.95 0.966 0.98
Imr90 0.971 0.965 0.963 0.967 0.967 0.957 0.963 0.963 0.966 0.943 0.964 0.977
K562 0.977 0.976 0.972 0.975 0.977 0.971 0.975 0.972 0.973 0.949 0.974 0.983
Mcf7 0.98 0.975 0.971 0.971 0.976 0.965 0.97 0.974 0.973 0.952 0.969 0.982
Sknshra 0.97 0.97 0.963 0.966 0.969 0.962 0.966 0.964 0.968 0.947 0.966 0.975
A549 Gm12878 H1hesc Hct116 Hek293 Helas3 Hepg2 Huvec Imr90 K562 Mcf7 Sknshr
CTCF (Non−interaction)
A549 0.966 0.96 0.961 0.964 0.969 0.959 0.963 0.962 0.969 0.94 0.964 0.977
Gm12878 0.964 0.96 0.961 0.965 0.967 0.955 0.96 0.958 0.967 0.941 0.963 0.976
H1hesc 0.969 0.965 0.958 0.965 0.969 0.958 0.962 0.962 0.967 0.941 0.962 0.978
Hct116 0.971 0.965 0.962 0.965 0.971 0.957 0.962 0.963 0.966 0.942 0.96 0.977
Hek293 0.969 0.963 0.962 0.965 0.962 0.957 0.962 0.962 0.966 0.942 0.962 0.972
Helas3 0.969 0.962 0.96 0.961 0.966 0.956 0.963 0.961 0.968 0.94 0.958 0.976
Hepg2 0.968 0.966 0.96 0.964 0.969 0.96 0.959 0.961 0.968 0.942 0.961 0.976
Huvec 0.97 0.962 0.96 0.967 0.968 0.956 0.96 0.961 0.966 0.941 0.961 0.977
Imr90 0.97 0.964 0.961 0.964 0.97 0.954 0.961 0.963 0.965 0.941 0.962 0.978
K562 0.97 0.966 0.962 0.966 0.973 0.959 0.963 0.965 0.969 0.943 0.963 0.978
Mcf7 0.972 0.965 0.962 0.967 0.97 0.959 0.962 0.966 0.968 0.941 0.965 0.979
Sknshra 0.967 0.962 0.957 0.961 0.964 0.954 0.957 0.959 0.963 0.936 0.956 0.972
A549 Gm12878 H1hesc Hct116 Hek293 Helas3 Hepg2 Huvec Imr90 K562 Mcf7 Sknshra
symmetry
1.00
0.98
0.96

]



EP300 (Interaction)

	A549	Gm12878	H1hesc	Helas3	Hepg2	Sknsh	T47d
A549	0.854	0.643	0.639	0.893	0.838	0.575	0.766
Gm12878	0.727	0.815	0.63	0.785	0.603	0.565	0.571
H1hesc	0.665	0.641	0.801	0.669	0.657	0.573	0.647
Helas3	0.807	0.645	0.605	0.925	0.719	0.558	0.673
Hepg2	0.767	0.565	0.596	0.774	0.896	0.541	0.702
Sknsh	0.61	0.591	0.593	0.63	0.597	0.693	0.644
T47d	0.786	0.572	0.615	0.783	0.793	0.604	0.815

EP300 (Non–interaction)

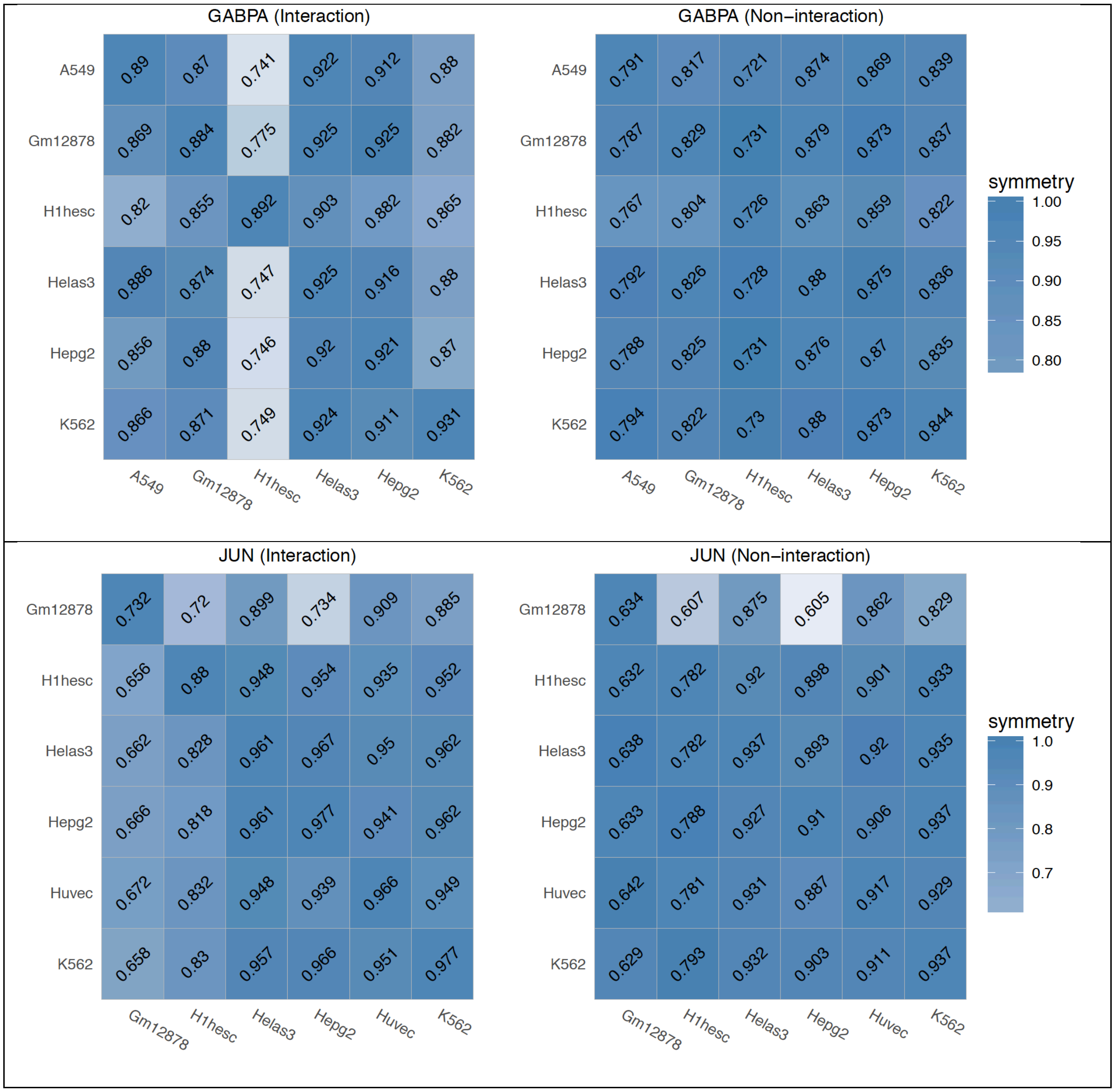
	A549	Gm12878	H1hesc	Helas3	Hepg2	Sknsh	T47d
A549	0.501	0.517	0.495	0.498	0.511	0.515	0.499
Gm12878	0.507	0.511	0.497	0.503	0.509	0.506	0.496
H1hesc	0.498	0.499	0.507	0.51	0.509	0.499	0.508
Helas3	0.501	0.506	0.49	0.493	0.506	0.501	0.489
Hepg2	0.509	0.497	0.516	0.504	0.512	0.506	0.517
Sknsh	0.511	0.494	0.5	0.505	0.51	0.513	0.522
T47d	0.511	0.501	0.497	0.506	0.506	0.502	0.518

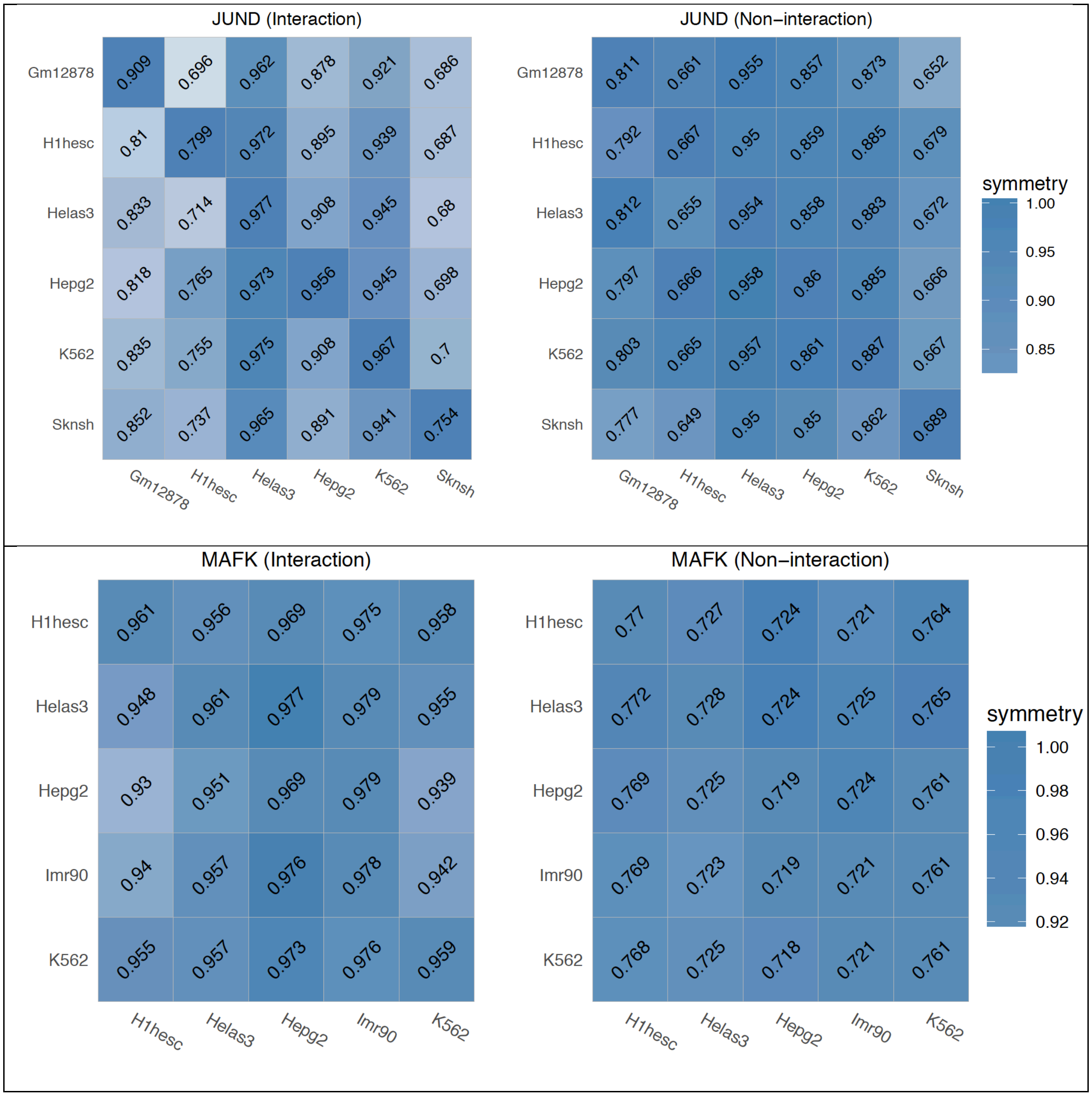
FOS (Interaction)

	Gm12878	Helas3	Huvec	K562	Mcf10a
Gm12878	0.965	0.882	0.855	0.953	0.873
Helas3	0.927	0.968	0.96	0.968	0.979
Huvec	0.585	0.917	0.974	0.773	0.97
K562	0.958	0.95	0.942	0.974	0.961
Mcf10a	0.582	0.925	0.951	0.755	0.979

FOS (Non–interaction)

	Gm12878	Helas3	Huvec	K562	Mcf10a
Gm12878	0.559	0.847	0.886	0.718	0.905
Helas3	0.562	0.901	0.947	0.731	0.966
Huvec	0.541	0.899	0.949	0.725	0.967
K562	0.58	0.9	0.936	0.741	0.959
Mcf10a	0.523	0.891	0.948	0.706	0.969


GABPA (Interaction)
A549 0.89 0.87 0.741 0.922 0.912 0.88
Gm12878 0.869 0.884 0.775 0.925 0.925 0.882
H1hesc 0.82 0.855 0.892 0.903 0.882 0.865
Helas3 0.886 0.874 0.747 0.925 0.916 0.88
Hepg2 0.856 0.88 0.746 0.92 0.921 0.87
K562 0.866 0.871 0.749 0.924 0.911 0.931
A549 Gm12878 H1hesc Helas3 Hepg2 K562
GABPA (Non−interaction)
A549 0.791 0.817 0.721 0.874 0.869 0.839
Gm12878 0.787 0.829 0.731 0.879 0.873 0.837
H1hesc 0.767 0.804 0.726 0.863 0.859 0.822
Helas3 0.792 0.826 0.728 0.88 0.875 0.836
Hepg2 0.788 0.825 0.731 0.876 0.87 0.835
K562 0.794 0.822 0.73 0.88 0.873 0.844
A549 Gm12878 H1hesc Helas3 Hepg2 K562
symmetry
1.00
0.95
0.90
0.85
0.80
JUN (Interaction)
Gm12878 0.732 0.72 0.899 0.734 0.909 0.885
H1hesc 0.656 0.88 0.948 0.954 0.935 0.952
Helas3 0.662 0.828 0.961 0.967 0.95 0.962
Hepg2 0.666 0.818 0.961 0.977 0.941 0.962
Huvec 0.672 0.832 0.948 0.939 0.966 0.949
K562 0.658 0.83 0.957 0.966 0.951 0.977
Gm12878 H1hesc Helas3 Hepg2 Huvec K562
JUN (Non−interaction)
Gm12878 0.634 0.607 0.875 0.605 0.862 0.829
H1hesc 0.632 0.782 0.92 0.898 0.901 0.933
Helas3 0.638 0.782 0.937 0.893 0.92 0.935
Hepg2 0.633 0.788 0.927 0.91 0.906 0.937
Huvec 0.642 0.781 0.931 0.887 0.917 0.929
K562 0.629 0.793 0.932 0.903 0.911 0.937
Gm12878 H1hesc Helas3 Hepg2 Huvec K562
symmetry
1.0
0.9
0.8
0.7



JUND (Interaction)

	Gm12878	H1hesc	Helas3	Hepg2	K562	Sknsh
Gm12878	0.909	0.696	0.962	0.878	0.921	0.686
H1hesc	0.81	0.799	0.972	0.895	0.939	0.687
Helas3	0.833	0.714	0.977	0.908	0.945	0.68
Hepg2	0.818	0.765	0.973	0.956	0.945	0.698
K562	0.835	0.755	0.975	0.908	0.967	0.7
Sknsh	0.852	0.737	0.965	0.891	0.941	0.754

JUND (Non−interaction)

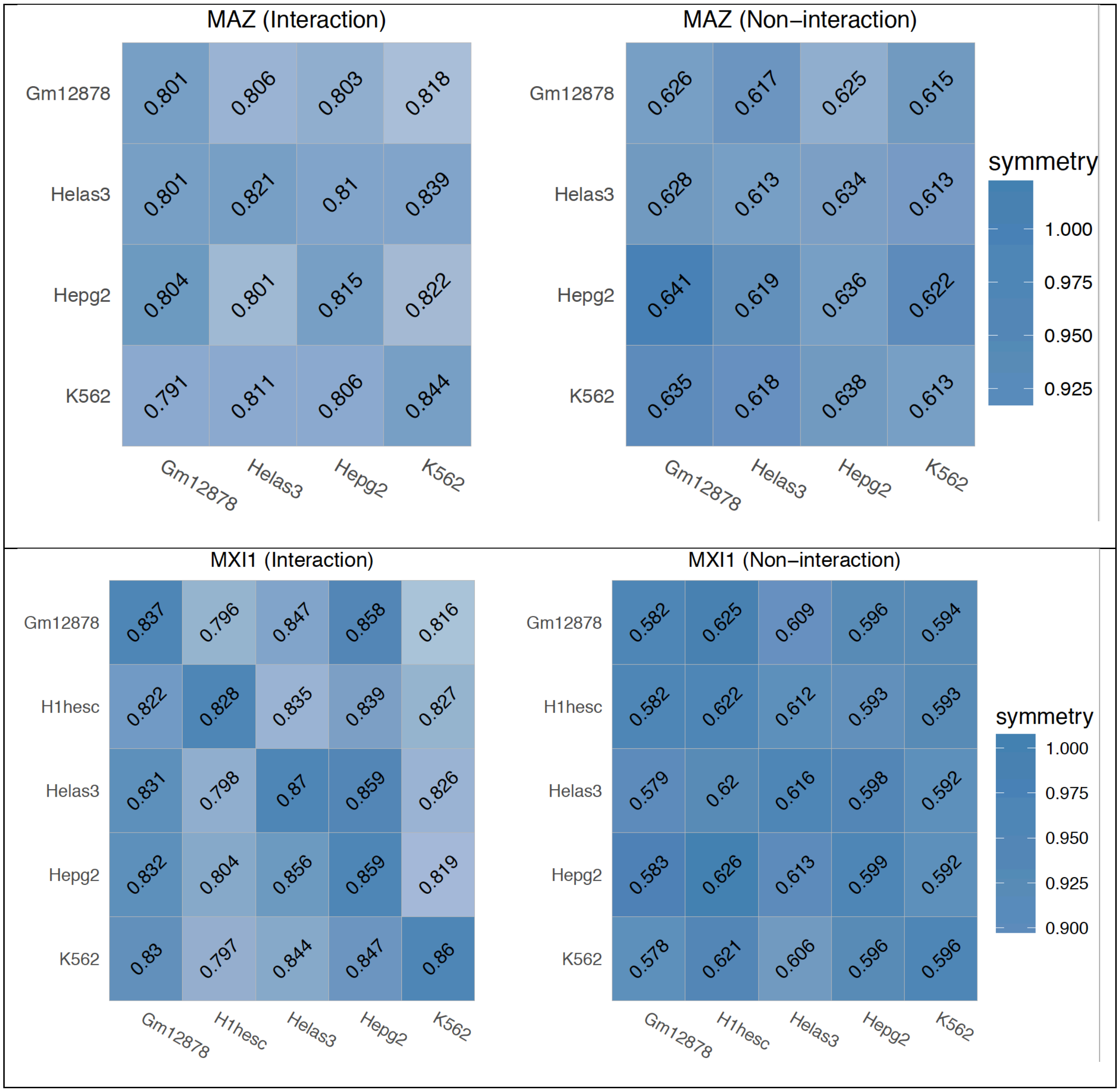
	Gm12878	H1hesc	Helas3	Hepg2	K562	Sknsh
Gm12878	0.811	0.661	0.955	0.857	0.873	0.652
H1hesc	0.792	0.667	0.95	0.859	0.885	0.679
Helas3	0.812	0.655	0.954	0.858	0.883	0.672
Hepg2	0.797	0.666	0.958	0.86	0.885	0.666
K562	0.803	0.665	0.957	0.861	0.887	0.667
Sknsh	0.777	0.649	0.95	0.85	0.862	0.689

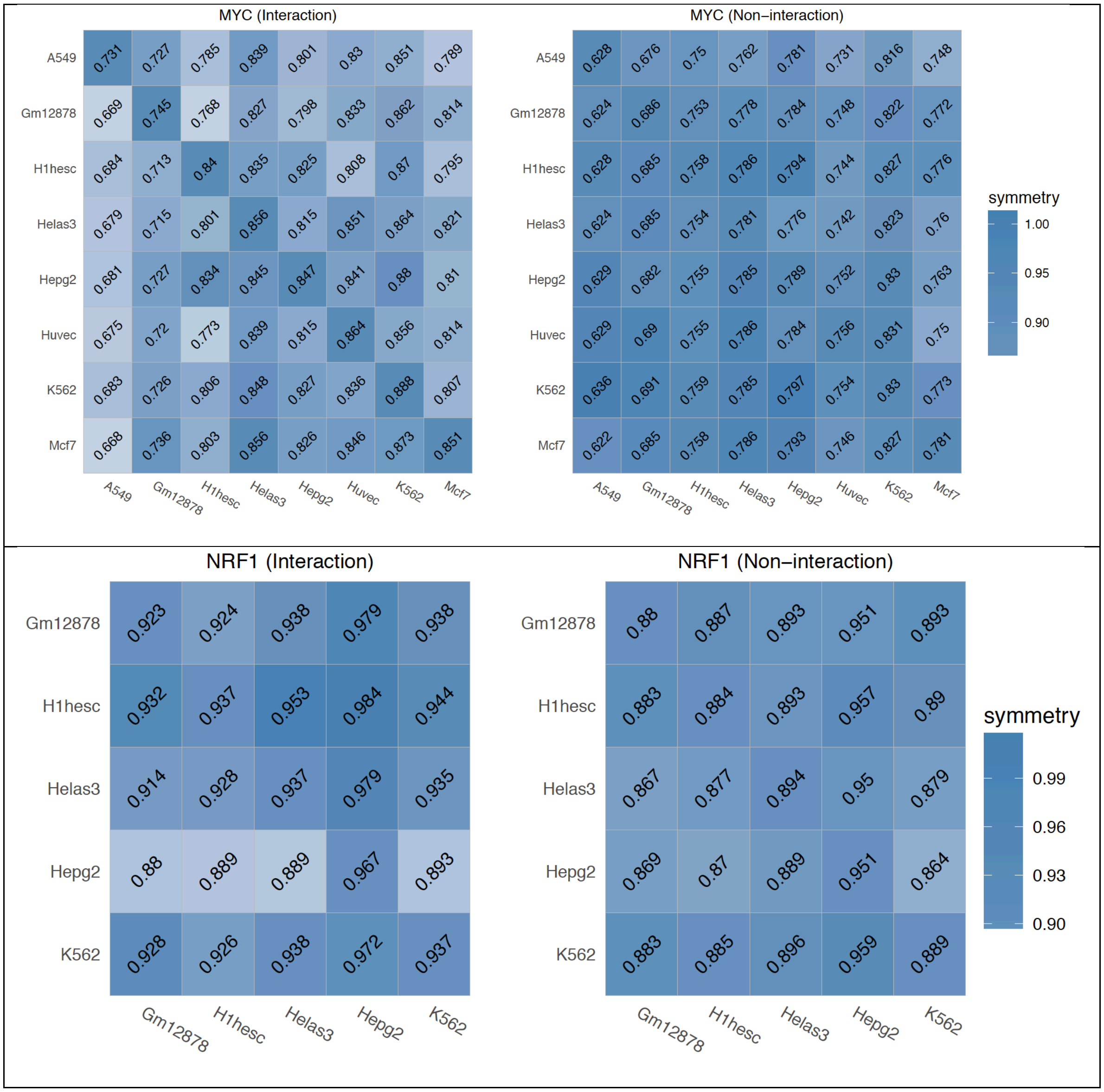
MAFK (Interaction)

	H1hesc	Helas3	Hepg2	Imr90	K562
H1hesc	0.961	0.956	0.969	0.975	0.958
Helas3	0.948	0.961	0.977	0.979	0.955
Hepg2	0.93	0.951	0.969	0.979	0.939
Imr90	0.94	0.957	0.976	0.978	0.942
K562	0.955	0.957	0.973	0.976	0.959

MAFK (Non−interaction)

	H1hesc	Helas3	Hepg2	Imr90	K562
H1hesc	0.77	0.727	0.724	0.721	0.764
Helas3	0.772	0.728	0.724	0.725	0.765
Hepg2	0.769	0.725	0.719	0.724	0.761
Imr90	0.769	0.723	0.719	0.721	0.761
K562	0.768	0.725	0.718	0.721	0.761


MAZ (Interaction)
Gm12878 0.801 0.806 0.803 0.818
Helas3 0.801 0.821 0.81 0.839
Hepg2 0.804 0.801 0.815 0.822
K562 0.791 0.811 0.806 0.844
Gm12878 Helas3 Hepg2 K562
MAZ (Non−interaction)
Gm12878 0.626 0.617 0.625 0.615
Helas3 0.628 0.613 0.634 0.613
Hepg2 0.641 0.619 0.636 0.622
K562 0.635 0.618 0.638 0.613
Gm12878 Helas3 Hepg2 K562
symmetry
1.000
0.975
0.950
0.925
MXI1 (Interaction)
Gm12878 0.837 0.796 0.847 0.858 0.816
H1hesc 0.822 0.828 0.835 0.839 0.827
Helas3 0.831 0.798 0.87 0.859 0.826
Hepg2 0.832 0.804 0.856 0.859 0.819
K562 0.83 0.797 0.844 0.847 0.86
Gm12878 H1hesc Helas3 Hepg2 K562
MXI1 (Non−interaction)
Gm12878 0.582 0.625 0.609 0.596 0.594
H1hesc 0.582 0.622 0.612 0.593 0.593
Helas3 0.579 0.62 0.616 0.598 0.592
Hepg2 0.583 0.626 0.613 0.599 0.592
K562 0.578 0.621 0.606 0.596 0.596
Gm12878 H1hesc Helas3 Hepg2 K562
symmetry
1.000
0.975
0.950
0.925
0.900



MYC (Interaction)

	A549	Gm12878	H1hesc	Helas3	Hepg2	Huvec	K562	Mcf7
A549	0.731	0.727	0.785	0.839	0.801	0.83	0.851	0.789
Gm12878	0.669	0.745	0.768	0.827	0.798	0.833	0.862	0.814
H1hesc	0.684	0.713	0.84	0.835	0.825	0.808	0.87	0.795
Helas3	0.679	0.715	0.801	0.856	0.815	0.851	0.864	0.821
Hepg2	0.681	0.727	0.834	0.845	0.847	0.841	0.88	0.81
Huvec	0.675	0.72	0.773	0.839	0.815	0.864	0.856	0.814
K562	0.683	0.726	0.806	0.848	0.827	0.836	0.888	0.807
Mcf7	0.668	0.736	0.803	0.856	0.826	0.846	0.873	0.851

MYC (Non–interaction)

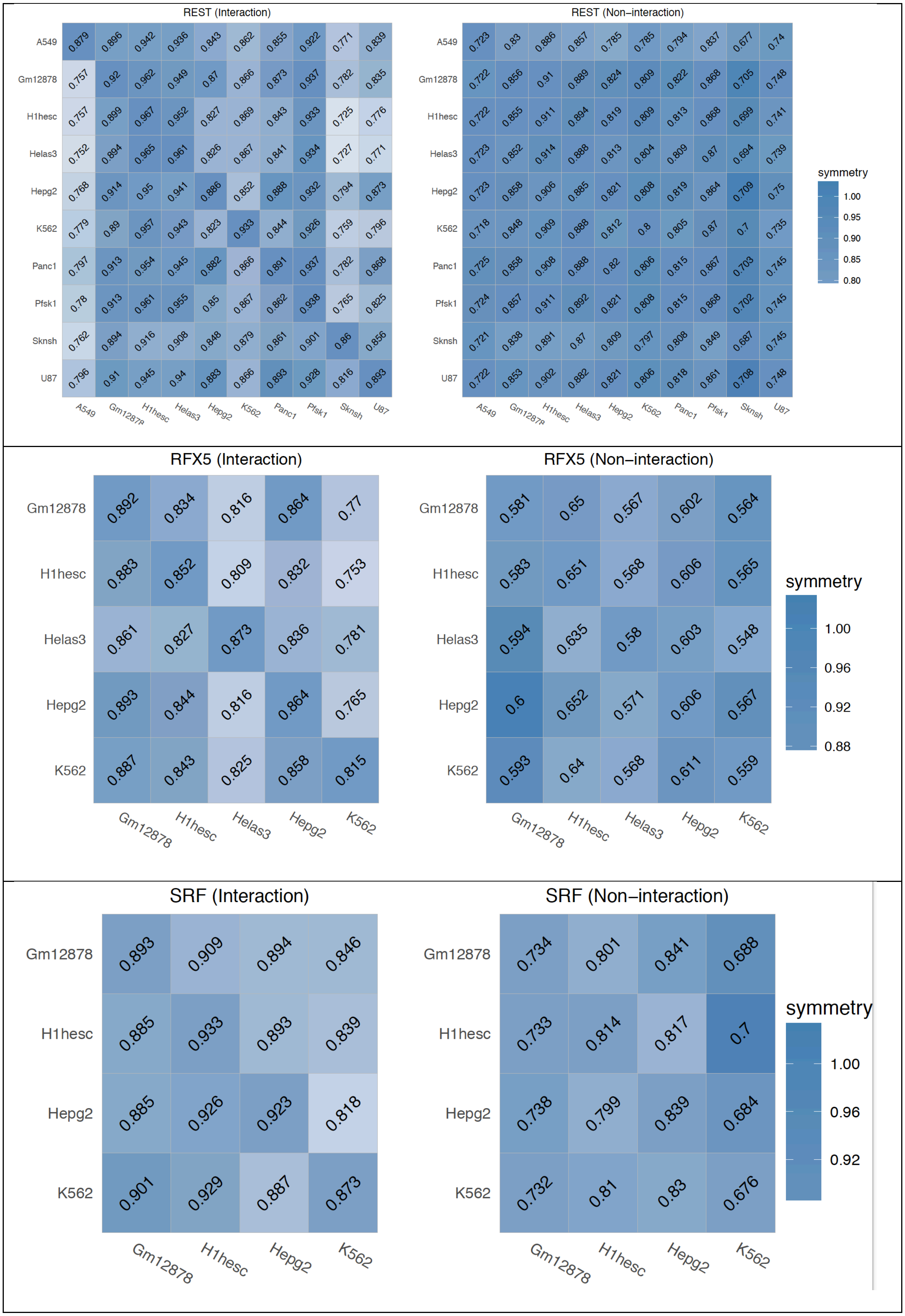
	A549	Gm12878	H1hesc	Helas3	Hepg2	Huvec	K562	Mcf7
A549	0.628	0.676	0.75	0.762	0.781	0.731	0.816	0.748
Gm12878	0.624	0.686	0.753	0.778	0.784	0.748	0.822	0.772
H1hesc	0.628	0.685	0.758	0.786	0.794	0.744	0.827	0.776
Helas3	0.624	0.685	0.754	0.781	0.776	0.742	0.823	0.76
Hepg2	0.629	0.682	0.755	0.785	0.789	0.752	0.83	0.763
Huvec	0.629	0.69	0.755	0.786	0.784	0.756	0.831	0.75
K562	0.636	0.691	0.759	0.785	0.797	0.754	0.83	0.773
Mcf7	0.622	0.685	0.758	0.786	0.793	0.746	0.827	0.781

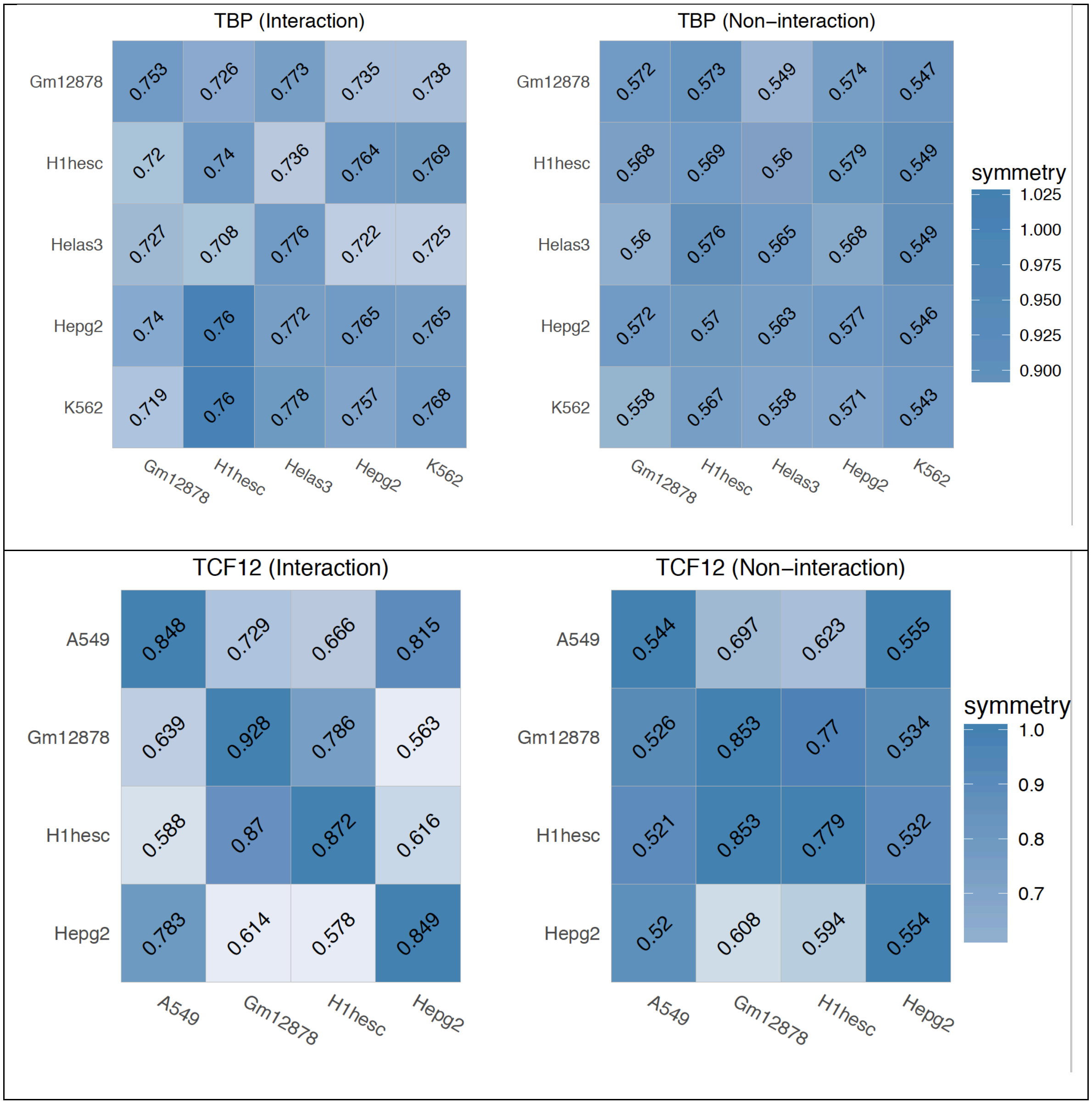
NRF1 (Interaction)

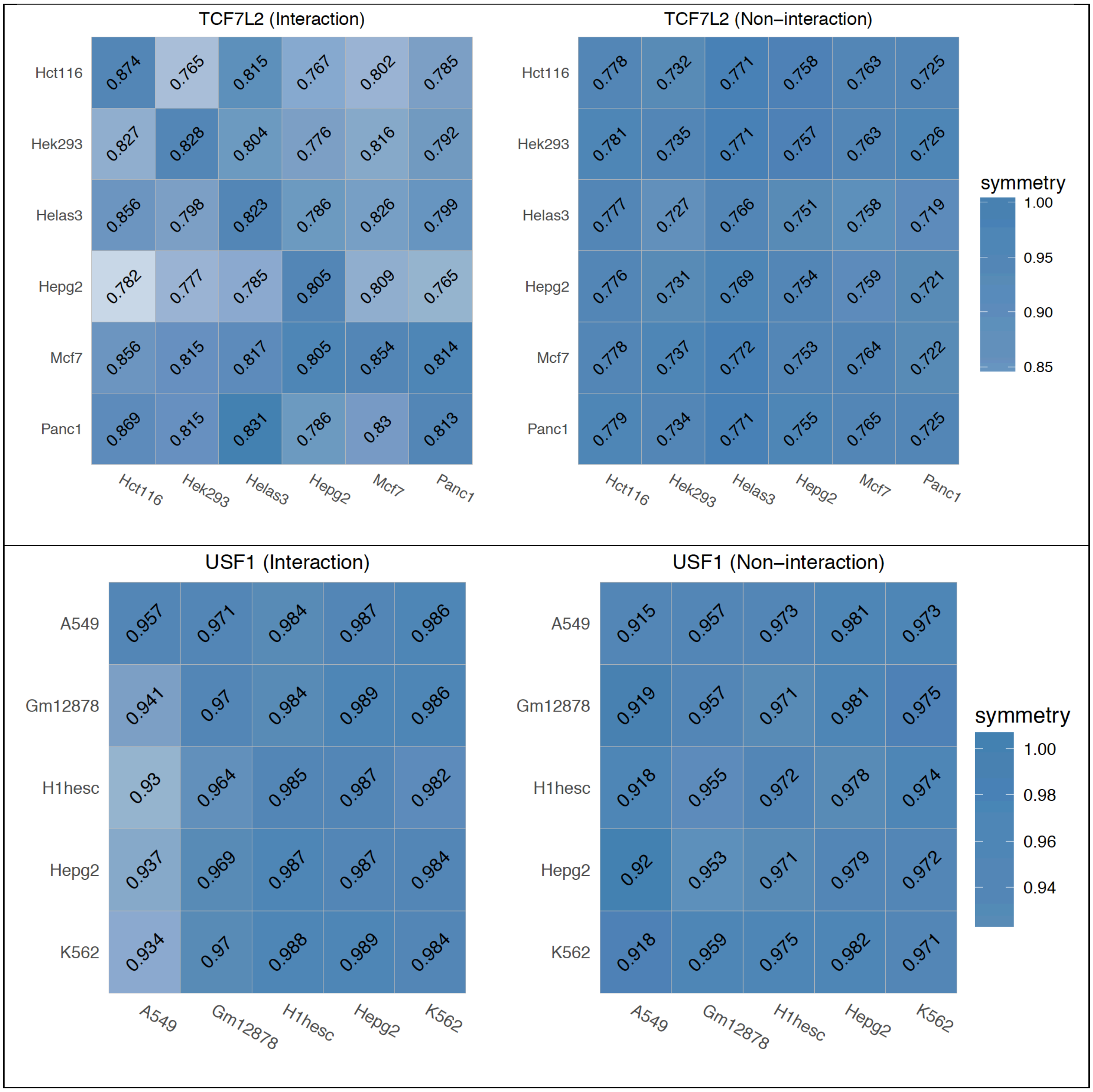
	Gm12878	H1hesc	Helas3	Hepg2	K562
Gm12878	0.923	0.924	0.938	0.979	0.938
H1hesc	0.932	0.937	0.953	0.984	0.944
Helas3	0.914	0.928	0.937	0.979	0.935
Hepg2	0.88	0.889	0.889	0.967	0.893
K562	0.928	0.926	0.938	0.972	0.937

NRF1 (Non–interaction)

	Gm12878	H1hesc	Helas3	Hepg2	K562
Gm12878	0.88	0.887	0.893	0.951	0.893
H1hesc	0.883	0.884	0.893	0.957	0.89
Helas3	0.867	0.877	0.894	0.95	0.879
Hepg2	0.869	0.87	0.889	0.951	0.864
K562	0.883	0.885	0.896	0.959	0.889


REST (Interaction)
A549 0.879 0.896 0.942 0.936 0.843 0.862 0.855 0.922 0.771 0.839
Gm12878 0.757 0.92 0.962 0.949 0.87 0.866 0.873 0.937 0.782 0.835
H1hesc 0.757 0.899 0.967 0.952 0.827 0.869 0.843 0.933 0.725 0.776
Helas3 0.752 0.894 0.965 0.961 0.826 0.867 0.841 0.934 0.727 0.771
Hepg2 0.768 0.914 0.95 0.941 0.886 0.852 0.888 0.932 0.794 0.873
K562 0.779 0.89 0.957 0.943 0.823 0.933 0.844 0.926 0.759 0.796
Panc1 0.797 0.913 0.954 0.945 0.882 0.866 0.891 0.937 0.782 0.868
Pfsk1 0.78 0.913 0.961 0.955 0.85 0.867 0.862 0.938 0.765 0.825
Sknsh 0.762 0.894 0.916 0.908 0.848 0.879 0.861 0.901 0.86 0.856
U87 0.796 0.91 0.945 0.94 0.883 0.866 0.893 0.928 0.816 0.893
A549 Gm12878 H1hesc Helas3 Hepg2 K562 Panc1 Pfsk1 Sknsh U87
REST (Non–interaction)
A549 0.723 0.83 0.886 0.857 0.785 0.785 0.794 0.837 0.677 0.74
Gm12878 0.722 0.856 0.91 0.889 0.824 0.809 0.822 0.868 0.705 0.748
H1hesc 0.722 0.855 0.911 0.894 0.819 0.809 0.813 0.868 0.699 0.741
Helas3 0.723 0.852 0.914 0.888 0.813 0.804 0.809 0.87 0.694 0.739
Hepg2 0.723 0.858 0.906 0.885 0.821 0.808 0.819 0.864 0.709 0.75
K562 0.718 0.848 0.909 0.888 0.812 0.8 0.805 0.87 0.7 0.735
Panc1 0.725 0.858 0.908 0.888 0.82 0.806 0.815 0.867 0.703 0.745
Pfsk1 0.724 0.857 0.911 0.892 0.821 0.808 0.815 0.868 0.702 0.745
Sknsh 0.721 0.838 0.891 0.87 0.809 0.797 0.808 0.849 0.687 0.745
U87 0.722 0.853 0.902 0.882 0.821 0.806 0.818 0.861 0.708 0.748
A549 Gm12878 H1hesc Helas3 Hepg2 K562 Panc1 Pfsk1 Sknsh U87
symmetry
1.00
0.95
0.90
0.85
0.80
RFX5 (Interaction)
Gm12878 0.892 0.834 0.816 0.864 0.77
H1hesc 0.883 0.852 0.809 0.832 0.753
Helas3 0.861 0.827 0.873 0.836 0.781
Hepg2 0.893 0.844 0.816 0.864 0.765
K562 0.887 0.843 0.825 0.858 0.815
Gm12878 H1hesc Helas3 Hepg2 K562
RFX5 (Non–interaction)
Gm12878 0.581 0.65 0.567 0.602 0.564
H1hesc 0.583 0.651 0.568 0.606 0.565
Helas3 0.594 0.635 0.58 0.603 0.548
Hepg2 0.6 0.652 0.571 0.606 0.567
K562 0.593 0.64 0.568 0.611 0.559
Gm12878 H1hesc Helas3 Hepg2 K562
symmetry
1.00
0.96
0.92
0.88
SRF (Interaction)
Gm12878 0.893 0.909 0.894 0.846
H1hesc 0.885 0.933 0.893 0.839
Hepg2 0.885 0.926 0.923 0.818
K562 0.901 0.929 0.887 0.873
Gm12878 H1hesc Hepg2 K562
SRF (Non–interaction)
Gm12878 0.734 0.801 0.841 0.688
H1hesc 0.733 0.814 0.817 0.7
Hepg2 0.738 0.799 0.839 0.684
K562 0.732 0.81 0.83 0.676
Gm12878 H1hesc Hepg2 K562
symmetry
1.00
0.96
0.92


TBP (Interaction)
Gm12878
H1hesc
Helas3
Hepg2
K562
0.753
0.726
0.773
0.735
0.738
0.72
0.74
0.736
0.764
0.769
0.727
0.708
0.776
0.722
0.725
0.74
0.76
0.772
0.765
0.765
0.719
0.76
0.778
0.757
0.768
Gm12878
H1hesc
Helas3
Hepg2
K562
TBP (Non–interaction)
Gm12878
H1hesc
Helas3
Hepg2
K562
0.572
0.573
0.549
0.574
0.547
0.568
0.569
0.56
0.579
0.549
0.56
0.576
0.565
0.568
0.549
0.572
0.57
0.563
0.577
0.546
0.558
0.567
0.558
0.571
0.543
Gm12878
H1hesc
Helas3
Hepg2
K562
symmetry
1.025
1.000
0.975
0.950
0.925
0.900
TCF12 (Interaction)
A549
Gm12878
H1hesc
Hepg2
0.848
0.729
0.666
0.815
0.639
0.928
0.786
0.563
0.588
0.87
0.872
0.616
0.783
0.614
0.578
0.849
A549
Gm12878
H1hesc
Hepg2
TCF12 (Non–interaction)
A549
Gm12878
H1hesc
Hepg2
0.544
0.697
0.623
0.555
0.526
0.853
0.77
0.534
0.521
0.853
0.779
0.532
0.52
0.608
0.594
0.554
A549
Gm12878
H1hesc
Hepg2
symmetry
1.0
0.9
0.8
0.7


TCF7L2 (Interaction)
Hct116 0.874 0.765 0.815 0.767 0.802 0.785
Hek293 0.827 0.828 0.804 0.776 0.816 0.792
Helas3 0.856 0.798 0.823 0.786 0.826 0.799
Hepg2 0.782 0.777 0.785 0.805 0.809 0.765
Mcf7 0.856 0.815 0.817 0.805 0.854 0.814
Panc1 0.869 0.815 0.831 0.786 0.83 0.813
Hct116 Hek293 Helas3 Hepg2 Mcf7 Panc1
TCF7L2 (Non–interaction)
Hct116 0.778 0.732 0.771 0.758 0.763 0.725
Hek293 0.781 0.735 0.771 0.757 0.763 0.726
Helas3 0.777 0.727 0.766 0.751 0.758 0.719
Hepg2 0.776 0.731 0.769 0.754 0.759 0.721
Mcf7 0.778 0.737 0.772 0.753 0.764 0.722
Panc1 0.779 0.734 0.771 0.755 0.765 0.725
Hct116 Hek293 Helas3 Hepg2 Mcf7 Panc1
symmetry
1.00
0.95
0.90
0.85
USF1 (Interaction)
A549 0.957 0.971 0.984 0.987 0.986
Gm12878 0.941 0.97 0.984 0.989 0.986
H1hesc 0.93 0.964 0.985 0.987 0.982
Hepg2 0.937 0.969 0.987 0.987 0.984
K562 0.934 0.97 0.988 0.989 0.984
A549 Gm12878 H1hesc Hepg2 K562
USF1 (Non–interaction)
A549 0.915 0.957 0.973 0.981 0.973
Gm12878 0.919 0.957 0.971 0.981 0.975
H1hesc 0.918 0.955 0.972 0.978 0.974
Hepg2 0.92 0.953 0.971 0.979 0.972
K562 0.918 0.959 0.975 0.982 0.971
A549 Gm12878 H1hesc Hepg2 K562
symmetry
1.00
0.98
0.96
0.94

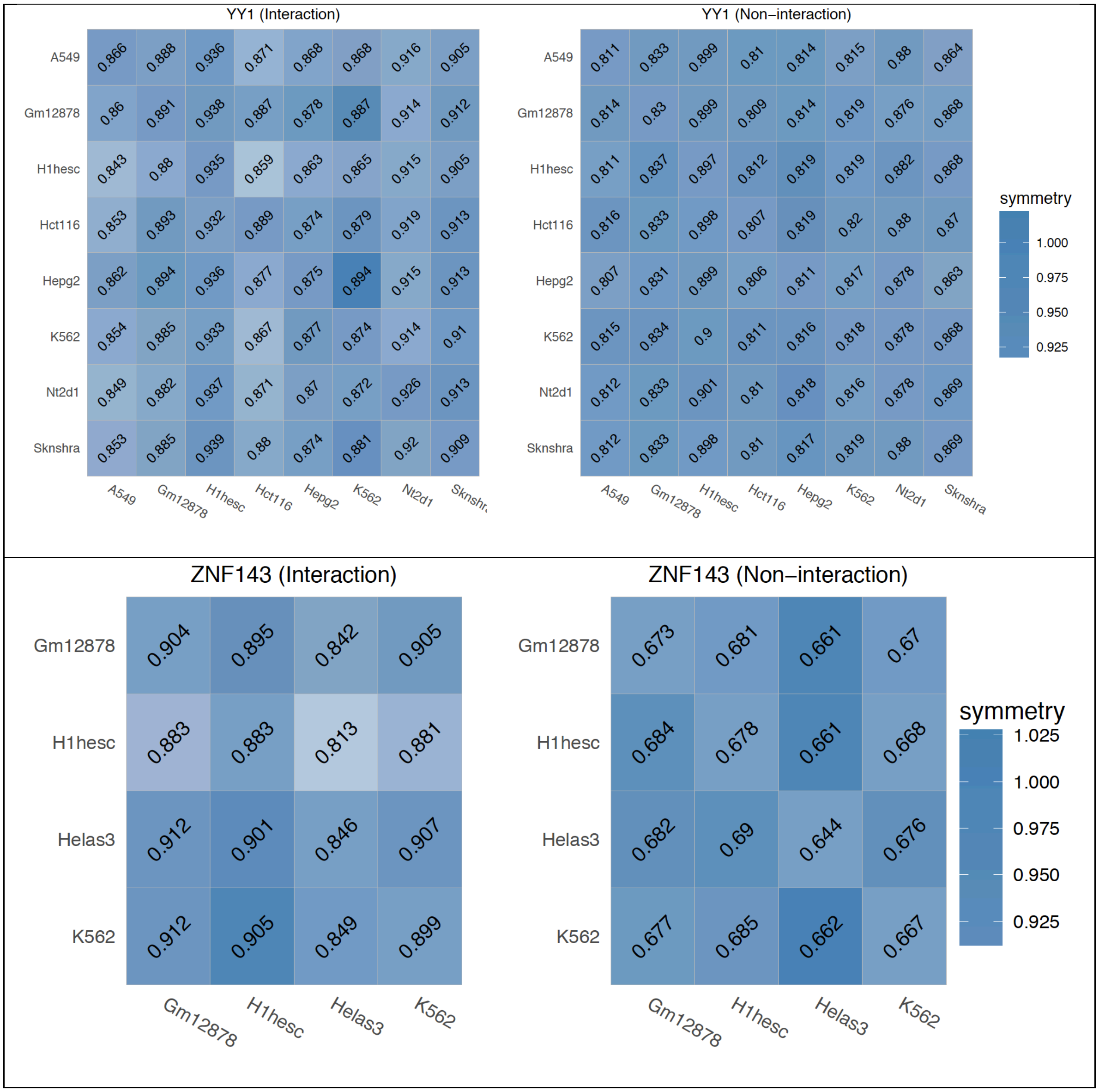
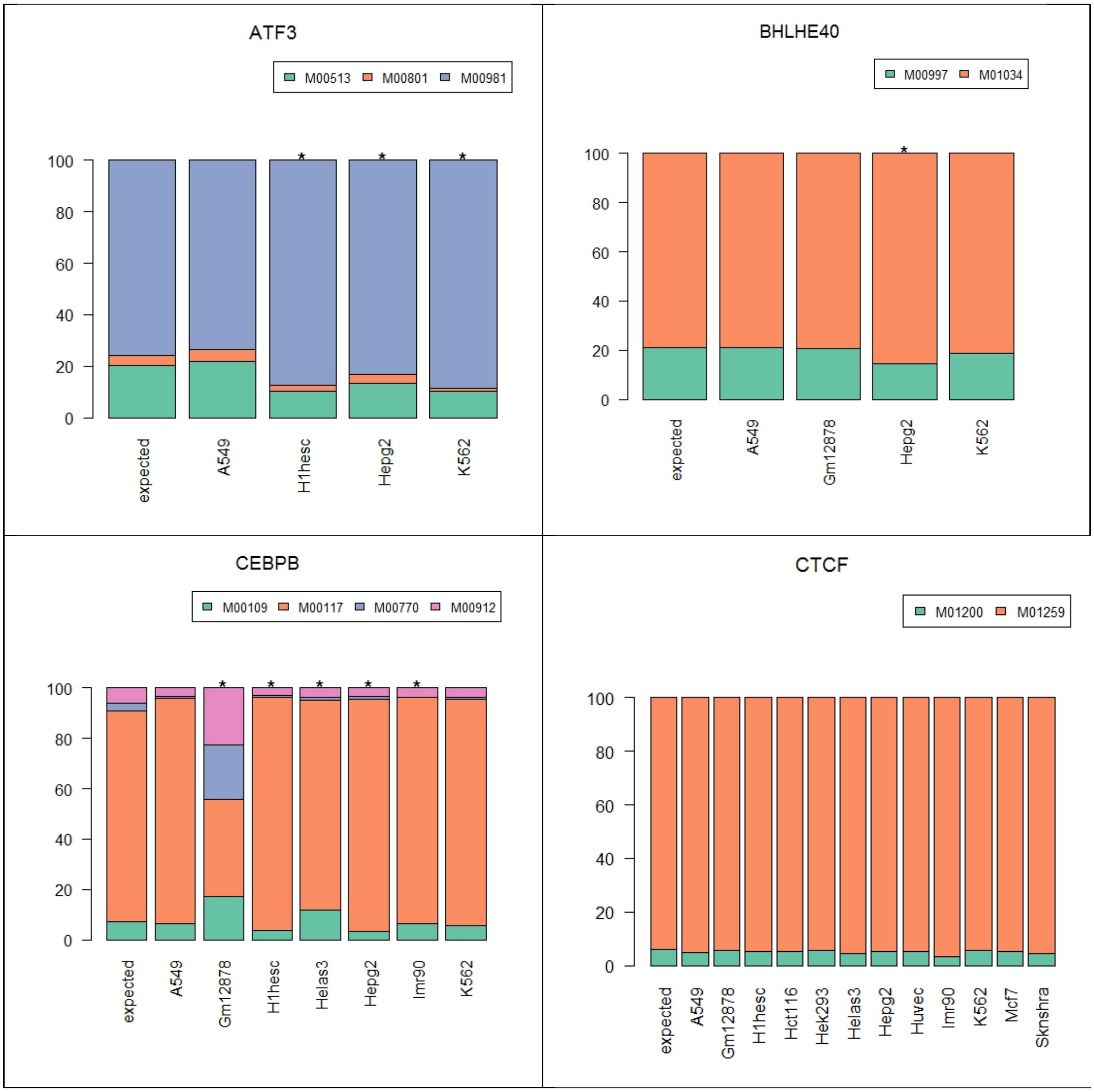
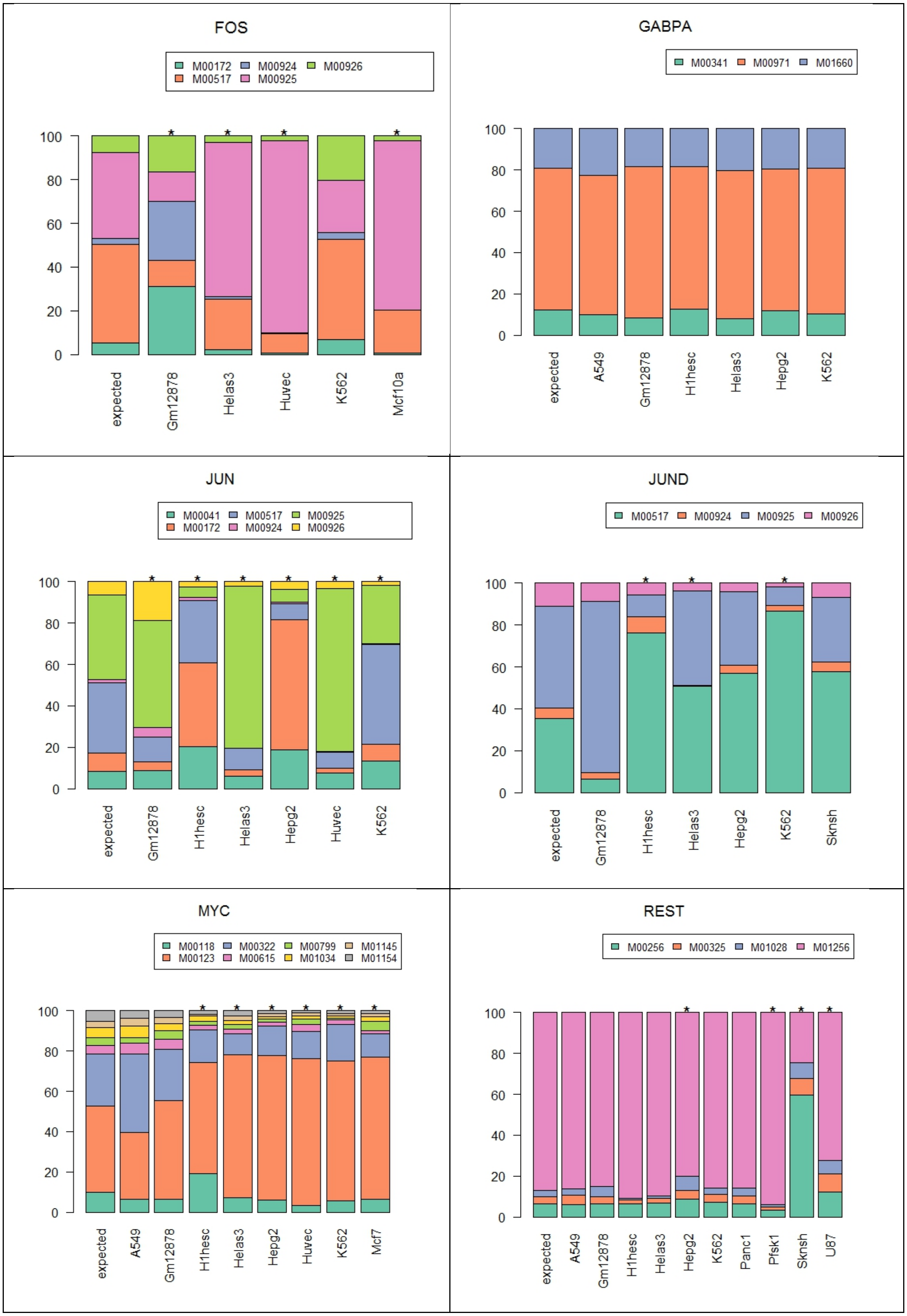
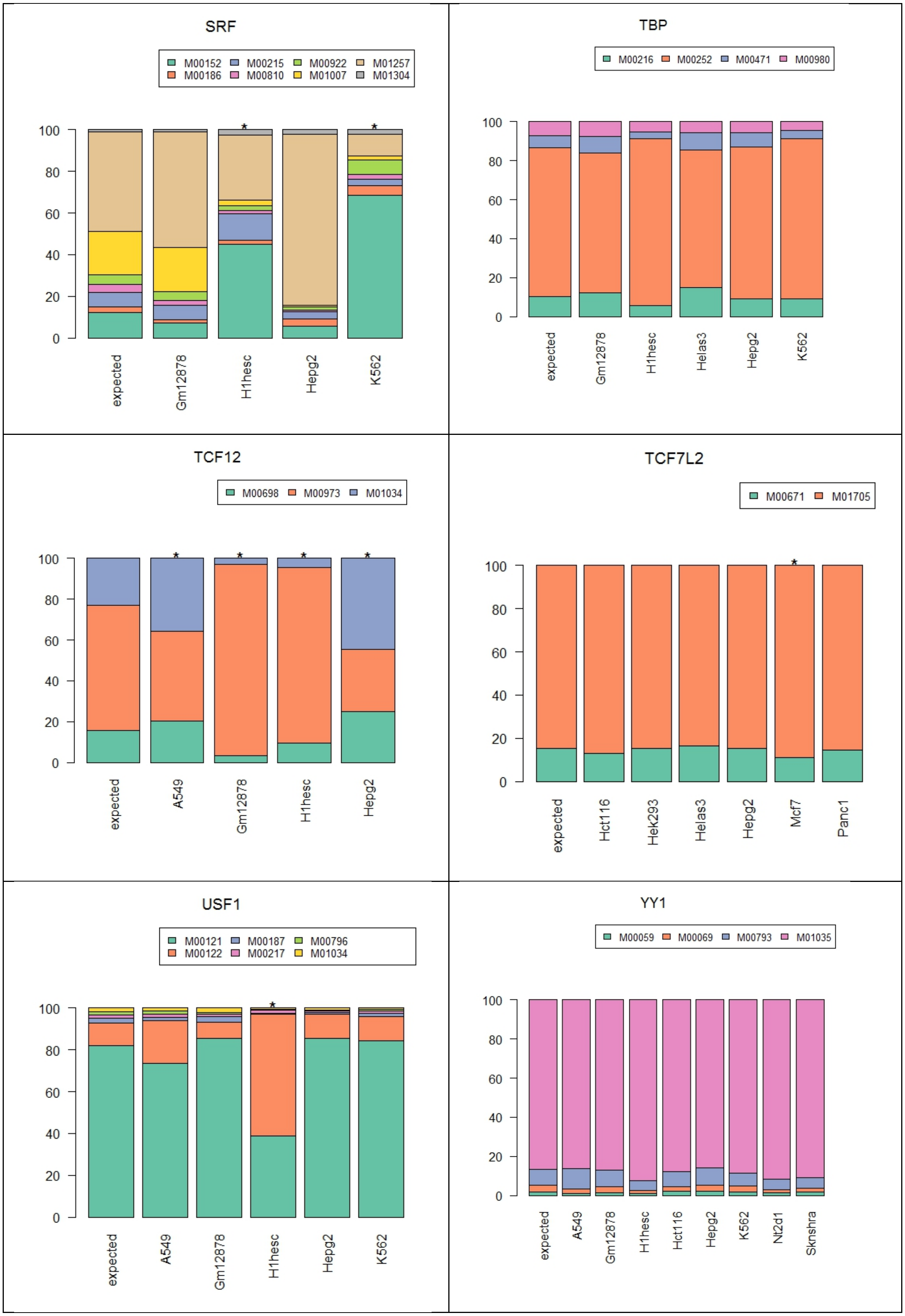


Figure 7.10 Same as Figure 7.8 of Supplemental Data, except the matrix is color coded according to the extent of symmetry of the non-diagonal elements. The symmetry is calculated by normalizing each row by the reference model (diagonal element).


ATF3
M00513
M00801
M00981
0
20
40
60
80
100
expected
A549
H1hesc
Hepg2
K562
*
*
*
BHLHE40
M00997
M01034
0
20
40
60
80
100
expected
A549
Gm12878
Hepg2
K562
*
CEBPB
M00109
M00117
M00770
M00912
0
20
40
60
80
100
expected
A549
Gm12878
H1hesc
Helas3
Hepg2
Imr90
K562
*
*
*
*
*
CTCF
M01200
M01259
0
20
40
60
80
100
expected
A549
Gm12878
H1hesc
Hct116
Hek293
Helas3
Hepg2
Huvec
Imr90
K562
Mcf7
Sknshra


FOS
M00172
M00924
M00926
M00517
M00925
100
80
60
40
20
0
*
expected
Gm12878
Helas3
Huvec
K562
Mcf10a
GABPA
M00341
M00971
M01660
expected
A549
Gm12878
H1hesc
Helas3
Hepg2
K562
JUN
M00041
M00517
M00925
M00172
M00924
M00926
expected
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562
JUND
M00517
M00924
M00925
M00926
expected
Gm12878
H1hesc
Helas3
Hepg2
K562
Sknsh
MYC
M00118
M00322
M00799
M01145
M00123
M00615
M01034
M01154
expected
A549
Gm12878
H1hesc
Helas3
Hepg2
Huvec
K562
Mcf7
REST
M00256
M00325
M01028
M01256
expected
A549
Gm12878
H1hesc
Helas3
Hepg2
K562
Panc1
Pfsk1
Sknsh
U87


SRF
M00152
M00186
M00215
M00810
M00922
M01007
M01257
M01304
100
80
60
40
20
0
*
*
expected
Gm12878
H1hesc
Hepg2
K562
TBP
M00216
M00252
M00471
M00980
100
80
60
40
20
0
expected
Gm12878
H1hesc
Helas3
Hepg2
K562
TCF12
M00698
M00973
M01034
100
80
60
40
20
0
*
*
*
*
expected
A549
Gm12878
H1hesc
Hepg2
TCF7L2
M00671
M01705
100
80
60
40
20
0
*
expected
Hct116
Hek293
Helas3
Hepg2
Mcf7
Panc1
USF1
M00121
M00122
M00187
M00217
M00796
M01034
100
80
60
40
20
0
*
expected
A549
Gm12878
H1hesc
Hepg2
K562
YY1
M00059
M00069
M00793
M01035
100
80
60
40
20
0
expected
A549
Gm12878
H1hesc
Hct116
Hepg2
K562
Nt2d1
Sknshra

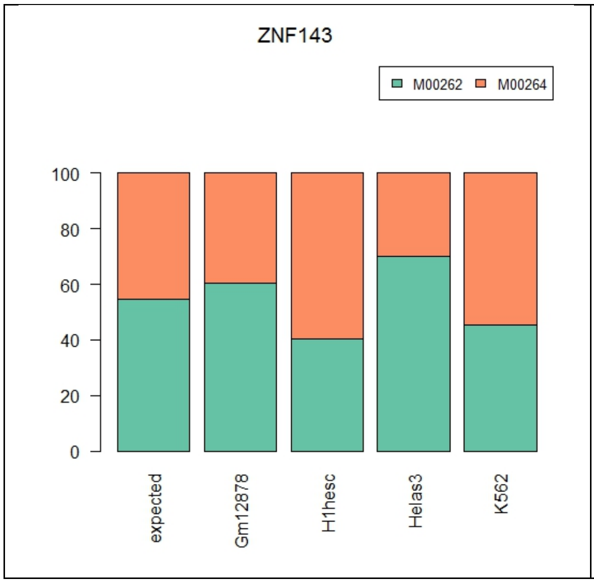


Figure 7.11 Motif usage for the reference TF in different cell types for the NonInteraction model. Y-axis denotes the feature importance of motif usage in the NonInteraction model. The sequence logos for the PWMs can be accessed from Figure 7.2 of Supplemental Data.

TF	Cell line	File name
ATF3	A549	wgEncodeAwgTfbsHaibA549Atf3V0422111Etoh02UniPk.narrowPeak.gz
ATF3	H1hesc	wgEncodeAwgTfbsHaibH1hescAtf3V0416102UniPk.narrowPeak.gz
ATF3	Hepg2	wgEncodeAwgTfbsHaibHepg2Atf3V0416101UniPk.narrowPeak.gz
ATF3	K562	wgEncodeAwgTfbsHaibK562Atf3V0416101UniPk.narrowPeak.gz
BHLHE40	A549	wgEncodeAwgTfbsSydhA549Bhlhe40IggrabUniPk.narrowPeak.gz
BHLHE40	Gm12878	wgEncodeAwgTfbsSydhGm12878Bhlhe40cIggmusUniPk.narrowPeak.gz
BHLHE40	Hepg2	wgEncodeAwgTfbsHaibHepg2Bhlhe40V0416101UniPk.narrowPeak.gz
BHLHE40	K562	wgEncodeAwgTfbsSydhK562Bhlhe40nb100IggrabUniPk.narrowPeak.gz
CEBPB	A549	wgEncodeAwgTfbsSydhA549CebpbIggrabUniPk.narrowPeak
CEBPB	Gm12878	wgEncodeAwgTfbsHaibGm12878Cebpbsc150V0422111UniPk.narrowPeak
CEBPB	H1hesc	wgEncodeAwgTfbsSydhH1hescCebpbIggrabUniPk.narrowPeak
CEBPB	Helas3	wgEncodeAwgTfbsSydhHelas3CebpbIggrabUniPk.narrowPeak
CEBPB	Hepg2	wgEncodeAwgTfbsSydhHepg2CebpbIggrabUniPk.narrowPeak
CEBPB	Imr90	wgEncodeAwgTfbsSydhImr90CebpbIggrabUniPk.narrowPeak
CEBPB	K562	wgEncodeAwgTfbsSydhK562CebpbIggrabUniPk.narrowPeak
CTCF	A549	wgEncodeAwgTfbsUtaA549CtcfUniPk.narrowPeak.gz
CTCF	Gm12878	wgEncodeAwgTfbsBroadGm12878CtcfUniPk.narrowPeak.gz
CTCF	H1hesc	wgEncodeAwgTfbsBroadH1hescCtcfUniPk.narrowPeak.gz
CTCF	Hct116	wgEncodeAwgTfbsUwHct116CtcfUniPk.narrowPeak.gz
CTCF	Hek293	wgEncodeAwgTfbsUwHek293CtcfUniPk.narrowPeak.gz
CTCF	Helas3	wgEncodeAwgTfbsBroadHelas3CtcfUniPk.narrowPeak.gz
CTCF	Hepg2	wgEncodeAwgTfbsBroadHepg2CtcfUniPk.narrowPeak.gz
CTCF	Huvec	wgEncodeAwgTfbsBroadHuvecCtcfUniPk.narrowPeak.gz

CTCF	Imr90	wgEncodeAwgTfbsSydhImr90CtcfbIggrabUniPk.narrowPeak.gz
CTCF	K562	wgEncodeAwgTfbsBroadK562CtcfUniPk.narrowPeak.gz
CTCF	Mcf7	wgEncodeAwgTfbsUtaMcf7CtcfUniPk.narrowPeak.gz
CTCF	Sknshra	wgEncodeAwgTfbsHaibSknshraCtcfV0416102UniPk.narrowPeak.gz
EP300	A549	wgEncodeAwgTfbsHaibA549P300V0422111Etoh02UniPk.narrowPeak
EP300	Gm12878	wgEncodeAwgTfbsHaibGm12878P300Pcr1xUniPk.narrowPeak
EP300	H1hesc	wgEncodeAwgTfbsHaibH1hescP300V0416102UniPk.narrowPeak
EP300	Helas3	wgEncodeAwgTfbsSydhHelas3P300sc584sc584IggrabUniPk.narrowPeak
EP300	Hepg2	wgEncodeAwgTfbsHaibHepg2P300V0416101UniPk.narrowPeak
EP300	Sknsh	wgEncodeSydhTfbsSknshP300bIggrabPk.narrowPeak
EP300	T47d	wgEncodeAwgTfbsHaibT47dP300V0416102Dm002p1hUniPk.narrowPeak
FOS	Gm12878	wgEncodeAwgTfbsSydhGm12878CfosUniPk.narrowPeak
FOS	Helas3	wgEncodeAwgTfbsSydhHelas3CfosUniPk.narrowPeak
FOS	Huvec	wgEncodeAwgTfbsSydhHuvecCfosUcdUniPk.narrowPeak
FOS	K562	wgEncodeAwgTfbsSydhK562CfosUniPk.narrowPeak
FOS	Mcf10	wgEncodeAwgTfbsSydhMcf10aesCfosEtoh01HvdUniPk.narrowPeak
GABPA	A549	wgEncodeAwgTfbsHaibA549GabpV0422111Etoh02UniPk.narrowPeak
GABPA	Gm12878	wgEncodeAwgTfbsHaibGm12878GabpPcr2xUniPk.narrowPeak
GABPA	H1hesc	wgEncodeAwgTfbsHaibH1hescGabpPcr1xUniPk.narrowPeak
GABPA	Helas3	wgEncodeAwgTfbsHaibHelas3GabpPcr1xUniPk.narrowPeak
GABPA	Hepg2	wgEncodeAwgTfbsHaibHepg2GabpPcr2xUniPk.narrowPeak
GABPA	K562	wgEncodeAwgTfbsHaibK562GabpV0416101UniPk.narrowPeak
JUN	Gm12878	wgEncodeYaleChIPseqGm12878Cjun.narrowPeak
JUN	H1hesc	wgEncodeAwgTfbsSydhH1hescCjunIggrabUniPk.narrowPeak
JUN	Helas3	wgEncodeAwgTfbsSydhHelas3CjunIggrabUniPk.narrowPeak
JUN	Hepg2	wgEncodeAwgTfbsSydhHepg2CjunIggrabUniPk.narrowPeak
JUN	Huvec	wgEncodeAwgTfbsSydhHuvecCjunUniPk.narrowPeak
JUN	K562	wgEncodeAwgTfbsSydhK562CjunUniPk.narrowPeak
JUND	Gm12878	wgEncodeAwgTfbsSydhGm12878JundUniPk.narrowPeak
JUND	H1hesc	wgEncodeAwgTfbsHaibH1hescJundV0416102UniPk.narrowPeak
JUND	Helas3	wgEncodeAwgTfbsSydhHelas3JundIggrabUniPk.narrowPeak
JUND	Hepg2	wgEncodeAwgTfbsHaibHepg2JundPcr1xUniPk.narrowPeak
JUND	K562	wgEncodeAwgTfbsSydhK562JundIggrabUniPk.narrowPeak
JUND	Sknsh	wgEncodeSydhTfbsSknshJundIggrabPk.narrowPeak
MAFK	H1hesc	wgEncodeAwgTfbsSydhH1hescMafkIggrabUniPk.narrowPeak.gz
MAFK	Helas3	wgEncodeAwgTfbsSydhHelas3MafkIggrabUniPk.narrowPeak.gz
MAFK	Hepg2	wgEncodeAwgTfbsSydhHepg2Mafkab50322IggrabUniPk.narrowPeak.gz
MAFK	Imr90	wgEncodeAwgTfbsSydhImr90MafkIggrabUniPk.narrowPeak.gz
MAFK	K562	wgEncodeAwgTfbsSydhK562Mafkab50322IggrabUniPk.narrowPeak.gz
MAZ	Gm12878	wgEncodeAwgTfbsSydhGm12878Mazab85725IggmusUniPk.narrowPeak.g
MAZ	Helas3	wgEncodeAwgTfbsSydhHelas3Mazab85725IggrabUniPk.narrowPeak.gz
MAZ	Hepg2	wgEncodeAwgTfbsSydhHepg2Mazab85725IggrabUniPk.narrowPeak.gz

MAZ	K562	wgEncodeAwgTfbsSydhK562Mazab85725IggrabUniPk.narrowPeak.gz
MXI1	Gm12878	wgEncodeAwgTfbsSydhGm12878Mxi1IggmusUniPk.narrowPeak.gz
MXI1	H1hesc	wgEncodeAwgTfbsSydhH1hescMxi1IggrabUniPk.narrowPeak.gz
MXI1	Helas3	wgEncodeAwgTfbsSydhHelas3Mxi1af4185IggrabUniPk.narrowPeak.gz
MXI1	Hepg2	wgEncodeAwgTfbsSydhHepg2Mxi1UniPk.narrowPeak.gz
MXI1	K562	wgEncodeAwgTfbsSydhK562Mxi1af4185IggrabUniPk.narrowPeak.gz
MYC	A549	wgEncodeSydhTfbsA549CmycIggrabPk.narrowPeak.gz
MYC	Gm12878	wgEncodeAwgTfbsUtaGm12878CmycUniPk.narrowPeak.gz
MYC	H1hesc	wgEncodeAwgTfbsSydhH1hescCmycIggrabUniPk.narrowPeak.gz
MYC	Helas3	wgEncodeAwgTfbsUtaHelas3CmycUniPk.narrowPeak.gz
MYC	Hepg2	wgEncodeAwgTfbsUtaHepg2CmycUniPk.narrowPeak.gz
MYC	Huvec	wgEncodeAwgTfbsUtaHuvecCmycUniPk.narrowPeak.gz
MYC	K562	wgEncodeAwgTfbsSydhK562CmycIggrabUniPk.narrowPeak.gz
MYC	Mcf7	wgEncodeAwgTfbsUtaMcf7CmycEstroUniPk.narrowPeak.gz
NRF1	Gm12878	wgEncodeAwgTfbsSydhGm12878Nrf1IggmusUniPk.narrowPeak.gz
NRF1	H1hesc	wgEncodeAwgTfbsSydhH1hescNrf1IggrabUniPk.narrowPeak.gz
NRF1	Helas3	wgEncodeAwgTfbsSydhHelas3Nrf1IggmusUniPk.narrowPeak.gz
NRF1	Hepg2	wgEncodeAwgTfbsSydhHepg2Nrf1IggrabUniPk.narrowPeak.gz
NRF1	K562	wgEncodeAwgTfbsSydhK562Nrf1IggrabUniPk.narrowPeak.gz
REST	A549	wgEncodeAwgTfbsHaibA549NrsfV0422111Etoh02UniPk.narrowPeak
REST	Gm12878	wgEncodeAwgTfbsHaibGm12878NrsfPcr1xUniPk.narrowPeak
REST	H1hesc	wgEncodeAwgTfbsHaibH1hescNrsfV0416102UniPk.narrowPeak
REST	Helas3	wgEncodeAwgTfbsHaibHelas3NrsfPcr1xUniPk.narrowPeak
REST	Hepg2	wgEncodeAwgTfbsHaibHepg2NrsfPcr2xUniPk.narrowPeak
REST	K562	wgEncodeAwgTfbsHaibK562NrsfV0416102UniPk.narrowPeak
REST	Panc1	wgEncodeAwgTfbsHaibPanc1NrsfPcr2xUniPk.narrowPeak
REST	Pfsk1	wgEncodeAwgTfbsHaibPfsk1NrsfPcr2xUniPk.narrowPeak
REST	Sknsh	wgEncodeAwgTfbsHaibSknshNrsfPcr2xUniPk.narrowPeak
REST	U87	wgEncodeAwgTfbsHaibU87NrsfPcr2xUniPk.narrowPeak
RFX5	Gm12878	wgEncodeAwgTfbsSydhGm12878Rfx5200401194IggmusUniPk.narrowPeak
RFX5	H1hesc	wgEncodeAwgTfbsSydhH1hescRfx5200401194IggrabUniPk.narrowPeak.gz
RFX5	Helas3	wgEncodeAwgTfbsSydhHelas3Rfx5200401194IggrabUniPk.narrowPeak.gz
RFX5	Hepg2	wgEncodeAwgTfbsSydhHepg2Rfx5200401194IggrabUniPk.narrowPeak.gz
RFX5	K562	wgEncodeAwgTfbsSydhK562Rfx5IggrabUniPk.narrowPeak.gz
SRF	Gm12878	wgEncodeAwgTfbsHaibGm12878SrfPcr2xUniPk.narrowPeak
SRF	H1hesc	wgEncodeAwgTfbsHaibH1hescSrfPcr1xUniPk.narrowPeak
SRF	Hepg2	wgEncodeAwgTfbsHaibHepg2SrfV0416101UniPk.narrowPeak
SRF	K562	wgEncodeAwgTfbsHaibK562SrfV0416101UniPk.narrowPeak
TBP	Gm12878	wgEncodeAwgTfbsSydhGm12878TbpIggmusUniPk.narrowPeak.gz
TBP	H1hesc	wgEncodeAwgTfbsSydhH1hescTbpIggrabUniPk.narrowPeak.gz
TBP	Helas3	wgEncodeAwgTfbsSydhHelas3TbpIggrabUniPk.narrowPeak.gz
TBP	Hepg2	wgEncodeAwgTfbsSydhHepg2TbpIggrabUniPk.narrowPeak.gz

TBP	K562	wgEncodeAwgTfbsSydhK562TbpIggmusUniPk.narrowPeak.gz
TCF12	A549	wgEncodeAwgTfbsHaibA549Tcf12V0422111Etoh02UniPk.narrowPeak.gz
TCF12	Gm12878	wgEncodeAwgTfbsHaibGm12878Tcf12Pcr1xUniPk.narrowPeak.gz
TCF12	H1hesc	wgEncodeAwgTfbsHaibH1hescTcf12Pcr1xUniPk.narrowPeak.gz
TCF12	Hepg2	wgEncodeAwgTfbsHaibHepg2Tcf12Pcr1xUniPk.narrowPeak.gz
TCF7L2	Hct116	wgEncodeAwgTfbsSydhHct116Tcf7l2UcdUniPk.narrowPeak.gz
TCF7L2	Hek293	wgEncodeAwgTfbsSydhHek293Tcf7l2UcdUniPk.narrowPeak.gz
TCF7L2	Helas3	wgEncodeAwgTfbsSydhHelas3Tcf7l2UcdUniPk.narrowPeak.gz
TCF7L2	Hepg2	wgEncodeAwgTfbsSydhHepg2Tcf7l2UcdUniPk.narrowPeak.gz
TCF7L2	Mcf7	wgEncodeAwgTfbsSydhMcf7Tcf7l2UcdUniPk.narrowPeak.gz
TCF7L2	Panc1	wgEncodeAwgTfbsSydhPanc1Tcf7l2UcdUniPk.narrowPeak.gz
USF1	A549	wgEncodeAwgTfbsHaibA549Usf1Pcr1xDex100nmUniPk.narrowPeak
USF1	Gm12878	wgEncodeAwgTfbsHaibGm12878Usf1Pcr2xUniPk.narrowPeak
USF1	H1hesc	wgEncodeAwgTfbsHaibH1hescUsf1Pcr1xUniPk.narrowPeak
USF1	Hepg2	wgEncodeAwgTfbsHaibHepg2Usf1Pcr1xUniPk.narrowPeak
USF1	K562	wgEncodeAwgTfbsHaibK562Usf1V0416101UniPk.narrowPeak
YY1	A549	wgEncodeAwgTfbsHaibA549Yy1cV0422111Etoh02UniPk.narrowPeak
YY1	Gm12878	wgEncodeAwgTfbsHaibGm12878Yy1sc281Pcr1xUniPk.narrowPeak
YY1	H1hesc	wgEncodeAwgTfbsHaibH1hescYy1sc281V0416102UniPk.narrowPeak
YY1	Hct116	wgEncodeAwgTfbsHaibHct116Yy1sc281V0416101UniPk.narrowPeak
YY1	Hepg2	wgEncodeAwgTfbsHaibHepg2Yy1sc281V0416101UniPk.narrowPeak
YY1	K562	wgEncodeAwgTfbsHaibK562Yy1V0416101UniPk.narrowPeak
YY1	Sknshra	wgEncodeAwgTfbsHaibSknshraYy1sc281V0416102UniPk.narrowPeak
YY1	Nt2d1	wgEncodeAwgTfbsSydhNt2d1Yy1UcdUniPk.narrowPeak
ZNF143	Gm12878	wgEncodeAwgTfbsSydhGm12878Znf143166181apUniPk.narrowPeak.gz
ZNF143	H1hesc	wgEncodeAwgTfbsSydhH1hescZnf143IggrabUniPk.narrowPeak.gz
ZNF143	Helas3	wgEncodeAwgTfbsSydhHelas3Znf143IggrabUniPk.narrowPeak.gz
ZNF143	K562	wgEncodeAwgTfbsSydhK562Znf143IggrabUniPk.narrowPeak.gz

Table 7.6 List of TF-cell pairs, the narrow peak file used for each pair.

	General Information for the TFs		
	# of motif ids	family name	PWMs
ATF3	3	bZIP, CH	M00513, M00801, M00981
BHLHE40	2	bHLH, bHLH-bZIP	M00997, M01034
CEBPB	4	bZIP	M00109, M00117, M00770, M00912
CTCF	2	CH	M01200, M01259
EP300	1	-	M00033
FOS	5	bZIP	M00172, M00517, M00924, M00925, M00926
GABPA	3	ETS	M00341, M00971, M01660

JUN	6	bZIP	M00041, M00172, M00517, M00924, M00925, M009
JUND	4	bZIP	M00517, M00924, M00925, M00926
MAFK	1	bZIP	M00517, M00924, M00925, M00926
MAZ	1	CH	M00649
MXI1	1	bHLH, bHLH-bZIP	M01034
MYC	8	bHLH, bHLH-bZIP	M00118, M00123, M00322, M00615, M00799, M010 M01154
NRF1	1	bZIP, unchar-DBD	M00652
REST	4	CH	M00256, M00325, M01028, M01256
RFX5	1	fork, unchar-DBD	M00975
SRF	8	Mads	M00152, M00186, M00215, M00810, M00922, M010 M01304
TBP	4	Tata	M00216, M00252, M00471, M00980
TCF12	3	bHLh	M00698, M00973, M01034
TCF7L2	2	HMG	M00671, M01705
USF1	6	bHLH, bHLH-bZIP	M00121, M00122, M00187, M00217, M00796, M010
YY1	4	CH	M00059, M00069, M00793, M01035
ZNF143	2	CH	M00262, M00264

family name	Explanation
bZIP	Basic Leucine Zipper Domain
bHLH	Basic Helix-Loop-Helix
ETS	E26 transformation-specific
MADS	MADS box
HMG	High Mobility Group
CH	C2H2 zinc finger
unchar-DBD	Uncharacterized DNA Binding Domain
Tata	TATA-box
fork	fork head domain

Table 7.7 List of TFs, their corresponding TRANSFAC ids, and family name. Explanation of family name abbreviation is also included.

		a) Number of submodels in each model					
TF	Cell	*K*-mer	*K*-merRC	Interaction (1k)	Interaction (2k)	Interaction (5k)	Interac (10k)
ATF3	A549	31	38	32	34	32	
ATF3	H1hesc	33	40	24	18	19	
ATF3	Hepg2	31	22	23	15	29	

ATF3	K562	28	30	29	14	28
BHLHE40	A549	30	31	18	31	27
BHLHE40	Gm12878	36	34	25	22	25
BHLHE40	Hepg2	34	28	30	30	20
BHLHE40	K562	31	31	26	26	29
CEBPB	A549	15	26	23	27	27
CEBPB	Gm12878	47	41	26	29	39
CEBPB	H1hesc	40	26	29	29	26
CEBPB	Helas3	51	32	27	29	30
CEBPB	Hepg2	14	28	18	22	21
CEBPB	Imr90	26	32	20	25	29
CEBPB	K562	18	24	29	29	28
CTCF	A549	30	26	18	29	29
CTCF	Gm12878	30	31	19	20	22
CTCF	H1hesc	30	30	23	21	29
CTCF	Hct116	27	30	17	29	29
CTCF	Hek293	29	27	26	29	29
CTCF	Helas3	30	27	19	20	29
CTCF	Hepg2	31	31	25	18	29
CTCF	Huvec	32	29	19	29	29
CTCF	Imr90	25	28	18	18	29
CTCF	K562	31	23	29	19	29
CTCF	Mcf7	32	27	27	22	18
CTCF	Sknshra	31	23	18	15	16
EP300	A549	31	38	28	35	35
EP300	Gm12878	42	41	32	30	37
EP300	H1hesc	48	42	34	33	27
EP300	Helas3	23	31	27	29	32
EP300	Hepg2	39	34	21	24	33
EP300	Sknsh	44	42	38	35	36
EP300	T47d	45	42	27	37	35
FOS	Gm12878	29	29	29	19	29
FOS	Helas3	31	31	28	28	29
FOS	Huvec	30	30	30	27	24
FOS	K562	20	29	24	14	22
FOS	Mcf10a	29	17	29	29	29
GABPA	A549	29	33	15	28	18
GABPA	Gm12878	32	33	25	25	29
GABPA	H1hesc	16	40	26	29	31
GABPA	Helas3	29	28	19	24	31
GABPA	Hepg2	23	30	23	23	22
GABPA	K562	28	26	28	28	27

JUN	Gm12878	29	35	31	33	31
JUN	H1hesc	27	34	15	28	29
JUN	Helas3	31	28	17	17	19
JUN	Hepg2	31	31	25	15	25
JUN	Huvec	31	30	18	26	24
JUN	K562	28	30	23	28	20
JUND	Gm12878	24	31	30	23	30
JUND	H1hesc	28	40	27	30	31
JUND	Helas3	26	1	25	20	25
JUND	Hepg2	32	33	26	30	31
JUND	K562	31	30	27	19	29
JUND	Sknsh	31	34	31	34	32
MAFK	H1hesc	27	31	24	29	30
MAFK	Helas3	29	24	30	30	30
MAFK	Hepg2	30	14	27	27	17
MAFK	Imr90	29	29	25	28	29
MAFK	K562	33	24	29	27	19
MAZ	Gm12878	45	39	31	40	36
MAZ	Helas3	48	37	36	36	37
MAZ	Hepg2	44	43	33	30	37
MAZ	K562	41	52	31	31	31
MXI1	Gm12878	35	32	30	28	32
MXI1	H1hesc	36	38	31	25	34
MXI1	Helas3	34	36	29	31	36
MXI1	Hepg2	32	36	25	23	28
MXI1	K562	37	31	29	31	34
MYC	A549	32	30	32	35	32
MYC	Gm12878	37	36	34	26	21
MYC	H1hesc	35	36	30	32	28
MYC	Helas3	36	36	27	31	33
MYC	Hepg2	36	36	36	30	34
MYC	Huvec	36	35	24	35	34
MYC	K562	33	35	28	31	32
MYC	Mcf7	34	32	36	36	33
NRF1	Gm12878	30	26	27	29	32
NRF1	H1hesc	31	27	30	30	30
NRF1	Helas3	31	13	18	29	30
NRF1	Hepg2	28	28	19	10	29
NRF1	K562	31	23	24	29	23
REST	A549	40	37	28	28	32
REST	Gm12878	24	24	25	31	24
REST	H1hesc	31	24	24	22	28

REST	Helas3	28	30	24	19	27
REST	Hepg2	39	38	32	27	32
REST	K562	32	31	26	22	30
REST	Panc1	36	32	26	23	31
REST	Pfsk1	32	19	28	21	30
REST	Sknsh	47	48	26	28	29
REST	U87	40	35	27	28	29
RFX5	Gm12878	25	31	29	32	31
RFX5	H1hesc	25	23	26	31	30
RFX5	Helas3	28	35	23	23	31
RFX5	Hepg2	22	31	23	30	31
RFX5	K562	26	32	31	30	30
SRF	Gm12878	33	27	19	15	15
SRF	H1hesc	30	32	26	26	29
SRF	Hepg2	35	34	26	26	27
SRF	K562	26	39	24	28	24
TBP	Gm12878	50	48	36	34	37
TBP	H1hesc	47	36	35	36	34
TBP	Helas3	40	39	31	31	33
TBP	Hepg2	45	45	39	39	31
TBP	K562	42	42	38	36	36
TCF12	A549	35	35	37	34	35
TCF12	Gm12878	34	29	25	19	32
TCF12	H1hesc	48	30	31	33	31
TCF12	Hepg2	37	37	25	14	31
TCF7L2	Hct116	30	34	22	33	33
TCF7L2	Hek293	40	37	25	37	37
TCF7L2	Helas3	35	31	29	26	30
TCF7L2	Hepg2	27	31	24	32	35
TCF7L2	Mcf7	32	34	34	33	37
TCF7L2	Panc1	38	38	32	36	32
USF1	A549	30	30	27	27	27
USF1	Gm12878	30	26	26	24	27
USF1	H1hesc	13	28	25	29	17
USF1	Hepg2	18	25	28	28	28
USF1	K562	30	26	24	28	28
YY1	A549	26	33	34	30	31
YY1	Gm12878	31	30	30	33	32
YY1	H1hesc	29	30	23	24	24
YY1	Hct116	29	33	32	33	31
YY1	Hepg2	32	31	34	24	33
YY1	K562	31	32	27	27	31

YY1	Nt2d1	29	28	32	22	31
YY1	Sknshra	31	32	28	28	19
ZNF143	Gm12878	29	31	31	29	31
ZNF143	H1hesc	51	53	17	32	31
ZNF143	Helas3	49	52	29	32	17
ZNF143	K562	30	31	31	27	25

		b) Area under ROC curve					
TF	Cell	*K*-mer	*K*-merRC	Interaction (1k)	Interaction (2k)	Interaction (5k)	Intera (10k)
ATF3	A549	0.782	0.805	0.822	0.826	0.817	
ATF3	H1hesc	0.886	0.907	0.923	0.922	0.925	
ATF3	Hepg2	0.900	0.898	0.919	0.918	0.925	
ATF3	K562	0.926	0.945	0.956	0.931	0.954	
BHLHE40	A549	0.903	0.902	0.899	0.898	0.888	
BHLHE40	Gm12878	0.918	0.929	0.918	0.908	0.903	
BHLHE40	Hepg2	0.918	0.925	0.937	0.933	0.932	
BHLHE40	K562	0.916	0.926	0.920	0.917	0.912	
CEBPB	A549	0.909	0.944	0.981	0.979	0.975	
CEBPB	Gm12878	0.773	0.794	0.787	0.778	0.781	
CEBPB	H1hesc	0.958	0.959	0.982	0.985	0.979	
CEBPB	Helas3	0.927	0.928	0.964	0.964	0.951	
CEBPB	Hepg2	0.925	0.963	0.981	0.983	0.979	
CEBPB	Imr90	0.924	0.953	0.980	0.977	0.974	
CEBPB	K562	0.919	0.940	0.980	0.977	0.975	
CTCF	A549	0.913	0.910	0.970	0.976	0.968	
CTCF	Gm12878	0.909	0.925	0.966	0.972	0.968	
CTCF	H1hesc	0.897	0.917	0.963	0.961	0.959	
CTCF	Hct116	0.906	0.926	0.962	0.969	0.967	
CTCF	Hek293	0.902	0.919	0.973	0.977	0.974	
CTCF	Helas3	0.914	0.921	0.962	0.963	0.954	
CTCF	Hepg2	0.898	0.916	0.967	0.969	0.958	
CTCF	Huvec	0.913	0.923	0.965	0.969	0.963	
CTCF	Imr90	0.890	0.916	0.966	0.970	0.963	
CTCF	K562	0.890	0.886	0.949	0.949	0.943	
CTCF	Mcf7	0.908	0.917	0.969	0.964	0.962	
CTCF	Sknshra	0.924	0.921	0.975	0.980	0.971	
EP300	A549	0.816	0.850	0.854	0.859	0.848	
EP300	Gm12878	0.802	0.824	0.815	0.807	0.799	
EP300	H1hesc	0.780	0.796	0.801	0.802	0.793	
EP300	Helas3	0.853	0.894	0.925	0.918	0.904	

EP300	Hepg2	0.870	0.872	0.896	0.895	0.893
EP300	Sknsh	0.715	0.729	0.693	0.693	0.677
EP300	T47d	0.832	0.848	0.815	0.814	0.794
FOS	Gm12878	0.945	0.949	0.965	0.956	0.951
FOS	Helas3	0.954	0.960	0.968	0.968	0.962
FOS	Huvec	0.972	0.972	0.974	0.974	0.962
FOS	K562	0.964	0.971	0.974	0.970	0.968
FOS	Mcf10a	0.976	0.975	0.979	0.981	0.977
GABPA	A549	0.887	0.909	0.890	0.900	0.885
GABPA	Gm12878	0.880	0.887	0.884	0.883	0.883
GABPA	H1hesc	0.811	0.860	0.892	0.889	0.886
GABPA	Helas3	0.922	0.925	0.925	0.927	0.928
GABPA	Hepg2	0.912	0.927	0.921	0.922	0.916
GABPA	K562	0.926	0.928	0.931	0.924	0.922
JUN	Gm12878	0.742	0.757	0.732	0.732	0.718
JUN	H1hesc	0.857	0.875	0.880	0.880	0.874
JUN	Helas3	0.958	0.959	0.961	0.956	0.950
JUN	Hepg2	0.961	0.972	0.977	0.975	0.974
JUN	Huvec	0.963	0.970	0.966	0.969	0.960
JUN	K562	0.964	0.972	0.977	0.977	0.970
JUND	Gm12878	0.910	0.928	0.909	0.901	0.888
JUND	H1hesc	0.779	0.809	0.799	0.811	0.806
JUND	Helas3	0.974	0.962	0.977	0.978	0.974
JUND	Hepg2	0.931	0.945	0.956	0.956	0.947
JUND	K562	0.953	0.960	0.967	0.964	0.962
JUND	Sknsh	0.740	0.755	0.754	0.749	0.743
MAFK	H1hesc	0.939	0.955	0.961	0.966	0.958
MAFK	Helas3	0.945	0.944	0.961	0.959	0.946
MAFK	Hepg2	0.961	0.953	0.969	0.971	0.959
MAFK	Imr90	0.970	0.976	0.979	0.977	0.966
MAFK	K562	0.936	0.939	0.959	0.954	0.937
MAZ	Gm12878	0.800	0.808	0.801	0.815	0.806
MAZ	Helas3	0.790	0.797	0.821	0.822	0.816
MAZ	Hepg2	0.794	0.802	0.815	0.816	0.816
MAZ	K562	0.799	0.827	0.844	0.841	0.844
MXI1	Gm12878	0.825	0.832	0.837	0.827	0.833
MXI1	H1hesc	0.818	0.829	0.828	0.825	0.822
MXI1	Helas3	0.862	0.872	0.870	0.872	0.862
MXI1	Hepg2	0.860	0.882	0.859	0.851	0.853

MXI1	K562	0.849	0.848	0.860	0.858	0.847
MYC	A549	0.716	0.735	0.731	0.734	0.734
MYC	Gm12878	0.749	0.764	0.745	0.724	0.700
MYC	H1hesc	0.818	0.837	0.840	0.847	0.840
MYC	Helas3	0.836	0.856	0.856	0.865	0.846
MYC	Hepg2	0.843	0.856	0.847	0.853	0.833
MYC	Huvec	0.876	0.885	0.864	0.874	0.862
MYC	K562	0.879	0.889	0.888	0.889	0.877
MYC	Mcf7	0.857	0.864	0.851	0.845	0.823
NRF1	Gm12878	0.941	0.943	0.923	0.940	0.937
NRF1	H1hesc	0.944	0.952	0.937	0.944	0.939
NRF1	Helas3	0.948	0.943	0.937	0.944	0.948
NRF1	Hepg2	0.971	0.982	0.967	0.981	0.973
NRF1	K562	0.938	0.934	0.937	0.944	0.926
REST	A549	0.849	0.862	0.879	0.876	0.871
REST	Gm12878	0.909	0.920	0.920	0.924	0.912
REST	H1hesc	0.965	0.965	0.967	0.964	0.961
REST	Helas3	0.955	0.968	0.961	0.959	0.947
REST	Hepg2	0.883	0.894	0.886	0.887	0.880
REST	K562	0.917	0.927	0.933	0.929	0.924
REST	Panc1	0.873	0.882	0.891	0.894	0.887
REST	Pfsk1	0.931	0.923	0.938	0.934	0.928
REST	Sknsh	0.847	0.868	0.860	0.858	0.840
REST	U87	0.867	0.879	0.893	0.896	0.895
RFX5	Gm12878	0.861	0.881	0.892	0.892	0.885
RFX5	H1hesc	0.830	0.823	0.852	0.860	0.841
RFX5	Helas3	0.816	0.839	0.873	0.875	0.879
RFX5	Hepg2	0.834	0.865	0.864	0.874	0.859
RFX5	K562	0.784	0.803	0.815	0.805	0.809
SRF	Gm12878	0.835	0.838	0.893	0.889	0.886
SRF	H1hesc	0.875	0.894	0.933	0.934	0.935
SRF	Hepg2	0.844	0.855	0.923	0.923	0.911
SRF	K562	0.823	0.848	0.873	0.879	0.876
TBP	Gm12878	0.749	0.768	0.753	0.753	0.740
TBP	H1hesc	0.745	0.758	0.740	0.742	0.733
TBP	Helas3	0.764	0.773	0.776	0.782	0.768
TBP	Hepg2	0.759	0.765	0.765	0.757	0.731
TBP	K562	0.765	0.779	0.768	0.759	0.753
TCF12	A549	0.819	0.831	0.848	0.848	0.832

TCF12	Gm12878	0.914	0.916	0.928	0.915	0.915
TCF12	H1hesc	0.859	0.856	0.872	0.874	0.864
TCF12	Hepg2	0.813	0.818	0.849	0.822	0.832
TCF7L2	Hct116	0.860	0.874	0.874	0.876	0.868
TCF7L2	Hek293	0.822	0.833	0.828	0.841	0.820
TCF7L2	Helas3	0.784	0.810	0.823	0.813	0.798
TCF7L2	Hepg2	0.783	0.801	0.805	0.813	0.792
TCF7L2	Mcf7	0.822	0.833	0.854	0.846	0.825
TCF7L2	Panc1	0.804	0.821	0.813	0.817	0.804
USF1	A549	0.942	0.947	0.957	0.957	0.953
USF1	Gm12878	0.970	0.969	0.970	0.969	0.966
USF1	H1hesc	0.974	0.984	0.985	0.981	0.979
USF1	Hepg2	0.975	0.985	0.987	0.986	0.985
USF1	K562	0.980	0.980	0.984	0.981	0.981
YY1	A549	0.857	0.873	0.866	0.858	0.849
YY1	Gm12878	0.885	0.893	0.891	0.889	0.885
YY1	H1hesc	0.932	0.935	0.935	0.928	0.925
YY1	Hct116	0.868	0.882	0.889	0.889	0.885
YY1	Hepg2	0.877	0.883	0.875	0.862	0.869
YY1	K562	0.876	0.886	0.874	0.875	0.875
YY1	Nt2d1	0.921	0.927	0.926	0.915	0.912
YY1	Sknshra	0.908	0.913	0.909	0.906	0.889
ZNF143	Gm12878	0.835	0.856	0.904	0.915	0.901
ZNF143	H1hesc	0.847	0.875	0.883	0.904	0.899
ZNF143	Helas3	0.818	0.829	0.846	0.860	0.845
ZNF143	K562	0.831	0.861	0.899	0.900	0.903

Table 7.8a&b Number of sub-models and performance of various EMT (Ensemble Model of TF).

TF	Cell	*K*-mer	*K*-merRC	Interaction	kmer-SVM
ATF3	A549	0.782	0.805	0.822	0.801
ATF3	H1hesc	0.886	0.907	0.923	0.907
ATF3	Hepg2	0.900	0.898	0.919	0.910
ATF3	K562	0.926	0.945	0.956	0.944
BHLHE40	A549	0.903	0.902	0.899	0.896
BHLHE40	Gm12878	0.918	0.929	0.918	0.924
BHLHE40	Hepg2	0.918	0.925	0.937	0.937
BHLHE40	K562	0.916	0.926	0.920	0.921
CEBPB	A549	0.909	0.944	0.981	0.956

CEBPB	Gm12878	0.773	0.794	0.787	0.778
CEBPB	H1hesc	0.958	0.959	0.982	0.968
CEBPB	Helas3	0.927	0.928	0.964	0.938
CEBPB	Hepg2	0.925	0.963	0.981	0.968
CEBPB	Imr90	0.924	0.953	0.980	0.960
CEBPB	K562	0.919	0.940	0.980	0.959
CTCF	A549	0.913	0.910	0.970	0.945
CTCF	Gm12878	0.909	0.925	0.966	0.941
CTCF	H1hesc	0.897	0.917	0.963	0.942
CTCF	Hct116	0.906	0.926	0.962	0.948
CTCF	Hek293	0.902	0.919	0.973	0.942
CTCF	Helas3	0.914	0.921	0.962	0.941
CTCF	Hepg2	0.898	0.916	0.967	0.934
CTCF	Huvec	0.913	0.923	0.965	0.946
CTCF	Imr90	0.890	0.916	0.966	0.945
CTCF	K562	0.890	0.886	0.949	0.925
CTCF	Mcf7	0.908	0.917	0.969	0.942
CTCF	Sknshra	0.924	0.921	0.975	0.955
EP300	A549	0.816	0.850	0.854	0.845
EP300	Gm12878	0.802	0.824	0.815	0.808
EP300	H1hesc	0.780	0.796	0.801	0.792
EP300	Helas3	0.853	0.894	0.925	0.914
EP300	Hepg2	0.870	0.872	0.896	0.894
EP300	Sknsh	0.715	0.729	0.693	0.713
EP300	T47d	0.832	0.848	0.815	0.832
FOS	Gm12878	0.945	0.949	0.965	0.962
FOS	Helas3	0.954	0.960	0.968	0.964
FOS	Huvec	0.972	0.972	0.974	0.976
FOS	K562	0.964	0.971	0.974	0.977
FOS	Mcf10a	0.976	0.975	0.979	0.980
GABPA	A549	0.887	0.909	0.890	0.900
GABPA	Gm12878	0.880	0.887	0.884	0.890
GABPA	H1hesc	0.811	0.860	0.892	0.859
GABPA	Helas3	0.922	0.925	0.925	0.931
GABPA	Hepg2	0.912	0.927	0.921	0.933
GABPA	K562	0.926	0.928	0.931	0.930
JUN	Gm12878	0.742	0.757	0.732	0.737
JUN	H1hesc	0.857	0.875	0.880	0.871
JUN	Helas3	0.958	0.959	0.961	0.966
JUN	Hepg2	0.961	0.972	0.977	0.974
JUN	Huvec	0.963	0.970	0.966	0.977
JUN	K562	0.964	0.972	0.977	0.975
JUND	Gm12878	0.910	0.928	0.909	0.948

JUND	H1hesc	0.779	0.809	0.799	0.802
JUND	Helas3	0.974	0.962	0.977	0.980
JUND	Hepg2	0.931	0.945	0.956	0.937
JUND	K562	0.953	0.960	0.967	0.954
JUND	Sknsh	0.740	0.755	0.754	0.738
MAFK	H1hesc	0.939	0.955	0.961	0.958
MAFK	Helas3	0.945	0.944	0.961	0.959
MAFK	Hepg2	0.961	0.953	0.969	0.975
MAFK	Imr90	0.970	0.976	0.979	0.980
MAFK	K562	0.936	0.939	0.959	0.947
MAZ	Gm12878	0.800	0.808	0.801	0.800
MAZ	Helas3	0.790	0.797	0.821	0.806
MAZ	Hepg2	0.794	0.802	0.815	0.800
MAZ	K562	0.799	0.827	0.844	0.833
MXI1	Gm12878	0.825	0.832	0.837	0.827
MXI1	H1hesc	0.818	0.829	0.828	0.830
MXI1	Helas3	0.862	0.872	0.870	0.863
MXI1	Hepg2	0.860	0.882	0.859	0.872
MXI1	K562	0.849	0.848	0.860	0.863
MYC	A549	0.716	0.735	0.731	0.696
MYC	Gm12878	0.749	0.764	0.745	0.733
MYC	H1hesc	0.818	0.837	0.840	0.835
MYC	Helas3	0.836	0.856	0.856	0.855
MYC	Hepg2	0.843	0.856	0.847	0.844
MYC	Huvec	0.876	0.885	0.864	0.878
MYC	K562	0.879	0.889	0.888	0.890
MYC	Mcf7	0.857	0.864	0.851	0.845
NRF1	Gm12878	0.941	0.943	0.923	0.951
NRF1	H1hesc	0.944	0.952	0.937	0.944
NRF1	Helas3	0.948	0.943	0.937	0.934
NRF1	Hepg2	0.971	0.982	0.967	0.988
NRF1	K562	0.938	0.934	0.937	0.938
REST	A549	0.849	0.862	0.879	0.866
REST	Gm12878	0.909	0.920	0.920	0.895
REST	H1hesc	0.965	0.965	0.967	0.980
REST	Helas3	0.955	0.968	0.961	0.977
REST	Hepg2	0.883	0.894	0.886	0.890
REST	K562	0.917	0.927	0.933	0.930
REST	Panc1	0.873	0.882	0.891	0.896
REST	Pfsk1	0.931	0.923	0.938	0.941
REST	Sknsh	0.847	0.868	0.860	0.840
REST	U87	0.867	0.879	0.893	0.885

RFX5	Gm12878	0.861	0.881	0.892	0.894
RFX5	H1hesc	0.830	0.823	0.852	0.855
RFX5	Helas3	0.816	0.839	0.873	0.858
RFX5	Hepg2	0.834	0.865	0.864	0.874
RFX5	K562	0.784	0.803	0.815	0.820
SRF	Gm12878	0.835	0.838	0.893	0.861
SRF	H1hesc	0.875	0.894	0.933	0.893
SRF	Hepg2	0.844	0.855	0.923	0.870
SRF	K562	0.823	0.848	0.873	0.845
TBP	Gm12878	0.749	0.768	0.753	0.739
TBP	H1hesc	0.745	0.758	0.740	0.749
TBP	Helas3	0.764	0.773	0.776	0.765
TBP	Hepg2	0.759	0.765	0.765	0.713
TBP	K562	0.765	0.779	0.768	0.752
TCF12	A549	0.819	0.831	0.848	0.831
TCF12	Gm12878	0.914	0.916	0.928	0.924
TCF12	H1hesc	0.859	0.856	0.872	0.874
TCF12	Hepg2	0.813	0.818	0.849	0.807
TCF7L2	Hct116	0.860	0.874	0.874	0.877
TCF7L2	Hek293	0.822	0.833	0.828	0.828
TCF7L2	Helas3	0.784	0.810	0.823	0.806
TCF7L2	Hepg2	0.783	0.801	0.805	0.806
TCF7L2	Mcf7	0.822	0.833	0.854	0.857
TCF7L2	Panc1	0.804	0.821	0.813	0.820
USF1	A549	0.942	0.947	0.957	0.951
USF1	Gm12878	0.970	0.969	0.970	0.972
USF1	H1hesc	0.974	0.984	0.985	0.986
USF1	Hepg2	0.975	0.985	0.987	0.983
USF1	K562	0.980	0.980	0.984	0.983
YY1	A549	0.857	0.873	0.866	0.866
YY1	Gm12878	0.885	0.893	0.891	0.892
YY1	H1hesc	0.932	0.935	0.935	0.940
YY1	Hct116	0.868	0.882	0.889	0.886
YY1	Hepg2	0.877	0.883	0.875	0.887
YY1	K562	0.876	0.886	0.874	0.885
YY1	Nt2d1	0.921	0.927	0.926	0.927
YY1	Sknshra	0.908	0.913	0.909	0.919
ZNF143	Gm12878	0.835	0.856	0.904	0.885
ZNF143	H1hesc	0.847	0.875	0.883	0.871
ZNF143	Helas3	0.818	0.829	0.846	0.821
ZNF143	K562	0.831	0.861	0.899	0.889

Table 7.9 Comparison of EMT (Ensemble Model of TF) with kmer-SVM (K-mer based Support Vector Machine).

Filename	Cell line
wgEncodeCshlLongRnaSeqA549CellLongnonpolyaAlnRep1.bam	A549
wgEncodeCshlLongRnaSeqA549CellLongnonpolyaAlnRep2.bam	
wgEncodeCshlLongRnaSeqA549CellPapAlnRep1.bam	
wgEncodeCshlLongRnaSeqA549CellPapAlnRep2.bam	
wgEncodeCshlLongRnaSeqGm12878CellLongnonpolyaAlnRep1.bam	Gm12878
wgEncodeCshlLongRnaSeqGm12878CellLongnonpolyaAlnRep2.bam	
wgEncodeCshlLongRnaSeqGm12878CellPapAlnRep1.bam	
wgEncodeCshlLongRnaSeqGm12878CellPapAlnRep2.bam	
wgEncodeCshlLongRnaSeqH1hescCellLongnonpolyaAlnRep1.bam	H1hesc
wgEncodeCshlLongRnaSeqH1hescCellLongnonpolyaAlnRep2.bam	
wgEncodeCshlLongRnaSeqH1hescCellPapAlnRep1.bam	
wgEncodeCshlLongRnaSeqH1hescCellPapAlnRep2.bam	
wgEncodeCaltechRnaSeqHct116R2x75Il200AlignsRep1V2.bam	Hct116
wgEncodeCaltechRnaSeqHct116R2x75Il200AlignsRep2V2.bam	
wgEncodeCshlLongRnaSeqHelas3CellLongnonpolyaAlnRep1.bam	Helas3
wgEncodeCshlLongRnaSeqHelas3CellLongnonpolyaAlnRep2.bam	
wgEncodeCshlLongRnaSeqHelas3CellPapAlnRep1.bam	
wgEncodeCshlLongRnaSeqHelas3CellPapAlnRep2.bam	
wgEncodeCshlLongRnaSeqHepg2CellLongnonpolyaAlnRep1.bam	Hepg2
wgEncodeCshlLongRnaSeqHepg2CellLongnonpolyaAlnRep2.bam	
wgEncodeCshlLongRnaSeqHepg2CellPapAlnRep1.bam	
wgEncodeCshlLongRnaSeqHepg2CellPapAlnRep2.bam	
wgEncodeCshlLongRnaSeqHuvecCellLongnonpolyaAlnRep1.bam	Huvec
wgEncodeCshlLongRnaSeqHuvecCellLongnonpolyaAlnRep2.bam	
wgEncodeCshlLongRnaSeqHuvecCellPapAlnRep1.bam	
wgEncodeCshlLongRnaSeqHuvecCellPapAlnRep2.bam	
wgEncodeCshlLongRnaSeqImr90CellPapAlnRep1.bam	Imr90
wgEncodeCshlLongRnaSeqImr90CellPapAlnRep2.bam	
wgEncodeCshlLongRnaSeqImr90CellTotalAlnRep1.bam	
wgEncodeCshlLongRnaSeqImr90CellTotalAlnRep2.bam	
wgEncodeCshlLongRnaSeqK562CellLongnonpolyaAlnRep1.bam	K562
wgEncodeCshlLongRnaSeqK562CellLongnonpolyaAlnRep2.bam	
wgEncodeCshlLongRnaSeqK562CellPapAlnRep1.bam	
wgEncodeCshlLongRnaSeqK562CellPapAlnRep2.bam	
wgEncodeCshlLongRnaSeqSknshCellPapAlnRep3.bam	Sknsh
wgEncodeCshlLongRnaSeqSknshCellPapAlnRep4.bam	
wgEncodeCshlLongRnaSeqSknshCytosolPapAlnRep3.bam	
wgEncodeCshlLongRnaSeqSknshCytosolPapAlnRep4.bam	
wgEncodeCshlLongRnaSeqSknshraCellLongnonpolyaAlnRep1.bam	Sknshra

wgEncodeCshlLongRnaSeqSknshraCellLongnonpolyaAlnRep2.bam wgEncodeCshlLongRnaSeqSknshraCellPapAlnRep1.bam wgEncodeCshlLongRnaSeqSknshraCellPapAlnRep2.bam	
wgEncodeHaibRnaSeqPanc1AlnRep1.bam wgEncodeHaibRnaSeqPanc1AlnRep2.bam	Panc1
wgEncodeHaibRnaSeqPfsk1AlnRep1.bam wgEncodeHaibRnaSeqPfsk1AlnRep2.bam	Pfsk1
wgEncodeHaibRnaSeqT47dBpa14hAlnRep1.bam wgEncodeHaibRnaSeqT47dBpa14hAlnRep2.bam wgEncodeHaibRnaSeqT47dDm002p4hAlnRep1.bam wgEncodeHaibRnaSeqT47dDm002p4hAlnRep2.bam	T47d
wgEncodeHaibRnaSeqU87AlnRep1V2.bam wgEncodeHaibRnaSeqU87AlnRep2V2.bam	U87

Table 7.10 List of RNASeq files for various cell lines obtained from ENCODE to measure log fold change (logFC) of the gene expression. Explanation of cell line is also included.

TRANSFAC id	Name
M00001	<NA>
M00002	TCF3
M00005	TFAP4
M00006	MEF2A
M00008	<NA>
M00017	<NA>
M00024	E2F1
M00026	MEF2A
M00034	<NA>
M00035	<NA>
M00036	<NA>
M00040	ATF2
M00041	JUN
M00050	E2F1
M00051	<NA>
M00052	RELA
M00053	REL
M00054	RELA
M00055	MYCN
M00056	NFIC
M00065	TCF3
M00066	TCF3
M00070	TAL1
M00071	TCF3

M00075	GATA1
M00096	<NA>
M00109	CEBPB
M00116	CEBPA
M00117	CEBPB
M00118	MYC
M00119	<NA>
M00121	USF1
M00122	USF1
M00123	MYC
M00124	<NA>
M00126	GATA1
M00127	GATA1
M00128	GATA1
M00132	<NA>
M00134	HNF4A
M00139	<NA>
M00146	HSF1
M00152	SRF
M00158	NR2F1
M00159	CEBPA
M00172	FOS
M00175	TFAP4
M00176	TFAP4
M00179	ATF2
M00184	<NA>
M00185	NFYB
M00186	SRF
M00187	USF1
M00190	CEBPA
M00191	ESR1
M00192	NR3C1
M00193	NFIC
M00194	NFKB1
M00196	<NA>
M00201	CEBPA
M00203	GATA1
M00205	NR3C1
M00206	<NA>
M00208	<NA>
M00215	SRF

M00217	USF1
M00222	TCF3
M00223	STAT1
M00224	STAT1
M00225	STAT3
M00231	MEF2A
M00232	MEF2A
M00233	MEF2A
M00235	<NA>
M00236	<NA>
M00237	<NA>
M00240	NKX2-5
M00241	NKX2-5
M00242	PPARA
M00243	EGR1
M00249	CEBPA
M00251	XBP1
M00260	<NA>
M00271	RUNX1
M00272	<NA>
M00280	RFX1
M00281	RFX1
M00284	NFE2L1
M00285	NFE2L1
M00302	NFATC1
M00322	MYC
M00327	PAX3
M00338	<NA>
M00341	GABPA
M00346	GATA1
M00347	GATA1
M00360	PAX3
M00403	MEF2A
M00405	MEF2A
M00406	MEF2A
M00407	MEF2A
M00411	HNF4A
M00416	<NA>
M00419	<NA>
M00420	HOXA9
M00421	HOXA9

M00425	<NA>
M00426	<NA>
M00427	<NA>
M00428	E2F1
M00430	E2F1
M00431	E2F1
M00444	VDR
M00457	STAT5A
M00460	STAT5A
M00466	HIF1A
M00490	<NA>
M00491	<NA>
M00492	STAT1
M00493	STAT5A
M00495	BACH1
M00496	STAT1
M00497	STAT3
M00498	STAT4
M00499	STAT5A
M00511	ESRRA
M00512	PPARG
M00513	ATF3
M00514	ATF4
M00515	PPARG
M00516	E2F1
M00517	FOS
M00518	RXRA
M00528	PPARG
M00538	XBP1
M00539	<NA>
M00615	MYC
M00619	<NA>
M00621	CEBPD
M00622	CEBPG
M00626	RFX1
M00631	NR1H4
M00638	HNF4A
M00641	HSF1
M00646	<NA>
M00647	NR1H3
M00691	<NA>

M00693	TCF3
M00698	TCF12
M00712	<NA>
M00726	USF2
M00731	RUNX2
M00736	<NA>
M00737	<NA>
M00738	<NA>
M00739	<NA>
M00744	POU1F1
M00750	HMGA1
M00761	TP63
M00762	NR2F2
M00763	PPARA
M00764	HNF4A
M00765	NR2F2
M00766	<NA>
M00767	<NA>
M00769	RUNX1
M00770	CEBPA
M00774	RELA
M00775	NFYB
M00777	STAT1
M00778	<NA>
M00789	GATA1
M00790	<NA>
M00792	<NA>
M00796	USF1
M00797	HIF1A
M00799	MYC
M00801	ATF7
M00802	POU1F1
M00803	E2F1
M00804	TCF3
M00806	NFIC
M00807	EGR1
M00808	PAX5
M00810	SRF
M00821	NFE2L2
M00912	CEBPA
M00916	CREB1

M00917	CREB1
M00918	E2F1
M00919	E2F1
M00920	E2F1
M00921	NR3C1
M00922	SRF
M00924	FOS
M00925	FOS
M00926	FOS
M00927	TFAP4
M00929	TCF3
M00931	SP3
M00932	SP3
M00933	SP3
M00935	NFATC1
M00938	E2F1
M00939	E2F1
M00940	E2F1
M00941	MEF2A
M00955	NR3C1
M00959	ESR1
M00960	NR3C1
M00961	VDR
M00963	RARB
M00964	NR1I3
M00965	NR2F2
M00966	VDR
M00967	NR2F2
M00971	ERG
M00973	TCF3
M00974	<NA>
M00975	RFX2
M00976	HIF1A
M00981	ATF7
M00982	EGR1
M00983	NFE2
M00984	RUNX1
M00993	TAL1
M00998	PBX2
M01007	SRF
M01009	<NA>

M01010	HMGA1
M01011	<NA>
M01013	<NA>
M01023	HSF1
M01029	TFE3
M01031	HNF4A
M01032	HNF4A
M01033	HNF4A
M01034	MYC
M01036	NR2F2
M01043	NKX2-5
M01075	ZBTB16
M01116	<NA>
M01145	MYC
M01154	MYC
M01196	NFIC
M01249	EPAS1
M01250	E2F1
M01251	E2F1
M01252	E2F6
M01257	SRF
M01260	STAT1
M01267	FOSL1
M01268	NR1H4
M01269	NR4A2
M01270	PPARG
M01281	NFATC2
M01282	PPARA
M01287	NEUROD1
M01288	NEUROD1
M01303	<NA>
M01304	SRF
M01339	PAX7
M01351	HOXA9
M01355	ALX3
M01357	<NA>
M01362	<NA>
M01379	<NA>
M01411	PKNOX2
M01414	NKX2-5
M01417	<NA>

M01419	<NA>
M01425	<NA>
M01453	<NA>
M01459	PKNOX1
M01465	POU1F1
M01588	KLF4
M01591	TAL1
M01595	STAT3
M01651	<NA>
M01652	<NA>
M01653	HMGA1
M01655	<NA>
M01658	RUNX1
M01666	STAT4
M01716	ATOH1
M01718	NFATC1
M01724	THRA
M01770	XBP1
M01801	ESR1
M01808	MYCN
M01820	CREM
M01823	STAT1
M01830	<NA>
M01835	KLF4
M01841	ESRRA

Table 7.11 List of Heterodimerizing TFs and their name.

a) Enrichment of PPI					
TF	Cell	Odds Ratio	p.value	Enriched?	family
ATF3	A549	6.5323844	9.20782E-30	TRUE	bZIP, CH
ATF3	H1hesc	6.2557296	2.19565E-25	TRUE	
ATF4	Hepg2	5.9925038	1.64231E-26	TRUE	
ATF4	K562	4.6587484	2.75264E-20	TRUE	
BHLHE40	A549	3.5067545	1.59104E-07	TRUE	bHLH, bZIP
BHLHE40	Gm12878	2.1649943	0.000643582	TRUE	
BHLHE40	Hepg2	3.391001	1.00576E-10	TRUE	
BHLHE40	K562	4.1999105	1.24572E-11	TRUE	
CEBPB	A549	5.4245242	2.86441E-14	TRUE	bzip
CEBPB	Gm12878	3.7084298	2.46229E-20	TRUE	
CEBPB	H1hesc	3.5505948	1.09844E-10	TRUE	
CEBPB	Helas3	3.4360464	1.04204E-14	TRUE	

CEBPB	Hepg2	3.7259682	2.22576E-08	TRUE	
CEBPB	Imr90	6.5404738	1.62673E-15	TRUE	
CEBPB	K562	3.6878416	1.22792E-18	TRUE	
CTCF	A549	2.9588052	0.01611419	TRUE	CH
CTCF	Gm12878	0.7890216	0.8102347	FALSE	
CTCF	H1hesc	2.3012978	0.0134669	TRUE	
CTCF	Hct116	2.9588053	0.001355871	TRUE	
CTCF	Hek293	1.5341975	0.1653797	FALSE	
CTCF	Helas3	4.8418404	3.44514E-06	TRUE	
CTCF	Hepg2	2.0583429	0.01744166	TRUE	
CTCF	Huvec	2.2760095	0.0315983	TRUE	
CTCF	Imr90	0.3945106	0.3094342	FALSE	
CTCF	K562	2.9588053	0.000125138	TRUE	
CTCF	Mcf7	1.0958909	0.7238471	FALSE	
CTCF	Sknshra	3.2277901	0.002031186	TRUE	
EP300	A549	9.4141739	5.07775E-16	TRUE	no entry in the file
EP300	Gm12878	12.7411179	2.95466E-28	TRUE	
EP300	H1hesc	7.8261245	7.4611E-17	TRUE	
EP300	Helas3	11.063288	8.20635E-21	TRUE	
EP300	Hepg2	9.2988925	4.23388E-15	TRUE	
EP300	Sknsh	7.4823913	3.5984E-17	TRUE	
EP300	T47d	5.5695443	5.59584E-11	TRUE	
FOS	Gm12878	9.90836	1.33966E-62	TRUE	bzip
FOS	Helas3	5.0831812	4.3867E-38	TRUE	
FOS	Huvec	14.5244274	2.87608E-69	TRUE	
FOS	K562	11.5441056	7.47071E-56	TRUE	
FOS	Mcf10a	4.5729256	1.09807E-22	TRUE	
GABPA	A549	2.0928447	0.001098074	TRUE	Ets
GABPA	Gm12878	1.1534148	0.5484333	FALSE	
GABPA	H1hesc	1.3214139	0.2524994	FALSE	
GABPA	Helas3	1.4937685	0.07821976	FALSE	
GABPA	Hepg2	0.9862751	1	FALSE	
GABPA	K562	1.4892028	0.03763997	TRUE	
JUN	Gm12878	8.5395417	1.19383E-95	TRUE	bzip
JUN	H1hesc	7.8495633	4.82001E-50	TRUE	
JUN	Helas3	6.8421842	6.50128E-27	TRUE	
JUN	Hepg2	4.8762612	1.11297E-28	TRUE	
JUN	Huvec	13.1447941	1.87963E-59	TRUE	
JUN	K562	18.2009724	5.42681E-90	TRUE	
JUND	Gm12878	7.6921435	6.47957E-40	TRUE	bzip
JUND	H1hesc	5.6358399	2.29014E-25	TRUE	

JUND	Helas3	8.1933641	3.64769E-29	TRUE	
JUND	Hepg2	7.1503854	2.14264E-42	TRUE	
JUND	K562	15.7798296	9.24009E-68	TRUE	
JUND	Sknsh	8.5802981	3.80358E-47	TRUE	
MAFK	H1hesc	5.2215017	1.21683E-05	TRUE	bzip
MAFK	Helas3	3.3533207	0.000183661	TRUE	
MAFK	Hepg2	2.4040323	0.01028812	TRUE	
MAFK	Imr90	4.0970453	1.74868E-05	TRUE	
MAFK	K562	5.1458426	5.15881E-07	TRUE	
MAZ	Gm12878	1.1623726	0.6034424	FALSE	CH
MAZ	Helas3	1.3038863	0.4001867	FALSE	
MAZ	Hepg2	2.1518754	0.004896951	TRUE	
MAZ	K562	1.7259724	0.07220691	FALSE	
MXI1	Gm12878	8.2080498	4.64414E-18	TRUE	bhlh-bzip
MXI1	H1hesc	9.8624885	2.27958E-16	TRUE	
MXI1	Helas3	7.9657566	9.68602E-15	TRUE	
MXI1	Hepg2	11.1776344	7.87708E-17	TRUE	
MXI1	K562	10.8488884	4.39108E-21	TRUE	
MYC	A549	3.0734878	1.41831E-22	TRUE	bhlh-bzip
MYC	Gm12878	3.0073102	2.30521E-20	TRUE	
MYC	H1hesc	2.6073074	1.5755E-15	TRUE	
MYC	Helas3	5.3259113	1.39955E-62	TRUE	
MYC	Hepg2	2.7020927	1.7356E-22	TRUE	
MYC	Huvec	3.5113022	3.03334E-28	TRUE	
MYC	K562	2.7836745	8.59437E-22	TRUE	
MYC	Mcf7	3.0276147	1.69524E-28	TRUE	
NRF1	Gm12878	0	0.00712004	TRUE	bzip
NRF1	H1hesc	0.1793236	0.05226093	FALSE	
NRF1	Helas3	0.268985	0.238341	FALSE	
NRF1	Hepg2	0	0.1573226	FALSE	
NRF1	K562	0.2276028	0.1663903	FALSE	
REST	A549	5.5888804	1.12679E-21	TRUE	CH
REST	Gm12878	0.6961964	0.4147254	FALSE	
REST	H1hesc	0.6575191	0.3386072	FALSE	
REST	Helas3	1.4794056	0.153898	FALSE	
REST	Hepg2	0	7.45439E-06	TRUE	
REST	K562	5.1286931	6.68219E-16	TRUE	
REST	Panc1	1.1835122	0.6138163	FALSE	
REST	Pfsk1	1.1271381	0.6689403	FALSE	
REST	Sknsh	1.8081631	0.001171466	TRUE	
REST	U87	2.1917188	8.68434E-05	TRUE	

RFX5	Gm12878	2.5794292	0.001964828	TRUE	fork
RFX5	H1hesc	3.1701497	0.000674995	TRUE	
RFX5	Helas3	4.0813381	6.30748E-06	TRUE	
RFX5	Hepg2	3.7259683	7.08709E-05	TRUE	
RFX5	K562	2.5979168	0.00136052	TRUE	
SRF	Gm12878	4.4384708	1.75931E-30	TRUE	Mads
SRF	H1hesc	4.5088989	7.26269E-33	TRUE	
SRF	Hepg2	4.6934709	3.70921E-48	TRUE	
SRF	K562	4.8219187	2.92653E-47	TRUE	
TBP	Gm12878	4.6934721	5.62827E-25	TRUE	tata
TBP	H1hesc	5.9176032	2.00955E-32	TRUE	
TBP	Helas3	5.0410418	1.6196E-26	TRUE	
TBP	Hepg2	4.1425978	1.6851E-20	TRUE	
TBP	K562	6.8720101	6.93555E-51	TRUE	
TCF12	A549	2.3605242	1.16708E-08	TRUE	bhlh
TCF12	Gm12878	3.8987878	2.04701E-13	TRUE	
TCF12	H1hesc	2.9588053	9.84801E-09	TRUE	
TCF12	Hepg2	1.9144307	1.29886E-05	TRUE	
TCF7L2	Hct116	1.2552501	0.4264807	FALSE	HMG
TCF7L2	Hek293	2.2330681	4.71099E-05	TRUE	
TCF7L2	Helas3	1.6907476	0.02149591	TRUE	
TCF7L2	Hepg2	3.8291406	1.09743E-10	TRUE	
TCF7L2	Mcf7	1.6610845	0.0133504	TRUE	
TCF7L2	Panc1	1.3398418	0.2107984	FALSE	
USF1	A549	3.7327137	2.92222E-26	TRUE	bhlh-bzip
USF1	Gm12878	3.4404862	2.82163E-15	TRUE	
USF1	H1hesc	6.5900213	2.9723E-34	TRUE	
USF1	Hepg2	2.7024762	9.33881E-12	TRUE	
USF1	K562	4.1737673	8.45784E-17	TRUE	
YY1	A549	4.3741677	9.04113E-18	TRUE	CH
YY1	Gm12878	2.6299531	2.52513E-07	TRUE	
YY1	H1hesc	3.1145323	7.04388E-08	TRUE	
YY1	Hct116	4.0685759	4.99682E-21	TRUE	
YY1	Hepg2	4.19195	4.28305E-17	TRUE	
YY1	K562	4.6028049	1.39009E-15	TRUE	
YY1	Nt2d1	5.7135667	3.44697E-35	TRUE	
YY1	Sknshra	3.7259682	3.07166E-15	TRUE	
ZNF143	Gm12878	1.2989938	0.3184913	FALSE	CH
ZNF143	H1hesc	0.8453854	1	FALSE	
ZNF143	Helas3	2.0998265	0.004612317	TRUE	
ZNF143	K562	0.8218973	0.6401558	FALSE	

b) Enrichment of heterodimerizing motifs

TF	Cell	Odds Ratio	p.value	Enriched?	family
ATF3	A549	1.5951375	5.10E-02	FALSE	bZIP, CH
ATF3	H1hesc	1.6686604	5.49E-02	FALSE	
ATF4	Hepg2	2.6046432	1.55E-04	TRUE	
ATF4	K562	2.574259	7.49E-05	TRUE	
BHLHE40	A549	1.6601464	6.22E-02	FALSE	bHLH, bZIP
BHLHE40	Gm12878	1.795937	1.56E-02	TRUE	
BHLHE40	Hepg2	1.8261008	6.33E-03	TRUE	
BHLHE40	K562	3.698435	2.15E-07	TRUE	
CEBPB	A549	2.4683057	1.21E-02	TRUE	bzip
CEBPB	Gm12878	1.8497733	7.95E-03	TRUE	
CEBPB	H1hesc	1.1999726	6.26E-01	FALSE	
CEBPB	Helas3	1.590322	9.26E-02	FALSE	
CEBPB	Hepg2	0.8090211	6.94E-01	FALSE	
CEBPB	Imr90	2.654741	8.91E-03	TRUE	
CEBPB	K562	1.5135991	1.10E-01	FALSE	
CTCF	A549	1.9013647	2.17E-01	FALSE	CH
CTCF	Gm12878	2.3232646	4.39E-02	TRUE	
CTCF	H1hesc	2.1023069	2.32E-02	TRUE	
CTCF	Hct116	2.941998	2.24E-03	TRUE	
CTCF	Hek293	1.8935869	3.81E-02	TRUE	
CTCF	Helas3	2.1782578	4.55E-02	TRUE	
CTCF	Hepg2	1.4721854	2.05E-01	FALSE	
CTCF	Huvec	3.7149106	1.81E-03	TRUE	
CTCF	Imr90	2.956021	8.02E-03	TRUE	
CTCF	K562	1.7242666	6.79E-02	FALSE	
CTCF	Mcf7	2.2305047	6.94E-03	TRUE	
CTCF	Sknshra	2.3232743	3.55E-02	TRUE	
EP300	A549	1.6463524	3.50E-02	TRUE	no entry in the file
EP300	Gm12878	2.2847931	2.90E-05	TRUE	
EP300	H1hesc	0.8243637	4.37E-01	FALSE	
EP300	Helas3	1.9366118	3.86E-03	TRUE	
EP300	Hepg2	1.9365985	6.99E-03	TRUE	
EP300	Sknsh	1.0173679	9.15E-01	FALSE	
EP300	T47d	1.2215659	4.03E-01	FALSE	
FOS	Gm12878	2.2128889	7.53E-04	TRUE	bzip
FOS	Helas3	1.487691	1.05E-01	FALSE	
FOS	Huvec	1.190842	5.69E-01	FALSE	
FOS	K562	2.8588065	1.25E-04	TRUE	
FOS	Mcf10a	1.843203	4.38E-02	TRUE	

GABPA	A549	1.4088713	2.83E-01	FALSE	Ets
GABPA	Gm12878	1.8347953	2.05E-02	TRUE	
GABPA	H1hesc	2.4092807	2.41E-03	TRUE	
GABPA	Helas3	2.7530465	2.35E-04	TRUE	
GABPA	Hepg2	2.5409111	3.31E-04	TRUE	
GABPA	K562	1.8882579	5.48E-03	TRUE	
JUN	Gm12878	0.8799802	6.50E-01	FALSE	bzip
JUN	H1hesc	1.1956941	5.42E-01	FALSE	
JUN	Helas3	1.6269539	1.85E-01	FALSE	
JUN	Hepg2	1.9918216	2.00E-02	TRUE	
JUN	Huvec	1.5159185	2.14E-01	FALSE	
JUN	K562	1.7429763	7.94E-02	FALSE	
JUND	Gm12878	1.1378199	6.86E-01	FALSE	bzip
JUND	H1hesc	1.3561083	3.01E-01	FALSE	
JUND	Helas3	2.2224252	1.23E-02	TRUE	
JUND	Hepg2	1.384935	1.95E-01	FALSE	
JUND	K562	1.590322	9.26E-02	FALSE	
JUND	Sknsh	1.5331356	9.67E-02	FALSE	
MAFK	H1hesc	1.9918027	4.96E-02	TRUE	bzip
MAFK	Helas3	1.2848257	3.83E-01	FALSE	
MAFK	Hepg2	1.3459269	3.12E-01	FALSE	
MAFK	Imr90	1.9365709	2.35E-02	TRUE	
MAFK	K562	1.815698	3.79E-02	TRUE	
MAZ	Gm12878	1.328489	2.09E-01	FALSE	CH
MAZ	Helas3	1.6869557	1.53E-02	TRUE	
MAZ	Hepg2	1.2115931	3.79E-01	FALSE	
MAZ	K562	1.2609199	3.41E-01	FALSE	
MXI1	Gm12878	1.7708446	6.73E-03	TRUE	bhlh-bzip
MXI1	H1hesc	1.7198	2.12E-02	TRUE	
MXI1	Helas3	1.9918752	2.09E-03	TRUE	
MXI1	Hepg2	2.6897975	2.20E-05	TRUE	
MXI1	K562	2.3234271	1.27E-04	TRUE	
MYC	A549	1.9541857	5.72E-03	TRUE	bhlh-bzip
MYC	Gm12878	2.001052	8.98E-03	TRUE	
MYC	H1hesc	1.2976181	3.38E-01	FALSE	
MYC	Helas3	1.9283793	5.01E-03	TRUE	
MYC	Hepg2	1.328489	2.09E-01	FALSE	
MYC	Huvec	1.8705538	1.28E-02	TRUE	
MYC	K562	1.9455965	5.41E-03	TRUE	
MYC	Mcf7	2.1300222	9.16E-04	TRUE	
NRF1	Gm12878	1.2992	3.62E-01	FALSE	bzip

NRF1	H1hesc	2.2340788	7.57E-03	TRUE	
NRF1	Helas3	1.9137884	7.74E-02	FALSE	
NRF1	Hepg2	2.9033042	9.47E-03	TRUE	
NRF1	K562	1.0231639	1.00E+00	FALSE	
REST	A549	1.6269698	1.18E-01	FALSE	CH
REST	Gm12878	0.9513565	1.00E+00	FALSE	
REST	H1hesc	1.9917834	8.66E-02	FALSE	
REST	Helas3	2.1447785	7.59E-02	FALSE	
REST	Hepg2	1.8347598	8.94E-02	FALSE	
REST	K562	2.4522036	1.00E-02	TRUE	
REST	Panc1	1.7770828	1.57E-01	FALSE	
REST	Pfsk1	1.5902947	2.41E-01	FALSE	
REST	Sknsh	1.5847777	6.76E-02	FALSE	
REST	U87	2.0747031	1.38E-02	TRUE	
RFX5	Gm12878	1.9056695	4.40E-03	TRUE	fork
RFX5	H1hesc	1.6850355	4.31E-02	TRUE	
RFX5	Helas3	1.5873465	7.90E-02	FALSE	
RFX5	Hepg2	1.2109026	5.09E-01	FALSE	
RFX5	K562	1.4971482	6.63E-02	FALSE	
SRF	Gm12878	1.4996472	2.11E-01	FALSE	Mads
SRF	H1hesc	1.5292215	1.21E-01	FALSE	
SRF	Hepg2	1.2485285	3.85E-01	FALSE	
SRF	K562	1.9271617	9.18E-03	TRUE	
TBP	Gm12878	0.9597338	9.04E-01	FALSE	tata
TBP	H1hesc	1.0569549	8.16E-01	FALSE	
TBP	Helas3	1.8188756	1.26E-02	TRUE	
TBP	Hepg2	0.855068	5.55E-01	FALSE	
TBP	K562	1.3813758	1.28E-01	FALSE	
TCF12	A549	1.8813786	1.95E-03	TRUE	bhlh
TCF12	Gm12878	1.8215875	2.69E-02	TRUE	
TCF12	H1hesc	0.7752908	4.13E-01	FALSE	
TCF12	Hepg2	1.7929622	2.92E-03	TRUE	
TCF7L2	Hct116	1.925531	1.69E-02	TRUE	HMG
TCF7L2	Hek293	0.54743	1.75E-02	FALSE	
TCF7L2	Helas3	1.9056695	4.40E-03	TRUE	
TCF7L2	Hepg2	0.750313	3.15E-01	FALSE	
TCF7L2	Mcf7	1.8755501	3.55E-03	TRUE	
TCF7L2	Panc1	1.1625291	5.50E-01	FALSE	
USF1	A549	2.3233887	5.39E-04	TRUE	bhlh-bzip
USF1	Gm12878	2.3233307	3.40E-03	TRUE	
USF1	H1hesc	3.8246313	2.04E-05	TRUE	

USF1	Hepg2	1.247473	4.01E-01	FALSE	
USF1	K562	2.5442986	4.00E-03	TRUE	
YY1	A549	1.1120383	6.84E-01	FALSE	CH
YY1	Gm12878	1.2152985	4.93E-01	FALSE	
YY1	H1hesc	1.4850919	1.49E-01	FALSE	
YY1	Hct116	1.394823	1.49E-01	FALSE	
YY1	Hepg2	1.1397586	6.06E-01	FALSE	
YY1	K562	1.1072289	7.76E-01	FALSE	
YY1	Nt2d1	1.5198428	8.75E-02	FALSE	
YY1	Sknshra	1.0285555	8.97E-01	FALSE	
ZNF143	Gm12878	1.4390338	1.74E-01	FALSE	CH
ZNF143	H1hesc	1.7429584	1.50E-01	FALSE	
ZNF143	Helas3	1.3709941	2.56E-01	FALSE	
ZNF143	K562	1.4740879	1.34E-01	FALSE	

c) Enrichment of same family TFs

TF	Cell	Odds Ratio	p.value	Enriched?	family
ATF3	A549	2.10E+00	8.75E-04	TRUE	bZIP, CH
ATF3	H1hesc	2.01E+00	5.49E-03	TRUE	
ATF4	Hepg2	2.19E+00	1.28E-03	TRUE	
ATF4	K562	1.72E+00	2.48E-02	TRUE	
BHLHE40	A549	3.21E+00	6.76E-04	TRUE	bHLH, bZIP
BHLHE40	Gm12878	3.55E+00	2.09E-05	TRUE	
BHLHE40	Hepg2	2.20E+00	1.06E-02	TRUE	
BHLHE40	K562	4.08E+00	3.74E-06	TRUE	
CEBPB	A549	8.08E+00	2.02E-10	TRUE	bzip
CEBPB	Gm12878	2.78E+00	1.19E-03	TRUE	
CEBPB	H1hesc	4.58E+00	6.17E-06	TRUE	
CEBPB	Helas3	5.70E+00	2.75E-10	TRUE	
CEBPB	Hepg2	5.01E+00	4.37E-05	TRUE	
CEBPB	Imr90	8.47E+00	2.37E-10	TRUE	
CEBPB	K562	5.03E+00	7.00E-09	TRUE	
CTCF	A549	1.06E+00	1.00E+00	FALSE	CH
CTCF	Gm12878	1.14E+00	7.74E-01	FALSE	
CTCF	H1hesc	9.04E-01	1.00E+00	FALSE	
CTCF	Hct116	8.48E-01	1.00E+00	FALSE	
CTCF	Hek293	9.08E-01	1.00E+00	FALSE	
CTCF	Helas3	2.75E-01	2.37E-01	FALSE	
CTCF	Hepg2	1.13E+00	8.26E-01	FALSE	
CTCF	Huvec	8.90E-01	1.00E+00	FALSE	
CTCF	Imr90	0.2969291	3.47E-01	FALSE	
CTCF	K562	0.9270318	1.00E+00	FALSE	

CTCF	Mcf7	0.7569474	6.67E-01	FALSE	
CTCF	Sknshra	1.3237119	5.82E-01	FALSE	
EP300	A549	0	1.00E+00	FALSE	no entry in the file
EP300	Gm12878	0	1.00E+00	FALSE	
EP300	H1hesc	0	1.00E+00	FALSE	
EP300	Helas3	0	1.00E+00	FALSE	
EP300	Hepg2	0	1.00E+00	FALSE	
EP300	Sknsh	0	1.00E+00	FALSE	
EP300	T47d	0	1.00E+00	FALSE	
FOS	Gm12878	2.1512599	2.67E-02	TRUE	bzip
FOS	Helas3	2.845051	6.38E-04	TRUE	
FOS	Huvec	1.6602691	1.80E-01	FALSE	
FOS	K562	2.2192407	5.39E-02	FALSE	
FOS	Mcf10a	4.3717434	6.28E-06	TRUE	
GABPA	A549	8.7215588	7.42E-06	TRUE	Ets
GABPA	Gm12878	5.6019798	3.37E-04	TRUE	
GABPA	H1hesc	8.3535803	4.38E-06	TRUE	
GABPA	Helas3	7.1992541	3.05E-05	TRUE	
GABPA	Hepg2	6.2222022	8.73E-05	TRUE	
GABPA	K562	7.011492	2.31E-06	TRUE	
JUN	Gm12878	1.3328249	4.48E-01	FALSE	bzip
JUN	H1hesc	4.6911107	1.69E-06	TRUE	
JUN	Helas3	6.6759432	4.78E-08	TRUE	
JUN	Hepg2	6.0934976	1.60E-09	TRUE	
JUN	Huvec	3.2357318	4.14E-03	TRUE	
JUN	K562	5.0886519	6.38E-07	TRUE	
JUND	Gm12878	3.0795399	5.04E-04	TRUE	bzip
JUND	H1hesc	3.9893934	2.69E-05	TRUE	
JUND	Helas3	7.7900861	1.84E-12	TRUE	
JUND	Hepg2	4.0485029	2.68E-06	TRUE	
JUND	K562	4.2593085	3.35E-06	TRUE	
JUND	Sknsh	3.3264881	8.72E-05	TRUE	
MAFK	H1hesc	7.8564785	6.38E-11	TRUE	bzip
MAFK	Helas3	6.0820785	2.73E-10	TRUE	
MAFK	Hepg2	5.0687937	4.60E-08	TRUE	
MAFK	Imr90	6.8608014	9.36E-12	TRUE	
MAFK	K562	6.5200645	4.42E-11	TRUE	
MAZ	Gm12878	2.1054598	3.71E-03	TRUE	CH
MAZ	Helas3	1.9264231	1.01E-02	TRUE	
MAZ	Hepg2	2.0538421	4.60E-03	TRUE	
MAZ	K562	2.2393291	2.51E-03	TRUE	

MXI1	Gm12878	1.8431646	5.07E-02	FALSE	bhlh-bzip
MXI1	H1hesc	3.2550065	7.71E-05	TRUE	
MXI1	Helas3	1.383034	4.22E-01	FALSE	
MXI1	Hepg2	2.2322579	1.60E-02	TRUE	
MXI1	K562	2.7028305	9.70E-04	TRUE	
MYC	A549	1.8246552	9.23E-02	FALSE	bhlh-bzip
MYC	Gm12878	3.4215318	9.17E-05	TRUE	
MYC	H1hesc	2.1630139	3.07E-02	TRUE	
MYC	Helas3	1.7565846	1.09E-01	FALSE	
MYC	Hepg2	1.5556397	1.83E-01	FALSE	
MYC	Huvec	2.4543355	7.81E-03	TRUE	
MYC	K562	1.7774064	1.41E-01	FALSE	
MYC	Mcf7	2.7627523	8.51E-04	TRUE	
NRF1	Gm12878	1.6528658	2.33E-01	FALSE	bzip
NRF1	H1hesc	1.8145005	1.41E-01	FALSE	
NRF1	Helas3	2.6418284	2.38E-02	TRUE	
NRF1	Hepg2	0.3996188	7.23E-01	FALSE	
NRF1	K562	2.7372159	1.47E-02	TRUE	
REST	A549	0.3029818	1.08E-01	FALSE	CH
REST	Gm12878	0.5935903	7.60E-01	FALSE	
REST	H1hesc	0.322736	3.45E-01	FALSE	
REST	Helas3	0.3092952	3.45E-01	FALSE	
REST	Hepg2	0.5496559	5.64E-01	FALSE	
REST	K562	0	4.08E-02	FALSE	
REST	Panc1	0.8899884	1.00E+00	FALSE	
REST	Pfsk1	0.2651286	2.38E-01	FALSE	
REST	Sknsh	0.7113629	4.68E-01	FALSE	
REST	U87	0.9794592	1.00E+00	FALSE	
RFX5	Gm12878	0.904439	1.00E+00	FALSE	fork
RFX5	H1hesc	0.990631	1.00E+00	FALSE	
RFX5	Helas3	1.9201193	1.28E-01	FALSE	
RFX5	Hepg2	2.2964117	3.09E-02	TRUE	
RFX5	K562	0.7237508	8.10E-01	FALSE	
SRF	Gm12878	0	6.36E-01	FALSE	Mads
SRF	H1hesc	0	3.93E-01	FALSE	
SRF	Hepg2	0	2.44E-01	FALSE	
SRF	K562	0	2.43E-01	FALSE	
TBP	Gm12878	0	1.00E+00	FALSE	tata
TBP	H1hesc	0	1.00E+00	FALSE	
TBP	Helas3	0	1.00E+00	FALSE	
TBP	Hepg2	0	1.00E+00	FALSE	

TBP	K562	0	1.00E+00	FALSE	
TCF12	A549	1.2660094	4.75E-01	FALSE	bhlh
TCF12	Gm12878	1.7228047	1.70E-01	FALSE	
TCF12	H1hesc	3.0128669	1.22E-03	TRUE	
TCF12	Hepg2	1.1705098	5.91E-01	FALSE	
TCF7L2	Hct116	4.9744678	4.65E-03	TRUE	HMG
TCF7L2	Hek293	3.0706106	3.23E-02	TRUE	
TCF7L2	Helas3	4.2184079	5.70E-03	TRUE	
TCF7L2	Hepg2	6.5642464	7.65E-05	TRUE	
TCF7L2	Mcf7	3.199521	2.78E-02	TRUE	
TCF7L2	Panc1	3.4528202	2.08E-02	TRUE	
USF1	A549	1.9642107	5.96E-02	FALSE	bhlh-bzip
USF1	Gm12878	3.5155867	3.14E-04	TRUE	
USF1	H1hesc	4.1389737	7.46E-05	TRUE	
USF1	Hepg2	2.6493886	6.56E-03	TRUE	
USF1	K562	2.7620515	1.01E-02	TRUE	
YY1	A549	1.2153855	5.68E-01	FALSE	CH
YY1	Gm12878	0.7418214	6.88E-01	FALSE	
YY1	H1hesc	0.8558185	8.35E-01	FALSE	
YY1	Hct116	0.7854063	6.10E-01	FALSE	
YY1	Hepg2	0.9961475	1.00E+00	FALSE	
YY1	K562	1.1706367	6.92E-01	FALSE	
YY1	Nt2d1	0.542949	2.12E-01	FALSE	
YY1	Sknshra	0.6744454	4.44E-01	FALSE	
ZNF143	Gm12878	0.7636152	7.05E-01	FALSE	CH
ZNF143	H1hesc	0.5496559	5.64E-01	FALSE	
ZNF143	Helas3	0.717915	5.56E-01	FALSE	
ZNF143	K562	0.5983648	4.22E-01	FALSE	

Table 7.12a-c. Hypergeometric test results: a) identified co-factors are enriched for heterodimerizing TFs, b) identified co-factors are enriched for same family as that of the reference TF

TF	Cell	Category	Term	Fold E
ATF3	A549	INTERPRO	bZIP transcription factor, bZIP-1	7.26
ATF3		SMART	BRLZ	4.99
ATF3		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.81
ATF3		INTERPRO	Kelch related	8.82
ATF3		SMART	FH	4.22
ATF3		INTERPRO	Transcription factor, fork head, conserved site	4.07
ATF3		INTERPRO	Transcription factor, fork head	4.07
ATF3	H1hesc	INTERPRO	Helix-loop-helix DNA-binding	5.05
ATF3		SMART	HLH	4.65

ATF3		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	4.43
ATF3		SMART	BRLZ	3.61
ATF3		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.44
ATF3		INTERPRO	bZIP transcription factor, bZIP-1	4.21
ATF3		INTERPRO	Basic leucine zipper	3.98
ATF3	Hepg2	SMART	HLH	4.68
ATF3		INTERPRO	Helix-loop-helix DNA-binding	5.05
ATF3		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	4.62
ATF3		SMART	BRLZ	2.93
ATF3		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.89
ATF3		INTERPRO	bZIP transcription factor, bZIP-1	3.86
ATF3	K562	SMART	HLH	4.91
ATF3		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	4.72
ATF3		INTERPRO	Helix-loop-helix DNA-binding	4.76
ATF3		SMART	BRLZ	2.84
ATF3		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.73
ATF3		INTERPRO	bZIP transcription factor, bZIP-1	3.64
ATF3		SMART	ZnF_GATA	5.85
ATF3		INTERPRO	Zinc finger, GATA-type	5.62
BHLHE40	A549	INTERPRO	bZIP transcription factor, bZIP-1	5.85
BHLHE40		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.62
BHLHE40		SMART	BRLZ	3.42
BHLHE40		INTERPRO	Fos transforming protein	7.96
BHLHE40	Gm12878	SMART	ETS	5.39
BHLHE40		INTERPRO	Ets	5.39
BHLHE40		INTERPRO	Winged helix repressor DNA-binding	2.59
BHLHE40		SMART	IRF	5.39
BHLHE40		INTERPRO	Interferon regulatory factor, conserved site	5.39
BHLHE40		INTERPRO	Interferon regulatory factor	5.39
BHLHE40		INTERPRO	bZIP transcription factor, bZIP-1	3.62
BHLHE40		SMART	BRLZ	2.61
BHLHE40		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.61
BHLHE40		SMART	HLH	2.58
BHLHE40		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	2.58
BHLHE40		SMART	SAM_PNT	4.62
BHLHE40		INTERPRO	Sterile alpha motif/pointed	4.62
BHLHE40		INTERPRO	Sterile alpha motif-type	4.11
BHLHE40		INTERPRO	Interferon regulatory factor-3	5.13
BHLHE40		INTERPRO	SMAD domain-like	5.13
BHLHE40	Hepg2	INTERPRO	bZIP transcription factor, bZIP-1	4.77
BHLHE40		INTERPRO	High mobility group, HMG1/HMG2	5.35
BHLHE40		SMART	HMG	4.99

BHLHE40		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.16
BHLHE40		SMART	BRLZ	2.95
BHLHE40		INTERPRO	Zinc finger, NHR/GATA-type	2.44
BHLHE40		SMART	ZnF_GATA	5.41
BHLHE40		INTERPRO	Fos transforming protein	5.79
BHLHE40		INTERPRO	Zinc finger, GATA-type	5.79
BHLHE40	K562	INTERPRO	bZIP transcription factor, bZIP-1	5.84
BHLHE40		SMART	BRLZ	3.99
BHLHE40		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.83
BHLHE40		SMART	ZnF_GATA	9.39
BHLHE40		INTERPRO	Fos transforming protein	9.02
BHLHE40		INTERPRO	Zinc finger, GATA-type	9.02
BHLHE40		SMART	HLH	2.73
CEBPB	A549	SMART	BRLZ	8.65
CEBPB		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	8.71
CEBPB		INTERPRO	bZIP transcription factor, bZIP-1	12.68
CEBPB		INTERPRO	Fos transforming protein	14.37
CEBPB		INTERPRO	Basic leucine zipper	7.19
CEBPB	Gm12878	INTERPRO	bZIP transcription factor, bZIP-1	5.07
CEBPB		SMART	ETS	5.00
CEBPB		INTERPRO	Ets	5.05
CEBPB		SMART	BRLZ	3.07
CEBPB		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.10
CEBPB		INTERPRO	Winged helix repressor DNA-binding	2.46
CEBPB		SMART	IRF	5.33
CEBPB		INTERPRO	Interferon regulatory factor, conserved site	5.39
CEBPB		INTERPRO	Interferon regulatory factor	5.39
CEBPB		SMART	SAM_PNT	4.66
CEBPB		INTERPRO	Sterile alpha motif/pointed	4.72
CEBPB		INTERPRO	Interferon regulatory factor-3	5.39
CEBPB		INTERPRO	SMAD domain-like	5.39
CEBPB		INTERPRO	Sterile alpha motif-type	4.19
CEBPB		INTERPRO	Fos transforming protein	5.39
CEBPB	H1hesc	SMART	BRLZ	7.33
CEBPB		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	6.86
CEBPB		INTERPRO	bZIP transcription factor, bZIP-1	8.88
CEBPB		INTERPRO	Basic leucine zipper	6.29
CEBPB		INTERPRO	Fos transforming protein	10.78
CEBPB	Helas3	SMART	BRLZ	5.92
CEBPB		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	6.15
CEBPB		INTERPRO	bZIP transcription factor, bZIP-1	8.82
CEBPB		INTERPRO	Fos transforming protein	8.82

CEBPB		INTERPRO	Basic leucine zipper	5.14
CEBPB		INTERPRO	Kelch related	8.82
CEBPB	Hepg2	SMART	BRLZ	6.39
CEBPB		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	6.65
CEBPB		INTERPRO	bZIP transcription factor, bZIP-1	6.95
CEBPB		INTERPRO	Basic leucine zipper	8.43
CEBPB	Imr90	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	10.29
CEBPB		SMART	BRLZ	9.47
CEBPB		INTERPRO	bZIP transcription factor, bZIP-1	14.26
CEBPB		INTERPRO	Basic leucine zipper	9.43
CEBPB		INTERPRO	Fos transforming protein	16.17
CEBPB		INTERPRO	Jun-like transcription factor	16.17
CEBPB		INTERPRO	Transcription factor Jun	16.17
CEBPB	K562	SMART	BRLZ	5.41
CEBPB		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	5.23
CEBPB		INTERPRO	bZIP transcription factor, bZIP-1	6.16
CEBPB		SMART	ZnF_GATA	6.38
CEBPB		SMART	POU	4.06
CEBPB		INTERPRO	Fos transforming protein	6.16
CEBPB		INTERPRO	Maf transcription factor	6.16
CEBPB		INTERPRO	Zinc finger, GATA-type	6.16
CEBPB		INTERPRO	POU-specific	3.92
CEBPB		INTERPRO	POU	3.92
CEBPB		INTERPRO	Basic leucine zipper	3.59
CEBPB		INTERPRO	DNA-binding RFX	5.13
CTCF	A549	INTERPRO	Helix-loop-helix DNA-binding	6.63
CTCF		SMART	HLH	5.76
CTCF		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	5.56
CTCF	Gm12878	INTERPRO	Helix-loop-helix DNA-binding	8.95
CTCF		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	7.87
CTCF		SMART	HLH	7.24
CTCF	H1hesc	SMART	ZnF_C4	3.53
CTCF		SMART	HOLI	3.43
CTCF		INTERPRO	Zinc finger, nuclear hormone receptor-type	3.56
CTCF		INTERPRO	Steroid hormone receptor	3.56
CTCF		INTERPRO	Nuclear hormone receptor, ligand-binding	3.46
CTCF		INTERPRO	Nuclear hormone receptor, ligand-binding, core	3.46
CTCF		INTERPRO	Helix-loop-helix DNA-binding	3.73
CTCF		SMART	HLH	3.10
CTCF		INTERPRO	Zinc finger, NHR/GATA-type	2.92
CTCF	Hct116	INTERPRO	Helix-loop-helix DNA-binding	8.64
CTCF		SMART	HLH	7.57

CTCF		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	7.58
CTCF	Hek293	SMART	HLH	5.62
CTCF		INTERPRO	Helix-loop-helix DNA-binding	6.09
CTCF		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	5.36
CTCF	Helas3	INTERPRO	Acute myeloid leukemia 1 protein (AML 1)/Runt	20.42
CTCF		INTERPRO	Runx inhibition	20.42
CTCF		INTERPRO	Acute myeloid leukemia 1 (AML 1)/Runt	20.42
CTCF		INTERPRO	Transcription factor, Runt-related, RUNX	20.42
CTCF	Hepg2	SMART	ZnF_C4	3.12
CTCF		SMART	HOLI	3.03
CTCF		INTERPRO	Zinc finger, nuclear hormone receptor-type	2.84
CTCF		INTERPRO	Nuclear hormone receptor, ligand-binding, core	2.76
CTCF		INTERPRO	Nuclear hormone receptor, ligand-binding	2.76
CTCF		INTERPRO	Steroid hormone receptor	2.58
CTCF	Huvec	SMART	ZnF_C4	4.17
CTCF		SMART	HOLI	4.06
CTCF		INTERPRO	Steroid hormone receptor	3.44
CTCF		INTERPRO	Zinc finger, nuclear hormone receptor-type	3.44
CTCF		INTERPRO	Nuclear hormone receptor, ligand-binding, core	3.34
CTCF		INTERPRO	Nuclear hormone receptor, ligand-binding	3.34
CTCF		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	8.92
CTCF	Imr90	SMART	HLH	6.91
CTCF		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	18.48
CTCF		INTERPRO	Helix-loop-helix DNA-binding	6.40
CTCF		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	5.36
CTCF		INTERPRO	E2F Family	16.63
CTCF	K562	INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	10.35
CTCF		INTERPRO	Transcription factor AP-2	15.52
CTCF		INTERPRO	Transcription factor AP-2, C-terminal	15.52
CTCF	Mcf7	SMART	HLH	7.78
CTCF		INTERPRO	Helix-loop-helix DNA-binding	7.62
CTCF		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	6.66
CTCF	Sknshra	INTERPRO	Helix-loop-helix DNA-binding	7.64
CTCF		SMART	HLH	6.67
CTCF		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	6.72
EP300	A549	SMART	BRLZ	5.18
EP300		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.51
EP300		INTERPRO	bZIP transcription factor, bZIP-1	5.71
EP300		INTERPRO	Basic leucine zipper	4.31
EP300		INTERPRO	Fos transforming protein	6.47

EP300		INTERPRO	DNA-binding RFX	5.39
EP300	Gm12878	INTERPRO	bZIP transcription factor, bZIP-1	4.17
EP300		SMART	BRLZ	2.49
EP300		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.53
EP300		INTERPRO	NF-kappa-B/Rel/dorsal	4.17
EP300		INTERPRO	Rel homology	4.17
EP300		SMART	IPT	3.28
EP300		INTERPRO	Cell surface receptor IPT/TIG	3.34
EP300		INTERPRO	Immunoglobulin-like fold	3.34
EP300		SMART	IRF	3.59
EP300		INTERPRO	Interferon regulatory factor, conserved site	3.65
EP300		INTERPRO	Interferon regulatory factor	3.65
EP300		SMART	ETS	2.56
EP300		INTERPRO	Ets	2.61
EP300	H1hesc	INTERPRO	bZIP transcription factor, bZIP-1	4.28
EP300		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.79
EP300		SMART	BRLZ	2.67
EP300		INTERPRO	High mobility group, HMG1/HMG2	3.73
EP300		SMART	HMG	3.57
EP300		INTERPRO	Fos transforming protein	4.85
EP300	Helas3	SMART	BRLZ	4.58
EP300		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.67
EP300		INTERPRO	bZIP transcription factor, bZIP-1	5.80
EP300		INTERPRO	Basic leucine zipper	5.13
EP300		SMART	ZnF_GATA	6.05
EP300		INTERPRO	Fos transforming protein	6.16
EP300		INTERPRO	Zinc finger, GATA-type	6.16
EP300	Hepg2	INTERPRO	Zinc finger, NHR/GATA-type	3.73
EP300		SMART	HOLI	3.10
EP300		INTERPRO	Steroid hormone receptor	3.20
EP300		INTERPRO	Nuclear hormone receptor, ligand-binding, core	3.11
EP300		INTERPRO	Nuclear hormone receptor, ligand-binding	3.11
EP300		SMART	ZnF_C4	2.98
EP300		INTERPRO	Vitamin D receptor	4.80
EP300		INTERPRO	Zinc finger, nuclear hormone receptor-type	2.98
EP300		SMART	ZnF_GATA	7.44
EP300		SMART	BRLZ	2.70
EP300		INTERPRO	Zinc finger, GATA-type	7.46
EP300		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.71
EP300		INTERPRO	bZIP transcription factor, bZIP-1	3.51
EP300	Sknsh	SMART	ZnF_GATA	5.17
EP300		INTERPRO	Zinc finger, GATA-type	5.04

EP300	T47d	SMART	BRLZ	3.93
EP300		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.71
EP300		INTERPRO	bZIP transcription factor, bZIP-1	4.50
EP300		INTERPRO	Basic leucine zipper	3.83
EP300		SMART	ZnF_GATA	5.41
EP300		INTERPRO	Zinc finger, GATA-type	5.11
EP300		INTERPRO	Fos transforming protein	5.11
FOS	Gm12878	SMART	HLH	4.83
FOS		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	4.79
FOS		INTERPRO	Helix-loop-helix DNA-binding	4.83
FOS	Helas3	SMART	BRLZ	3.25
FOS		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.94
FOS		SMART	ZnF_GATA	7.14
FOS		INTERPRO	Zinc finger, GATA-type	6.47
FOS		INTERPRO	DNA-binding RFX	5.39
FOS		INTERPRO	bZIP transcription factor, bZIP-1	3.04
FOS	Huvec	INTERPRO	Ets	6.81
FOS		SMART	ETS	6.38
FOS		INTERPRO	Winged helix repressor DNA-binding	2.87
FOS		INTERPRO	Sterile alpha motif/pointed	6.81
FOS		SMART	SAM_PNT	6.38
FOS		INTERPRO	Sterile alpha motif-type	6.05
FOS		INTERPRO	Interferon regulatory factor	5.96
FOS		INTERPRO	Interferon regulatory factor, conserved site	5.96
FOS		SMART	IRF	5.58
FOS		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.48
FOS		SMART	BRLZ	2.32
FOS		INTERPRO	SMAD domain-like	5.67
FOS		INTERPRO	Interferon regulatory factor-3	5.67
FOS		INTERPRO	bZIP transcription factor, bZIP-1	3.20
FOS	K562	SMART	BRLZ	4.06
FOS		INTERPRO	bZIP transcription factor, bZIP-1	5.27
FOS		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.62
FOS	Mcf10a	SMART	BRLZ	4.56
FOS		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.76
FOS		INTERPRO	Basic leucine zipper	5.17
FOS		INTERPRO	DNA-binding RFX	6.47
FOS		INTERPRO	Winged helix repressor DNA-binding	2.04
GABPA	A549	INTERPRO	bZIP transcription factor, bZIP-1	7.96
GABPA		SMART	BRLZ	4.84
GABPA		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.65
GABPA		SMART	SH2	9.39

GABPA		INTERPRO	STAT transcription factor, all-alpha	9.02
GABPA		INTERPRO	SH2 motif	9.02
GABPA		INTERPRO	STAT transcription factor, DNA-binding	9.02
GABPA		INTERPRO	STAT transcription factor, core	9.02
GABPA		INTERPRO	STAT transcription factor, protein interaction	9.02
GABPA		INTERPRO	STAT transcription factor, DNA-binding, subdomain	9.02
GABPA		INTERPRO	EF-Hand type	7.90
GABPA		INTERPRO	Fos transforming protein	9.02
GABPA		INTERPRO	E2F Family	7.22
GABPA	Gm12878	SMART	BRLZ	4.57
GABPA		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.56
GABPA		INTERPRO	bZIP transcription factor, bZIP-1	6.52
GABPA		INTERPRO	Fos transforming protein	7.92
GABPA	H1hesc	SMART	SH2	9.92
GABPA		INTERPRO	Winged helix repressor DNA-binding	2.69
GABPA		INTERPRO	SH2 motif	8.08
GABPA		INTERPRO	STAT transcription factor, protein interaction	8.08
GABPA		INTERPRO	STAT transcription factor, all-alpha	8.08
GABPA		INTERPRO	STAT transcription factor, DNA-binding, subdomain	8.08
GABPA		INTERPRO	STAT transcription factor, core	8.08
GABPA		INTERPRO	STAT transcription factor, DNA-binding	8.08
GABPA		SMART	ETS	5.58
GABPA		INTERPRO	EF-Hand type	7.07
GABPA		INTERPRO	Ets	4.55
GABPA		SMART	DWA	9.92
GABPA		INTERPRO	DNA-binding RFX	6.74
GABPA		SMART	SAM_PNT	6.20
GABPA		INTERPRO	CTF transcription factor/nuclear factor 1, N-terminal	8.08
GABPA		INTERPRO	CTF transcription factor/nuclear factor 1, conserved site	8.08
GABPA		INTERPRO	CTF transcription factor/nuclear factor 1	8.08
GABPA		INTERPRO	MAD homology 1, Dwarfin-type	8.08
GABPA		INTERPRO	Sterile alpha motif/pointed	5.05
GABPA	Helas3	INTERPRO	bZIP transcription factor, bZIP-1	7.99
GABPA		SMART	BRLZ	5.41
GABPA		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	5.00
GABPA		SMART	SH2	10.50
GABPA		INTERPRO	STAT transcription factor, DNA-binding, subdomain	9.70
GABPA		INTERPRO	STAT transcription factor, DNA-binding	9.70

GABPA		INTERPRO	STAT transcription factor, core	9.70
GABPA		INTERPRO	SH2 motif	9.70
GABPA		INTERPRO	STAT transcription factor, all-alpha	9.70
GABPA		INTERPRO	STAT transcription factor, protein interaction	9.70
GABPA		INTERPRO	EF-Hand type	8.49
GABPA		INTERPRO	Fos transforming protein	9.70
GABPA		INTERPRO	E2F Family	7.76
GABPA	Hepg2	SMART	SH2	10.82
GABPA		INTERPRO	STAT transcription factor, all-alpha	9.70
GABPA		INTERPRO	SH2 motif	9.70
GABPA		INTERPRO	STAT transcription factor, protein interaction	9.70
GABPA		INTERPRO	STAT transcription factor, DNA-binding, subdomain	9.70
GABPA		INTERPRO	STAT transcription factor, DNA-binding	9.70
GABPA		INTERPRO	STAT transcription factor, core	9.70
GABPA		INTERPRO	EF-Hand type	8.49
GABPA		SMART	BRLZ	3.28
GABPA		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.94
GABPA		INTERPRO	bZIP transcription factor, bZIP-1	3.99
GABPA		SMART	ETS	4.06
GABPA	K562	SMART	SH2	8.11
GABPA		INTERPRO	SH2 motif	7.46
GABPA		INTERPRO	STAT transcription factor, DNA-binding, subdomain	7.46
GABPA		INTERPRO	STAT transcription factor, core	7.46
GABPA		INTERPRO	STAT transcription factor, protein interaction	7.46
GABPA		INTERPRO	STAT transcription factor, all-alpha	7.46
GABPA		INTERPRO	STAT transcription factor, DNA-binding	7.46
GABPA		INTERPRO	EF-Hand type	6.53
GABPA		SMART	ZnF_GATA	8.11
GABPA		INTERPRO	Zinc finger, GATA-type	7.46
GABPA		SMART	ETS	3.55
GABPA		SMART	BRLZ	2.46
GABPA		INTERPRO	Winged helix repressor DNA-binding	1.96
JUN	Gm12878	SMART	ETS	4.81
JUN		INTERPRO	Ets	4.65
JUN		SMART	IRF	5.49
JUN		INTERPRO	Interferon regulatory factor	5.32
JUN		INTERPRO	Interferon regulatory factor, conserved site	5.32
JUN		INTERPRO	Winged helix repressor DNA-binding	2.14
JUN		SMART	SH2	5.49
JUN		INTERPRO	SH2 motif	5.32

JUN		INTERPRO	STAT transcription factor, DNA-binding	5.32
JUN		INTERPRO	STAT transcription factor, core	5.32
JUN		INTERPRO	STAT transcription factor, protein interaction	5.32
JUN		INTERPRO	STAT transcription factor, all-alpha	5.32
JUN		INTERPRO	STAT transcription factor, DNA-binding, subdomain	5.32
JUN		INTERPRO	EF-Hand type	4.65
JUN		INTERPRO	Interferon regulatory factor-3	5.32
JUN		INTERPRO	SMAD domain-like	5.32
JUN		SMART	SAM_PNT	4.12
JUN		INTERPRO	Maf transcription factor	5.32
JUN		INTERPRO	Sterile alpha motif/pointed	3.99
JUN	H1hesc	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	5.53
JUN		SMART	BRLZ	5.25
JUN		INTERPRO	bZIP transcription factor, bZIP-1	6.04
JUN		INTERPRO	Basic leucine zipper	6.66
JUN	Helas3	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	8.55
JUN		SMART	BRLZ	7.87
JUN		INTERPRO	Basic leucine zipper	11.76
JUN		INTERPRO	bZIP transcription factor, bZIP-1	7.26
JUN		INTERPRO	CCAAT/enhancer-binding	17.64
JUN	Hepg2	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	6.60
JUN		SMART	BRLZ	6.22
JUN		INTERPRO	bZIP transcription factor, bZIP-1	6.12
JUN		INTERPRO	Basic leucine zipper	6.31
JUN		INTERPRO	Kelch related	9.46
JUN		INTERPRO	Maf transcription factor	7.57
JUN	Huvec	SMART	ETS	8.37
JUN		INTERPRO	Ets	8.66
JUN		SMART	SAM_PNT	7.81
JUN		INTERPRO	Sterile alpha motif/pointed	8.08
JUN		INTERPRO	Winged helix repressor DNA-binding	2.76
JUN		INTERPRO	Sterile alpha motif-type	7.19
JUN		SMART	BRLZ	2.70
JUN		INTERPRO	Nuclear factor of activated T cells (NFAT)	2.80
JUN		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	9.24
JUN		INTERPRO	Nuclear factor of activated T cells (NFAT), subgroup	9.24
JUN		INTERPRO	bZIP transcription factor, bZIP-1	3.80
JUN	K562	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.78
JUN		SMART	BRLZ	4.54
JUN		SMART	ZnF_GATA	11.52
JUN		INTERPRO	Zinc finger, GATA-type	12.13

JUN		INTERPRO	bZIP transcription factor, bZIP-1	4.99
JUN		INTERPRO	Basic leucine zipper	5.05
JUND	Gm12878	SMART	ETS	6.04
JUND		INTERPRO	Ets	5.95
JUND		INTERPRO	Winged helix repressor DNA-binding	3.08
JUND		SMART	BRLZ	3.61
JUND		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.55
JUND		SMART	IRF	7.44
JUND		INTERPRO	Interferon regulatory factor	7.32
JUND		INTERPRO	Interferon regulatory factor, conserved site	7.32
JUND		INTERPRO	bZIP transcription factor, bZIP-1	4.31
JUND		INTERPRO	Interferon regulatory factor-3	7.32
JUND		INTERPRO	SMAD domain-like	7.32
JUND		SMART	SAM_PNT	4.65
JUND	H1hesc	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.55
JUND		SMART	BRLZ	3.35
JUND		INTERPRO	POU-specific	5.74
JUND		INTERPRO	POU	5.74
JUND		SMART	POU	5.41
JUND		INTERPRO	bZIP transcription factor, bZIP-1	4.25
JUND		INTERPRO	Helix-loop-helix DNA-binding	3.12
JUND	Helas3	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	8.33
JUND		SMART	BRLZ	7.66
JUND		INTERPRO	Basic leucine zipper	12.13
JUND		INTERPRO	bZIP transcription factor, bZIP-1	6.66
JUND		INTERPRO	CCAAT/enhancer-binding	16.17
JUND	Hepg2	SMART	HMG	7.01
JUND		INTERPRO	High mobility group, HMG1/HMG2	6.89
JUND		SMART	BRLZ	2.99
JUND		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.94
JUND		INTERPRO	Basic leucine zipper	4.35
JUND		INTERPRO	DNA-binding RFX	6.22
JUND	K562	SMART	BRLZ	3.61
JUND		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.53
JUND		SMART	ZnF_GATA	7.93
JUND		SMART	ETS	3.97
JUND		INTERPRO	Zinc finger, GATA-type	7.76
JUND		INTERPRO	Ets	3.88
JUND		INTERPRO	bZIP transcription factor, bZIP-1	3.65
JUND		SMART	SAM_PNT	4.96
JUND	Sknsh	SMART	ETS	5.12
JUND		SMART	BRLZ	3.55

JUND		INTERPRO	Ets	4.99
JUND		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.46
JUND		INTERPRO	bZIP transcription factor, bZIP-1	4.36
JUND		SMART	SH2	5.85
JUND		INTERPRO	STAT transcription factor, protein interaction	5.71
JUND		INTERPRO	STAT transcription factor, all-alpha	5.71
JUND		INTERPRO	SH2 motif	5.71
JUND		INTERPRO	STAT transcription factor, DNA-binding, subdomain	5.71
JUND		INTERPRO	STAT transcription factor, core	5.71
JUND		INTERPRO	STAT transcription factor, DNA-binding	5.71
JUND		INTERPRO	EF-Hand type	4.99
JUND		SMART	SAM_PNT	4.39
JUND		INTERPRO	Sterile alpha motif/pointed	4.28
JUND		INTERPRO	Sterile alpha motif-type	3.80
MAFK	H1hesc	SMART	BRLZ	7.46
MAFK		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	7.13
MAFK		INTERPRO	bZIP transcription factor, bZIP-1	8.99
MAFK		INTERPRO	Kelch related	11.76
MAFK		INTERPRO	Fos transforming protein	9.41
MAFK		INTERPRO	Maf transcription factor	9.41
MAFK		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	7.84
MAFK	Helas3	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	5.75
MAFK		SMART	BRLZ	5.41
MAFK		INTERPRO	bZIP transcription factor, bZIP-1	6.59
MAFK		SMART	POU	5.16
MAFK		INTERPRO	POU	5.49
MAFK		INTERPRO	POU-specific	5.49
MAFK		INTERPRO	Kelch related	8.62
MAFK		INTERPRO	Basic leucine zipper	4.31
MAFK	Hepg2	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	5.60
MAFK		SMART	BRLZ	5.28
MAFK		INTERPRO	bZIP transcription factor, bZIP-1	6.52
MAFK		SMART	POU	5.54
MAFK		INTERPRO	POU	5.88
MAFK		INTERPRO	POU-specific	5.88
MAFK		INTERPRO	Fos transforming protein	7.39
MAFK		INTERPRO	Maf transcription factor	7.39
MAFK	Imr90	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	7.35
MAFK		SMART	BRLZ	6.76
MAFK		INTERPRO	bZIP transcription factor, bZIP-1	7.99

MAFK		INTERPRO	Basic leucine zipper	6.47
MAFK		INTERPRO	Kelch related	9.70
MAFK		INTERPRO	Fos transforming protein	7.76
MAFK		INTERPRO	Maf transcription factor	7.76
MAFK	K562	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	6.19
MAFK		SMART	BRLZ	5.69
MAFK		INTERPRO	bZIP transcription factor, bZIP-1	7.81
MAFK		INTERPRO	Kelch related	10.21
MAFK		INTERPRO	Fos transforming protein	8.17
MAFK		INTERPRO	Maf transcription factor	8.17
MAFK		SMART	ETS	3.52
MAZ	Gm12878	INTERPRO	bZIP transcription factor, bZIP-1	3.99
MAZ		SMART	BRLZ	2.85
MAZ		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.79
MAZ		SMART	IRF	4.96
MAZ		INTERPRO	Interferon regulatory factor	4.85
MAZ		INTERPRO	Interferon regulatory factor, conserved site	4.85
MAZ		SMART	ETS	3.41
MAZ		INTERPRO	Ets	3.33
MAZ		INTERPRO	Winged helix repressor DNA-binding	1.96
MAZ		INTERPRO	SMAD domain-like	4.85
MAZ		INTERPRO	Interferon regulatory factor-3	4.85
MAZ		SMART	ZnF_C2H2	1.81
MAZ		SMART	SAM_PNT	3.72
MAZ		INTERPRO	Fos transforming protein	4.85
MAZ		INTERPRO	Zinc finger, C2H2-like	1.77
MAZ		INTERPRO	Zinc finger, C2H2-type	1.74
MAZ	Helas3	SMART	BRLZ	3.50
MAZ		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.59
MAZ		INTERPRO	bZIP transcription factor, bZIP-1	4.75
MAZ		SMART	ETS	4.27
MAZ		INTERPRO	Ets	4.38
MAZ		SMART	ZnF_C2H2	1.92
MAZ		INTERPRO	Zinc finger, C2H2-like	1.97
MAZ		INTERPRO	Basic leucine zipper	3.59
MAZ		INTERPRO	Zinc finger, C2H2-type	1.93
MAZ		INTERPRO	Fos transforming protein	5.39
MAZ		SMART	SAM_PNT	3.94
MAZ		INTERPRO	Sterile alpha motif/pointed	4.04
MAZ	Hepg2	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.25
MAZ		SMART	BRLZ	3.16
MAZ		INTERPRO	bZIP transcription factor, bZIP-1	4.20

MAZ		INTERPRO	Zinc finger, C2H2-like	2.16
MAZ		SMART	ZnF_C2H2	2.10
MAZ		INTERPRO	Zinc finger, C2H2-type	2.12
MAZ		INTERPRO	Zinc finger, C2H2-type/integrase, DNA-binding	2.02
MAZ		INTERPRO	Basic leucine zipper	3.40
MAZ		INTERPRO	Fos transforming protein	5.11
MAZ	K562	INTERPRO	bZIP transcription factor, bZIP-1	5.43
MAZ		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.73
MAZ		SMART	BRLZ	3.61
MAZ		INTERPRO	Zinc finger, C2H2-like	2.49
MAZ		SMART	ZnF_C2H2	2.40
MAZ		INTERPRO	Zinc finger, C2H2-type	2.44
MAZ		INTERPRO	Zinc finger, C2H2-type/integrase, DNA-binding	2.31
MAZ		SMART	ETS	3.35
MAZ		INTERPRO	Ets	3.46
MAZ		SMART	ZnF_GATA	5.95
MAZ		INTERPRO	Zinc finger, GATA-type	6.16
MAZ		INTERPRO	Fos transforming protein	6.16
MXI1	Gm12878	INTERPRO	bZIP transcription factor, bZIP-1	4.68
MXI1		SMART	BRLZ	3.39
MXI1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.17
MXI1		INTERPRO	Winged helix repressor DNA-binding	2.18
MXI1		SMART	IRF	4.66
MXI1		INTERPRO	Interferon regulatory factor, conserved site	4.35
MXI1		INTERPRO	Interferon regulatory factor	4.35
MXI1		SMART	ETS	3.00
MXI1		SMART	HLH	2.23
MXI1		INTERPRO	Ets	2.80
MXI1		INTERPRO	Fos transforming protein	4.97
MXI1		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	2.09
MXI1	H1hesc	SMART	HLH	3.16
MXI1		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	2.83
MXI1		SMART	BRLZ	2.76
MXI1		INTERPRO	bZIP transcription factor, bZIP-1	3.31
MXI1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.47
MXI1		SMART	DWA	7.00
MXI1		INTERPRO	DNA-binding RFX	5.22
MXI1	Helas3	SMART	BRLZ	4.44
MXI1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.76
MXI1		INTERPRO	bZIP transcription factor, bZIP-1	5.07
MXI1		INTERPRO	Fos transforming protein	5.39
MXI1	Hepg2	INTERPRO	bZIP transcription factor, bZIP-1	6.30

MXI1		SMART	BRLZ	4.54
MXI1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.26
MXI1		SMART	HLH	2.76
MXI1		INTERPRO	Fos transforming protein	6.69
MXI1		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	2.59
MXI1	K562	INTERPRO	bZIP transcription factor, bZIP-1	4.50
MXI1		SMART	ETS	4.46
MXI1		INTERPRO	Ets	4.47
MXI1		SMART	BRLZ	2.94
MXI1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.94
MXI1		SMART	HLH	2.47
MXI1		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	2.47
MXI1		INTERPRO	Winged helix repressor DNA-binding	1.97
MXI1		SMART	SAM_PNT	4.46
MXI1		INTERPRO	Sterile alpha motif/pointed	4.47
MXI1		INTERPRO	Sterile alpha motif-type	3.97
MXI1		SMART	ZnF_GATA	5.10
MXI1		INTERPRO	Zinc finger, GATA-type	5.11
MXI1		INTERPRO	Fos transforming protein	5.11
MYC	A549	SMART	SH2	7.76
MYC		INTERPRO	STAT transcription factor, protein interaction	6.81
MYC		INTERPRO	STAT transcription factor, DNA-binding	6.81
MYC		INTERPRO	STAT transcription factor, DNA-binding, subdomain	6.81
MYC		INTERPRO	SH2 motif	6.81
MYC		INTERPRO	STAT transcription factor, core	6.81
MYC		INTERPRO	STAT transcription factor, all-alpha	6.81
MYC		INTERPRO	EF-Hand type	5.96
MYC		SMART	BRLZ	2.35
MYC	Gm12878	SMART	BRLZ	4.19
MYC		INTERPRO	bZIP transcription factor, bZIP-1	5.51
MYC		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.85
MYC		INTERPRO	Fos transforming protein	6.69
MYC		SMART	HLH	2.59
MYC		SMART	ETS	3.19
MYC		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	2.37
MYC	H1hesc	SMART	HMG	6.39
MYC		INTERPRO	High mobility group, HMG1/HMG2	5.85
MYC		INTERPRO	Basic leucine zipper	4.44
MYC		INTERPRO	DNA-binding RFX	6.34
MYC		SMART	HLH	2.68
MYC		SMART	BRLZ	2.52

MYC		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	2.45
MYC	Helas3	SMART	BRLZ	4.81
MYC		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.41
MYC		INTERPRO	bZIP transcription factor, bZIP-1	6.06
MYC		INTERPRO	Fos transforming protein	6.06
MYC		INTERPRO	Basic leucine zipper	3.54
MYC		INTERPRO	DNA-binding RFX	5.05
MYC	Hepg2	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.23
MYC		SMART	BRLZ	3.95
MYC		INTERPRO	Basic leucine zipper	5.56
MYC		INTERPRO	bZIP transcription factor, bZIP-1	4.28
MYC		INTERPRO	Ets	3.41
MYC		SMART	ETS	3.19
MYC		SMART	SAM_PNT	4.25
MYC		INTERPRO	Sterile alpha motif/pointed	4.55
MYC		INTERPRO	Sterile alpha motif-type	4.04
MYC	Huvec	SMART	BRLZ	4.26
MYC		INTERPRO	bZIP transcription factor, bZIP-1	5.79
MYC		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.21
MYC		SMART	ETS	5.49
MYC		INTERPRO	Ets	5.43
MYC		SMART	SAM_PNT	5.12
MYC		INTERPRO	Sterile alpha motif/pointed	5.07
MYC		INTERPRO	Sterile alpha motif-type	4.50
MYC		INTERPRO	Fos transforming protein	5.79
MYC		INTERPRO	Winged helix repressor DNA-binding	1.83
MYC	K562	SMART	BRLZ	4.20
MYC		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.25
MYC		INTERPRO	bZIP transcription factor, bZIP-1	5.07
MYC		SMART	ETS	4.33
MYC		INTERPRO	Ets	4.38
MYC		INTERPRO	Basic leucine zipper	4.04
MYC		SMART	ZnF_GATA	5.33
MYC		INTERPRO	Fos transforming protein	5.39
MYC		INTERPRO	Zinc finger, GATA-type	5.39
MYC		SMART	SAM_PNT	4.00
MYC		INTERPRO	Winged helix repressor DNA-binding	1.80
MYC		INTERPRO	Sterile alpha motif/pointed	4.04
MYC	Mcf7	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.18
MYC		SMART	BRLZ	4.06
MYC		INTERPRO	bZIP transcription factor, bZIP-1	4.64
MYC		SMART	HLH	2.88

MYC		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	2.97
MYC		INTERPRO	Maf transcription factor	6.58
MYC		INTERPRO	Helix-loop-helix DNA-binding	2.53
NRF1	Gm12878	INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	8.34
NRF1	H1hesc	INTERPRO	Winged helix repressor DNA-binding	2.43
NRF1		SMART	HLH	3.57
NRF1		INTERPRO	DNA-binding RFX	7.70
NRF1		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	7.70
NRF1		INTERPRO	E2F Family	7.39
NRF1		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	2.68
NRF1	Helas3	INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	11.76
NRF1	Hepg2	SMART	DWA	23.80
NRF1		INTERPRO	CTF transcription factor/nuclear factor 1, conserved site	18.48
NRF1		INTERPRO	CTF transcription factor/nuclear factor 1	18.48
NRF1		INTERPRO	CTF transcription factor/nuclear factor 1, N-terminal	18.48
NRF1		INTERPRO	MAD homology 1, Dwarfin-type	18.48
NRF1		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	12.32
NRF1	K562	SMART	BRLZ	5.41
NRF1		INTERPRO	bZIP transcription factor, bZIP-1	7.61
NRF1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	5.09
NRF1		INTERPRO	Fos transforming protein	10.35
NRF1		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	8.62
REST	A549	INTERPRO	bZIP transcription factor, bZIP-1	9.87
REST		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	6.36
REST		SMART	BRLZ	6.01
REST		INTERPRO	Fos transforming protein	10.49
REST		SMART	FH	4.58
REST		INTERPRO	Transcription factor, fork head	4.84
REST		INTERPRO	Transcription factor, fork head, conserved site	4.84
REST	Gm12878	INTERPRO	Nuclear hormone receptor, ligand-binding, core	4.49
REST		INTERPRO	Nuclear hormone receptor, ligand-binding	4.49
REST		SMART	HOLI	4.13
REST		INTERPRO	Zinc finger, nuclear hormone receptor-type	4.16
REST		INTERPRO	Steroid hormone receptor	4.16
REST		SMART	ZnF_C4	3.83
REST		SMART	DWA	14.88
REST		INTERPRO	Zinc finger, NHR/GATA-type	3.83

REST		INTERPRO	CTF transcription factor/nuclear factor 1, N-terminal	16.17
REST		INTERPRO	MAD homology 1, Dwarfin-type	16.17
REST		INTERPRO	CTF transcription factor/nuclear factor 1	16.17
REST		INTERPRO	CTF transcription factor/nuclear factor 1, conserved site	16.17
REST	Helas3	INTERPRO	Zinc finger, nuclear hormone receptor-type	4.75
REST		SMART	ZnF_C4	4.37
REST		INTERPRO	Nuclear hormone receptor, ligand-binding, core	4.62
REST		INTERPRO	Nuclear hormone receptor, ligand-binding	4.62
REST		SMART	HOLI	4.25
REST		INTERPRO	Zinc finger, NHR/GATA-type	4.38
REST		INTERPRO	Steroid hormone receptor	4.22
REST		INTERPRO	Vitamin D receptor	6.60
REST	K562	SMART	ZnF_GATA	14.88
REST		INTERPRO	Zinc finger, GATA-type	15.52
REST		INTERPRO	bZIP transcription factor, bZIP-1	6.39
REST		SMART	BRLZ	4.06
REST		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.23
REST		INTERPRO	Fos transforming protein	12.42
REST		INTERPRO	Transcription factor Jun	15.52
REST		INTERPRO	Jun-like transcription factor	15.52
REST		INTERPRO	Transcription factor, GATA-1/2/3	15.52
REST	Panc1	INTERPRO	Helix-loop-helix DNA-binding	5.22
REST		SMART	HLH	4.24
REST		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	4.38
REST	Pfsk1	INTERPRO	Zinc finger, nuclear hormone receptor-type	3.41
REST		INTERPRO	Steroid hormone receptor	3.41
REST		SMART	ZnF_C4	3.14
REST		INTERPRO	Nuclear hormone receptor, ligand-binding	3.32
REST		INTERPRO	Nuclear hormone receptor, ligand-binding, core	3.32
REST		SMART	HOLI	3.05
REST		INTERPRO	Zinc finger, NHR/GATA-type	3.14
REST	Sknsh	INTERPRO	DNA-binding RFX	6.74
REST		SMART	DWA	8.71
REST		SMART	BRLZ	2.64
REST		INTERPRO	MAD homology 1, Dwarfin-type	8.08
REST		INTERPRO	CTF transcription factor/nuclear factor 1, conserved site	8.08
REST		INTERPRO	CTF transcription factor/nuclear factor 1	8.08
REST		INTERPRO	CTF transcription factor/nuclear factor 1, N-terminal	8.08
REST		INTERPRO	Helix-loop-helix DNA-binding	2.80

REST		SMART	ZnF_GATA	6.97
REST		SMART	HLH	2.53
REST		INTERPRO	bZIP transcription factor, bZIP-1	3.33
REST	U87	INTERPRO	bZIP transcription factor, bZIP-1	8.49
REST		SMART	BRLZ	5.85
REST		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	5.47
REST		INTERPRO	Fos transforming protein	9.02
REST		INTERPRO	DNA-binding RFX	7.52
RFX5	A549	INTERPRO	bZIP transcription factor, bZIP-1	9.9
RFX5		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	6.4
RFX5		SMART	BRLZ	6
RFX5		INTERPRO	Fos transforming protein	10.5
RFX5		INTERPRO	Transcription factor, fork head, conserved site	4.8
RFX5		INTERPRO	Transcription factor, fork head	4.8
RFX5		SMART	FH	4.6
RFX5	Gm12878	SMART	BRLZ	3.79
RFX5		INTERPRO	bZIP transcription factor, bZIP-1	5.35
RFX5		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.86
RFX5		SMART	ETS	4.83
RFX5		INTERPRO	Ets	4.93
RFX5		SMART	SAM_PNT	4.46
RFX5		INTERPRO	Fos transforming protein	6.06
RFX5		INTERPRO	Sterile alpha motif/pointed	4.55
RFX5		SMART	HLH	2.30
RFX5		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	2.35
RFX5		INTERPRO	Sterile alpha motif-type	4.04
RFX5	H1hesc	INTERPRO	bZIP transcription factor, bZIP-1	6.39
RFX5		SMART	BRLZ	4.53
RFX5		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.23
RFX5		INTERPRO	Fos transforming protein	7.76
RFX5		SMART	DWA	8.30
RFX5		INTERPRO	CTF transcription factor/nuclear factor 1, conserved site	7.76
RFX5		INTERPRO	MAD homology 1, Dwarfin-type	7.76
RFX5		INTERPRO	CTF transcription factor/nuclear factor 1, N-terminal	7.76
RFX5		INTERPRO	CTF transcription factor/nuclear factor 1	7.76
RFX5	Helas3	SMART	BRLZ	5.03
RFX5		INTERPRO	bZIP transcription factor, bZIP-1	6.85
RFX5		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.70
RFX5		SMART	SH2	8.30
RFX5		INTERPRO	SH2 motif	7.76

RFX5		INTERPRO	STAT transcription factor, DNA-binding	7.76
RFX5		INTERPRO	STAT transcription factor, core	7.76
RFX5		INTERPRO	STAT transcription factor, protein interaction	7.76
RFX5		INTERPRO	STAT transcription factor, all-alpha	7.76
RFX5		INTERPRO	STAT transcription factor, DNA-binding, subdomain	7.76
RFX5		INTERPRO	EF-Hand type	6.79
RFX5		INTERPRO	Fos transforming protein	7.76
RFX5	Hepg2	SMART	BRLZ	5.53
RFX5		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	5.50
RFX5		INTERPRO	bZIP transcription factor, bZIP-1	7.77
RFX5		INTERPRO	Fos transforming protein	8.26
RFX5		SMART	FH	4.47
RFX5		INTERPRO	Transcription factor, fork head	4.45
RFX5		INTERPRO	Transcription factor, fork head, conserved site	4.45
RFX5	K562	SMART	BRLZ	3.88
RFX5		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.85
RFX5		INTERPRO	bZIP transcription factor, bZIP-1	4.72
RFX5		SMART	POU	4.90
RFX5		INTERPRO	POU-specific	4.87
RFX5		INTERPRO	POU	4.87
RFX5		SMART	ZnF_GATA	6.74
RFX5		INTERPRO	Zinc finger, GATA-type	6.69
SRF	Gm12878	INTERPRO	bZIP transcription factor, bZIP-1	8.41
SRF		SMART	BRLZ	5.25
SRF		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.95
SRF		INTERPRO	Fos transforming protein	10.21
SRF	H1hesc	INTERPRO	Winged helix repressor DNA-binding	2.87
SRF		SMART	ETS	5.91
SRF		INTERPRO	Ets	4.85
SRF		INTERPRO	Sterile alpha motif-type	5.75
SRF		INTERPRO	DNA-binding RFX	7.19
SRF		SMART	SAM_PNT	6.56
SRF		SMART	BRLZ	2.86
SRF		INTERPRO	bZIP transcription factor, bZIP-1	3.55
SRF		INTERPRO	Sterile alpha motif/pointed	5.39
SRF	Hepg2	SMART	SH2	7.93
SRF		INTERPRO	STAT transcription factor, protein interaction	7.32
SRF		INTERPRO	SH2 motif	7.32
SRF		INTERPRO	STAT transcription factor, DNA-binding	7.32
SRF		INTERPRO	STAT transcription factor, core	7.32
SRF		INTERPRO	STAT transcription factor, DNA-binding,	7.32

			subdomain	
SRF		INTERPRO	STAT transcription factor, all-alpha	7.32
SRF		INTERPRO	EF-Hand type	6.41
SRF		SMART	BRLZ	2.88
SRF		SMART	ETS	3.97
SRF		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.66
SRF		INTERPRO	Ets	3.66
SRF		INTERPRO	Winged helix repressor DNA-binding	2.05
SRF		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	6.10
SRF	K562	INTERPRO	bZIP transcription factor, bZIP-1	7.46
SRF		SMART	BRLZ	4.94
SRF		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.75
SRF		SMART	ETS	6.79
SRF		INTERPRO	Ets	6.53
SRF		INTERPRO	Winged helix repressor DNA-binding	2.62
SRF		SMART	SAM_PNT	6.79
SRF		INTERPRO	Sterile alpha motif/pointed	6.53
SRF		INTERPRO	Sterile alpha motif-type	5.80
SRF		INTERPRO	Fos transforming protein	7.46
SRF		INTERPRO	Kelch related	7.46
TBP	Gm12878	INTERPRO	bZIP transcription factor, bZIP-1	4.92
TBP		SMART	BRLZ	3.48
TBP		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.26
TBP		SMART	IRF	5.58
TBP		INTERPRO	Interferon regulatory factor, conserved site	5.22
TBP		INTERPRO	Interferon regulatory factor	5.22
TBP		INTERPRO	Winged helix repressor DNA-binding	2.09
TBP		SMART	POU	4.06
TBP		INTERPRO	Fos transforming protein	5.97
TBP		INTERPRO	POU	3.80
TBP		INTERPRO	POU-specific	3.80
TBP		INTERPRO	Interferon regulatory factor-3	4.97
TBP		INTERPRO	SMAD domain-like	4.97
TBP	H1hesc	INTERPRO	bZIP transcription factor, bZIP-1	6.27
TBP		SMART	BRLZ	4.03
TBP		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.69
TBP		SMART	ETS	5.19
TBP		INTERPRO	Ets	4.75
TBP		INTERPRO	Winged helix repressor DNA-binding	2.27
TBP		INTERPRO	Fos transforming protein	7.61
TBP		SMART	SAM_PNT	5.19

TBP	Helas3	SMART	BRLZ	4.95
TBP		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.81
TBP		INTERPRO	bZIP transcription factor, bZIP-1	5.88
TBP		SMART	SH2	6.05
TBP		INTERPRO	STAT transcription factor, DNA-binding, subdomain	5.88
TBP		INTERPRO	STAT transcription factor, DNA-binding	5.88
TBP		INTERPRO	STAT transcription factor, core	5.88
TBP		INTERPRO	STAT transcription factor, protein interaction	5.88
TBP		INTERPRO	SH2 motif	5.88
TBP		INTERPRO	STAT transcription factor, all-alpha	5.88
TBP		INTERPRO	EF-Hand type	5.14
TBP		SMART	ETS	3.40
TBP		INTERPRO	Ets	3.31
TBP		SMART	SAM_PNT	4.54
TBP		INTERPRO	Fos transforming protein	5.88
TBP		INTERPRO	Sterile alpha motif/pointed	4.41
TBP		INTERPRO	Basic leucine zipper	3.43
TBP		INTERPRO	Sterile alpha motif-type	3.92
TBP	Hepg2	SMART	BRLZ	4.94
TBP		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.66
TBP		INTERPRO	bZIP transcription factor, bZIP-1	6.46
TBP		INTERPRO	Fos transforming protein	7.32
TBP		SMART	ETS	3.40
TBP		SMART	SAM_PNT	4.85
TBP	K562	INTERPRO	bZIP transcription factor, bZIP-1	5.37
TBP		SMART	ETS	5.29
TBP		INTERPRO	Ets	4.99
TBP		SMART	BRLZ	3.48
TBP		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.29
TBP		SMART	SAM_PNT	5.29
TBP		INTERPRO	Sterile alpha motif/pointed	4.99
TBP		INTERPRO	Sterile alpha motif-type	4.44
TBP		INTERPRO	Winged helix repressor DNA-binding	1.90
TBP		INTERPRO	Fos transforming protein	5.71
TCF12	A549	SMART	BRLZ	3.81
TCF12		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.46
TCF12		INTERPRO	bZIP transcription factor, bZIP-1	4.03
TCF12		INTERPRO	Basic leucine zipper	3.42
TCF12		SMART	FH	3.09
TCF12		INTERPRO	Fos transforming protein	4.56
TCF12	Gm12878	SMART	IRF	7.76

TCF12		INTERPRO	Interferon regulatory factor, conserved site	7.76
TCF12		INTERPRO	Interferon regulatory factor	7.76
TCF12		INTERPRO	Winged helix repressor DNA-binding	2.45
TCF12		SMART	ETS	4.37
TCF12		INTERPRO	Interferon regulatory factor-3	7.76
TCF12		INTERPRO	SMAD domain-like	7.76
TCF12		INTERPRO	Ets	4.37
TCF12		INTERPRO	bZIP transcription factor, bZIP-1	4.11
TCF12		SMART	BRLZ	2.35
TCF12	H1hesc	SMART	POU	5.83
TCF12		INTERPRO	POU-specific	6.17
TCF12		INTERPRO	POU	6.17
TCF12		SMART	HMG	4.22
TCF12		INTERPRO	High mobility group, HMG1/HMG2	4.48
TCF12		INTERPRO	Helix-loop-helix DNA-binding	2.98
TCF12	Hepg2	INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.83
TCF12		SMART	BRLZ	3.73
TCF12		INTERPRO	bZIP transcription factor, bZIP-1	4.36
TCF12		INTERPRO	Zinc finger, NHR/GATA-type	2.29
TCF12		INTERPRO	Steroid hormone receptor	2.24
TCF12		SMART	HOLI	2.13
TCF12		INTERPRO	Nuclear hormone receptor, ligand-binding, core	2.18
TCF12		INTERPRO	Nuclear hormone receptor, ligand-binding	2.18
TCF12		SMART	ZnF_C4	2.06
TCF12		INTERPRO	Zinc finger, nuclear hormone receptor-type	2.12
TCF12		INTERPRO	Vitamin D receptor	2.80
TCF12		INTERPRO	Basic leucine zipper	2.91
TCF7L2	Hct116	INTERPRO	bZIP transcription factor, bZIP-1	7.44
TCF7L2		SMART	BRLZ	5.25
TCF7L2		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.35
TCF7L2		INTERPRO	Fos transforming protein	8.43
TCF7L2		SMART	DWA	10.20
TCF7L2		INTERPRO	CTF transcription factor/nuclear factor 1, conserved site	8.43
TCF7L2		INTERPRO	CTF transcription factor/nuclear factor 1, N-terminal	8.43
TCF7L2		INTERPRO	MAD homology 1, Dwarfin-type	8.43
TCF7L2		INTERPRO	CTF transcription factor/nuclear factor 1	8.43
TCF7L2	Hek293	SMART	HMG	4.87
TCF7L2		INTERPRO	High mobility group, HMG1/HMG2	4.56
TCF7L2		SMART	ZnF_GATA	5.76
TCF7L2		INTERPRO	Zinc finger, GATA-type	5.39
TCF7L2		SMART	BRLZ	2.09

TCF7L2	Helas3	SMART	BRLZ	4.54
TCF7L2		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.26
TCF7L2		INTERPRO	bZIP transcription factor, bZIP-1	5.90
TCF7L2		SMART	ZnF_GATA	7.14
TCF7L2		INTERPRO	Zinc finger, GATA-type	6.69
TCF7L2		INTERPRO	Fos transforming protein	6.69
TCF7L2		INTERPRO	Basic leucine zipper	3.90
TCF7L2		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	5.57
TCF7L2	Hepg2	INTERPRO	Zinc finger, NHR/GATA-type	3.99
TCF7L2		SMART	ZnF_C4	3.94
TCF7L2		SMART	HOLI	3.83
TCF7L2		INTERPRO	Zinc finger, nuclear hormone receptor-type	3.81
TCF7L2		INTERPRO	Steroid hormone receptor	3.81
TCF7L2		INTERPRO	Nuclear hormone receptor, ligand-binding	3.70
TCF7L2		INTERPRO	Nuclear hormone receptor, ligand-binding, core	3.70
TCF7L2		SMART	HMG	5.78
TCF7L2		INTERPRO	High mobility group, HMG1/HMG2	5.60
TCF7L2		SMART	FH	4.82
TCF7L2		INTERPRO	Transcription factor, fork head, conserved site	4.66
TCF7L2		INTERPRO	Transcription factor, fork head	4.66
TCF7L2		INTERPRO	Retinoid X receptor	5.30
TCF7L2		SMART	ZnF_GATA	6.26
TCF7L2		INTERPRO	Zinc finger, GATA-type	6.06
TCF7L2	Mcf7	INTERPRO	bZIP transcription factor, bZIP-1	5.29
TCF7L2		SMART	BRLZ	3.25
TCF7L2		SMART	HMG	5.03
TCF7L2		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.07
TCF7L2		INTERPRO	High mobility group, HMG1/HMG2	4.76
TCF7L2		INTERPRO	Zinc finger, NHR/GATA-type	2.52
TCF7L2		SMART	ZnF_GATA	5.95
TCF7L2		INTERPRO	Fos transforming protein	5.62
TCF7L2		INTERPRO	Zinc finger, GATA-type	5.62
TCF7L2	Panc1	INTERPRO	bZIP transcription factor, bZIP-1	5.35
TCF7L2		SMART	BRLZ	3.61
TCF7L2		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.94
TCF7L2		INTERPRO	Fos transforming protein	6.06
USF1	A549	INTERPRO	bZIP transcription factor, bZIP-1	7.32
USF1		SMART	BRLZ	4.73
USF1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.66
USF1		INTERPRO	EF-Hand type	7.32
USF1		SMART	SH2	7.44

USF1		INTERPRO	STAT transcription factor, all-alpha	7.32
USF1		INTERPRO	STAT transcription factor, protein interaction	7.32
USF1		INTERPRO	STAT transcription factor, core	7.32
USF1		INTERPRO	SH2 motif	7.32
USF1		INTERPRO	STAT transcription factor, DNA-binding	7.32
USF1		INTERPRO	STAT transcription factor, DNA-binding, subdomain	7.32
USF1		INTERPRO	Fos transforming protein	7.32
USF1		INTERPRO	Kelch related	7.32
USF1	Gm12878	INTERPRO	bZIP transcription factor, bZIP-1	7.10
USF1		SMART	BRLZ	4.38
USF1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.44
USF1		INTERPRO	Fos transforming protein	8.62
USF1	H1hesc	SMART	HLH	3.24
USF1		INTERPRO	bZIP transcription factor, bZIP-1	4.56
USF1		SMART	BRLZ	3.04
USF1		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	3.22
USF1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.02
USF1		INTERPRO	Fos transforming protein	8.87
USF1	Hepg2	INTERPRO	bZIP transcription factor, bZIP-1	4.19
USF1		SMART	HLH	3.00
USF1		INTERPRO	Basic helix-loop-helix dimerisation region bHLH	3.07
USF1		SMART	BRLZ	2.82
USF1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.88
USF1		INTERPRO	Helix-loop-helix DNA-binding	2.74
USF1	K562	SMART	BRLZ	4.10
USF1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.04
USF1		INTERPRO	bZIP transcription factor, bZIP-1	5.71
USF1		INTERPRO	Fos transforming protein	9.70
YY1	A549	INTERPRO	bZIP transcription factor, bZIP-1	5.81
YY1		SMART	ETS	6.04
YY1		INTERPRO	Ets	5.73
YY1		SMART	BRLZ	3.83
YY1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.63
YY1		INTERPRO	Winged helix repressor DNA-binding	2.23
YY1		INTERPRO	Fos transforming protein	7.05
YY1		SMART	SAM_PNT	4.65
YY1	Gm12878	SMART	SH2	9.15
YY1		INTERPRO	STAT transcription factor, DNA-binding	9.02
YY1		INTERPRO	SH2 motif	9.02
YY1		INTERPRO	STAT transcription factor, all-alpha	9.02
YY1		INTERPRO	STAT transcription factor, DNA-binding,	9.02

			subdomain	
YY1		INTERPRO	STAT transcription factor, core	9.02
YY1		INTERPRO	STAT transcription factor, protein interaction	9.02
YY1		INTERPRO	EF-Hand type	7.90
YY1		SMART	BRLZ	3.33
YY1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.28
YY1		INTERPRO	bZIP transcription factor, bZIP-1	3.72
YY1	H1hesc	SMART	BRLZ	4.99
YY1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.41
YY1		INTERPRO	bZIP transcription factor, bZIP-1	4.99
YY1		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	8.08
YY1	Hct116	SMART	BRLZ	4.64
YY1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.41
YY1		INTERPRO	bZIP transcription factor, bZIP-1	6.11
YY1		SMART	SH2	7.29
YY1		INTERPRO	STAT transcription factor, protein interaction	6.93
YY1		INTERPRO	STAT transcription factor, all-alpha	6.93
YY1		INTERPRO	STAT transcription factor, core	6.93
YY1		INTERPRO	SH2 motif	6.93
YY1		INTERPRO	STAT transcription factor, DNA-binding	6.93
YY1		INTERPRO	STAT transcription factor, DNA-binding, subdomain	6.93
YY1		INTERPRO	EF-Hand type	6.06
YY1		INTERPRO	Fos transforming protein	6.93
YY1	Hepg2	SMART	HMG	6.87
YY1		INTERPRO	High mobility group, HMG1/HMG2	6.44
YY1		SMART	DWA	8.11
YY1		INTERPRO	CTF transcription factor/nuclear factor 1, N-terminal	7.61
YY1		INTERPRO	CTF transcription factor/nuclear factor 1	7.61
YY1		INTERPRO	MAD homology 1, Dwarfin-type	7.61
YY1		INTERPRO	CTF transcription factor/nuclear factor 1, conserved site	7.61
YY1	K562	SMART	BRLZ	5.15
YY1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	4.20
YY1		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	11.55
YY1		INTERPRO	bZIP transcription factor, bZIP-1	5.71
YY1		INTERPRO	E2F Family	11.09
YY1	Nt2d1	SMART	BRLZ	3.09
YY1		INTERPRO	bZIP transcription factor, bZIP-1	4.04
YY1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	2.98

YY1		SMART	ETS	4.14
YY1		INTERPRO	Ets	3.99
YY1		SMART	HMG	3.92
YY1		INTERPRO	High mobility group, HMG1/HMG2	3.78
YY1		SMART	SH2	5.10
YY1		INTERPRO	STAT transcription factor, DNA-binding, subdomain	4.91
YY1		INTERPRO	STAT transcription factor, all-alpha	4.91
YY1		INTERPRO	SH2 motif	4.91
YY1		INTERPRO	STAT transcription factor, protein interaction	4.91
YY1		INTERPRO	STAT transcription factor, core	4.91
YY1		INTERPRO	STAT transcription factor, DNA-binding	4.91
YY1		INTERPRO	Winged helix repressor DNA-binding	1.90
YY1		INTERPRO	EF-Hand type	4.30
YY1		SMART	SAM_PNT	3.83
YY1		INTERPRO	Fos transforming protein	4.91
YY1	Sknshra	INTERPRO	bZIP transcription factor, bZIP-1	6.39
YY1		SMART	BRLZ	3.93
YY1		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.76
YY1		SMART	SH2	8.11
YY1		INTERPRO	STAT transcription factor, DNA-binding	7.76
YY1		INTERPRO	STAT transcription factor, all-alpha	7.76
YY1		INTERPRO	STAT transcription factor, DNA-binding, subdomain	7.76
YY1		INTERPRO	STAT transcription factor, protein interaction	7.76
YY1		INTERPRO	STAT transcription factor, core	7.76
YY1		INTERPRO	SH2 motif	7.76
YY1		INTERPRO	EF-Hand type	6.79
YY1		INTERPRO	Fos transforming protein	7.76
YY1		INTERPRO	Transcription factor E2F/dimerisation partner (TDP)	6.47
ZNF143	Gm12878	INTERPRO	Winged helix repressor DNA-binding	2.85
ZNF143		SMART	ETS	5.27
ZNF143		INTERPRO	Ets	5.08
ZNF143		SMART	SH2	6.49
ZNF143		INTERPRO	STAT transcription factor, core	6.26
ZNF143		INTERPRO	STAT transcription factor, protein interaction	6.26
ZNF143		INTERPRO	STAT transcription factor, all-alpha	6.26
ZNF143		INTERPRO	STAT transcription factor, DNA-binding, subdomain	6.26
ZNF143		INTERPRO	STAT transcription factor, DNA-binding	6.26
ZNF143		INTERPRO	SH2 motif	6.26
ZNF143		SMART	PAX	5.68

ZNF143		INTERPRO	Paired box protein, N-terminal	5.48
ZNF143		INTERPRO	EF-Hand type	5.48
ZNF143		SMART	SAM_PNT	4.87
ZNF143		INTERPRO	Sterile alpha motif/pointed	4.69
ZNF143		INTERPRO	Sterile alpha motif-type	4.17
ZNF143	H1hesc	SMART	SH2	35.70
ZNF143		INTERPRO	STAT transcription factor, protein interaction	27.71
ZNF143		INTERPRO	STAT transcription factor, all-alpha	27.71
ZNF143		INTERPRO	STAT transcription factor, core	27.71
ZNF143		INTERPRO	SH2 motif	27.71
ZNF143		INTERPRO	STAT transcription factor, DNA-binding, subdomain	27.71
ZNF143		INTERPRO	STAT transcription factor, DNA-binding	27.71
ZNF143		INTERPRO	EF-Hand type	24.25
ZNF143	Helas3	SMART	BRLZ	4.57
ZNF143		INTERPRO	bZIP transcription factor, bZIP-1	5.51
ZNF143		INTERPRO	Basic-leucine zipper (bZIP) transcription factor	3.85
ZNF143		SMART	SH2	7.93
ZNF143		INTERPRO	STAT transcription factor, all-alpha	6.69
ZNF143		INTERPRO	SH2 motif	6.69
ZNF143		INTERPRO	STAT transcription factor, DNA-binding	6.69
ZNF143		INTERPRO	STAT transcription factor, core	6.69
ZNF143		INTERPRO	STAT transcription factor, DNA-binding, subdomain	6.69
ZNF143		INTERPRO	STAT transcription factor, protein interaction	6.69
ZNF143		INTERPRO	EF-Hand type	5.85
ZNF143		INTERPRO	Fos transforming protein	6.69
ZNF143		INTERPRO	DNA-binding RFX	5.57
ZNF143		INTERPRO	Winged helix repressor DNA-binding	1.88
ZNF143	K562	SMART	SH2	12.31
ZNF143		INTERPRO	STAT transcription factor, all-alpha	10.49
ZNF143		INTERPRO	STAT transcription factor, DNA-binding	10.49
ZNF143		INTERPRO	STAT transcription factor, core	10.49
ZNF143		INTERPRO	STAT transcription factor, DNA-binding, subdomain	10.49
ZNF143		INTERPRO	STAT transcription factor, protein interaction	10.49
ZNF143		INTERPRO	SH2 motif	10.49
ZNF143		INTERPRO	EF-Hand type	9.18

Table 7.13 List of protein domains found as enriched in the identified co-factors. The analysis was done using DAVID tool.

TF	Cell line	GO-term	Biological Process (BP) name

ATF3	A549	GO:0007346	regulation of mitotic cell cycle
ATF3	A549	GO:0065008	regulation of biological quality
ATF3	H1hesc	GO:0006953	acute-phase response
ATF3	H1hesc	GO:0002526	acute inflammatory response
ATF3	H1hesc	GO:0001889	liver development
ATF3	H1hesc	GO:0044281	small molecule metabolic process
ATF3	Hepg2	GO:0070345	negative regulation of fat cell proliferation
ATF3	Hepg2	GO:0070344	regulation of fat cell proliferation
ATF3	Hepg2	GO:1900739	regulation of protein insertion into mitochondrial membrane
ATF3	Hepg2	GO:1900740	positive regulation of protein insertion into mitochondrial me
ATF3	Hepg2	GO:1901028	regulation of mitochondrial outer membrane permeabilizatio
ATF3	Hepg2	GO:1901030	positive regulation of mitochondrial outer membrane permea pathway
ATF3	Hepg2	GO:0000278	mitotic cell cycle
ATF3	Hepg2	GO:0007049	cell cycle
ATF3	Hepg2	GO:0007219	Notch signaling pathway
ATF3	K562	GO:0000429	carbon catabolite regulation of transcription from RNA polyr
ATF3	K562	GO:0000430	regulation of transcription from RNA polymerase II promote
ATF3	K562	GO:0000432	positive regulation of transcription from RNA polymerase II
ATF3	K562	GO:0000436	carbon catabolite activation of transcription from RNA polym
ATF3	K562	GO:0022037	metencephalon development
ATF3	K562	GO:0045672	positive regulation of osteoclast differentiation
ATF3	K562	GO:0019086	late viral transcription
ATF3	K562	GO:0019083	viral transcription
ATF3	K562	GO:0060430	lung saccule development
ATF3	K562	GO:0060575	intestinal epithelial cell differentiation
ATF3	K562	GO:0045766	positive regulation of angiogenesis
ATF3	K562	GO:0002763	positive regulation of myeloid leukocyte differentiation
ATF3	K562	GO:1904018	positive regulation of vasculature development
ATF3	K562	GO:0060395	SMAD protein signal transduction
ATF3	K562	GO:0045639	positive regulation of myeloid cell differentiation
ATF3	K562	GO:0032941	secretion by tissue
ATF3	K562	GO:0031098	stress-activated protein kinase signaling cascade
ATF3	K562	GO:0007595	lactation
ATF3	K562	GO:0007612	learning
ATF3	K562	GO:0051403	stress-activated MAPK cascade
ATF3	K562	GO:0007589	body fluid secretion
ATF3	K562	GO:0090287	regulation of cellular response to growth factor stimulus
ATF3	K562	GO:0007611	learning or memory
ATF3	K562	GO:0003208	cardiac ventricle morphogenesis
ATF3	K562	GO:0023014	signal transduction by protein phosphorylation
ATF3	K562	GO:0000165	MAPK cascade
ATF3	K562	GO:0050890	cognition

ATF3	K562	GO:0046903	secretion
ATF3	K562	GO:0050878	regulation of body fluid levels
ATF3	K562	GO:1902107	positive regulation of leukocyte differentiation
ATF3	K562	GO:1903708	positive regulation of hemopoiesis
ATF3	K562	GO:0009314	response to radiation
ATF3	K562	GO:0002768	immune response-regulating cell surface receptor signaling
ATF3	K562	GO:0002253	activation of immune response
ATF3	K562	GO:0002757	immune response-activating signal transduction
ATF3	K562	GO:0042493	response to drug
ATF3	K562	GO:0044765	single-organism transport
ATF3	K562	GO:0009725	response to hormone
ATF3	K562	GO:0009628	response to abiotic stimulus
ATF3	K562	GO:0071495	cellular response to endogenous stimulus
ATF3	K562	GO:0003008	system process
ATF3	K562	GO:0065008	regulation of biological quality
ATF3	K562	GO:0009719	response to endogenous stimulus
ATF3	K562	GO:0048584	positive regulation of response to stimulus
ATF3	K562	GO:0070887	cellular response to chemical stimulus
ATF3	K562	GO:0051240	positive regulation of multicellular organismal process
BHLHE40	Gm12878	GO:0032647	regulation of interferon-alpha production
BHLHE40	Gm12878	GO:0032727	positive regulation of interferon-alpha production
BHLHE40	Gm12878	GO:0032728	positive regulation of interferon-beta production
BHLHE40	Gm12878	GO:0051385	response to mineralocorticoid
BHLHE40	Gm12878	GO:0051412	response to corticosterone
BHLHE40	Gm12878	GO:0060333	interferon-gamma-mediated signaling pathway
BHLHE40	Gm12878	GO:0032648	regulation of interferon-beta production
BHLHE40	Gm12878	GO:0071277	cellular response to calcium ion
BHLHE40	Gm12878	GO:0051592	response to calcium ion
BHLHE40	Gm12878	GO:0060337	type I interferon signaling pathway
BHLHE40	Gm12878	GO:0002761	regulation of myeloid leukocyte differentiation
BHLHE40	Gm12878	GO:0019221	cytokine-mediated signaling pathway
BHLHE40	Gm12878	GO:0009612	response to mechanical stimulus
BHLHE40	Gm12878	GO:0045637	regulation of myeloid cell differentiation
BHLHE40	Gm12878	GO:0043207	response to external biotic stimulus
BHLHE40	Gm12878	GO:0009607	response to biotic stimulus
BHLHE40	Gm12878	GO:0006464	cellular protein modification process
BHLHE40	Gm12878	GO:0036211	protein modification process
BHLHE40	Gm12878	GO:1903706	regulation of hemopoiesis
BHLHE40	Hepg2	GO:0032922	circadian regulation of gene expression
BHLHE40	Hepg2	GO:1901564	organonitrogen compound metabolic process
BHLHE40	Hepg2	GO:0033500	carbohydrate homeostasis
BHLHE40	Hepg2	GO:0042593	glucose homeostasis
BHLHE40	Hepg2	GO:0048608	reproductive structure development

BHLHE40	Hepg2	GO:0048878	chemical homeostasis
BHLHE40	Hepg2	GO:0003006	developmental process involved in reproduction
BHLHE40	Hepg2	GO:0044710	single-organism metabolic process
BHLHE40	Hepg2	GO:0044702	single organism reproductive process
BHLHE40	Hepg2	GO:0022414	reproductive process
BHLHE40	Hepg2	GO:0065008	regulation of biological quality
BHLHE40	K562	GO:0071773	cellular response to BMP stimulus
BHLHE40	K562	GO:0071772	response to BMP
BHLHE40	K562	GO:0030097	hemopoiesis
CEBPB	Gm12878	GO:0060333	interferon-gamma-mediated signaling pathway
CEBPB	Gm12878	GO:0045075	regulation of interleukin-12 biosynthetic process
CEBPB	Gm12878	GO:0038061	NIK/NF-kappaB signaling
CEBPB	Gm12878	GO:0007249	I-kappaB kinase/NF-kappaB signaling
CEBPB	Gm12878	GO:0032648	regulation of interferon-beta production
CEBPB	Gm12878	GO:0043122	regulation of I-kappaB kinase/NF-kappaB signaling
CEBPB	Gm12878	GO:0060337	type I interferon signaling pathway
CEBPB	Gm12878	GO:0032735	positive regulation of interleukin-12 production
CEBPB	Gm12878	GO:0032728	positive regulation of interferon-beta production
CEBPB	Gm12878	GO:0032479	regulation of type I interferon production
CEBPB	Gm12878	GO:0002221	pattern recognition receptor signaling pathway
CEBPB	Gm12878	GO:0002224	toll-like receptor signaling pathway
CEBPB	Gm12878	GO:0002756	MyD88-independent toll-like receptor signaling pathway
CEBPB	Gm12878	GO:0034142	toll-like receptor 4 signaling pathway
CEBPB	Gm12878	GO:0034138	toll-like receptor 3 signaling pathway
CEBPB	Gm12878	GO:0035666	TRIF-dependent toll-like receptor signaling pathway
CEBPB	Gm12878	GO:0032655	regulation of interleukin-12 production
CEBPB	Gm12878	GO:0051607	defense response to virus
CEBPB	Gm12878	GO:0032481	positive regulation of type I interferon production
CEBPB	Gm12878	GO:0034162	toll-like receptor 9 signaling pathway
CEBPB	Gm12878	GO:0048011	neurotrophin TRK receptor signaling pathway
CEBPB	Gm12878	GO:0034166	toll-like receptor 10 signaling pathway
CEBPB	Gm12878	GO:0002755	MyD88-dependent toll-like receptor signaling pathway
CEBPB	Gm12878	GO:0034134	toll-like receptor 2 signaling pathway
CEBPB	Gm12878	GO:0034146	toll-like receptor 5 signaling pathway
CEBPB	Gm12878	GO:0038179	neurotrophin signaling pathway
CEBPB	Gm12878	GO:0038124	toll-like receptor TLR6:TLR2 signaling pathway
CEBPB	Gm12878	GO:0038123	toll-like receptor TLR1:TLR2 signaling pathway
CEBPB	Gm12878	GO:0019221	cytokine-mediated signaling pathway
CEBPB	Gm12878	GO:0098542	defense response to other organism
CEBPB	Gm12878	GO:0002218	activation of innate immune response
CEBPB	Gm12878	GO:0002758	innate immune response-activating signal transduction
CEBPB	Gm12878	GO:0002253	activation of immune response
CEBPB	Gm12878	GO:0002757	immune response-activating signal transduction

CEBPB	Gm12878	GO:0002252	immune effector process
CEBPB	Gm12878	GO:0007169	transmembrane receptor protein tyrosine kinase signaling p
CEBPB	Gm12878	GO:0009615	response to virus
CEBPB	Gm12878	GO:0045089	positive regulation of innate immune response
CEBPB	Gm12878	GO:0031349	positive regulation of defense response
CEBPB	Gm12878	GO:0001819	positive regulation of cytokine production
CEBPB	Gm12878	GO:0045088	regulation of innate immune response
CEBPB	Gm12878	GO:0050778	positive regulation of immune response
CEBPB	Gm12878	GO:0002764	immune response-regulating signaling pathway
CEBPB	Gm12878	GO:0051707	response to other organism
CEBPB	Gm12878	GO:0006915	apoptotic process
CEBPB	Gm12878	GO:0012501	programmed cell death
CEBPB	Gm12878	GO:0008219	cell death
CEBPB	Gm12878	GO:0071407	cellular response to organic cyclic compound
CEBPB	Gm12878	GO:0016265	death
CEBPB	Gm12878	GO:0001817	regulation of cytokine production
CEBPB	Gm12878	GO:0031347	regulation of defense response
CEBPB	Gm12878	GO:0050776	regulation of immune response
CEBPB	Gm12878	GO:0006952	defense response
CEBPB	Gm12878	GO:0019538	protein metabolic process
CEBPB	Gm12878	GO:0002682	regulation of immune system process
CEBPB	Gm12878	GO:0002376	immune system process
CEBPB	Gm12878	GO:0007166	cell surface receptor signaling pathway
CEBPB	Gm12878	GO:0006950	response to stress
CEBPB	Gm12878	GO:0006366	transcription from RNA polymerase II promoter
CEBPB	Gm12878	GO:0007165	signal transduction
CEBPB	H1hesc	GO:0070345	negative regulation of fat cell proliferation
CEBPB	H1hesc	GO:0070344	regulation of fat cell proliferation
CEBPB	H1hesc	GO:1900739	regulation of protein insertion into mitochondrial membrane
CEBPB	H1hesc	GO:1900740	positive regulation of protein insertion into mitochondrial me
CEBPB	H1hesc	GO:1901028	regulation of mitochondrial outer membrane permeabilizatio
CEBPB	H1hesc	GO:1901030	positive regulation of mitochondrial outer membrane permea pathway
CEBPB	H1hesc	GO:0000278	mitotic cell cycle
CEBPB	H1hesc	GO:0007049	cell cycle
CEBPB	H1hesc	GO:0032388	positive regulation of intracellular transport
CEBPB	H1hesc	GO:1903829	positive regulation of cellular protein localization
CEBPB	H1hesc	GO:0033043	regulation of organelle organization
CEBPB	Helas3	GO:0036499	PERK-mediated unfolded protein response
CEBPB	Helas3	GO:0034976	response to endoplasmic reticulum stress
CEBPB	K562	GO:0050817	coagulation
CEBPB	K562	GO:0007596	blood coagulation
CEBPB	K562	GO:0007599	hemostasis

CTCF	Hct116	GO:1901741	positive regulation of myoblast fusion
CTCF	Hct116	GO:1901739	regulation of myoblast fusion
CTCF	Hct116	GO:0051149	positive regulation of muscle cell differentiation
CTCF	Hct116	GO:0010720	positive regulation of cell development
CTCF	Hek293	GO:0009719	response to endogenous stimulus
CTCF	Hepg2	GO:0030513	positive regulation of BMP signaling pathway
CTCF	Hepg2	GO:0090100	positive regulation of transmembrane receptor protein serin
CTCF	Hepg2	GO:0030510	regulation of BMP signaling pathway
CTCF	Hepg2	GO:0090092	regulation of transmembrane receptor protein serine/threoni
CTCF	Huvec	GO:0006974	cellular response to DNA damage stimulus
CTCF	K562	GO:0070345	negative regulation of fat cell proliferation
CTCF	K562	GO:0070344	regulation of fat cell proliferation
CTCF	K562	GO:1900739	regulation of protein insertion into mitochondrial membrane
CTCF	K562	GO:1900740	positive regulation of protein insertion into mitochondrial me
CTCF	K562	GO:1901028	regulation of mitochondrial outer membrane permeabilizatio
CTCF	K562	GO:1901030	positive regulation of mitochondrial outer membrane perme pathway
CTCF	K562	GO:0003334	keratinocyte development
CTCF	K562	GO:0072210	metanephric nephron development
CTCF	K562	GO:0000278	mitotic cell cycle
CTCF	K562	GO:0007049	cell cycle
CTCF	K562	GO:0042127	regulation of cell proliferation
CTCF	Mcf7	GO:0030097	hemopoiesis
CTCF	Sknshra	GO:1901522	positive regulation of transcription from RNA polymerase II chemical stimulus
EP300	A549	GO:0007267	cell-cell signaling
EP300	A549	GO:0044700	single organism signaling
EP300	A549	GO:0023052	signaling
EP300	A549	GO:0007154	cell communication
EP300	Gm12878	GO:0060333	interferon-gamma-mediated signaling pathway
EP300	Gm12878	GO:0002223	stimulatory C-type lectin receptor signaling pathway
EP300	Gm12878	GO:0002220	innate immune response activating cell surface receptor sig
EP300	Gm12878	GO:0032481	positive regulation of type I interferon production
EP300	Gm12878	GO:0060337	type I interferon signaling pathway
EP300	Gm12878	GO:0002260	lymphocyte homeostasis
EP300	Gm12878	GO:0071260	cellular response to mechanical stimulus
EP300	Gm12878	GO:0032728	positive regulation of interferon-beta production
EP300	Gm12878	GO:0050851	antigen receptor-mediated signaling pathway
EP300	Gm12878	GO:0002429	immune response-activating cell surface receptor signaling
EP300	Gm12878	GO:0032479	regulation of type I interferon production
EP300	Gm12878	GO:0051607	defense response to virus
EP300	Gm12878	GO:0050778	positive regulation of immune response
EP300	Gm12878	GO:0002218	activation of innate immune response

EP300	Gm12878	GO:0002758	innate immune response-activating signal transduction
EP300	Gm12878	GO:0050871	positive regulation of B cell activation
EP300	Gm12878	GO:0045089	positive regulation of innate immune response
EP300	Gm12878	GO:0019221	cytokine-mediated signaling pathway
EP300	Gm12878	GO:0031349	positive regulation of defense response
EP300	Gm12878	GO:0002221	pattern recognition receptor signaling pathway
EP300	Gm12878	GO:0002224	toll-like receptor signaling pathway
EP300	Gm12878	GO:0002756	MyD88-independent toll-like receptor signaling pathway
EP300	Gm12878	GO:0034142	toll-like receptor 4 signaling pathway
EP300	Gm12878	GO:0034138	toll-like receptor 3 signaling pathway
EP300	Gm12878	GO:0035666	TRIF-dependent toll-like receptor signaling pathway
EP300	Gm12878	GO:0002253	activation of immune response
EP300	Gm12878	GO:0002757	immune response-activating signal transduction
EP300	Gm12878	GO:0034162	toll-like receptor 9 signaling pathway
EP300	Gm12878	GO:0050864	regulation of B cell activation
EP300	Gm12878	GO:0045088	regulation of innate immune response
EP300	Gm12878	GO:0071375	cellular response to peptide hormone stimulus
EP300	Gm12878	GO:0002697	regulation of immune effector process
EP300	Gm12878	GO:0098542	defense response to other organism
EP300	Gm12878	GO:0002764	immune response-regulating signaling pathway
EP300	Gm12878	GO:0050776	regulation of immune response
EP300	Gm12878	GO:0001819	positive regulation of cytokine production
EP300	Gm12878	GO:1901699	cellular response to nitrogen compound
EP300	Gm12878	GO:0002696	positive regulation of leukocyte activation
EP300	Gm12878	GO:0071345	cellular response to cytokine stimulus
EP300	Gm12878	GO:0051251	positive regulation of lymphocyte activation
EP300	Gm12878	GO:0009615	response to virus
EP300	Gm12878	GO:0051249	regulation of lymphocyte activation
EP300	Gm12878	GO:0002684	positive regulation of immune system process
EP300	Gm12878	GO:0050867	positive regulation of cell activation
EP300	Gm12878	GO:0002768	immune response-regulating cell surface receptor signaling
EP300	Gm12878	GO:0006915	apoptotic process
EP300	Gm12878	GO:0012501	programmed cell death
EP300	Gm12878	GO:0071417	cellular response to organonitrogen compound
EP300	Gm12878	GO:0045087	innate immune response
EP300	Gm12878	GO:0008219	cell death
EP300	Gm12878	GO:0045580	regulation of T cell differentiation
EP300	Gm12878	GO:0002694	regulation of leukocyte activation
EP300	Gm12878	GO:0016265	death
EP300	Gm12878	GO:0051707	response to other organism
EP300	Gm12878	GO:0031347	regulation of defense response
EP300	Gm12878	GO:0048534	hematopoietic or lymphoid organ development
EP300	Gm12878	GO:0050865	regulation of cell activation

EP300	Gm12878	GO:0001817	regulation of cytokine production
EP300	Gm12878	GO:0002682	regulation of immune system process
EP300	Gm12878	GO:0006952	defense response
EP300	Gm12878	GO:0043207	response to external biotic stimulus
EP300	Gm12878	GO:0034097	response to cytokine
EP300	Gm12878	GO:0009607	response to biotic stimulus
EP300	Gm12878	GO:1903706	regulation of hemopoiesis
EP300	Gm12878	GO:0080134	regulation of response to stress
EP300	Gm12878	GO:0002376	immune system process
EP300	Gm12878	GO:0050678	regulation of epithelial cell proliferation
EP300	Gm12878	GO:1901698	response to nitrogen compound
EP300	Gm12878	GO:0035556	intracellular signal transduction
EP300	Gm12878	GO:0006955	immune response
EP300	Gm12878	GO:0007166	cell surface receptor signaling pathway
EP300	Gm12878	GO:0048584	positive regulation of response to stimulus
EP300	Gm12878	GO:0070887	cellular response to chemical stimulus
EP300	Gm12878	GO:0019538	protein metabolic process
EP300	Gm12878	GO:0006950	response to stress
EP300	Gm12878	GO:0042221	response to chemical
EP300	Gm12878	GO:0051240	positive regulation of multicellular organismal process
EP300	Gm12878	GO:0008284	positive regulation of cell proliferation
EP300	Gm12878	GO:0071310	cellular response to organic substance
EP300	Gm12878	GO:0048583	regulation of response to stimulus
EP300	Gm12878	GO:0051716	cellular response to stimulus
EP300	Gm12878	GO:0010033	response to organic substance
EP300	Gm12878	GO:0050896	response to stimulus
EP300	Gm12878	GO:0007165	signal transduction
EP300	Gm12878	GO:0051239	regulation of multicellular organismal process
EP300	Gm12878	GO:0045595	regulation of cell differentiation
EP300	Gm12878	GO:0006366	transcription from RNA polymerase II promoter
EP300	Gm12878	GO:0048522	positive regulation of cellular process
EP300	Gm12878	GO:0048518	positive regulation of biological process
EP300	H1hesc	GO:0060174	limb bud formation
EP300	Helas3	GO:0008637	apoptotic mitochondrial changes
EP300	Helas3	GO:0070059	intrinsic apoptotic signaling pathway in response to endopla
EP300	Helas3	GO:0034976	response to endoplasmic reticulum stress
EP300	Helas3	GO:1990440	positive regulation of transcription from RNA polymerase II stress
EP300	Helas3	GO:0043620	regulation of DNA-templated transcription in response to str
EP300	Helas3	GO:0043618	regulation of transcription from RNA polymerase II promoter
EP300	Helas3	GO:0001889	liver development
EP300	Helas3	GO:0001819	positive regulation of cytokine production
EP300	Hepg2	GO:0051385	response to mineralocorticoid

EP300	Hepg2	GO:0051412	response to corticosterone
EP300	Hepg2	GO:0006805	xenobiotic metabolic process
EP300	Hepg2	GO:0035902	response to immobilization stress
EP300	Hepg2	GO:0048145	regulation of fibroblast proliferation
EP300	Hepg2	GO:0048146	positive regulation of fibroblast proliferation
EP300	Hepg2	GO:0032570	response to progesterone
EP300	Hepg2	GO:0071277	cellular response to calcium ion
EP300	Hepg2	GO:0051592	response to calcium ion
EP300	Hepg2	GO:0007565	female pregnancy
EP300	Hepg2	GO:0051384	response to glucocorticoid
EP300	Hepg2	GO:0031960	response to corticosteroid
EP300	Hepg2	GO:0014074	response to purine-containing compound
EP300	Hepg2	GO:0071248	cellular response to metal ion
EP300	Hepg2	GO:1901654	response to ketone
EP300	Hepg2	GO:0009612	response to mechanical stimulus
EP300	Hepg2	GO:0051591	response to cAMP
EP300	Hepg2	GO:0046683	response to organophosphorus
EP300	Hepg2	GO:0043401	steroid hormone mediated signaling pathway
EP300	Hepg2	GO:0071241	cellular response to inorganic substance
EP300	Hepg2	GO:0097305	response to alcohol
EP300	Hepg2	GO:0031668	cellular response to extracellular stimulus
EP300	Hepg2	GO:0009755	hormone-mediated signaling pathway
EP300	Hepg2	GO:0030522	intracellular receptor signaling pathway
EP300	Hepg2	GO:0010038	response to metal ion
EP300	Hepg2	GO:0006629	lipid metabolic process
EP300	Hepg2	GO:0006367	transcription initiation from RNA polymerase II promoter
EP300	Hepg2	GO:0006352	DNA-templated transcription, initiation
EP300	Hepg2	GO:0048545	response to steroid hormone
EP300	Hepg2	GO:0010467	gene expression
EP300	Hepg2	GO:0042493	response to drug
EP300	Hepg2	GO:0051090	regulation of sequence-specific DNA binding transcription fa
EP300	Hepg2	GO:0014070	response to organic cyclic compound
EP300	Hepg2	GO:0010243	response to organonitrogen compound
EP300	Hepg2	GO:0032870	cellular response to hormone stimulus
EP300	Hepg2	GO:0033993	response to lipid
EP300	Hepg2	GO:1901698	response to nitrogen compound
EP300	Hepg2	GO:0071495	cellular response to endogenous stimulus
EP300	Hepg2	GO:1901700	response to oxygen-containing compound
EP300	Hepg2	GO:0009725	response to hormone
EP300	Hepg2	GO:0009628	response to abiotic stimulus
EP300	Hepg2	GO:0003006	developmental process involved in reproduction
EP300	Hepg2	GO:0022414	reproductive process
EP300	Hepg2	GO:0071310	cellular response to organic substance

EP300	Hepg2	GO:0044710	single-organism metabolic process
EP300	Hepg2	GO:0009719	response to endogenous stimulus
EP300	Hepg2	GO:0010033	response to organic substance
EP300	Hepg2	GO:0007165	signal transduction
EP300	Hepg2	GO:0042221	response to chemical
EP300	Hepg2	GO:0051716	cellular response to stimulus
EP300	T47d	GO:0022037	metencephalon development
FOS	Gm12878	GO:0007219	Notch signaling pathway
FOS	Helas3	GO:0007267	cell-cell signaling
FOS	Helas3	GO:0044700	single organism signaling
FOS	Helas3	GO:0023052	signaling
FOS	Helas3	GO:0009653	anatomical structure morphogenesis
FOS	Huvec	GO:0032647	regulation of interferon-alpha production
FOS	Huvec	GO:0032727	positive regulation of interferon-alpha production
FOS	Huvec	GO:0060333	interferon-gamma-mediated signaling pathway
FOS	Huvec	GO:0032728	positive regulation of interferon-beta production
FOS	Huvec	GO:0060337	type I interferon signaling pathway
FOS	Huvec	GO:0032648	regulation of interferon-beta production
FOS	Huvec	GO:0019221	cytokine-mediated signaling pathway
FOS	Huvec	GO:0032481	positive regulation of type I interferon production
FOS	Huvec	GO:0045088	regulation of innate immune response
FOS	Huvec	GO:0050776	regulation of immune response
FOS	Huvec	GO:0006952	defense response
FOS	K562	GO:0070345	negative regulation of fat cell proliferation
FOS	K562	GO:0070344	regulation of fat cell proliferation
FOS	K562	GO:1900739	regulation of protein insertion into mitochondrial membrane
FOS	K562	GO:1900740	positive regulation of protein insertion into mitochondrial me
FOS	K562	GO:1901028	regulation of mitochondrial outer membrane permeabilizatio
FOS	K562	GO:1901030	positive regulation of mitochondrial outer membrane permea pathway
FOS	K562	GO:1901724	positive regulation of cell proliferation involved in kidney dev
FOS	K562	GO:0000278	mitotic cell cycle
FOS	K562	GO:0007049	cell cycle
FOS	K562	GO:0007219	Notch signaling pathway
JUN	Mcf10a	GO:0031100	organ regeneration
JUN	Gm12878	GO:0055072	iron ion homeostasis
JUN	Gm12878	GO:0006879	cellular iron ion homeostasis
JUN	Gm12878	GO:0046916	cellular transition metal ion homeostasis
JUN	Gm12878	GO:0055076	transition metal ion homeostasis
JUN	Gm12878	GO:0098771	inorganic ion homeostasis
JUN	Gm12878	GO:0055080	cation homeostasis
JUN	Gm12878	GO:0006873	cellular ion homeostasis
JUN	Gm12878	GO:0030003	cellular cation homeostasis

JUN	Gm12878	GO:0060249	anatomical structure homeostasis
JUN	Gm12878	GO:0055065	metal ion homeostasis
JUN	Gm12878	GO:0050801	ion homeostasis
JUN	Gm12878	GO:0042592	homeostatic process
JUN	H1hesc	GO:0006260	DNA replication
JUN	Helas3	GO:0045672	positive regulation of osteoclast differentiation
JUN	Helas3	GO:0002763	positive regulation of myeloid leukocyte differentiation
JUN	Helas3	GO:0045639	positive regulation of myeloid cell differentiation
JUN	Helas3	GO:0031098	stress-activated protein kinase signaling cascade
JUN	Helas3	GO:0051403	stress-activated MAPK cascade
JUN	Helas3	GO:0023014	signal transduction by protein phosphorylation
JUN	Helas3	GO:0000165	MAPK cascade
JUN	Helas3	GO:0034166	toll-like receptor 10 signaling pathway
JUN	Helas3	GO:0002755	MyD88-dependent toll-like receptor signaling pathway
JUN	Helas3	GO:0034134	toll-like receptor 2 signaling pathway
JUN	Helas3	GO:0034146	toll-like receptor 5 signaling pathway
JUN	Helas3	GO:0038124	toll-like receptor TLR6:TLR2 signaling pathway
JUN	Helas3	GO:0038123	toll-like receptor TLR1:TLR2 signaling pathway
JUN	Helas3	GO:0034162	toll-like receptor 9 signaling pathway
JUN	Helas3	GO:0007611	learning or memory
JUN	Helas3	GO:0002761	regulation of myeloid leukocyte differentiation
JUN	Helas3	GO:1902107	positive regulation of leukocyte differentiation
JUN	Helas3	GO:0000302	response to reactive oxygen species
JUN	Helas3	GO:0002221	pattern recognition receptor signaling pathway
JUN	Helas3	GO:0002224	toll-like receptor signaling pathway
JUN	Helas3	GO:0002756	MyD88-independent toll-like receptor signaling pathway
JUN	Helas3	GO:0034142	toll-like receptor 4 signaling pathway
JUN	Helas3	GO:0034138	toll-like receptor 3 signaling pathway
JUN	Helas3	GO:0038093	Fc receptor signaling pathway
JUN	Helas3	GO:0038095	Fc-epsilon receptor signaling pathway
JUN	Helas3	GO:0050890	cognition
JUN	Helas3	GO:0035666	TRIF-dependent toll-like receptor signaling pathway
JUN	Helas3	GO:0006468	protein phosphorylation
JUN	Helas3	GO:0016310	phosphorylation
JUN	Helas3	GO:0034097	response to cytokine
JUN	Helas3	GO:0045089	positive regulation of innate immune response
JUN	Helas3	GO:0031349	positive regulation of defense response
JUN	Helas3	GO:1903708	positive regulation of hemopoiesis
JUN	Helas3	GO:0045088	regulation of innate immune response
JUN	Helas3	GO:0006796	phosphate-containing compound metabolic process
JUN	Helas3	GO:0006793	phosphorus metabolic process
JUN	Helas3	GO:0050778	positive regulation of immune response
JUN	Helas3	GO:0045637	regulation of myeloid cell differentiation

JUN	Helas3	GO:0031347	regulation of defense response
JUN	Helas3	GO:0045087	innate immune response
JUN	Helas3	GO:0010035	response to inorganic substance
JUN	Helas3	GO:0050776	regulation of immune response
JUN	Helas3	GO:1902105	regulation of leukocyte differentiation
JUN	Helas3	GO:0048511	rhythmic process
JUN	Helas3	GO:0006955	immune response
JUN	Helas3	GO:0035556	intracellular signal transduction
JUN	Helas3	GO:0006952	defense response
JUN	Helas3	GO:0051704	multi-organism process
JUN	Helas3	GO:0002376	immune system process
JUN	Helas3	GO:0080134	regulation of response to stress
JUN	Helas3	GO:0070887	cellular response to chemical stimulus
JUN	Helas3	GO:0010033	response to organic substance
JUN	Helas3	GO:0042221	response to chemical
JUN	Hepg2	GO:1901700	response to oxygen-containing compound
JUN	K562	GO:0070345	negative regulation of fat cell proliferation
JUN	K562	GO:0070344	regulation of fat cell proliferation
JUN	K562	GO:1904018	positive regulation of vasculature development
JUN	K562	GO:0000278	mitotic cell cycle
JUN	K562	GO:0007049	cell cycle
JUND	Gm12878	GO:0060333	interferon-gamma-mediated signaling pathway
JUND	Gm12878	GO:0007259	JAK-STAT cascade
JUND	Gm12878	GO:0060338	regulation of type I interferon-mediated signaling pathway
JUND	Gm12878	GO:0060397	JAK-STAT cascade involved in growth hormone signaling p
JUND	Gm12878	GO:0060337	type I interferon signaling pathway
JUND	Gm12878	GO:0019221	cytokine-mediated signaling pathway
JUND	Gm12878	GO:0051607	defense response to virus
JUND	Gm12878	GO:0002697	regulation of immune effector process
JUND	Gm12878	GO:0098542	defense response to other organism
JUND	Gm12878	GO:0032481	positive regulation of type I interferon production
JUND	Gm12878	GO:0007166	cell surface receptor signaling pathway
JUND	H1hesc	GO:0008340	determination of adult lifespan
JUND	H1hesc	GO:1901983	regulation of protein acetylation
JUND	Helas3	GO:0001759	organ induction
JUND	Helas3	GO:0060272	embryonic skeletal joint morphogenesis
JUND	Helas3	GO:0009954	proximal/distal pattern formation
JUND	Hepg2	GO:0055088	lipid homeostasis
JUND	Hepg2	GO:0033500	carbohydrate homeostasis
JUND	Hepg2	GO:0042593	glucose homeostasis
JUND	Hepg2	GO:0048878	chemical homeostasis
JUND	Hepg2	GO:0042592	homeostatic process
JUND	Hepg2	GO:0003006	developmental process involved in reproduction

JUND	Huvec	GO:0045333	cellular respiration
JUND	Huvec	GO:0002223	stimulatory C-type lectin receptor signaling pathway
JUND	Huvec	GO:0002220	innate immune response activating cell surface receptor sig
JUND	Huvec	GO:0050778	positive regulation of immune response
JUND	K562	GO:0045766	positive regulation of angiogenesis
JUND	K562	GO:0035162	embryonic hemopoiesis
JUND	K562	GO:1904018	positive regulation of vasculature development
JUND	K562	GO:0001776	leukocyte homeostasis
JUND	K562	GO:0043627	response to estrogen
JUND	Gm12878	GO:0032647	regulation of interferon-alpha production
JUND	Gm12878	GO:0032727	positive regulation of interferon-alpha production
JUND	Gm12878	GO:0060333	interferon-gamma-mediated signaling pathway
JUND	Gm12878	GO:0032728	positive regulation of interferon-beta production
JUND	Gm12878	GO:0060337	type I interferon signaling pathway
JUND	Gm12878	GO:0032648	regulation of interferon-beta production
JUND	Gm12878	GO:0019221	cytokine-mediated signaling pathway
JUND	Gm12878	GO:0002252	immune effector process
JUND	H1hesc	GO:0031647	regulation of protein stability
JUND	Helas3	GO:0006094	gluconeogenesis
JUND	Helas3	GO:0061394	regulation of transcription from RNA polymerase II promoter
JUND	Helas3	GO:0016051	carbohydrate biosynthetic process
JUND	Helas3	GO:0019319	hexose biosynthetic process
JUND	Helas3	GO:0046364	monosaccharide biosynthetic process
JUND	Hepg2	GO:0060174	limb bud formation
JUND	Hepg2	GO:0060441	epithelial tube branching involved in lung morphogenesis
JUND	Hepg2	GO:0031018	endocrine pancreas development
JUND	Hepg2	GO:0033500	carbohydrate homeostasis
JUND	Hepg2	GO:0042593	glucose homeostasis
JUND	K562	GO:0070345	negative regulation of fat cell proliferation
JUND	K562	GO:0070344	regulation of fat cell proliferation
JUND	K562	GO:0000278	mitotic cell cycle
JUND	Sknsh	GO:0007259	JAK-STAT cascade
JUND	Sknsh	GO:0060397	JAK-STAT cascade involved in growth hormone signaling p
JUND	Sknsh	GO:0042542	response to hydrogen peroxide
JUND	Sknsh	GO:0000302	response to reactive oxygen species
JUND	Sknsh	GO:0006979	response to oxidative stress
MAFK	H1hesc	GO:0070345	negative regulation of fat cell proliferation
MAFK	H1hesc	GO:0070344	regulation of fat cell proliferation
MAFK	H1hesc	GO:1900739	regulation of protein insertion into mitochondrial membrane
MAFK	H1hesc	GO:1900740	positive regulation of protein insertion into mitochondrial me
MAFK	H1hesc	GO:1901028	regulation of mitochondrial outer membrane permeabilizatio
MAFK	H1hesc	GO:1901030	positive regulation of mitochondrial outer membrane permea pathway

MAFK	H1hesc	GO:0044843	cell cycle G1/S phase transition
MAFK	H1hesc	GO:0000082	G1/S transition of mitotic cell cycle
MAFK	H1hesc	GO:0000278	mitotic cell cycle
MAFK	H1hesc	GO:0044772	mitotic cell cycle phase transition
MAFK	H1hesc	GO:0044770	cell cycle phase transition
MAFK	H1hesc	GO:0007049	cell cycle
MAFK	H1hesc	GO:2001235	positive regulation of apoptotic signaling pathway
MAFK	H1hesc	GO:1903047	mitotic cell cycle process
MAFK	Hepg2	GO:0040020	regulation of meiotic nuclear division
MAFK	Hepg2	GO:0045836	positive regulation of meiotic nuclear division
MAFK	Hepg2	GO:0090427	activation of meiosis
MAFK	Hepg2	GO:0051446	positive regulation of meiotic cell cycle
MAFK	Hepg2	GO:0051445	regulation of meiotic cell cycle
MAFK	Hepg2	GO:0035880	embryonic nail plate morphogenesis
MAFK	Hepg2	GO:0071407	cellular response to organic cyclic compound
MAFK	Imr90	GO:0045945	positive regulation of transcription from RNA polymerase III
MAFK	K562	GO:0051385	response to mineralocorticoid
MAFK	K562	GO:0051412	response to corticosterone
MAFK	K562	GO:0032570	response to progesterone
MAFK	K562	GO:0071277	cellular response to calcium ion
MAFK	K562	GO:0051592	response to calcium ion
MAFK	K562	GO:0007565	female pregnancy
MAFK	K562	GO:0051384	response to glucocorticoid
MAFK	K562	GO:0031960	response to corticosteroid
MAFK	K562	GO:0051591	response to cAMP
MAFK	K562	GO:0046683	response to organophosphorus
MAFK	K562	GO:0014074	response to purine-containing compound
MAFK	K562	GO:0009612	response to mechanical stimulus
MAFK	K562	GO:0071248	cellular response to metal ion
MAFK	K562	GO:1901654	response to ketone
MAFK	K562	GO:0071241	cellular response to inorganic substance
MAFK	K562	GO:0010038	response to metal ion
MAFK	K562	GO:0042493	response to drug
MAFK	K562	GO:0034097	response to cytokine
MAFK	K562	GO:0032496	response to lipopolysaccharide
MAFK	K562	GO:0032870	cellular response to hormone stimulus
MAFK	K562	GO:0010035	response to inorganic substance
MAFK	K562	GO:0010243	response to organonitrogen compound
MAFK	K562	GO:0009628	response to abiotic stimulus
MAFK	K562	GO:1901698	response to nitrogen compound
MAFK	K562	GO:0044267	cellular protein metabolic process
MAFK	K562	GO:0009725	response to hormone
MAFK	K562	GO:0071495	cellular response to endogenous stimulus

MAFK	K562	GO:0033993	response to lipid
MAFK	K562	GO:0014070	response to organic cyclic compound
MAFK	K562	GO:0006366	transcription from RNA polymerase II promoter
MAZ	Gm12878	GO:0060333	interferon-gamma-mediated signaling pathway
MAZ	Gm12878	GO:0032647	regulation of interferon-alpha production
MAZ	Gm12878	GO:0032727	positive regulation of interferon-alpha production
MAZ	Gm12878	GO:0070345	negative regulation of fat cell proliferation
MAZ	Gm12878	GO:0070344	regulation of fat cell proliferation
MAZ	Gm12878	GO:0060337	type I interferon signaling pathway
MAZ	Gm12878	GO:0032728	positive regulation of interferon-beta production
MAZ	Gm12878	GO:0032648	regulation of interferon-beta production
MAZ	Gm12878	GO:0019221	cytokine-mediated signaling pathway
MAZ	Gm12878	GO:0051607	defense response to virus
MAZ	Gm12878	GO:0098542	defense response to other organism
MAZ	Gm12878	GO:0001819	positive regulation of cytokine production
MAZ	Gm12878	GO:0016032	viral process
MAZ	Gm12878	GO:0044764	multi-organism cellular process
MAZ	Gm12878	GO:0044403	symbiosis, encompassing mutualism through parasitism
MAZ	Gm12878	GO:0045088	regulation of innate immune response
MAZ	Gm12878	GO:0001817	regulation of cytokine production
MAZ	Gm12878	GO:0050776	regulation of immune response
MAZ	Gm12878	GO:0035556	intracellular signal transduction
MAZ	Gm12878	GO:0007166	cell surface receptor signaling pathway
MAZ	Gm12878	GO:0051704	multi-organism process
MAZ	Gm12878	GO:0007165	signal transduction
MAZ	Gm12878	GO:0006950	response to stress
MAZ	Hepg2	GO:0030335	positive regulation of cell migration
MAZ	Hepg2	GO:0040017	positive regulation of locomotion
MAZ	Hepg2	GO:2000147	positive regulation of cell motility
MAZ	K562	GO:2000352	negative regulation of endothelial cell apoptotic process
MAZ	K562	GO:0060575	intestinal epithelial cell differentiation
MAZ	K562	GO:0050817	coagulation
MAZ	K562	GO:0007596	blood coagulation
MAZ	K562	GO:0007599	hemostasis
MAZ	K562	GO:0001701	in utero embryonic development
MAZ	K562	GO:0043009	chordate embryonic development
MAZ	K562	GO:0050878	regulation of body fluid levels
MAZ	K562	GO:0009792	embryo development ending in birth or egg hatching
MAZ	K562	GO:0009790	embryo development
MXI1	Gm12878	GO:0032647	regulation of interferon-alpha production
MXI1	Gm12878	GO:0032727	positive regulation of interferon-alpha production
MXI1	Gm12878	GO:0032728	positive regulation of interferon-beta production
MXI1	Gm12878	GO:0060333	interferon-gamma-mediated signaling pathway

MXI1	Gm12878	GO:0060337	type I interferon signaling pathway
MXI1	Gm12878	GO:0032648	regulation of interferon-beta production
MXI1	Gm12878	GO:0098542	defense response to other organism
MXI1	Gm12878	GO:0019221	cytokine-mediated signaling pathway
MXI1	Gm12878	GO:0051707	response to other organism
MXI1	Gm12878	GO:0033554	cellular response to stress
MXI1	H1hesc	GO:0006260	DNA replication
MXI1	H1hesc	GO:0033044	regulation of chromosome organization
MXI1	H1hesc	GO:0006259	DNA metabolic process
MXI1	Helas3	GO:0006260	DNA replication
MXI1	Helas3	GO:0045600	positive regulation of fat cell differentiation
MXI1	Helas3	GO:0045598	regulation of fat cell differentiation
MXI1	Helas3	GO:0016032	viral process
MXI1	Helas3	GO:0044764	multi-organism cellular process
MXI1	Helas3	GO:0044403	symbiosis, encompassing mutualism through parasitism
MXI1	Helas3	GO:0044419	interspecies interaction between organisms
MXI1	Helas3	GO:1901698	response to nitrogen compound
MXI1	Hepg2	GO:0070345	negative regulation of fat cell proliferation
MXI1	Hepg2	GO:0070344	regulation of fat cell proliferation
MXI1	Hepg2	GO:0006953	acute-phase response
MXI1	Hepg2	GO:0002526	acute inflammatory response
MXI1	Hepg2	GO:0007005	mitochondrion organization
MXI1	Hepg2	GO:0000278	mitotic cell cycle
MXI1	Hepg2	GO:0006952	defense response
MXI1	K562	GO:0051385	response to mineralocorticoid
MXI1	K562	GO:0051412	response to corticosterone
MXI1	K562	GO:0051384	response to glucocorticoid
MXI1	K562	GO:0031960	response to corticosteroid
MXI1	K562	GO:0032570	response to progesterone
MXI1	K562	GO:0071277	cellular response to calcium ion
MXI1	K562	GO:0051592	response to calcium ion
MXI1	K562	GO:0009612	response to mechanical stimulus
MXI1	K562	GO:0051591	response to cAMP
MXI1	K562	GO:0046683	response to organophosphorus
MXI1	K562	GO:0071241	cellular response to inorganic substance
MXI1	K562	GO:0014074	response to purine-containing compound
MXI1	K562	GO:0071248	cellular response to metal ion
MXI1	K562	GO:0042493	response to drug
MXI1	K562	GO:0010035	response to inorganic substance
MXI1	K562	GO:0032870	cellular response to hormone stimulus
MXI1	K562	GO:0048545	response to steroid hormone
MXI1	K562	GO:0009725	response to hormone
MXI1	K562	GO:0009628	response to abiotic stimulus

MXI1	K562	GO:0010243	response to organonitrogen compound
MXI1	K562	GO:0071495	cellular response to endogenous stimulus
MXI1	K562	GO:1901698	response to nitrogen compound
MXI1	K562	GO:0014070	response to organic cyclic compound
MXI1	K562	GO:0009605	response to external stimulus
MXI1	K562	GO:0033993	response to lipid
MXI1	K562	GO:0009719	response to endogenous stimulus
MXI1	K562	GO:0071310	cellular response to organic substance
MXI1	K562	GO:0006366	transcription from RNA polymerase II promoter
MYC	A549	GO:0007259	JAK-STAT cascade
MYC	A549	GO:0060397	JAK-STAT cascade involved in growth hormone signaling p
MYC	A549	GO:0002705	positive regulation of leukocyte mediated immunity
MYC	A549	GO:0002708	positive regulation of lymphocyte mediated immunity
MYC	A549	GO:0002699	positive regulation of immune effector process
MYC	A549	GO:0040014	regulation of multicellular organism growth
MYC	A549	GO:0002697	regulation of immune effector process
MYC	A549	GO:0019221	cytokine-mediated signaling pathway
MYC	Gm12878	GO:0070345	negative regulation of fat cell proliferation
MYC	Gm12878	GO:0070344	regulation of fat cell proliferation
MYC	Gm12878	GO:1900739	regulation of protein insertion into mitochondrial membrane
MYC	Gm12878	GO:1900740	positive regulation of protein insertion into mitochondrial me
MYC	Gm12878	GO:1901028	regulation of mitochondrial outer membrane permeabilizatio
MYC	Gm12878	GO:1901030	positive regulation of mitochondrial outer membrane permea pathway
MYC	Gm12878	GO:0000278	mitotic cell cycle
MYC	Gm12878	GO:0007049	cell cycle
MYC	Gm12878	GO:0007219	Notch signaling pathway
MYC	H1hesc	GO:0060174	limb bud formation
MYC	H1hesc	GO:0042246	tissue regeneration
MYC	H1hesc	GO:0021781	glial cell fate commitment
MYC	H1hesc	GO:0021510	spinal cord development
MYC	H1hesc	GO:0045165	cell fate commitment
MYC	Hepg2	GO:0048732	gland development
MYC	Huvec	GO:0000075	cell cycle checkpoint
MYC	Huvec	GO:0000077	DNA damage checkpoint
MYC	Huvec	GO:0031570	DNA integrity checkpoint
MYC	Huvec	GO:0006974	cellular response to DNA damage stimulus
MYC	Huvec	GO:0035556	intracellular signal transduction
MYC	K562	GO:0060575	intestinal epithelial cell differentiation
MYC	K562	GO:0034698	response to gonadotropin
MYC	K562	GO:0071371	cellular response to gonadotropin stimulus
MYC	K562	GO:0007267	cell-cell signaling
MYC	K562	GO:0044700	single organism signaling

MYC	K562	GO:0023052	signaling
MYC	Mcf7	GO:0000429	carbon catabolite regulation of transcription from RNA polyr
MYC	Mcf7	GO:0000430	regulation of transcription from RNA polymerase II promote
MYC	Mcf7	GO:0000432	positive regulation of transcription from RNA polymerase II
MYC	Mcf7	GO:0000436	carbon catabolite activation of transcription from RNA polym
MYC	Mcf7	GO:0036342	post-anal tail morphogenesis
MYC	Mcf7	GO:0019086	late viral transcription
MYC	Mcf7	GO:0019083	viral transcription
NRF1	H1hesc	GO:0000278	mitotic cell cycle
NRF1	H1hesc	GO:0010927	cellular component assembly involved in morphogenesis
NRF1	H1hesc	GO:0030030	cell projection organization
NRF1	H1hesc	GO:0022607	cellular component assembly
NRF1	H1hesc	GO:0071840	cellular component organization or biogenesis
NRF1	H1hesc	GO:0016043	cellular component organization
NRF1	Helas3	GO:1901566	organonitrogen compound biosynthetic process
NRF1	Hepg2	GO:0070345	negative regulation of fat cell proliferation
NRF1	Hepg2	GO:0070344	regulation of fat cell proliferation
NRF1	Hepg2	GO:1900739	regulation of protein insertion into mitochondrial membrane
NRF1	Hepg2	GO:1900740	positive regulation of protein insertion into mitochondrial me
NRF1	Hepg2	GO:1901028	regulation of mitochondrial outer membrane permeabilizatio
NRF1	Hepg2	GO:1901030	positive regulation of mitochondrial outer membrane perme pathway
NRF1	Hepg2	GO:0006260	DNA replication
NRF1	Hepg2	GO:0000278	mitotic cell cycle
NRF1	Hepg2	GO:0007049	cell cycle
NRF1	K562	GO:0051385	response to mineralocorticoid
NRF1	K562	GO:0051412	response to corticosterone
NRF1	K562	GO:0045672	positive regulation of osteoclast differentiation
NRF1	K562	GO:0009629	response to gravity
NRF1	K562	GO:0060430	lung saccule development
NRF1	K562	GO:0032570	response to progesterone
NRF1	K562	GO:0071277	cellular response to calcium ion
NRF1	K562	GO:0051592	response to calcium ion
NRF1	K562	GO:0002763	positive regulation of myeloid leukocyte differentiation
NRF1	K562	GO:0007565	female pregnancy
NRF1	K562	GO:0051384	response to glucocorticoid
NRF1	K562	GO:0031960	response to corticosteroid
NRF1	K562	GO:0051591	response to cAMP
NRF1	K562	GO:0046683	response to organophosphorus
NRF1	K562	GO:0031098	stress-activated protein kinase signaling cascade
NRF1	K562	GO:0007612	learning
NRF1	K562	GO:0051403	stress-activated MAPK cascade
NRF1	K562	GO:0014074	response to purine-containing compound

NRF1	K562	GO:0071248	cellular response to metal ion
NRF1	K562	GO:0071241	cellular response to inorganic substance
NRF1	K562	GO:0007611	learning or memory
NRF1	K562	GO:0009612	response to mechanical stimulus
NRF1	K562	GO:1901654	response to ketone
NRF1	K562	GO:0045639	positive regulation of myeloid cell differentiation
NRF1	K562	GO:0023014	signal transduction by protein phosphorylation
NRF1	K562	GO:0000165	MAPK cascade
NRF1	K562	GO:0044706	multi-multicellular organism process
NRF1	K562	GO:0050890	cognition
NRF1	K562	GO:0010038	response to metal ion
NRF1	K562	GO:0044703	multi-organism reproductive process
NRF1	K562	GO:0042493	response to drug
NRF1	K562	GO:0097305	response to alcohol
NRF1	K562	GO:0010035	response to inorganic substance
NRF1	K562	GO:0010243	response to organonitrogen compound
NRF1	K562	GO:0007623	circadian rhythm
NRF1	K562	GO:1901698	response to nitrogen compound
NRF1	K562	GO:0048545	response to steroid hormone
NRF1	K562	GO:0032870	cellular response to hormone stimulus
NRF1	K562	GO:0034097	response to cytokine
NRF1	K562	GO:0009725	response to hormone
NRF1	K562	GO:0014070	response to organic cyclic compound
NRF1	K562	GO:0009628	response to abiotic stimulus
NRF1	K562	GO:0071495	cellular response to endogenous stimulus
NRF1	K562	GO:0009719	response to endogenous stimulus
NRF1	K562	GO:0033993	response to lipid
NRF1	K562	GO:0051704	multi-organism process
NRF1	K562	GO:1901700	response to oxygen-containing compound
NRF1	K562	GO:0009605	response to external stimulus
NRF1	K562	GO:0010033	response to organic substance
NRF1	K562	GO:0070887	cellular response to chemical stimulus
NRF1	K562	GO:0042221	response to chemical
NRF1	K562	GO:0051716	cellular response to stimulus
REST	A549	GO:0051385	response to mineralocorticoid
REST	A549	GO:0051412	response to corticosterone
REST	A549	GO:0051384	response to glucocorticoid
REST	A549	GO:0031960	response to corticosteroid
REST	A549	GO:0031098	stress-activated protein kinase signaling cascade
REST	A549	GO:0007612	learning
REST	A549	GO:0051403	stress-activated MAPK cascade
REST	A549	GO:0032570	response to progesterone
REST	A549	GO:0071277	cellular response to calcium ion

REST	A549	GO:0051592	response to calcium ion
REST	A549	GO:0071241	cellular response to inorganic substance
REST	A549	GO:0007565	female pregnancy
REST	A549	GO:0071248	cellular response to metal ion
REST	A549	GO:0007611	learning or memory
REST	A549	GO:0023014	signal transduction by protein phosphorylation
REST	A549	GO:0000165	MAPK cascade
REST	A549	GO:0050890	cognition
REST	A549	GO:0034166	toll-like receptor 10 signaling pathway
REST	A549	GO:0002755	MyD88-dependent toll-like receptor signaling pathway
REST	A549	GO:0034134	toll-like receptor 2 signaling pathway
REST	A549	GO:0034146	toll-like receptor 5 signaling pathway
REST	A549	GO:0038124	toll-like receptor TLR6:TLR2 signaling pathway
REST	A549	GO:0038123	toll-like receptor TLR1:TLR2 signaling pathway
REST	A549	GO:0051591	response to cAMP
REST	A549	GO:0046683	response to organophosphorus
REST	A549	GO:0010038	response to metal ion
REST	A549	GO:0014074	response to purine-containing compound
REST	A549	GO:0009612	response to mechanical stimulus
REST	A549	GO:0045598	regulation of fat cell differentiation
REST	A549	GO:1901654	response to ketone
REST	A549	GO:0010035	response to inorganic substance
REST	A549	GO:0097305	response to alcohol
REST	A549	GO:0010243	response to organonitrogen compound
REST	A549	GO:1901698	response to nitrogen compound
REST	A549	GO:0042493	response to drug
REST	A549	GO:0048545	response to steroid hormone
REST	A549	GO:0009725	response to hormone
REST	A549	GO:0009628	response to abiotic stimulus
REST	A549	GO:0014070	response to organic cyclic compound
REST	A549	GO:0009605	response to external stimulus
REST	A549	GO:0033993	response to lipid
REST	A549	GO:0006952	defense response
REST	A549	GO:0051726	regulation of cell cycle
REST	A549	GO:1901700	response to oxygen-containing compound
REST	A549	GO:0009719	response to endogenous stimulus
REST	A549	GO:0003006	developmental process involved in reproduction
REST	A549	GO:0022414	reproductive process
REST	A549	GO:0071310	cellular response to organic substance
REST	A549	GO:0010033	response to organic substance
REST	A549	GO:0070887	cellular response to chemical stimulus
REST	A549	GO:0042221	response to chemical
REST	A549	GO:0051716	cellular response to stimulus

REST	A549	GO:0050896	response to stimulus
REST	K562	GO:0060575	intestinal epithelial cell differentiation
REST	K562	GO:0014888	striated muscle adaptation
REST	K562	GO:0014887	cardiac muscle adaptation
REST	K562	GO:0014897	striated muscle hypertrophy
REST	K562	GO:0014896	muscle hypertrophy
REST	K562	GO:0014898	cardiac muscle hypertrophy in response to stress
REST	K562	GO:0043500	muscle adaptation
REST	K562	GO:0003299	muscle hypertrophy in response to stress
REST	K562	GO:0003300	cardiac muscle hypertrophy
REST	K562	GO:0051890	regulation of cardioblast differentiation
REST	K562	GO:0051891	positive regulation of cardioblast differentiation
REST	K562	GO:0045766	positive regulation of angiogenesis
REST	K562	GO:0003281	ventricular septum development
REST	K562	GO:0003012	muscle system process
REST	K562	GO:0071773	cellular response to BMP stimulus
REST	K562	GO:0071772	response to BMP
REST	K562	GO:1904018	positive regulation of vasculature development
REST	K562	GO:0055023	positive regulation of cardiac muscle tissue growth
REST	K562	GO:0055025	positive regulation of cardiac muscle tissue development
REST	K562	GO:0003215	cardiac right ventricle morphogenesis
REST	K562	GO:0003279	cardiac septum development
REST	K562	GO:0060045	positive regulation of cardiac muscle cell proliferation
REST	K562	GO:0050817	coagulation
REST	K562	GO:0007596	blood coagulation
REST	K562	GO:0007599	hemostasis
REST	K562	GO:0043627	response to estrogen
REST	K562	GO:0008584	male gonad development
REST	K562	GO:0008406	gonad development
REST	K562	GO:0001701	in utero embryonic development
REST	K562	GO:0043009	chordate embryonic development
REST	K562	GO:0050878	regulation of body fluid levels
REST	K562	GO:0009792	embryo development ending in birth or egg hatching
REST	K562	GO:0009790	embryo development
REST	K562	GO:0042493	response to drug
REST	K562	GO:0022603	regulation of anatomical structure morphogenesis
REST	K562	GO:0048646	anatomical structure formation involved in morphogenesis
REST	Panc1	GO:1901741	positive regulation of myoblast fusion
REST	Panc1	GO:1901739	regulation of myoblast fusion
REST	Panc1	GO:0051149	positive regulation of muscle cell differentiation
REST	Panc1	GO:0030183	B cell differentiation
REST	Panc1	GO:0042113	B cell activation
REST	Panc1	GO:0045666	positive regulation of neuron differentiation

REST	Panc1	GO:0010720	positive regulation of cell development
REST	Panc1	GO:0060284	regulation of cell development
REST	Panc1	GO:0045597	positive regulation of cell differentiation
REST	Panc1	GO:0051094	positive regulation of developmental process
REST	Panc1	GO:0051240	positive regulation of multicellular organismal process
REST	Pfsk1	GO:0042445	hormone metabolic process
REST	Pfsk1	GO:0010817	regulation of hormone levels
REST	Pfsk1	GO:0007154	cell communication
REST	U87	GO:0045672	positive regulation of osteoclast differentiation
REST	U87	GO:0055072	iron ion homeostasis
REST	U87	GO:0002763	positive regulation of myeloid leukocyte differentiation
REST	U87	GO:0023014	signal transduction by protein phosphorylation
REST	U87	GO:0000165	MAPK cascade
REST	U87	GO:0031098	stress-activated protein kinase signaling cascade
REST	U87	GO:0051403	stress-activated MAPK cascade
REST	U87	GO:0045639	positive regulation of myeloid cell differentiation
REST	U87	GO:0034166	toll-like receptor 10 signaling pathway
REST	U87	GO:0002755	MyD88-dependent toll-like receptor signaling pathway
REST	U87	GO:0034134	toll-like receptor 2 signaling pathway
REST	U87	GO:0034146	toll-like receptor 5 signaling pathway
REST	U87	GO:0038124	toll-like receptor TLR6:TLR2 signaling pathway
REST	U87	GO:0038123	toll-like receptor TLR1:TLR2 signaling pathway
REST	U87	GO:0034162	toll-like receptor 9 signaling pathway
REST	U87	GO:0007611	learning or memory
REST	U87	GO:0002761	regulation of myeloid leukocyte differentiation
REST	U87	GO:0000302	response to reactive oxygen species
REST	U87	GO:0002221	pattern recognition receptor signaling pathway
REST	U87	GO:0002224	toll-like receptor signaling pathway
REST	U87	GO:0002756	MyD88-independent toll-like receptor signaling pathway
REST	U87	GO:0034142	toll-like receptor 4 signaling pathway
REST	U87	GO:0034138	toll-like receptor 3 signaling pathway
REST	U87	GO:0038093	Fc receptor signaling pathway
REST	U87	GO:0038095	Fc-epsilon receptor signaling pathway
REST	U87	GO:0050890	cognition
REST	U87	GO:0007179	transforming growth factor beta receptor signaling pathway
REST	U87	GO:0035666	TRIF-dependent toll-like receptor signaling pathway
REST	U87	GO:0006468	protein phosphorylation
REST	U87	GO:0016310	phosphorylation
REST	U87	GO:0043620	regulation of DNA-templated transcription in response to str
REST	U87	GO:0043618	regulation of transcription from RNA polymerase II promote
REST	U87	GO:0051345	positive regulation of hydrolase activity
REST	U87	GO:0006796	phosphate-containing compound metabolic process
REST	U87	GO:0006793	phosphorus metabolic process

REST	U87	GO:0045637	regulation of myeloid cell differentiation
REST	U87	GO:0006464	cellular protein modification process
REST	U87	GO:0036211	protein modification process
REST	U87	GO:0043412	macromolecule modification
REST	U87	GO:0010035	response to inorganic substance
REST	U87	GO:1902105	regulation of leukocyte differentiation
REST	U87	GO:0044267	cellular protein metabolic process
REST	U87	GO:0033554	cellular response to stress
REST	U87	GO:0019538	protein metabolic process
REST	U87	GO:0080134	regulation of response to stress
REST	U87	GO:0070887	cellular response to chemical stimulus
REST	U87	GO:0042221	response to chemical
REST	U87	GO:0007165	signal transduction
RFX5	Gm12878	GO:0051385	response to mineralocorticoid
RFX5	Gm12878	GO:0051412	response to corticosterone
RFX5	Gm12878	GO:0002762	negative regulation of myeloid leukocyte differentiation
RFX5	Gm12878	GO:0048146	positive regulation of fibroblast proliferation
RFX5	Gm12878	GO:0032570	response to progesterone
RFX5	Gm12878	GO:0071277	cellular response to calcium ion
RFX5	Gm12878	GO:0051592	response to calcium ion
RFX5	Gm12878	GO:0051384	response to glucocorticoid
RFX5	Gm12878	GO:0031960	response to corticosteroid
RFX5	Gm12878	GO:0051591	response to cAMP
RFX5	Gm12878	GO:0046683	response to organophosphorus
RFX5	Gm12878	GO:0002761	regulation of myeloid leukocyte differentiation
RFX5	Gm12878	GO:0014074	response to purine-containing compound
RFX5	Gm12878	GO:0009612	response to mechanical stimulus
RFX5	Gm12878	GO:0007178	transmembrane receptor protein serine/threonine kinase sig
RFX5	Gm12878	GO:0045637	regulation of myeloid cell differentiation
RFX5	Gm12878	GO:0042493	response to drug
RFX5	H1hesc	GO:0051385	response to mineralocorticoid
RFX5	H1hesc	GO:0051412	response to corticosterone
RFX5	H1hesc	GO:0070345	negative regulation of fat cell proliferation
RFX5	H1hesc	GO:0070344	regulation of fat cell proliferation
RFX5	H1hesc	GO:0006260	DNA replication
RFX5	H1hesc	GO:0051384	response to glucocorticoid
RFX5	H1hesc	GO:0031960	response to corticosteroid
RFX5	H1hesc	GO:0032570	response to progesterone
RFX5	H1hesc	GO:0071277	cellular response to calcium ion
RFX5	H1hesc	GO:0051592	response to calcium ion
RFX5	H1hesc	GO:0007565	female pregnancy
RFX5	H1hesc	GO:0051591	response to cAMP
RFX5	H1hesc	GO:0046683	response to organophosphorus

RFX5	H1hesc	GO:0000278	mitotic cell cycle
RFX5	H1hesc	GO:0014074	response to purine-containing compound
RFX5	H1hesc	GO:1901654	response to ketone
RFX5	H1hesc	GO:0007179	transforming growth factor beta receptor signaling pathway
RFX5	H1hesc	GO:0007049	cell cycle
RFX5	H1hesc	GO:0009612	response to mechanical stimulus
RFX5	H1hesc	GO:0071248	cellular response to metal ion
RFX5	H1hesc	GO:0071241	cellular response to inorganic substance
RFX5	H1hesc	GO:0006259	DNA metabolic process
RFX5	H1hesc	GO:0097305	response to alcohol
RFX5	H1hesc	GO:0042493	response to drug
RFX5	H1hesc	GO:0010035	response to inorganic substance
RFX5	H1hesc	GO:0032870	cellular response to hormone stimulus
RFX5	H1hesc	GO:0014070	response to organic cyclic compound
RFX5	H1hesc	GO:0033993	response to lipid
RFX5	Helas3	GO:0007259	JAK-STAT cascade
RFX5	Helas3	GO:0006953	acute-phase response
RFX5	Helas3	GO:0060397	JAK-STAT cascade involved in growth hormone signaling p
RFX5	Helas3	GO:0043603	cellular amide metabolic process
RFX5	Helas3	GO:0002705	positive regulation of leukocyte mediated immunity
RFX5	Helas3	GO:0002708	positive regulation of lymphocyte mediated immunity
RFX5	Helas3	GO:0070345	negative regulation of fat cell proliferation
RFX5	Helas3	GO:0070344	regulation of fat cell proliferation
RFX5	Helas3	GO:0002699	positive regulation of immune effector process
RFX5	Helas3	GO:0002526	acute inflammatory response
RFX5	Helas3	GO:0051385	response to mineralocorticoid
RFX5	Helas3	GO:0051412	response to corticosterone
RFX5	Helas3	GO:0007565	female pregnancy
RFX5	Helas3	GO:0002703	regulation of leukocyte mediated immunity
RFX5	Helas3	GO:0071277	cellular response to calcium ion
RFX5	Helas3	GO:0051592	response to calcium ion
RFX5	Helas3	GO:0051384	response to glucocorticoid
RFX5	Helas3	GO:0031960	response to corticosteroid
RFX5	Helas3	GO:0000302	response to reactive oxygen species
RFX5	Helas3	GO:0002697	regulation of immune effector process
RFX5	Helas3	GO:0071241	cellular response to inorganic substance
RFX5	Helas3	GO:0044703	multi-organism reproductive process
RFX5	Helas3	GO:0019221	cytokine-mediated signaling pathway
RFX5	Helas3	GO:0071248	cellular response to metal ion
RFX5	Helas3	GO:0044706	multi-multicellular organism process
RFX5	Helas3	GO:0007179	transforming growth factor beta receptor signaling pathway
RFX5	Helas3	GO:0051591	response to cAMP
RFX5	Helas3	GO:0046683	response to organophosphorus

RFX5	Helas3	GO:0014074	response to purine-containing compound
RFX5	Helas3	GO:0034097	response to cytokine
RFX5	Helas3	GO:0009612	response to mechanical stimulus
RFX5	Helas3	GO:0006979	response to oxidative stress
RFX5	Helas3	GO:0016310	phosphorylation
RFX5	Helas3	GO:0010035	response to inorganic substance
RFX5	Helas3	GO:0010038	response to metal ion
RFX5	Helas3	GO:0050778	positive regulation of immune response
RFX5	Helas3	GO:0050776	regulation of immune response
RFX5	Helas3	GO:0045088	regulation of innate immune response
RFX5	Helas3	GO:1901698	response to nitrogen compound
RFX5	Helas3	GO:0043207	response to external biotic stimulus
RFX5	Helas3	GO:0032870	cellular response to hormone stimulus
RFX5	Helas3	GO:0009607	response to biotic stimulus
RFX5	Helas3	GO:0010243	response to organonitrogen compound
RFX5	Helas3	GO:0006952	defense response
RFX5	Helas3	GO:0051704	multi-organism process
RFX5	Helas3	GO:0042493	response to drug
RFX5	Helas3	GO:0031347	regulation of defense response
RFX5	Helas3	GO:0035556	intracellular signal transduction
RFX5	Helas3	GO:0014070	response to organic cyclic compound
RFX5	Helas3	GO:0009605	response to external stimulus
RFX5	Helas3	GO:0009725	response to hormone
RFX5	Helas3	GO:0009628	response to abiotic stimulus
RFX5	Helas3	GO:0007166	cell surface receptor signaling pathway
RFX5	Helas3	GO:0071495	cellular response to endogenous stimulus
RFX5	Helas3	GO:0071310	cellular response to organic substance
RFX5	Helas3	GO:0033993	response to lipid
RFX5	Helas3	GO:0002682	regulation of immune system process
RFX5	Helas3	GO:0070887	cellular response to chemical stimulus
RFX5	Helas3	GO:1901700	response to oxygen-containing compound
RFX5	Helas3	GO:0010033	response to organic substance
RFX5	Helas3	GO:0051716	cellular response to stimulus
RFX5	Helas3	GO:0042221	response to chemical
RFX5	Helas3	GO:0006950	response to stress
RFX5	Helas3	GO:0042127	regulation of cell proliferation
RFX5	Helas3	GO:0007165	signal transduction
RFX5	Helas3	GO:0050896	response to stimulus
RFX5	Hepg2	GO:0060441	epithelial tube branching involved in lung morphogenesis
RFX5	K562	GO:0045766	positive regulation of angiogenesis
RFX5	K562	GO:0045765	regulation of angiogenesis
SRF	Gm12878	GO:0055072	iron ion homeostasis
SRF	Gm12878	GO:0071456	cellular response to hypoxia

SRF	Gm12878	GO:0036294	cellular response to decreased oxygen levels
SRF	Gm12878	GO:0071453	cellular response to oxygen levels
SRF	Gm12878	GO:0007219	Notch signaling pathway
SRF	Gm12878	GO:0001666	response to hypoxia
SRF	Gm12878	GO:0036293	response to decreased oxygen levels
SRF	Gm12878	GO:0050877	neurological system process
SRF	Hepg2	GO:0007259	JAK-STAT cascade
SRF	Hepg2	GO:0006953	acute-phase response
SRF	Hepg2	GO:0060397	JAK-STAT cascade involved in growth hormone signaling p
SRF	Hepg2	GO:0043603	cellular amide metabolic process
SRF	Hepg2	GO:0002705	positive regulation of leukocyte mediated immunity
SRF	Hepg2	GO:0002708	positive regulation of lymphocyte mediated immunity
SRF	Hepg2	GO:1901605	alpha-amino acid metabolic process
SRF	Hepg2	GO:0002526	acute inflammatory response
SRF	Hepg2	GO:0019221	cytokine-mediated signaling pathway
SRF	Hepg2	GO:0044281	small molecule metabolic process
SRF	Hepg2	GO:0006952	defense response
SRF	K562	GO:0070345	negative regulation of fat cell proliferation
SRF	K562	GO:0070344	regulation of fat cell proliferation
SRF	K562	GO:0051385	response to mineralocorticoid
SRF	K562	GO:0051412	response to corticosterone
SRF	K562	GO:0071277	cellular response to calcium ion
SRF	K562	GO:0051592	response to calcium ion
SRF	K562	GO:0051384	response to glucocorticoid
SRF	K562	GO:0031960	response to corticosteroid
SRF	K562	GO:0007179	transforming growth factor beta receptor signaling pathway
SRF	K562	GO:0006979	response to oxidative stress
SRF	K562	GO:0051591	response to cAMP
SRF	K562	GO:0046683	response to organophosphorus
SRF	K562	GO:0006366	transcription from RNA polymerase II promoter
TBP	Gm12878	GO:0032647	regulation of interferon-alpha production
TBP	Gm12878	GO:0032727	positive regulation of interferon-alpha production
TBP	Gm12878	GO:0032728	positive regulation of interferon-beta production
TBP	Gm12878	GO:0060333	interferon-gamma-mediated signaling pathway
TBP	Gm12878	GO:0032648	regulation of interferon-beta production
TBP	Gm12878	GO:0060337	type I interferon signaling pathway
TBP	Gm12878	GO:0032481	positive regulation of type I interferon production
TBP	Gm12878	GO:0019221	cytokine-mediated signaling pathway
TBP	H1hesc	GO:0051606	detection of stimulus
TBP	Helas3	GO:0007259	JAK-STAT cascade
TBP	Helas3	GO:0060397	JAK-STAT cascade involved in growth hormone signaling p
TBP	Helas3	GO:0002705	positive regulation of leukocyte mediated immunity
TBP	Helas3	GO:0002708	positive regulation of lymphocyte mediated immunity

TBP	Helas3	GO:0071499	cellular response to laminar fluid shear stress
TBP	Helas3	GO:0034616	response to laminar fluid shear stress
TBP	Helas3	GO:0006953	acute-phase response
TBP	Helas3	GO:0045577	regulation of B cell differentiation
TBP	Helas3	GO:0045579	positive regulation of B cell differentiation
TBP	Helas3	GO:0002699	positive regulation of immune effector process
TBP	Helas3	GO:0042542	response to hydrogen peroxide
TBP	Helas3	GO:0034599	cellular response to oxidative stress
TBP	Helas3	GO:0002697	regulation of immune effector process
TBP	Helas3	GO:0019221	cytokine-mediated signaling pathway
TBP	Helas3	GO:0006979	response to oxidative stress
TBP	Helas3	GO:2000026	regulation of multicellular organismal development
TBP	Hepg2	GO:0065008	regulation of biological quality
TBP	K562	GO:0070345	negative regulation of fat cell proliferation
TBP	K562	GO:0070344	regulation of fat cell proliferation
TCF12	Gm12878	GO:0060333	interferon-gamma-mediated signaling pathway
TCF12	Gm12878	GO:0045075	regulation of interleukin-12 biosynthetic process
TCF12	Gm12878	GO:0032647	regulation of interferon-alpha production
TCF12	Gm12878	GO:0097028	dendritic cell differentiation
TCF12	Gm12878	GO:0002753	cytoplasmic pattern recognition receptor signaling pathway
TCF12	Gm12878	GO:0045084	positive regulation of interleukin-12 biosynthetic process
TCF12	Gm12878	GO:0001773	myeloid dendritic cell activation
TCF12	Gm12878	GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB signaling
TCF12	Gm12878	GO:0043011	myeloid dendritic cell differentiation
TCF12	Gm12878	GO:0032727	positive regulation of interferon-alpha production
TCF12	Gm12878	GO:0032648	regulation of interferon-beta production
TCF12	Gm12878	GO:0060337	type I interferon signaling pathway
TCF12	Gm12878	GO:0032735	positive regulation of interleukin-12 production
TCF12	Gm12878	GO:0032728	positive regulation of interferon-beta production
TCF12	Gm12878	GO:0032655	regulation of interleukin-12 production
TCF12	Gm12878	GO:0042108	positive regulation of cytokine biosynthetic process
TCF12	Gm12878	GO:0051607	defense response to virus
TCF12	Gm12878	GO:0043122	regulation of I-kappaB kinase/NF-kappaB signaling
TCF12	Gm12878	GO:0032479	regulation of type I interferon production
TCF12	Gm12878	GO:0098542	defense response to other organism
TCF12	Gm12878	GO:0032481	positive regulation of type I interferon production
TCF12	Gm12878	GO:0019221	cytokine-mediated signaling pathway
TCF12	Gm12878	GO:0042035	regulation of cytokine biosynthetic process
TCF12	Gm12878	GO:0002252	immune effector process
TCF12	Gm12878	GO:0009615	response to virus
TCF12	Gm12878	GO:0051707	response to other organism
TCF12	Gm12878	GO:0071345	cellular response to cytokine stimulus
TCF12	Gm12878	GO:0001819	positive regulation of cytokine production

TCF12	Gm12878	GO:0045088	regulation of innate immune response
TCF12	Gm12878	GO:0001817	regulation of cytokine production
TCF12	Gm12878	GO:0043207	response to external biotic stimulus
TCF12	Gm12878	GO:0009607	response to biotic stimulus
TCF12	Gm12878	GO:0006464	cellular protein modification process
TCF12	Gm12878	GO:0036211	protein modification process
TCF12	Gm12878	GO:0019538	protein metabolic process
TCF12	Gm12878	GO:0006952	defense response
TCF12	Gm12878	GO:0002376	immune system process
TCF12	Gm12878	GO:0007166	cell surface receptor signaling pathway
TCF12	Hepg2	GO:0006869	lipid transport
TCF12	Hepg2	GO:0015718	monocarboxylic acid transport
TCF12	Hepg2	GO:0015711	organic anion transport
TCF12	Hepg2	GO:0001938	positive regulation of endothelial cell proliferation
TCF12	Hepg2	GO:0015849	organic acid transport
TCF12	Hepg2	GO:0046942	carboxylic acid transport
TCF12	Hepg2	GO:0006873	cellular ion homeostasis
TCF12	Hepg2	GO:0030003	cellular cation homeostasis
TCF12	Hepg2	GO:0001936	regulation of endothelial cell proliferation
TCF12	Hepg2	GO:0098771	inorganic ion homeostasis
TCF12	Hepg2	GO:0055080	cation homeostasis
TCF12	Hepg2	GO:0055065	metal ion homeostasis
TCF12	Hepg2	GO:0050801	ion homeostasis
TCF12	Hepg2	GO:0048145	regulation of fibroblast proliferation
TCF12	Hepg2	GO:0006811	ion transport
TCF12	Hepg2	GO:0050817	coagulation
TCF12	Hepg2	GO:0007596	blood coagulation
TCF12	Hepg2	GO:0071456	cellular response to hypoxia
TCF12	Hepg2	GO:0036294	cellular response to decreased oxygen levels
TCF12	Hepg2	GO:0071453	cellular response to oxygen levels
TCF12	Hepg2	GO:0043436	oxoacid metabolic process
TCF12	Hepg2	GO:0006082	organic acid metabolic process
TCF12	Hepg2	GO:0007599	hemostasis
TCF12	Hepg2	GO:0019752	carboxylic acid metabolic process
TCF12	Hepg2	GO:0050878	regulation of body fluid levels
TCF12	Hepg2	GO:0001666	response to hypoxia
TCF12	Hepg2	GO:0036293	response to decreased oxygen levels
TCF12	Hepg2	GO:0070482	response to oxygen levels
TCF12	Hepg2	GO:0043401	steroid hormone mediated signaling pathway
TCF12	Hepg2	GO:0009755	hormone-mediated signaling pathway
TCF12	Hepg2	GO:0043085	positive regulation of catalytic activity
TCF12	Hepg2	GO:0007623	circadian rhythm
TCF12	Hepg2	GO:0030522	intracellular receptor signaling pathway

TCF12	Hepg2	GO:0048878	chemical homeostasis
TCF12	Hepg2	GO:0006629	lipid metabolic process
TCF12	Hepg2	GO:0019216	regulation of lipid metabolic process
TCF12	Hepg2	GO:0007267	cell-cell signaling
TCF12	Hepg2	GO:0007154	cell communication
TCF12	Hepg2	GO:0044765	single-organism transport
TCF12	Hepg2	GO:0048511	rhythmic process
TCF12	Hepg2	GO:0050790	regulation of catalytic activity
TCF12	Hepg2	GO:0006810	transport
TCF12	Hepg2	GO:0051234	establishment of localization
TCF12	Hepg2	GO:1902578	single-organism localization
TCF12	Hepg2	GO:0050678	regulation of epithelial cell proliferation
TCF12	Hepg2	GO:0065008	regulation of biological quality
TCF12	Hepg2	GO:0051179	localization
TCF12	Hepg2	GO:0006367	transcription initiation from RNA polymerase II promoter
TCF12	Hepg2	GO:0006352	DNA-templated transcription, initiation
TCF12	Hepg2	GO:0044710	single-organism metabolic process
TCF12	Hepg2	GO:0042592	homeostatic process
TCF12	Hepg2	GO:0010467	gene expression
TCF12	Hepg2	GO:0009725	response to hormone
TCF12	Hepg2	GO:0070887	cellular response to chemical stimulus
TCF12	Hepg2	GO:0044702	single organism reproductive process
TCF12	Hepg2	GO:0071310	cellular response to organic substance
TCF12	Hepg2	GO:1901700	response to oxygen-containing compound
TCF12	Hepg2	GO:0042221	response to chemical
TCF12	Hepg2	GO:0022414	reproductive process
TCF12	Hepg2	GO:0010033	response to organic substance
TCF12	Hepg2	GO:0051716	cellular response to stimulus
TCF12	Hepg2	GO:2000026	regulation of multicellular organismal development
TCF12	Hepg2	GO:0048583	regulation of response to stimulus
TCF12	Hepg2	GO:0050793	regulation of developmental process
TCF12	Hepg2	GO:0051239	regulation of multicellular organismal process
TCF7L2	Hct116	GO:0006260	DNA replication
TCF7L2	Hct116	GO:0030097	hemopoiesis
TCF7L2	Hek293	GO:0021546	rhombomere development
TCF7L2	Hepg2	GO:0048384	retinoic acid receptor signaling pathway
TCF7L2	Hepg2	GO:0051348	negative regulation of transferase activity
TCF7L2	Hepg2	GO:0009409	response to cold
TCF7L2	Hepg2	GO:0031331	positive regulation of cellular catabolic process
TCF7L2	Hepg2	GO:0009896	positive regulation of catabolic process
TCF7L2	Hepg2	GO:0043401	steroid hormone mediated signaling pathway
TCF7L2	Hepg2	GO:0009755	hormone-mediated signaling pathway
TCF7L2	Hepg2	GO:0030522	intracellular receptor signaling pathway

TCF7L2	Hepg2	GO:0009894	regulation of catabolic process
TCF7L2	Hepg2	GO:0006367	transcription initiation from RNA polymerase II promoter
TCF7L2	Hepg2	GO:0006352	DNA-templated transcription, initiation
TCF7L2	Hepg2	GO:0010467	gene expression
TCF7L2	Hepg2	GO:0022414	reproductive process
TCF7L2	Hepg2	GO:0007165	signal transduction
TCF7L2	Hepg2	GO:0044763	single-organism cellular process
TCF7L2	Mcf7	GO:0051348	negative regulation of transferase activity
TCF7L2	Mcf7	GO:0000278	mitotic cell cycle
TCF7L2	Mcf7	GO:0051726	regulation of cell cycle
TCF7L2	Mcf7	GO:0006367	transcription initiation from RNA polymerase II promoter
TCF7L2	Mcf7	GO:0006352	DNA-templated transcription, initiation
TCF7L2	Mcf7	GO:0065008	regulation of biological quality
TCF7L2	Mcf7	GO:0007165	signal transduction
TCF7L2	Panc1	GO:0006368	transcription elongation from RNA polymerase II promoter
TCF7L2	Panc1	GO:0006354	DNA-templated transcription, elongation
TCF7L2	Panc1	GO:0034622	cellular macromolecular complex assembly
TCF7L2	Panc1	GO:0065004	protein-DNA complex assembly
TCF7L2	Panc1	GO:0016032	viral process
TCF7L2	Panc1	GO:0044764	multi-organism cellular process
TCF7L2	Panc1	GO:0044403	symbiosis, encompassing mutualism through parasitism
TCF7L2	Panc1	GO:0044419	interspecies interaction between organisms
USF1	A549	GO:0007259	JAK-STAT cascade
USF1	A549	GO:0006953	acute-phase response
USF1	A549	GO:0060397	JAK-STAT cascade involved in growth hormone signaling p
USF1	A549	GO:0043603	cellular amide metabolic process
USF1	A549	GO:0002705	positive regulation of leukocyte mediated immunity
USF1	A549	GO:0002708	positive regulation of lymphocyte mediated immunity
USF1	A549	GO:0002699	positive regulation of immune effector process
USF1	A549	GO:0002526	acute inflammatory response
USF1	A549	GO:0051385	response to mineralocorticoid
USF1	A549	GO:0051412	response to corticosterone
USF1	A549	GO:0007565	female pregnancy
USF1	A549	GO:0002703	regulation of leukocyte mediated immunity
USF1	A549	GO:0071277	cellular response to calcium ion
USF1	A549	GO:0042542	response to hydrogen peroxide
USF1	A549	GO:0051592	response to calcium ion
USF1	A549	GO:0022602	ovulation cycle process
USF1	A549	GO:0051384	response to glucocorticoid
USF1	A549	GO:0031960	response to corticosteroid
USF1	A549	GO:0000302	response to reactive oxygen species
USF1	A549	GO:0002697	regulation of immune effector process
USF1	A549	GO:0044703	multi-organism reproductive process

USF1	A549	GO:0034097	response to cytokine
USF1	A549	GO:0019221	cytokine-mediated signaling pathway
USF1	A549	GO:0071248	cellular response to metal ion
USF1	A549	GO:0044706	multi-multicellular organism process
USF1	A549	GO:0009612	response to mechanical stimulus
USF1	A549	GO:0006979	response to oxidative stress
USF1	A549	GO:0016310	phosphorylation
USF1	A549	GO:0051591	response to cAMP
USF1	A549	GO:0046683	response to organophosphorus
USF1	A549	GO:0071241	cellular response to inorganic substance
USF1	A549	GO:0014074	response to purine-containing compound
USF1	A549	GO:0006468	protein phosphorylation
USF1	A549	GO:0032496	response to lipopolysaccharide
USF1	A549	GO:0071345	cellular response to cytokine stimulus
USF1	A549	GO:0010035	response to inorganic substance
USF1	A549	GO:0006796	phosphate-containing compound metabolic process
USF1	A549	GO:0006793	phosphorus metabolic process
USF1	A549	GO:0002237	response to molecule of bacterial origin
USF1	A549	GO:0019220	regulation of phosphate metabolic process
USF1	A549	GO:0051174	regulation of phosphorus metabolic process
USF1	A549	GO:0042493	response to drug
USF1	A549	GO:0043207	response to external biotic stimulus
USF1	A549	GO:1901698	response to nitrogen compound
USF1	A549	GO:0009607	response to biotic stimulus
USF1	A549	GO:0010243	response to organonitrogen compound
USF1	A549	GO:0032870	cellular response to hormone stimulus
USF1	A549	GO:0050776	regulation of immune response
USF1	A549	GO:0035556	intracellular signal transduction
USF1	A549	GO:0009605	response to external stimulus
USF1	A549	GO:0009628	response to abiotic stimulus
USF1	A549	GO:0051726	regulation of cell cycle
USF1	A549	GO:0071495	cellular response to endogenous stimulus
USF1	A549	GO:0051704	multi-organism process
USF1	A549	GO:0033993	response to lipid
USF1	A549	GO:0071310	cellular response to organic substance
USF1	A549	GO:0009725	response to hormone
USF1	A549	GO:0014070	response to organic cyclic compound
USF1	A549	GO:0070887	cellular response to chemical stimulus
USF1	A549	GO:1901700	response to oxygen-containing compound
USF1	A549	GO:0002682	regulation of immune system process
USF1	A549	GO:0007166	cell surface receptor signaling pathway
USF1	A549	GO:0009719	response to endogenous stimulus
USF1	A549	GO:0010941	regulation of cell death

USF1	A549	GO:0022414	reproductive process
USF1	A549	GO:0042981	regulation of apoptotic process
USF1	A549	GO:0010033	response to organic substance
USF1	A549	GO:0051716	cellular response to stimulus
USF1	A549	GO:0042221	response to chemical
USF1	H1hesc	GO:0043401	steroid hormone mediated signaling pathway
USF1	Hepg2	GO:0010831	positive regulation of myotube differentiation
USF1	Hepg2	GO:0048384	retinoic acid receptor signaling pathway
USF1	Hepg2	GO:0060509	Type I pneumocyte differentiation
USF1	Hepg2	GO:0060479	lung cell differentiation
USF1	Hepg2	GO:0060487	lung epithelial cell differentiation
USF1	Hepg2	GO:0051149	positive regulation of muscle cell differentiation
USF1	Hepg2	GO:0043401	steroid hormone mediated signaling pathway
USF1	Hepg2	GO:0009755	hormone-mediated signaling pathway
USF1	Hepg2	GO:0010720	positive regulation of cell development
USF1	K562	GO:0060430	lung saccule development
YY1	Gm12878	GO:0060743	epithelial cell maturation involved in prostate gland develop
YY1	Hct116	GO:0051385	response to mineralocorticoid
YY1	Hct116	GO:0051412	response to corticosterone
YY1	Hct116	GO:0032570	response to progesterone
YY1	Hct116	GO:0071277	cellular response to calcium ion
YY1	Hct116	GO:0051592	response to calcium ion
YY1	Hct116	GO:0007565	female pregnancy
YY1	Hct116	GO:0051591	response to cAMP
YY1	Hct116	GO:0046683	response to organophosphorus
YY1	Hct116	GO:0051384	response to glucocorticoid
YY1	Hct116	GO:0031960	response to corticosteroid
YY1	Hct116	GO:0014074	response to purine-containing compound
YY1	Hct116	GO:1901654	response to ketone
YY1	Hct116	GO:0009612	response to mechanical stimulus
YY1	Hct116	GO:0044703	multi-organism reproductive process
YY1	Hct116	GO:0031668	cellular response to extracellular stimulus
YY1	Hct116	GO:0042493	response to drug
YY1	Hct116	GO:0071248	cellular response to metal ion
YY1	Hct116	GO:0043434	response to peptide hormone
YY1	Hct116	GO:0071241	cellular response to inorganic substance
YY1	Hct116	GO:1901652	response to peptide
YY1	Hct116	GO:0034097	response to cytokine
YY1	Hct116	GO:0097305	response to alcohol
YY1	Hct116	GO:0006979	response to oxidative stress
YY1	Hct116	GO:0032870	cellular response to hormone stimulus
YY1	Hct116	GO:0071496	cellular response to external stimulus
YY1	Hct116	GO:0010038	response to metal ion

YY1	Hct116	GO:0010243	response to organonitrogen compound
YY1	Hct116	GO:0009991	response to extracellular stimulus
YY1	Hct116	GO:0032496	response to lipopolysaccharide
YY1	Hct116	GO:1901698	response to nitrogen compound
YY1	Hct116	GO:0010035	response to inorganic substance
YY1	Hct116	GO:0009725	response to hormone
YY1	Hct116	GO:0048545	response to steroid hormone
YY1	Hct116	GO:0014070	response to organic cyclic compound
YY1	Hct116	GO:0071495	cellular response to endogenous stimulus
YY1	Hct116	GO:0043207	response to external biotic stimulus
YY1	Hct116	GO:0009607	response to biotic stimulus
YY1	Hct116	GO:0009628	response to abiotic stimulus
YY1	Hct116	GO:0009719	response to endogenous stimulus
YY1	Hct116	GO:0009605	response to external stimulus
YY1	Hct116	GO:0033993	response to lipid
YY1	Hct116	GO:0051704	multi-organism process
YY1	Hct116	GO:1901700	response to oxygen-containing compound
YY1	Hct116	GO:0071310	cellular response to organic substance
YY1	Hct116	GO:0070887	cellular response to chemical stimulus
YY1	Hct116	GO:0051716	cellular response to stimulus
YY1	Hct116	GO:0010033	response to organic substance
YY1	Hepg2	GO:0006260	DNA replication
YY1	Hepg2	GO:0006259	DNA metabolic process
YY1	K562	GO:0006094	gluconeogenesis
YY1	K562	GO:0061394	regulation of transcription from RNA polymerase II promoter
YY1	K562	GO:0016051	carbohydrate biosynthetic process
YY1	K562	GO:0019319	hexose biosynthetic process
YY1	K562	GO:0046364	monosaccharide biosynthetic process
YY1	K562	GO:0044710	single-organism metabolic process
YY1	Nt2d1	GO:0007259	JAK-STAT cascade
YY1	Nt2d1	GO:0060397	JAK-STAT cascade involved in growth hormone signaling p
YY1	Nt2d1	GO:0002705	positive regulation of leukocyte mediated immunity
YY1	Nt2d1	GO:0002708	positive regulation of lymphocyte mediated immunity
YY1	Nt2d1	GO:0006953	acute-phase response
YY1	Nt2d1	GO:0002526	acute inflammatory response
YY1	Nt2d1	GO:0002706	regulation of lymphocyte mediated immunity
YY1	Nt2d1	GO:0042129	regulation of T cell proliferation
YY1	Nt2d1	GO:0040014	regulation of multicellular organism growth
YY1	Nt2d1	GO:0019221	cytokine-mediated signaling pathway
YY1	Nt2d1	GO:0016310	phosphorylation
YY1	Nt2d1	GO:0007346	regulation of mitotic cell cycle
YY1	Nt2d1	GO:0016032	viral process
YY1	Nt2d1	GO:0044764	multi-organism cellular process

YY1	Nt2d1	GO:0044403	symbiosis, encompassing mutualism through parasitism
YY1	Nt2d1	GO:0044419	interspecies interaction between organisms
YY1	Nt2d1	GO:0050776	regulation of immune response
YY1	Nt2d1	GO:0051704	multi-organism process
YY1	Nt2d1	GO:0006952	defense response
YY1	Nt2d1	GO:0035556	intracellular signal transduction
ZNF143	Gm12878	GO:0070345	negative regulation of fat cell proliferation
ZNF143	Gm12878	GO:0070344	regulation of fat cell proliferation
ZNF143	Gm12878	GO:0032269	negative regulation of cellular protein metabolic process
ZNF143	Gm12878	GO:0051248	negative regulation of protein metabolic process
ZNF143	Helas3	GO:0051385	response to mineralocorticoid
ZNF143	Helas3	GO:0051412	response to corticosterone
ZNF143	Helas3	GO:0070345	negative regulation of fat cell proliferation
ZNF143	Helas3	GO:0070344	regulation of fat cell proliferation
ZNF143	Helas3	GO:0051384	response to glucocorticoid
ZNF143	Helas3	GO:0031960	response to corticosteroid
ZNF143	Helas3	GO:0032570	response to progesterone
ZNF143	Helas3	GO:0071277	cellular response to calcium ion
ZNF143	Helas3	GO:0051592	response to calcium ion
ZNF143	Helas3	GO:0007565	female pregnancy
ZNF143	Helas3	GO:0071248	cellular response to metal ion
ZNF143	Helas3	GO:0007179	transforming growth factor beta receptor signaling pathway
ZNF143	Helas3	GO:0051591	response to cAMP
ZNF143	Helas3	GO:0046683	response to organophosphorus
ZNF143	Helas3	GO:0071241	cellular response to inorganic substance
ZNF143	Helas3	GO:0014074	response to purine-containing compound
ZNF143	Helas3	GO:0009612	response to mechanical stimulus
ZNF143	Helas3	GO:0010038	response to metal ion
ZNF143	Helas3	GO:0010243	response to organonitrogen compound
ZNF143	Helas3	GO:1901698	response to nitrogen compound
ZNF143	K562	GO:0098602	single organism cell adhesion
ZNF143	K562	GO:0098609	cell-cell adhesion
ZNF143	K562	GO:0016337	single organismal cell-cell adhesion
ZNF143	K562	GO:0035556	intracellular signal transduction
ZNF143	K562	GO:0065008	regulation of biological quality

Table 7.14 List of Biological Processes (BP) found as enriched among the identified co-factors. The analysis was done using GOrilla tool.

7.3 Supplementary for Chapter 4

Supplemental Note 1. Features were selected based on all samples without considering the partition of training and testing. The performance of the model can be vulnerable due to overfitting. However, we confirmed that our models are not vulnerable to overfitting by three measure. First, we determined the fraction of common features selected considering with or without test data. The 80-90% common features across conditions and across fold indicate that the models are not over-trained. Second, we compared the expression variability taking all samples with the same measure of taking only training samples. The significantly high correlation (0.95-0.97, p.value < 10^{-16}) of expression variability of two conditions confirms that the set of selected genes are not going to vary much leaving the test data or not. Third, we repeated the feature selection using only training samples and rechecked the model performance. The insignificant difference of two sets of models using two sets of features affirms the absence of overfitting of our models (Figure 7.12).

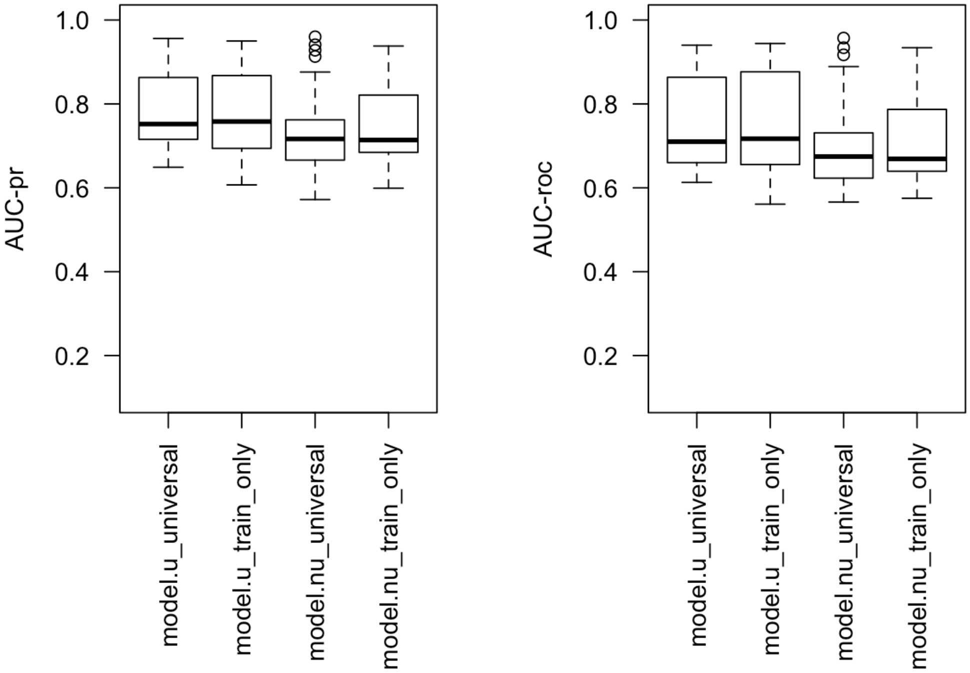


Figure 7.12 Model performance using feature selection using all samples vs. training samples.

Supplemental Note 2. We have shown high performance of models build by Adaboost method on a set of features selected by all samples. We perform the following 4 comparisons to show that the models are learning useful information. First, we have used another machine learning model, Support Vector Machine (SVM) to confirm the accuracy of models (Figure 7.13). Second, we randomized the expression data and rebuild the model expecting that the accuracy will diminish. In particular, along each feature the expression values are randomized and the diminished accuracy as shown in Figure 7.14, indicates previous higher performance was not any random event. Third, we conducted a block permutation: randomizing the expression value within each platform instead of across all samples. Figure 7.15 conveys the same message of Figure 7.14.

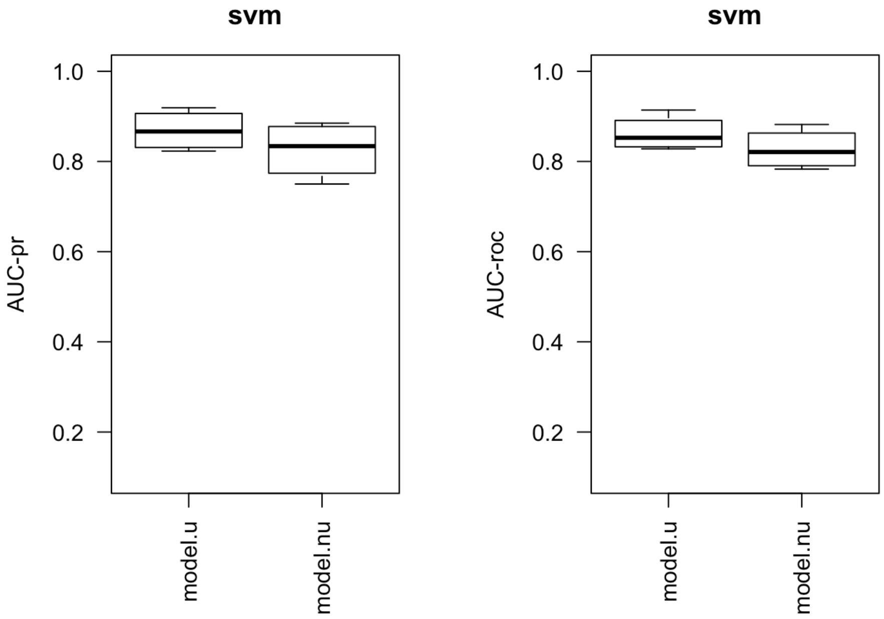


Figure 7.13 Model performance using SVM.

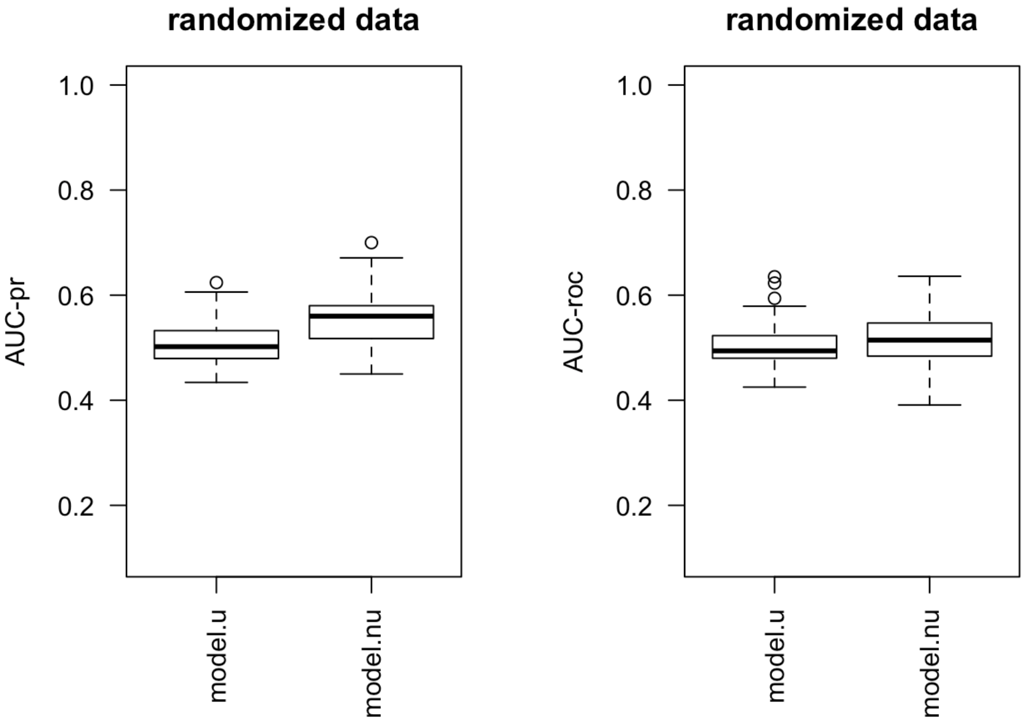


Figure 7.14 Model performance after randomizing the expression data.

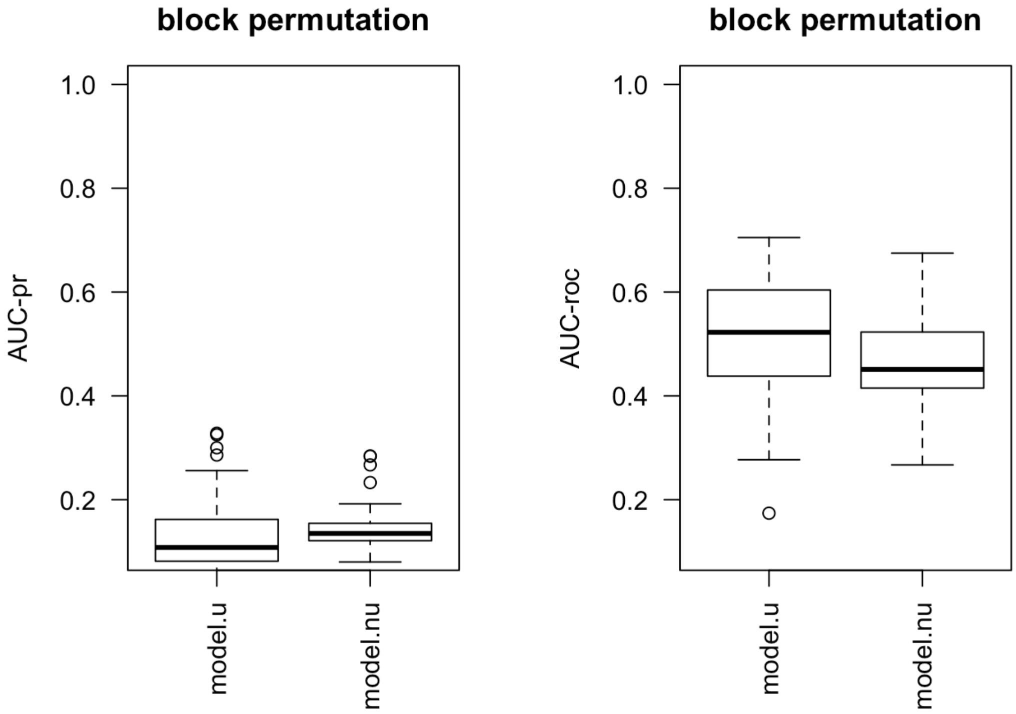


Figure 7.15 Model performance after block permutation

7.4 Supplementary for Chapter 5

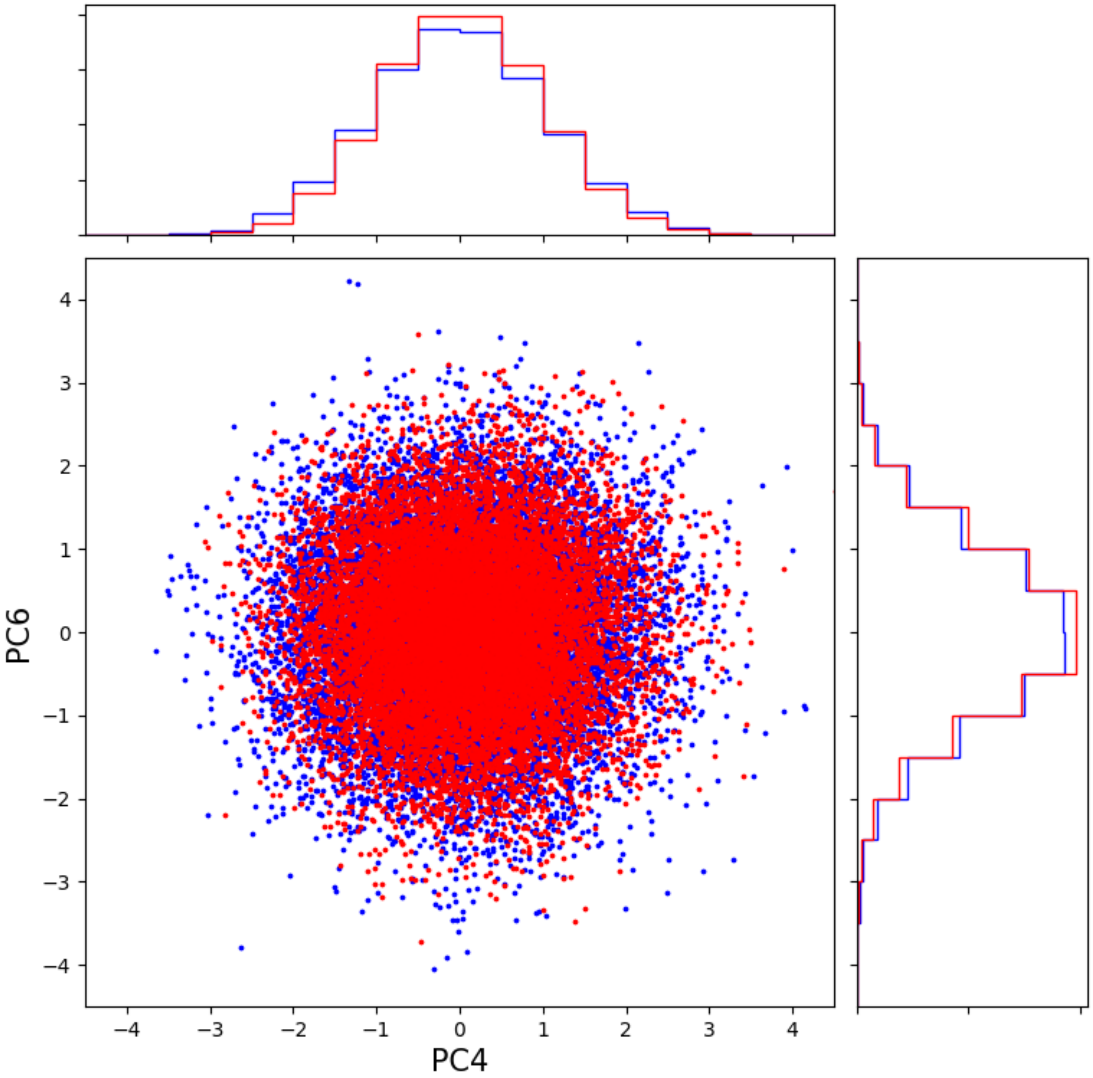


Figure 7.16 Principal component analysis of the mouse_retina data before calibration.

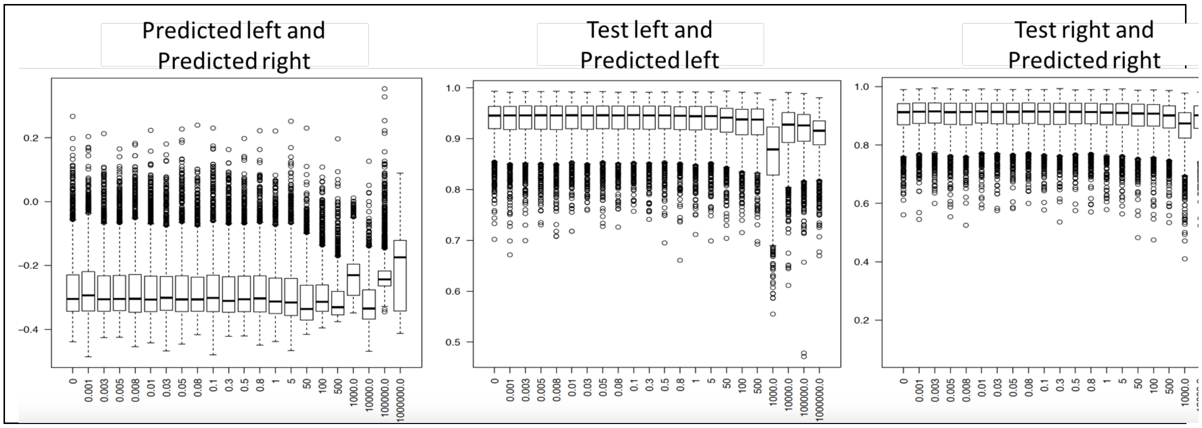


Figure 7.17 Effect of lamda in predicted output.

8 Bibliography

[1] P. Satir and S. T. Christensen, “Structure and function of mammalian cilia,” *Histochem. Biol.*, 2008.

[2] G. Griffiths, “Cell evolution and the problem of membrane topology.,” *Nat. Rev. Mol. Cell Biol.*, 2007.

[3] G. Griffiths, “NATURE REVIEWS MOLECULAR CELL BIOLOGY,” *Cell evolution and the problem of membrane topology*. 2007.

[4] M. B. Clark and J. S. Mattick, “Long noncoding RNAs in cell biology.,” *Semin. Cell Dev. Biol.*, 2011.

[5] 239. doi:10.1186/1471-2105-10-239 Arrial, R. T., Togawa, R. C., & Brigido, M. de M. (2009). Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. BMC Bioinformatics, 10 *et al.*, “Genome Regulation by Long Noncoding RNAs,” *Annu. Rev. Biochem.*, 2012.

[6] J. D. Watson and F. H. C. Crick, “Molecular structure of nucleic aids: A structure for deoxyribose nucleic acid,” *Nature*, 1953.

[7] J. D. WATSON *et al.*, “Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.,” *Nature*, 1953.

[8] A. T. Annuziato and A. Annunziato, “DNA Packaging: Nucleosomes and Chromatin,” *Nat. Educ.*, 2008.

[9] F. Crick, “Central dogma of molecular biology.,” *Nature*, 1970.

[10] B. S. Zhao, I. A. Roundtree, and C. He, “Post-transcriptional gene regulation by mRNA modifications.,” *Nat. Rev. Mol. Cell Biol.*, 2017.

[11] G. A. Wray *et al.*, “The evolution of transcriptional regulation in eukaryotes,” *Molecular Biology and Evolution*. 2003.

[12] T. Kohlsdorf *et al.*, “The evolution of transcriptional regulation in eukaryotes.,” *Molecular biology and evolution*. 2003.

[13] F. Spitz and E. E. M. Furlong, “Transcription factors: from enhancer binding to developmental control,” *Nat. Rev. Genet.*, vol. 13, no. 9, pp. 613–626, 2012.

[14] C. D. Allis and T. Jenuwein, “The molecular hallmarks of epigenetic control,” *Nat. Rev. Genet.*, 2016.

[15] T. O. Tollefsbol, *Handbook of Epigenetics*. 2011.

[16] A. Eccleston, F. Cesari, and M. Skipper, “Transcription and epigenetics,” *Nature*, 2013.

[17] P. Boyle and B. Levin, “World Cancer Report 2014,” in *World Cancer Report 2014*, 2014.

[18] GBD 2015 Disease and Injury Incidence and Prevalence Collaborators, “Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015,” *Lancet*, 2016.

[19] a G. Knudson, “Two genetic hits (more or less) to cancer.,” *Nat. Rev. Cancer*, 2001.

[20] K. D. Hansen *et al.*, “Increased methylation variation in epigenetic domains across cancer types.,” *Nat. Genet.*, vol. 43, no. 8, pp. 768–775, 2011.

[21] R. Lister *et al.*, “Human DNA methylomes at base resolution show widespread epigenomic differences.,” *Nature*, vol. 462, no. 7271, pp. 315–322, 2009.

[22] D. Hanahan and R. A. Weinberg, “The hallmarks of cancer.,” *Cell*, 2000.

[23] G. I. Evan and K. H. Vousden, “Proliferation, cell cycle and apoptosis in cancer,”

Nat. Rev., 2001.

[24] M. Andrea and A. Yap, “Contact inhibition (of proliferation) redux.,” *Curr. Opin. Cell Biol.*, 2012.

[25] S. Elmore, “Apoptosis: A Review of Programmed Cell Death,” *Toxicol. Pathol.*, 2007.

[26] R. Greenberg, “Telomeres, Crisis and Cancer,” *Curr. Mol. Med.*, 2005.

[27] A. Wicki and G. Christofori, “The angiogenic switch in tumorigenesis,” in *Tumor Angiogenesis: Basic Mechanisms and Cancer Therapy*, 2008.

[28] G. Bergers and L. E. Benjamin, “Angiogenesis: Tumorigenesis and the angiogenic switch,” *Nat. Rev. Cancer*, 2003.

[29] F. Van Zijl, G. Krupitza, and W. Mikulits, “Initial steps of metastasis: Cell invasion and endothelial transmigration,” *Mutation Research - Reviews in Mutation Research*. 2011.

[30] D. X. Nguyen and J. Massagué, “Genetic determinants of cancer metastasis,” *Nat. Rev. Genet.*, 2007.

[31] B. A. Yoshida, “Metastasis-Suppressor Genes: a Review and Perspective on an Emerging Field,” *J. Natl. Cancer Inst.*, 2000.

[32] I. R. Hart, “‘Seed and soil’ revisited: mechanisms of site-specific metastasis,” *Cancer Metastasis Rev.*, 1982.

[33] J. G. Scott, P. Kuhn, and a R. a Anderson, “Unifying metastasis--Integrating intravasation, circulation and end organ colonization,” *Nat. Rev. Cancer*, 2012.

[34] N. Syn, L. Wang, G. Sethi, J. P. Thiery, and B. C. Goh, “Exosome-Mediated Metastasis: From Epithelial-Mesenchymal Transition to Escape from Immunosurveillance,” *Trends in Pharmacological Sciences*. 2016.

[35] D. S. Ettinger *et al.*, “NCCN Clinical Practice Guidelines Occult primary.,” *J. Natl. Compr. Canc. Netw.*, 2011.

[36] L. Breiman and A. Cutler, “Breiman and Cutler’s random forests for classification and regression,” *Packag. “randomForest,”* 2012.

[37] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. 2013.

[38] T. K. Ho, “The random subspace method for constructing decision forests,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998.

[39] P. Giudici and S. Figini, *Applied Data Mining for Business and Industry*. 2009.

[40] R. E. Schapire, “A Short Introduction to Boosting,” *Society*, 2009.

[41] R. E. Schapire and Y. Singer, “Improved boosting algorithms using confidence-rated predictions,” *Mach. Learn.*, 1999.

[42] T. Zhang, “Statistical behavior and consistency of classification methods based on convex risk minimization,” *Ann. Stat.*, 2004.

[43] T. G. Margineant and D. D. Dietterich, “Pruning Adaptive Boosting,” *14th Int’l Conf. Mach. Learn.*, 1997.

[44] P. L. Bartlett, “Discussions of boosting papers, and rejoinders,” *Ann. Stat.*, 2004.

[45] J. T. Leek *et al.*, “Tackling the widespread and critical impact of batch effects in high-throughput data,” *Nat. Rev. Genet.*, 2010.

[46] O. Alter, P. O. Brown, and D. Botstein, “Singular value decomposition for genome-wide expression data processing and modeling. Supplementary data,” *Proc. Natl. Acad. Sci. U. S. A.*, 2000.

[47] M. Benito *et al.*, “Adjustment of systematic microarray data biases,” *Bioinformatics*, 2004.

[48] W. E. Johnson and C. Li, “Adjusting batch effects in microarray expression data

using empirical Bayes methods.," *Biostatistics*, 2007.

[49] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genet.*, 2007.

[50] G. K. Smyth, "Limma: linear models fro microarray data.," in *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, 2005.

[51] D. Tanikic, M. Manic, G. Radenkovic, and D. Mancic, "Metal cutting process parameters modeling: An artificial intelligence approach," *J. Sci. Ind. Res. (India).*, 2009.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[54] C. Y. Liou, W. C. Cheng, J. W. Liou, and D. R. Liou, "Autoencoder for words," *Neurocomputing*, 2014.

[55] A. Bird, "Perceptions of epigenetics.," *Nature*, vol. 447, no. 7143, pp. 396–8, 2007.

[56] M. M. Suzuki and A. Bird, "DNA methylation landscapes: provocative insights from epigenomics.," *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 465–76, 2008.

[57] S. Horvath *et al.*, "Aging effects on DNA methylation modules in human brain and blood tissue," *Genome Biol*, vol. 13, no. 10, p. R97, 2012.

[58] J. Sharif *et al.*, "The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA.," *Nature*, vol. 450, no. 7171, pp. 908–912, 2007.

[59] A. P. Feinberg and B. Tycko, "The history of cancer epigenetics.," *Nat. Rev. Cancer*, vol. 4, no. 2, pp. 143–153, 2004.

[60] M. Esteller, "Molecular Origins of Cancer Epigenetics in Cancer," *N Engl J Med*, vol. 358, pp. 1148–59, 2008.

[61] M. Ehrlich, "DNA methylation in cancer: too much, but also too little.," *Oncogene*, vol. 21, no. 35, pp. 5400–5413, 2002.

[62] C. Hong *et al.*, "Epigenome scans and cancer genome sequencing converge on WNK2, a kinase-independent suppressor of cell growth.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, pp. 10974–10979, 2007.

[63] A. Doi *et al.*, "Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts," *Nat Genet*, vol. 41, no. 12, pp. 1350–1353, 2009.

[64] B. Wen, H. Wu, Y. Shinkai, R. a Irizarry, and A. P. Feinberg, "Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells.," *Nat. Genet.*, vol. 41, no. 2, pp. 246–250, 2009.

[65] J. R. Hesselberth *et al.*, "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.," *Nat. Methods*, vol. 6, no. 4, pp. 283–289, 2009.

[66] X. He, R. Chatterjee, D. Tillo, A. Smith, P. FitzGerald, and C. Vinson, "Nucleosomes are enriched at the boundaries of hypomethylated regions (HMRs) in mouse dermal fibroblasts and keratinocytes.," *Epigenetics Chromatin*, vol. 7, no. 1, p. 34, 2014.

[67] T. K. Kelly, D. D. De Carvalho, and P. A. Jones, "Epigenetic modifications as therapeutic targets.," *Nat. Biotechnol.*, vol. 28, no. 10, pp. 1069–1078, 2010.

[68] J. R. Dixon *et al.*, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, no. 7398, pp. 376–380, 2012.

[69] H. Cedar and Y. Bergman, “Linking DNA methylation and histone modification: patterns and paradigms.,” *Nat. Rev. Genet.*, vol. 10, no. 5, pp. 295–304, 2009.

[70] J. Harrow *et al.*, “GENCODE: the reference human genome annotation for The ENCODE Project.,” *Genome Res.*, vol. 22, no. 9, pp. 1760–74, 2012.

[71] C. E. Grant, T. L. Bailey, and W. S. Noble, “FIMO: scanning for occurrences of a given motif.,” *Bioinformatics*, vol. 27, no. 7, pp. 1017–8, 2011.

[72] J. T. Kadonaga, “Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines.,” *Cell*, vol. 92, no. 3, pp. 307–313, Feb. 1998.

[73] C. von Mering *et al.*, “STRING 7 - Recent developments in the integration and prediction of protein interactions,” *Nucleic Acids Res.*, vol. 35, no. SUPPL. 1, pp. 358–362, 2007.

[74] C.-T. Ong and V. G. Corces, “CTCF: an architectural protein bridging genome topology and function.,” *Nat. Rev. Genet.*, vol. 15, no. 4, pp. 234–46, 2014.

[75] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[76] V. Matys *et al.*, “TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.,” *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D108--10, 2006.

[77] A. Liaw and M. Wiener, “Package ’ randomForest‘. Breiman and Cutler’s random forests for classification and regression,” *CRAN Reference manual*. CRAN, 2015.

[78] N. L. van Berkum *et al.*, “Hi-C: a method to study the three-dimensional architecture of genomes.,” *J. Vis. Exp.*, no. 39, pp. 1–7, 2010.

[79] E. Deniaud *et al.*, “Overexpression of transcription factor Sp1 leads to gene expression perturbations and cell cycle inhibition.,” *PLoS One*, vol. 4, no. 9, p. e7035, 2009.

[80] S. Safe and M. Abdelrahim, “Sp transcription factor family and its role in cancer,” *Eur. J. Cancer*, vol. 41, no. 16, pp. 2438–2448, 2005.

[81] S. Harel *et al.*, “ZFX Controls the Self-Renewal of Human Embryonic Stem Cells,” *PLoS One*, vol. 7, no. 8, p. e42302, 2012.

[82] T. L. Arenzana, M. R. Smith-Raska, and B. Reizis, “Transcription factor Zfx controls BCR-induced proliferation and survival of B lymphocytes,” *Blood*, vol. 113, no. 23, pp. 5857–5867, 2009.

[83] H. Daitoku, J. I. Sakamaki, and A. Fukamizu, “Regulation of FoxO transcription factors by acetylation and protein-protein interactions,” *Biochim. Biophys. Acta - Mol. Cell Res.*, vol. 1813, no. 11, pp. 1954–1960, 2011.

[84] S.-T. Lee *et al.*, “A global DNA methylation and gene expression analysis of early human B-cell development reveals a demethylation signature and transcription factor network.,” *Nucleic Acids Res.*, vol. 40, no. 22, pp. 11339–51, 2012.

[85] Z.-X. Wang *et al.*, “The transcription factor Zfp281 controls embryonic stem cell pluripotency by direct activation and repression of target genes.,” *Stem Cells*, vol. 26, no. 11, pp. 2791–2799, 2008.

[86] B. Keenen and I. L. De La Serna, “Chromatin remodeling in Embryonic stem cells: regulating the balance between pluripotency and differentiation,” *J. Cell. Physiol.*, vol. 219, no. 1, pp. 1–7, 2009.

[87] M. Fidalgo *et al.*, “Zfp281 mediates Nanog autorepression through recruitment of the NuRD complex and inhibits somatic cell reprogramming.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 40, pp. 16202–7, 2012.

[88] W. Timp *et al.*, “Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors.,” *Genome Med.*, vol. 6, no. 8, p. 61, 2014.

[89] J. W. Whitaker, Z. Chen, and W. Wang, “Predicting the human epigenome from DNA motifs.,” *Nat. Methods*, vol. 12, no. 3, p. 265–72, 7 p following 272, Mar. 2015.

[90] F. Jacob and J. Monod, “Genetic regulatory mechanisms in the synthesis of proteins.,” *J. Mol. Biol.*, vol. 3, pp. 318–356, 1961.

[91] S. Busby and R. H. Ebright, “Promoter structure, promoter recognition, and transcription activation in prokaryotes.,” *Cell*, vol. 79, no. 5, pp. 743–746, Dec. 1994.

[92] R. P. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E. E. M. Furlong, “Combinatorial binding predicts spatio-temporal cis-regulatory activity.,” *Nature*, vol. 462, no. 7269, pp. 65–70, 2009.

[93] J. O. Yáñez-Cuna, H. Q. Dinh, E. Z. Kvon, D. Shlyueva, and A. Stark, “Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding,” *Genome Res.*, vol. 22, no. 10, pp. 2018–2030, 2012.

[94] S. Hannenhalli and S. Levy, “Predicting transcription factor synergism.,” *Nucleic Acids Res.*, vol. 30, no. 19, pp. 4278–4284, 2002.

[95] A. Mathelier and W. W. Wasserman, “The Next Generation of Transcription Factor Binding Site Prediction,” *PLoS Comput. Biol.*, vol. 9, no. 9, 2013.

[96] I. Dror, T. Golan, C. Levy, R. Rohs, and Y. Mandel-Gutfreund, “A widespread role of the motif environment in transcription factor binding across diverse protein families.,” *Genome Res.*, Jul. 2015.

[97] L. Liu, W. Zhao, and X. Zhou, “Modeling co-occupancy of transcription factors using chromatin features,” *Nucleic Acids Res.*, p. gkv1281, 2015.

[98] L. Wang, S. Jensen, and S. Hannenhalli, “An interaction-dependent model for transcription factor binding,” *Syst. Biol. Regul. Genomics*, vol. 4023, pp. 225–234, 2006.

[99] S. Kumar and P. Bucher, “Predicting transcription factor site occupancy using DNA sequence intrinsic and cell-type specific chromatin features,” *BMC Bioinformatics*, vol. 17, no. S1, p. 4, 2016.

[100] N. Gheldof *et al.*, “Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene,” *Nucleic Acids Res.*, vol. 38, no. 13, pp. 4325–4336, 2010.

[101] N. D. Heintzman *et al.*, “Histone modifications at human enhancers reflect global cell-type-specific gene expression.,” *Nature*, vol. 459, no. 7243, pp. 108–112, 2009.

[102] A. Arvey, P. Agius, W. S. Noble, and C. Leslie, “Sequence and chromatin determinants of cell-type-specific transcription factor binding,” *Genome Res.*, vol. 22, no. 9, pp. 1723–1734, 2012.

[103] D. Benveniste, H.-J. Sonntag, G. Sanguinetti, and D. Sproul, “Transcription factor binding predicts histone modifications in human cell lines.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 37, pp. 13367–13372, Sep. 2014.

[104] S. Frietze and P. J. Farnham, “Transcription factor effector domains.,” *Subcell. Biochem.*, vol. 52, pp. 261–277, 2011.

[105] J. Friedman, T. Hastie, and R. Tibshirani, “Additive Logistic Regression: a Statistical View of Boosting,” *Ann. Stat.*, vol. 28, no. 2, pp. 337–407, 2000.

[106] J. H. Friedman, “Stochastic gradient boosting,” *Comput. Stat. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002.

[107] J. H. Freidman, “Greedy Function Approximation : A Gradient Boosting Machine Author (s): Jerome H . Friedman Source : The Annals of Statistics , Vol . 29 , No

. 5 (Oct ., 2001), pp . 1189-1232 Published by : Institute of Mathematical Statistics Stable URL : http://www," *Institue Math. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2008.

[108] S. Levy and S. Hannenhalli, "Identification of transcription factor binding sites in the human genome sequence.," *Mamm. Genome*, vol. 13, no. 9, pp. 510–514, 2002.

[109] R. Worsley Hunt and W. W. Wasserman, *Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets.*, vol. 15, no. 7. 2014.

[110] M. L. Bulyk, P. L. F. Johnson, and G. M. Church, "Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors.," *Nucleic Acids Res.*, vol. 30, no. 5, pp. 1255–1261, Mar. 2002.

[111] S. Hannenhalli, "Eukaryotic transcription factor binding sites--modeling and integrative search methods.," *Bioinformatics*, vol. 24, no. 11, pp. 1325–1331, Jun. 2008.

[112] C. Linhart, Y. Halperin, and R. Shamir, "Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets," *Genome Res.*, vol. 18, no. 7, pp. 1180–1189, 2008.

[113] G. D. Amoutzias, D. L. Robertson, Y. Van de Peer, and S. G. Oliver, "Choose your partners: dimerization in eukaryotic transcription factors," *Trends in Biochemical Sciences*, vol. 33, no. 5. pp. 220–229, 2008.

[114] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, 2009.

[115] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.," *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, 2009.

[116] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.," *BMC Bioinformatics*, vol. 10, p. 48, 2009.

[117] A. E. Allen-Jennings, M. G. Hartman, G. J. Kociba, and T. Hai, "The roles of ATF3 in glucose homeostasis. A transgenic mouse model with liver dysfunction and defects in endocrine pancreas.," *J. Biol. Chem.*, vol. 276, no. 31, pp. 29507–29514, 2001.

[118] Y. Tanaka *et al.*, "Systems analysis of ATF3 in stress response and cancer reveals opposing effects on pro-apoptotic genes in p53 pathway," *PLoS One*, vol. 6, no. 10, 2011.

[119] B. P. Chen, C. D. Wolfgang, and T. Hai, "Analysis of ATF3, a transcription factor induced by physiological stresses and modulated by gadd153/Chop10.," *Mol. Cell. Biol.*, vol. 16, no. 3, pp. 1157–1168, 1996.

[120] A. I. Su, L. G. Guidotti, J. P. Pezacki, F. V Chisari, and P. G. Schultz, "Gene expression during the priming phase of liver regeneration after partial hepatectomy in mice.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 17, pp. 11181–11186, 2002.

[121] J. Lotem, H. Benjamin, D. Netanely, E. Domany, and L. Sachs, "Induction in myeloid leukemic cells of genes that are expressed in different normal tissues.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 45, pp. 16022–16027, 2004.

[122] J. Lotem, D. Netanely, E. Domany, and L. Sachs, "Human cancers overexpress genes that are specific to a variety of normal human tissues.," *Proc. Natl. Acad.*

Sci. U. S. A., vol. 102, no. 51, pp. 18556–18561, 2005.

[123] B. Mayr and M. Montminy, "Transcriptional regulation by the phosphorylation-dependent factor CREB.," *Nat. Rev. Mol. Cell Biol.*, vol. 2, no. 8, pp. 599–609, 2001.

[124] S. Rockowitz *et al.*, "Comparison of REST Cistromes across Human Cell Types Reveals Common and Context-Specific Functions," *PLoS Comput. Biol.*, vol. 10, no. 6, 2014.

[125] D. A. Liebermann, B. Gregory, and B. Huffman, "AP-1 (FOS/JUN) transcription factors in hematopoietic differentiation and apoptosis (Review)," *Int. J. Oncol.*, vol. 12, no. 3, pp. 685–700, 1998.

[126] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[127] M. S. Handcock and M. Morris, "Relative Distribution Methods," *Sociol. Methodol.*, vol. 28, no. 1998, pp. 53–97, 1998.

[128] T. Siggers, M. H. Duyzend, J. Reddy, S. Khan, and M. L. Bulyk, "Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex," *Molecular Systems Biology*, vol. 7. 2011.

[129] M. Slattery *et al.*, "Cofactor binding evokes latent differences in DNA binding specificity between hox proteins," *Cell*, vol. 147, no. 6, pp. 1270–1282, 2011.

[130] The ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2013.

[131] J. D. Storey, "qvalue: Q-value estimation for false discovery rate control." 2015.

[132] V. Gotea, A. Visel, J. M. Westlund, M. A. Nobrega, L. A. Pennacchio, and I. Ovcharenko, "Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers," *Genome Res.*, vol. 20, no. 5, pp. 565–577, 2010.

[133] X. He, T. S. P. C. Duque, and S. Sinha, "Evolutionary origins of transcription factor binding site clusters," *Mol. Biol. Evol.*, vol. 29, no. 3, pp. 1059–1070, 2012.

[134] G. Ridgeway, "Generalized Boosted Regression Models." 2015.

[135] C. Henning, "fpc: Flexible Procedures for Clustering." 2015.

[136] R. C. Dubes and G. Zeng, "A test for spatial homogeneity in cluster analysis," *J. Classif.*, vol. 4, no. 1, pp. 33–56, 1987.

[137] G. Zeng and R. C. Dubes, "A test for spatial randomness based on k-NN distances," *Pattern Recognit. Lett.*, vol. 3, no. 2, pp. 85–91, Mar. 1985.

[138] G. Zeng and R. C. Dubes, "A comparison of tests for randomness," *Pattern Recognit.*, vol. 18, no. 2, pp. 191–198, Jan. 1985.

[139] W. Melssen, R. Wehrens, and L. Buydens, "Supervised Kohonen networks for classification problems," *Chemom. Intell. Lab. Syst.*, vol. 83, no. 2, pp. 99–113, 2006.

[140] R. and L. M. C. B. Wehrens, "Supervised and Unsupervised Self-Organising Maps," *J. Stat. Softw.*, vol. 21, no. 5, 2007.

[141] L. J. Zhu *et al.*, "ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data.," *BMC Bioinformatics*, vol. 11, p. 237, 2010.

[142] S. Pietrokovski, "Searching databases of conserved sequence regions by aligning protein multiple-alignments," *Nucleic Acids Res.*, vol. 24, no. 19, pp. 3836–3845, 1996.

[143] D. Szklarczyk *et al.*, "The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res.*, vol. 39, no. SUPPL. 1, 2011.

[144] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, “The SVA package for removing batch effects and other unwanted variation in high-throughput experiments,” *Bioinformatics*, vol. 28, no. 6, pp. 882–883, 2012.

[145] M. E. Ritchie *et al.*, “Limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, 2015.

[146] M. Slotani, “Tolerance regions for a multivariate normal population,” *Ann. Inst. Stat. Math.*, vol. 16, no. 1, pp. 135–153, 1964.

[147] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, “Misc functions of the Department of Statistics (e1071), TU Wien,” *R package version 1.6-2*. p. http://cran.r-project.org/package=e1071, 2014.

[148] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard, “Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data,” *Genome Res.*, vol. 21, no. 3, pp. 447–455, 2011.

[149] M. A. el Hassan and C. R. Calladine, “Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA.,” *J. Mol. Biol.*, vol. 259, no. 1, pp. 95–103, 1996.

[150] S. P. Hancock, T. Ghane, D. Cascio, R. Rohs, R. Di Felice, and R. C. Johnson, “Control of DNA minor groove width and Fis protein binding by the purine 2-amino group.,” *Nucleic Acids Res.*, vol. 41, no. 13, pp. 6750–6760, Jul. 2013.

[151] B. Prud’homme, N. Gompel, and S. B. Carroll, “Emerging principles of regulatory evolution.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104 Suppl, pp. 8605–8612, 2007.

[152] S. H. Meijsing, M. A. Pufall, A. Y. So, D. L. Bates, L. Chen, and K. R. Yamamoto, “DNA binding site sequence directs glucocorticoid receptor structure and activity.,” *Science*, vol. 324, no. 5925, pp. 407–410, 2009.

[153] T. U. Wien, “an Examination of Indexes for Determining,” vol. 67, no. 3, 2002.

[154] B. Lenhard and W. W. Wasserman, “TFBS: Computational framework for transcription factor binding site analysis.,” *Bioinformatics*, vol. 18, no. 8, pp. 1135–6, 2002.

[155] R. Duda, P. Hart, and D. Stork, “Pattern Classification,” *New York John Wiley, Sect.*, p. 680, 2001.

[156] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. 1988.

[157] A. Nakashima *et al.*, “DEC1 modulates the circadian phase of clock gene expression.,” *Mol. Cell. Biol.*, vol. 28, no. 12, pp. 4080–4092, 2008.

[158] S. Honma *et al.*, “Dec1 and Dec2 are regulators of the mammalian molecular clock.,” *Nature*, vol. 419, no. 6909, pp. 841–844, 2002.

[159] M. Shen *et al.*, “Molecular characterization of the novel basic helix-loop-helix protein DEC1 expressed in differentiated human embryo chondrocytes.,” *Biochem. Biophys. Res. Commun.*, vol. 236, no. 2, pp. 294–298, 1997.

[160] H. Sun and R. Taneja, “Stra13 expression is associated with growth arrest and represses transcription through histone deacetylase (HDAC)-dependent and HDAC-independent mechanisms.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 8, pp. 4058–4063, Apr. 2000.

[161] Y. Li *et al.*, “Abundant expression of Dec1/stra13/sharp2 in colon carcinoma: its antagonizing role in serum deprivation-induced apoptosis and selective inhibition of procaspase activation.,” *Biochem. J.*, vol. 367, no. Pt 2, pp. 413–422, 2002.

[162] L. Shen *et al.*, “Hepatic differentiated embryo-chondrocyte-expressed gene 1 (Dec1) inhibits sterol regulatory element-binding protein-1c (Srebp-1c) expression and alleviates fatty liver phenotype.,” *J. Biol. Chem.*, vol. 289, no. 34, pp. 23332–

23342, Aug. 2014.

[163] M. Noshiro *et al.*, “Liver X receptors (LXRalpha and LXRbeta) are potent regulators for hepatic Dec1 expression.,” *Genes Cells*, vol. 14, no. 1, pp. 29–40, 2009.

[164] J. D. Crispino and M. M. Le Beau, “BMP Meets AML: Induction of BMP Signaling by a Novel Fusion Gene Promotes Pediatric Acute Leukemia,” *Cancer Cell*, vol. 22, no. 5. pp. 567–568, 2012.

[165] R. Chinery, J. A. Brockman, D. T. Dransfield, and R. J. Coffey, “Antioxidant-induced nuclear translocation of CCAAT/enhancer-binding protein beta. A critical role for protein kinase A-mediated phosphorylation of Ser299.,” *J. Biol. Chem.*, vol. 272, no. 48, pp. 30356–30361, Nov. 1997.

[166] S. Akira *et al.*, “A nuclear factor for IL-6 expression (NF-IL6) is a member of a C/EBP family.,” *EMBO J.*, vol. 9, no. 6, pp. 1897–1906, Jun. 1990.

[167] L. W. Harries *et al.*, “CCAAT-enhancer-binding protein-beta expression in vivo is associated with muscle strength,” *Aging Cell*, vol. 11, no. 2, pp. 262–268, 2012.

[168] S. Gery, A. F. Gombart, W. S. Yi, C. Koeffler, W. K. Hofmann, and H. P. Koeffler, “Transcription profiling of C/EBP targets identifies Per2 as a gene implicated in myeloid leukemia,” *Blood*, vol. 106, no. 8, pp. 2827–2836, 2005.

[169] G. J. Brem, I. Mylonas, and A. Brüning, “Eeyarestatin causes cervical cancer cell sensitization to bortezomib treatment by augmenting ER stress and CHOP expression,” *Gynecol. Oncol.*, vol. 128, no. 2, pp. 383–390, 2013.

[170] H. Mujcic *et al.*, “Hypoxic activation of the PERK/eIF2α arm of the unfolded protein response promotes metastasis through induction of LAMP3,” *Clin. Cancer Res.*, vol. 19, no. 22, pp. 6126–6137, 2013.

[171] S. Sukumaran, W. J. Jusko, D. C. Dubois, and R. R. Almon, “Light-dark oscillations in the lung transcriptome: implications for lung homeostasis, repair, metabolism, disease, and drug action.,” *J. Appl. Physiol.*, vol. 110, no. 6, pp. 1732–1747, Jun. 2011.

[172] C. Vollmers, S. Gill, L. DiTacchio, S. R. Pulivarthy, H. D. Le, and S. Panda, “Time of feeding and the intrinsic circadian clock drive rhythms in hepatic gene expression.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 50, pp. 21453–21458, Dec. 2009.

[173] I. G. Campbell, D. Choong, and G. Chenevix-Trench, “No germline mutations in the histone acetyltransferase gene EP300 in BRCA1 and BRCA2 negative families with breast cancer and gastric, pancreatic, or colorectal cancer.,” *Breast Cancer Res.*, vol. 6, no. 4, pp. R366-71, 2004.

[174] S. A. Gayther *et al.*, “Mutations truncating the EP300 acetylase in human cancers.,” *Nat. Genet.*, vol. 24, no. 3, pp. 300–303, 2000.

[175] D. P. Steensma *et al.*, “More on Myb in myelofibrosis: molecular analyses of MYB and EP300 in 55 patients with myeloproliferative disorders.,” *Blood*, vol. 107, no. 4. United States, p. 1733–5; author reply 1735, Feb-2006.

[176] E. A. Kimbrel and A. L. Kung, “The F-box protein beta-TrCp1/Fbw1a interacts with p300 to enhance beta-catenin transcriptional activity.,” *J. Biol. Chem.*, vol. 284, no. 19, pp. 13033–13044, 2009.

[177] G. Dieci, A. Conti, A. Pagano, and D. Carnevali, “Identification of RNA polymerase III-transcribed genes in eukaryotic genomes,” *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, vol. 1829, no. 3–4. pp. 296–305, 2013.

[178] D. Krubasik *et al.*, “Absence of p300 induces cellular phenotypic changes characteristic of epithelial to mesenchyme transition.,” *Br. J. Cancer*, vol. 94, no.

9, pp. 1326–1332, 2006.

[179] D. F. Salisbury *et al.*, “First-episode schizophrenic psychosis differs from first-episode affective psychosis and controls in P300 amplitude over left temporal lobe.,” *Arch. Gen. Psychiatry*, vol. 55, no. 2, pp. 173–180, Feb. 1998.

[180] M. E. Drake Jr., S. J. Huber, A. Pakalnis, and B. B. Phillips, “Neuropsychological and event-related potential correlates of nonepileptic seizures,” *J.Neuropsychiatry Clin.Neurosci.*, vol. 5, no. 0895–0172 SB–IM, pp. 102–104, 1993.

[181] S. P. Kutcher, D. H. Blackwood, D. St Clair, D. F. Gaskell, and W. J. Muir, “Auditory P300 in borderline personality disorder and schizophrenia.,” *Arch. Gen. Psychiatry*, vol. 44, no. 7, pp. 645–650, 1987.

[182] B. Kalayam and G. S. Alexopoulos, “Prefrontal dysfunction and treatment response in geriatric depression.,” *Arch. Gen. Psychiatry*, vol. 56, no. 8, pp. 713–718, 1999.

[183] C. F. Kügler *et al.*, “Visual event-related P300 potentials in early portosystemic encephalopathy.,” *Gastroenterology*, vol. 103, no. 1, pp. 302–310, 1992.

[184] F. E. Henken *et al.*, “The functional role of Notch signaling in HPV-mediated transformation is dose-dependent and linked to AP-1 alterations.,” *Cell. Oncol. (Dordr).*, vol. 35, no. 2, pp. 77–84, Apr. 2012.

[185] R. Chiu, W. J. Boyle, J. Meek, T. Smeal, T. Hunter, and M. Karin, “The c-Fos protein interacts with c-Jun/AP-1 to stimulate transcription of AP-1 responsive genes.,” *Cell*, vol. 54, no. 4, pp. 541–552, 1988.

[186] E. Tulchinsky, “Fos family members: Regulation, structure and role in oncogenic transformation,” *Histology and Histopathology*, vol. 15, no. 3. pp. 921–928, 2000.

[187] S. Langer *et al.*, “Jun and Fos family protein expression in human breast cancer: Correlation of protein expression and clinicopathological parameters,” *Eur. J. Gynaecol. Oncol.*, vol. 27, no. 4, pp. 345–352, 2006.

[188] I. Fialka, H. Schwarz, E. Reichmann, M. Oft, M. Busslinger, and H. Beug, “The estrogen-dependent C-junER protein causes a reversible loss of mammary epithelial cell polarity involving a destabilization of adherens junctions,” *J. Cell Biol.*, vol. 132, no. 6, pp. 1115–1132, 1996.

[189] K. I. Bland, M. M. Konstadoulakis, M. P. Vezeridis, and H. J. Wanebo, “Oncogene protein co-expression. Value of Ha-ras, c-myc, c-fos, and p53 as prognostic discriminants for breast carcinoma.,” *Ann. Surg.*, vol. 221, no. 6, pp. 706–720, Jun. 1995.

[190] A. Hoeben, B. Landuyt, M. S. Highley, H. Wildiers, A. T. Van Oosterom, and E. A. De Bruijn, “Vascular endothelial growth factor and angiogenesis.,” *Pharmacol. Rev.*, vol. 56, no. 4, pp. 549–580, 2004.

[191] V. Giguere, “Transcriptional control of energy homeostasis by the estrogen-related receptors,” *Endocr. Rev.*, vol. 29, no. 6, pp. 677–696, 2008.

[192] Z.-F. Yang, K. Drumea, S. Mott, J. Wang, and A. G. Rosmarin, “GABP transcription factor (nuclear respiratory factor 2) is required for mitochondrial biogenesis.,” *Mol. Cell. Biol.*, vol. 34, no. 17, pp. 3194–3201, Sep. 2014.

[193] M. F. Yueh and R. H. Tukey, “Nrf2-Keap1 Signaling pathway regulates human UGT1A1 expression in vitro and in transgenic UGT1 mice,” *J. Biol. Chem.*, vol. 282, no. 12, pp. 8749–8758, 2007.

[194] D. D. Zhang, “Mechanistic studies of the Nrf2-Keap1 signaling pathway.,” *Drug Metab. Rev.*, vol. 38, no. 4, pp. 769–789, 2006.

[195] T. Nguyen, P. J. Sherratt, and C. B. Pickett, “Regulatory mechanisms controlling gene expression mediated by the antioxidant response element.,” *Annu. Rev.*

Pharmacol. Toxicol., vol. 43, pp. 233–260, 2003.

[196] Y. Sato, "Role of ETS family transcription factors in vascular development and angiogenesis.," *Cell Struct. Funct.*, vol. 26, no. 1, pp. 19–24, 2001.

[197] A. Behrens, M. Sibilia, and E. F. Wagner, "Amino-terminal phosphorylation of c-Jun regulates stress-induced apoptosis and cellular proliferation.," *Nat. Genet.*, vol. 21, no. 3, pp. 326–329, 1999.

[198] R. Wisdom, R. S. Johnson, and C. Moore, "c-Jun regulates cell cycle progression and apoptosis by distinct mechanisms.," *EMBO J.*, vol. 18, no. 1, pp. 188–197, 1999.

[199] M. M. Vleugel, A. E. Greijer, R. Bos, E. van der Wall, and P. J. van Diest, "c-Jun activation is associated with proliferation and angiogenesis in invasive breast cancer," *Hum. Pathol.*, vol. 37, no. 6, pp. 668–674, 2006.

[200] I. A. Vasilevskaya and P. J. O'Dwyer, "Effects of geldanamycin on signaling through activator-protein 1 in hypoxic HT29 human colon adenocarcinoma cells," *Cancer Res.*, vol. 59, no. 16, pp. 3935–3940, 1999.

[201] R. Eferl *et al.*, "Liver tumor development. c-Jun antagonizes the proapoptotic activity of p53.," *Cell*, vol. 112, no. 2, pp. 181–192, Jan. 2003.

[202] M. Naumann, T. Rudel, B. Wieland, C. Bartsch, and T. F. Meyer, "Coordinate activation of activator protein 1 and inflammatory cytokines in response to Neisseria gonorrhoeae epithelial cell contact involves stress response kinases.," *J. Exp. Med.*, vol. 188, no. 7, pp. 1277–1286, 1998.

[203] B. Isermann and P. P. Nawroth, "The role of platelets during reproduction.," *Pathophysiol. Haemost. Thromb.*, vol. 35, no. 1–2, pp. 23–27, 2006.

[204] M. B. Kannan, V. Solovieva, and V. Blank, "The small MAF transcription factors MAFF, MAFG and MAFK: Current knowledge and perspectives," *Biochimica et Biophysica Acta - Molecular Cell Research*, vol. 1823, no. 10. pp. 1841–1846, 2012.

[205] K. Igarashi, K. Itoh, N. Hayashi, M. Nishizawa, and M. Yamamoto, "Conditional expression of the ubiquitous transcription factor MafK induces erythroleukemia cell differentiation.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 16, pp. 7445–7449, 1995.

[206] Y.-C. Shyu *et al.*, "Sumoylation of p45/NF-E2: nuclear positioning and transcriptional activation of the mammalian beta-like globin gene locus.," *Mol. Cell. Biol.*, vol. 25, no. 23, pp. 10365–10378, 2005.

[207] Y.-J. Hwang, E.-W. Lee, J. Song, H.-R. Kim, Y.-C. Jun, and K.-A. Hwang, "MafK positively regulates NF-κB activity by enhancing CBP-mediated p65 acetylation.," *Sci. Rep.*, vol. 3, p. 3242, 2013.

[208] S. J. Lu, S. Rowan, M. R. Bani, and Y. Ben-David, "Retroviral integration within the Fli-2 locus results in inactivation of the erythroid transcription factor NF-E2 in Friend erythroleukemias: evidence that NF-E2 is essential for globin expression.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, no. 18, pp. 8398–8402, Aug. 1994.

[209] A. Ray *et al.*, "Induction of matrix metalloproteinase 1 gene expression is regulated by inflammation-responsive transcription factor SAF-1 in osteoarthritis," *Arthritis Rheum.*, vol. 48, no. 1, pp. 134–145, 2003.

[210] A. Ray, A. Shakya, D. Kumar, and B. K. Ray, "Overexpression of serum amyloid A-activating factor 1 inhibits cell proliferation by the induction of cyclin-dependent protein kinase inhibitor p21WAF-1/Cip-1/Sdi-1 expression.," *J. Immunol.*, vol. 172, no. 8, pp. 5006–5015, 2004.

[211] D. D. Duncan, A. Stupakoff, S. M. Hedrick, K. B. Marcu, and G. Siu, "A Myc-

associated zinc finger protein binding site is one of four important functional regions in the CD4 promoter.," *Mol. Cell. Biol.*, vol. 15, no. 6, pp. 3179–3186, 1995.

[212] W. Zhong *et al.*, "Hypertrophic growth in cardiac myocytes is mediated by Myc through a Cyclin D2-dependent pathway.," *EMBO J.*, vol. 25, no. 16, pp. 3869–3879, 2006.

[213] H. Kim, J. Shin, S. Kim, J. Poling, H. C. Park, and B. Appel, "Notch-regulated oligodendrocyte specification from radial glia in the spinal cord of zebrafish embryos," *Dev. Dyn.*, vol. 237, no. 8, pp. 2081–2089, 2008.

[214] K. D. Yokoyama, Y. Zhang, and J. Ma, "Tracing the Evolution of Lineage-Specific Transcription Factor Binding Sites in a Birth-Death Framework.," *PLoS Comput. Biol.*, vol. 10, no. 8, p. e1003771, 2014.

[215] S. Ota, Z.-Q. Zhou, D. R. Keene, P. Knoepfler, and P. J. Hurlin, "Activities of N-Myc in the developing limb link control of skeletal size with digit separation.," *Development*, vol. 134, no. 8, pp. 1583–1592, 2007.

[216] S. U. Chen *et al.*, "Human chorionic gonadotropin up-regulates expression of myeloid cell leukemia-1 protein in human granulosa-lutein cells: Implication of corpus luteum rescue and ovarian hyperstimulation syndrome," *J. Clin. Endocrinol. Metab.*, vol. 95, no. 8, pp. 3982–3992, 2010.

[217] B. C. Delidow, B. A. White, and J. J. Peluso, "Gonadotropin induction of c-fos and c-myc expression and deoxyribonucleic acid synthesis in rat granulosa cells," *Endocrinology*, vol. 126, no. 5, pp. 2302–2306, 1990.

[218] M. D. Bettess *et al.*, "c-Myc is required for the formation of intestinal crypts but dispensable for homeostasis of the adult intestinal epithelium.," *Mol. Cell. Biol.*, vol. 25, no. 17, pp. 7868–7878, Sep. 2005.

[219] W. Jiang, I. Ferrero, E. Laurenti, A. Trumpp, and H. R. MacDonald, "c-Myc controls the development of CD8alphaalpha TCRalphabeta intestinal intraepithelial lymphocytes from thymic precursors by regulating IL-15-dependent survival.," *Blood*, vol. 115, no. 22, pp. 4431–4438, Jun. 2010.

[220] J. Milner, "RNA interference for treating cancers caused by viral infection.," *Expert Opin. Biol. Ther.*, vol. 3, no. 3, pp. 459–467, 2003.

[221] Y. Wang *et al.*, "Knockdown of c-Myc expression by RNAi inhibits MCF-7 breast tumor cells growth in vitro and in vivo.," *Breast Cancer Res.*, vol. 7, no. 2, pp. R220–R228, 2005.

[222] B. Li, J. O. Holloszy, and C. F. Semenkovich, "Respiratory uncoupling induces ??-aminolevulinate synthase expression through a nuclear respiratory factor-1-dependent mechanism in HeLa cells," *J. Biol. Chem.*, vol. 274, no. 25, pp. 17534–17540, 1999.

[223] M. J. Evans and R. C. Scarpulla, "Interaction of nuclear factors with multiple sites in the somatic cytochrome c promoter. Characterization of upstream NRF-1, ATF, and intron Sp1 recognition sequences," *J. Biol. Chem.*, vol. 264, no. 24, pp. 14361–14368, 1989.

[224] Y. S. Choi, S. Kim, H. K. Lee, K. U. Lee, and Y. K. Pak, "In vitro methylation of nuclear respiratory factor-1 binding site suppresses the promoter activity of mitochondrial transcription factor A," *Biochem. Biophys. Res. Commun.*, vol. 314, no. 1, pp. 118–122, 2004.

[225] K. Vercauteren, R. A. Pasko, N. Gleyzer, V. M. Marino, and R. C. Scarpulla, "PGC-1-related coactivator: immediate early expression and characterization of a CREB/NRF-1 binding domain associated with cytochrome c promoter occupancy

and respiratory growth.," *Mol. Cell. Biol.*, vol. 26, no. 20, pp. 7409–7419, 2006.

[226] W.-T. Chang, H. Chen, R.-J. Chiou, C.-Y. Chen, and A.-M. Huang, "A novel function of transcription factor alpha-Pal/NRF-1: increasing neurite outgrowth.," *Biochem. Biophys. Res. Commun.*, vol. 334, no. 1, pp. 199–206, 2005.

[227] M. Biswas and J. Y. Chan, "Role of Nrf1 in antioxidant response element-mediated gene expression and beyond," *Toxicology and Applied Pharmacology*, vol. 244, no. 1. pp. 16–20, 2010.

[228] J. S. Carew *et al.*, "Increased mitochondrial biogenesis in primary leukemia cells: the role of endogenous nitric oxide and impact on sensitivity to fludarabine.," *Leuk. Off. J. Leuk. Soc. Am. Leuk. Res. Fund, U.K*, vol. 18, no. 12, pp. 1934–1940, 2004.

[229] T. Kuwabara, J. Hsieh, K. Nakashima, K. Taira, and F. H. Gage, "A small modulatory dsRNA specifies the fate of adult neural stem cells," *Cell*, vol. 116, no. 6, pp. 779–793, 2004.

[230] Y. Huang, S. J. Myers, and R. Dingledine, "Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes.," *Nat. Neurosci.*, vol. 2, no. 10, pp. 867–872, 1999.

[231] Y. Naruse, T. Aoki, T. Kojima, and N. Mori, "Neural restrictive silencer factor recruits mSin3 and histone deacetylase complex to repress neuron-specific target genes.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 24, pp. 13691–13696, 1999.

[232] D. M. Kemp, J. C. Lin, and J. F. Habener, "Regulation of Pax4 paired homeodomain gene by neuron-restrictive silencer factor," *J. Biol. Chem.*, vol. 278, no. 37, pp. 35057–35062, 2003.

[233] X. Su *et al.*, "Abnormal expression of REST/NRSF and Myc in neural stem/progenitor cells causes cerebellar tumors by blocking neuronal differentiation.," *Mol. Cell. Biol.*, vol. 26, no. 5, pp. 1666–1678, 2006.

[234] J. Villard *et al.*, "A functionally essential domain of RFX5 mediates activation of major histocompatibility complex class II promoters by promoting cooperative binding between RFX and NF-Y.," *Mol. Cell. Biol.*, vol. 20, no. 10, pp. 3364–3376, 2000.

[235] Y. Xu, L. Wang, G. Buttice, P. K. Sengupta, and B. D. Smith, "Interferon gamma repression of collagen (COL1A2) transcription is mediated by the RFX5 complex.," *J. Biol. Chem.*, vol. 278, no. 49, pp. 49134–49144, 2003.

[236] T. Twardowski, A. Fertala, J. P. R. O. Orgel, and J. D. San Antonio, "Type I collagen and collagen mimetics as angiogenesis promoting superpolymers.," *Curr. Pharm. Des.*, vol. 13, no. 35, pp. 3608–3621, 2007.

[237] G. Millien *et al.*, "Characterization of the mid-foregut transcriptome identifies genes regulated during lung bud induction," *Gene Expr. Patterns*, vol. 8, no. 2, pp. 124–139, 2008.

[238] P. K. Sengupta, J. Fargo, and B. D. Smith, "The RFX family interacts at the collagen (COL1A2) start site and represses transcription," *J. Biol. Chem.*, vol. 277, no. 28, pp. 24926–24937, 2002.

[239] J. Naukkarinen *et al.*, "USF1 and dyslipidemias: Converging evidence for a functional intronic variant," *Hum. Mol. Genet.*, vol. 14, no. 17, pp. 2595–2605, 2005.

[240] P. Pajukanta *et al.*, "Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1).," *Nat. Genet.*, vol. 36, no. 4, pp. 371–376, 2004.

[241] L. T. Putowski, W. J. Schillings, C. M. Lee, E. P. Reddy, and J. A. Jakowicki, "Human follicle-stimulating hormone receptor (FSH-R) promoter/enhancer activity

is inhibited by transcriptional factors, from the upstream stimulating factors family, via E-box and newly identified initiator element (Inr) in FSH-R non-expressing cells.," *Gynecol. Endocrinol.*, vol. 19, no. 1, pp. 9–17, Jul. 2004.

[242] V. Kashyap and B. Bonavida, "Role of YY1 in the pathogenesis of prostate cancer and correlation with bioinformatic data sets of gene expression.," *Genes Cancer*, vol. 5, no. 3–4, pp. 71–83, Mar. 2014.

[243] M. K. Bennett, T. T. Ngo, J. N. Athanikar, J. M. Rosenfeld, and T. F. Osborne, "Co-stimulation of promoter for low density lipoprotein receptor gene by sterol regulatory element-binding protein and Sp1 is specifically disrupted by the yin yang 1 protein.," *J. Biol. Chem.*, vol. 274, no. 19, pp. 13025–13032, May 1999.

[244] A. Villagra, N. Ulloa, X. Zhang, Z. Yuan, E. Sotomayor, and E. Seto, "Histone deacetylase 3 down-regulates cholesterol synthesis through repression of lanosterol synthase gene expression," *J. Biol. Chem.*, vol. 282, no. 49, pp. 35457–35470, 2007.

[245] R. Rizkallah and M. M. Hurt, "Regulation of the transcription factor YY1 in mitosis through phosphorylation of its DNA-binding domain.," *Mol. Biol. Cell*, vol. 20, no. 22, pp. 4766–4776, Nov. 2009.

[246] E. B. Affar *et al.*, "Essential dosage-dependent functions of the transcription factor yin yang 1 in late embryonic development and cell cycle progression.," *Mol. Cell. Biol.*, vol. 26, no. 9, pp. 3565–3581, 2006.

[247] S. D. Bailey *et al.*, "ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters.," *Nat. Commun.*, vol. 2, p. 6186, 2015.

[248] N. Heidari *et al.*, "Genome-wide map of regulatory interactions in the human genome.," *Genome Res.*, vol. 24, no. 12, pp. 1905–1917, Dec. 2014.

[249] W. Lu, Z. Chen, H. Zhang, Y. Wang, Y. Luo, and P. Huang, "ZNF143 transcription factor mediates cell survival through upregulation of the GPX1 activity in the mitochondrial respiratory dysfunction.," *Cell Death Dis.*, vol. 3, p. e422, 2012.

[250] H. Izumi *et al.*, "Role of ZNF143 in tumor growth through transcriptional regulation of DNA replication and cell-cycle-associated genes," *Cancer Sci.*, vol. 101, no. 12, pp. 2538–2545, 2010.

[251] M. Fujii *et al.*, "Convergent signaling in the regulation of connective tissue growth factor in malignant mesothelioma: TGFbeta signaling and defects in the Hippo signaling cascade.," *Cell Cycle*, vol. 11, no. 18, pp. 3373–3379, Sep. 2012.

[252] C. Schwartz *et al.*, "Recruitment of p300 by C/EBP?? triggers phosphorylation of p300 and modulates coactivator activity," *EMBO J.*, vol. 22, no. 4, pp. 882–892, 2003.

[253] B. Karanam *et al.*, "Multiple roles for acetylation in the interaction of p300 HAT with ATF-2.," *Biochemistry*, vol. 46, no. 28, pp. 8207–8216, Jul. 2007.

[254] P. Zhang *et al.*, "Expression of COUP-TFII in metabolic tissues during development," *Mech. Dev.*, vol. 119, no. 1, pp. 109–114, 2002.

[255] P. Bailey, V. Sartorelli, Y. Hamamori, and G. E. Muscat, "The orphan nuclear receptor, COUP-TF II, inhibits myogenesis by post-transcriptional regulation of MyoD function: COUP-TF II directly interacts with p300 and myoD.," *Nucleic Acids Res.*, vol. 26, no. 23, pp. 5501–5510, 1998.

[256] Y. S. Dai and B. E. Markham, "p300 Functions as a Coactivator of Transcription Factor GATA-4," *J. Biol. Chem.*, vol. 276, no. 40, pp. 37178–37185, 2001.

[257] K. Sun, M. a Battle, R. P. Misra, and S. a Duncan, "Hepatocyte expression of serum response factor is essential for liver function, hepatocyte proliferation and survival, and postnatal body growth in mice.," *Hepatology*, vol. 49, no. 5, pp.

1645–54, 2009.

[258] C. S. Lee, J. R. Friedman, J. T. Fulmer, and K. H. Kaestner, “The initiation of liver development is dependent on Foxa transcription factors.,” *Nature*, vol. 435, no. 7044, pp. 944–947, 2005.

[259] P. Shore and A. D. Sharrocks, “The transcription factors Elk-1 and serum response factor interact by direct protein-protein contacts mediated by a short region of Elk-1.,” *Mol. Cell. Biol.*, vol. 14, no. 5, pp. 3283–3291, May 1994.

[260] a Minn *et al.*, “Genes that mediate breast cancer metastasis to lung,” *Nature*, 2005.

[261] P. D. Bos *et al.*, “Genes that mediate breast cancer metastasis to the brain,” *Nature*, 2009.

[262] J. C. Harrell *et al.*, “Genomic analysis identifies unique signatures predictive of brain, lung, and liver relapse,” *Breast Cancer Res. Treat.*, 2012.

[263] A. Prat, B. Adamo, M. C. U. Cheang, C. K. Anders, L. A. Carey, and C. M. Perou, “Molecular Characterization of Basal-Like and Non-Basal-Like Triple-Negative Breast Cancer,” *Oncologist*, 2013.

[264] A. Prat *et al.*, “Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer,” *Breast Cancer Res.*, 2010.

[265] P. S. Bernard *et al.*, “Supervised risk predictor of breast cancer based on intrinsic subtypes,” *J. Clin. Oncol.*, 2009.

[266] J. Tan, J. H. Hammond, D. A. Hogan, and C. S. Greene, “ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions,” *mSystems*, 2016.

[267] U. Shaham *et al.*, “Removal of Batch Effects using Distribution-Matching Residual Networks.,” *Bioinformatics*, Apr. 2017.

[268] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal Deep Learning,” *Proc. 28th Int. Conf. Mach. Learn.*, 2011.