Deciphering Transcriptome: Transcription Factors as Master Mind Behind Gene Expression

Deciphering Transcriptome: Transcription Factors as Master Mind Behind Gene Expression

A DISSERTATION PRESENTED
BY
WENTING CAI
TO
THE DEPARTMENT OF CHEMISTRY AND CHEMICAL BIOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
CHEMISTRY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
JUNE 2019

Thesis advisor: Professor Xiaoliang S. Xie Wenting Cai

Deciphering Transcriptome: Transcription Factors as Master Mind Behind Gene Expression

Abstract

Transcription Factors (TFs) are key players in orchestrating diversified transcriptomes across cell types. Known to function in co-operative pairs, TFs are dynamic and hard to be experimentally captured in motion. With an innovated scRNA-Seq technique, MALBAC-DT, the steady-state measurements were dissected deeper into the dynamic fluctuations. Provided with precise and accurate expression correlations, co-activated gene pairs were revealed to be expressed in the same cell at the same time, in contrast to the traditional terminology 'co-expression', which refers to the similar expression patterns across different cell populations. Further, in combination with motif analysis on open chromatin, we inferred the co-localized and co-activated combinatorial TF pairs (TF3C), which shed light onto the co-operative regulation of TFs on the target genes. Differential across cell types yet highly preserved within co-regulated gene clusters, TF3C led us to propose a transcription regulation scheme: gene looping of the target gene's promoter, enhancer, and TTS, assisted by the shared TF3Cs. Finally, we present a TF combinatorial regulation map that is unique to each cell-type, capturing the essence of the dynamic fluctuations otherwise convoluted in each cell population.

Contents

Listing of figures

Listing of tables

Dedicated to Prof. Xiaoliang Xie, who taught me to fearlessly explore the scientific world by thinking across boundaries.

Acknowledgments

Just like transcription factors which work together to bring manic to protein, humans cannot work alone either. Throughout my six years of graduate study, I have received help from numerous seniors, colleagues, friends, and families, without whom I would never have been able to finish this arduous journey.

First and foremost, I would like to express my sincere gratitude towards my advisor, Prof. Sunney Xie, who offered me the opportunity to work in such a world-class laboratory and to work on cutting edge researches in the frontier of biology. Sunney has provided me with continuous support and guidance throughout my graduate study, and he never stops to inspire me with fresh perspectives towards tough scientific questions and to affect me with his intense excitement of scientific discovery. His unique framework for problem-solving and distilled enthusiasm for science has already left distinct markers on me. I am blessed to carry these markers on myself for the rest of my life's journey as I'm very confident they will continue to support and guide me for whatever challenges awaits.

I would also like to thank my committe members, Prof. Adam Cohen and Prof. Hongkun Park. The insightful comments given by the two great minds really broadened my vision and benefited my research.

I would like to thank Dr. Yan Gu for her rigorous molecular biology training on me. For the first three years of my graduate study, I have spent so much time with her together from pipetting to western blot. I still vividly remember all the nervousness and excitement I had after every batch of the experiment and all the brain-storming and soul-searching after every failure. Although the experiments didn't all go as we expected, I truly cherished the mentorship I received and enjoyed the friendship we develop.

I would like to thank Dr. Alec Chapman, Dr. David Lee, Mr. Wenping Ma, Mr. Xiang Li, Dr. Yinghui Zheng, Dr. Yi Yin, and Dr. Sabin Mulapeti, for our collaborative projects. I switched to bio-informatics in my 3rd year, and I am incredibly grateful towards my experimental colleagues for their patience for my gradual start. Data crunching is not an easy task, and I would have wasted a lot more fertile effort if it's not for the help of Alec. He shared with me years of experience without any reservation and carried me through the initial transition period, and I found myself always learning something new whenever I have a discussion with him. There's little doubt that experimental work is even harder than the data-crunching tasks, especially when you are working with finicky single cells. My experimental collaborators have been working day and night to perfect the protocols and collect the important data, which I relied upon for my analysis. Working with those brightest minds in biology have greatly inspired my understandings of gene regulations, and I'm extremely pleased to see that years of hard collaborative work is culminating into significant scientific discoveries.

The Xie lab is not only a group but also a family. I enjoyed every moment in the lab with all the wonderful members. I would like to thank Dr. Patricia Purcell for her bits of advice both in academics and in life. My son's favorite sleeping spot is still the old crib from Aunt Patty. I would like

to thank Dr. Longzhi Tan, Dr. Dong Xing, Dr. Chongyi Chen, Dr. Dan Fu, Dr. Shasha Chong, Mr. Zi He, Dr. Ziqing Zhao, Dr. Xu Zhang, Dr. Minbiao Ji, Dr. Fa-ke Lu, Dr. Asaf Tal, Dr. Lei Huang, Dr. Bo Zhao, Dr. Wenlong Yang, Mr. Chi-Han Chang, Mr. Yunlong Cao, for all the scientific discussion we had and all the laughter we shared.

Finally, I would like to thank my parent for their continuous support and understanding. For the pursuit of academic careers and scientific advancement, I wasn't able to stay by their side, and owing to the sophisticated visa screening process I wasn't even able to come back to China to visit them. But in my heart, they will always be the people I can turn to, and the harbor I can come back to from the outside storm. I'm happy that I have been able to extend my family during my graduate study with my loving husband Tianyang Ye and my lovely son Boya Ye, as well as my beloved cat Taxue and bunny Gene. And here with your unconditional love and support, we will together build a warm home and will work together for a better future!

Genome: bought the book; hard to read.

Eric Lander [1]

1

Introduction

Despite having thousands of different cell types, an organism shares the same genome. Although we have the genome blueprint spelled out word by word, we still can not confidently to say that it has been decoded.

In order to cope with different environments and effectuate different function a human requires, cells need to be as diversified as possible, by producing different proteins. How can we go from the

same genome to diverging proteomes? A review at the central dogma (Figure 1.1□ directs our focus to the middle level between DNA and protein, transcripts. Transcriptome represents the upper-stream regulation of proteomics. If we can decode the first step in the realization of genome information by figuring out how transcription is regulated, we are one step further to decode the human genome.

Figure 1.1: Central Dogma. The process of how information inherited and stored in the genome flows into proteins *via* a two-step process: first a transcription resulting in RNA from DNA, then a translation from RNA to protein. Here, the transcription process is highlighted to show the importance of it being one of the first steps to differentiate cells from cells.

1.1 Transcription factors are essential to regulate transcriptome

Transcription Factors (TF) are DNA binding proteins that regulate gene expression.[2] They bind to regulatory elements, such as promoters and enhancers, to activate or suppress the transcription of the target gene. In Chapter 2, it will be explored how TFs are specific to each tissue, with the number of expressed TFs in a linear relationship with the total number of genes expressed. As the diversified expression patterns of TFs characterizes tissue-specific states, a TF regulation hierarchy unique to each cell type is in need. In order to build such regulation hierarchy, the combinatorial function of transcription factors in connection with the diversified cell states as a result of tissue differentiation needs to be uncovered. However, the deduction of such cell-type specific network is convoluted by the bulk measurements of transcriptomes at the tissue level.

1.2 Advancement of the transcriptomics technologies

With an illustrated need to dissect the steady state to reveal the hidden dynamic system, a high accuracy single-cell RNA-Seq that can capture transcription fluctuations is required. Nevertheless, the current methods are only designed to differentiate cells within a heterogeneous population.[3–8] The accuracy would only be enough to do cell-typing and identify steady states, but would not be sufficient to probe into the more delicate fluctuations around the equilibrium of each state. Designed with high accuracy, MALBAC-DT will be demonstrated in Chapter 3 about how it can reveal the vibrant gene-gene interactions from a steady state population. In contrast to 'co-expression' that describes two genes being expressed in the same population, 'co-activation' is proposed as a more suitable word to represent the close association of two genes that are being expressed in the same cell at the same time. The superiority of co-activation analysis as compared to co-expression is then established by the better inference of protein-protein interaction, implying a more accurate representation of functional relationships between gene pairs. Expanding a whole new realm to the transcriptome analysis in addition to simply compare expression levels, the correlational analysis of co-activation shows excellent potential in shedding light on elucidating transcription regulation.

1.3 Inference of co-operative TF pairs

Integrating the knowledge of TF from Chapter 2 and gene co-activation from Chapter 3, Chapter 4 is dedicated to revealing the co-operative TF pairs that play a crucial role in helping the limited number of TFs regulating an expansive transcriptome. The combinatorial function of transcrip-

tion factors is untangled first by inferring TF interacting pairs from them being co-activation, then the confidence of the inference is further strengthened by extrapolating pairs co-localized to the same open window on the genome. Together, the inference of co-localized and co-activated combinatorial TF pairs (TF3C) is used to elucidate the otherwise complex regulation of a specific gene, rendering a genome-wide combinatorial regulation map.

1.4 References

[1] Lander, E. The genome. https://www.improbable.com/airchives/paperair/volume9/v9i6/nano/nano_6.php (2003). Online; accessed 9 May 2019.

[2] Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650 – 665 (2018).

[3] Klein, A. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

[4] Macosko, E. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

[5] Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236 – 240 (2013).

[6] Grün, D. *et al.* Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).

[7] Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science* **347**, 1138–1142 (2015).

[8] Usoskin, D. *et al.* Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature Neuroscience* **18**, 145–153 (2014).

2

Characteristic TF Expressions Among Various Human Cells

"Qin zei xian qin wang". This is an old Chinese saying noting that to defeat an enemy, you start by catching the King. In order to unravel the complex transcription regulation network, we start by

interrogating the director behind this show, the Transcription Factors.

2.1 What are Transcription Factors and why are they important?

Transcription Factors (TFs), as indicated by its name, are protein factors that can regulate transcription, either by suppressing or enhancing target genes. [1,2] TFs are distinguished from the rest of the proteins that have an effect on transcription by having a strong binding specificity to the DNA, which suffices to be the very definition of a protein being a TF, as argued by Lambert et al. [2] Such high binding specificity is usually measured by the dissociation constant of TFs binding to DNA, with its binding to a preferred sequence exhibiting a dissociation constant that can be as high as 1,000 fold slower as compared to its binding to a non-preferred sequence, while their association constants are of about the same magnitude. [3]

The strong binding specificity of TF is effectuated by part of its protein structure, referred to as the DNA-binding domain (DBD). [4] Large protein families tend to share similar DBDs, thus making it possible to identify and categorize TFs based on the DBD family. [2] Lambert et al. cataloged 1,600 human TFs in their recently published database, which is used as the criteria of TF in this dissertation.

The functions of TFs upon binding to DNA are diversified, and can either act as a repressor or activator of their target gene. Most commonly, the role of a TF in regulating its targeting gene is by recruiting the necessary cofactors, including Mediators and RNA Polymerase II, thereby enhancing or activating the expression of the target gene. [5] However, efforts to catalog the interactions among

cofactors have failed to see a significant number of TF involvements due to the weak nature of the TF-TF and TF-cofactor interactions.[2,6] This is in accordance with the short residence time observed for TFs, leading to their dynamic characteristics.[7] On the other hand, repression is achieved by blocking the binding of such proteins necessary for the ongoing transcription.[8] Interestingly and counter-intuitively, many TFs have been found to have a dual role at the same time: Nuclear transcription factor Y (NFY), a transcription factor known for promoting the chromatin accessibility for other binding proteins, can either activate or repress one of its target genes, VMF (Von Willebrand factor), when binding to the promoter on different consensus sequences depending on which cofactors are involved.[9] This further amplified the complexity when decoding the function and role of TF in regulating gene expression.

Another dimension to the unsolved puzzle regarding TF function is where do they bind. There are two major classes of *cis*-regulatory elements that TFs tend to bind and exert effect: promoters, and enhancers.[10] A promoter is a sequence around the transcription start site (TSS) where the RNA polymerase II, along with other factors, are recruited to initiate and maintain the downstream genetic transcription. The field has been focusing on TF binding for decades due to its convenience in the determination of the target gene only by proximity.[11] On the other hand is the enhancers, which are distal sequences that have a long-range effect on the expression of a target gene.[12] TFs are believed to form protein complexes that bridge the enhancer-promoter pair, yielding a stable loop with access to polymerase and factors bound on both sites.[10] The existence of such loop is also experimentally observed in two recent works with improved precision on inferring chromatin structure[13,14]. However, it is still a great challenge to determine the target of enhancer elements,

nevertheless to say the function of the TFs bound on them. The uncertainty in regulation targets further adds to the complexity of deciphering the transcription regulation network.

With its sophisticated functions and binding sites, TFs are capable of having highly diversified roles to accommodate the massive number of different tissues and cell types developed from the same genome. During the hematopoiesis differentiation process, TAL1 (TAL BHLH Transcription Factor 1, Erythroid Differentiation Factor) serves as one of the key members to initiate the lineage specification, with different targets genes at different stages and with different partners or cofactors, while different TFs are regulating the expression of itself at different stages.[15] The five 'different' indicated in the last sentences summarize the exact extent to which the regulatory network operated by transcription factors is entangled, thus extremely hard to be fully decoded.

Logically, one would expect the natural balance of such an intricate system to be tipped over if the validity of TF is perturbed: about 20% of all oncogenes identified for human so far are transcription factors.[16] With more TF functions unraveled, drugs targetting TFs are being developed, but only a handful of the developed drugs have entered into clinical trials.[16] Therefore, as hard as it is to unveil the truth behind a transcription factor regulation network, it is essential and in urgent need to keep advancing the field. To start, we can look at the differential expression of a TF across different cell types in development and diseases, hoping that this will shed light on its role in organo-development and oncogenesis.

2.2 Differential TFs hidden behind differential transcriptome across various cell types.

Over the past decades, a handful of endeavors have been dedicated to gain knowledge of the transcriptome across different cells: Human Protein Atlas (HPA)[17] (www.proteinatlas.org), The Genotype-Tissue Expression (GTEx) Project[12], and The functional annotation of the mammalian genome 5 (FANTOM5) project[18].

In this work, the focus will be on HPA, a database comprised of 37 human tissues and 64 cell lines. The genes that are annotated to be TFs and analyzed here are based on Lambert et al.[2] and the Cis-BP database (catalog of inferred sequence binding preferences)[19]. There are in total 1,639 TFs cataloged under *Homo sapiens* in Cis-BP, and out of those, there are 1,595 expressed in any of the 101 cell types from the tissue panel.

As pointed out by the data publisher, the transcriptome panel has shown a great disparity between immortalized cell lines and normal human tissues, since nearly all the *in vitro* cultured cell lines are cancerous and do not have biologically normal behavior.[20] There are 1,217 genes uniquely expressed in some human tissues and not in the studied cell lines, whereas there are only 254 genes characteristic to cell lines. On the other hand, each type of human tissue expresses 16% more type of genes on average than a cell line does, as illustrated in Figure 2.1(b). All these differences can be attributed to the more vibrant functions possessed by the 37 analyzed tissues, each of which is comprised of a highly heterogeneous collection of cells. Although the cancerous cell lines are known to be heterogeneous as well[21], it is not comparable to tissues, as the latter requires a great complexity to

be able to perform a diverse set of functions, for example, the perfect coordination of developmental processes all controlled by signaling molecules. [17,20].

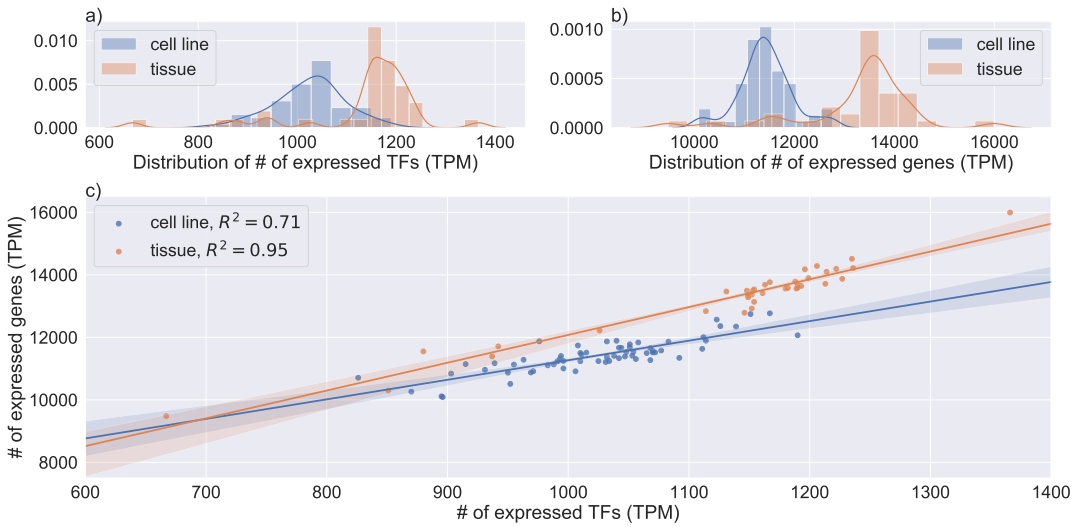


Figure 2.1: Number of genes and TFs expressed in each cell type. The RNA-seq expression levels are from Human Protein Atlas [17] (www.proteinatlas.org), including 37 human tissues and 64 cell lines. a) and b) each depicts the distribution histogram of the number of genes and TFs in each cell type, separated by two categories: immortalized cell line, or human tissues. In both genes and TFs, human tissues show a much higher count than the cell lines. c) illustrates the relationship between the number of genes expressed vs. number of TFs expressed, which is linear in each category, with human tissues have a much steeper slope, indicating more targets per TF.

The highly diversified tissue transcriptome is not possible without a simultaneously diversified collection of TFs. Interestingly, the ratio of the number of genes expressed *vs.* the number of TFs expressed showed a linear relationship, with a R^2 being 0.95 for human tissues: for each additional TF, 8.8 more genes would be expressed. This linear relationship also exists, albeit weaker, for the cancerous cell lines: a smaller slope of 6.3 genes for each additional TF, with a weaker $R^2 = 0.71$. (2.1(c)) The fundamental difference in the gene:TF ratio among the two cell/tissue types outlines the possibility of TF playing a vibrant role in facilitating the developmental need in tissue differentiation, as

every single TF is responsible for regulating more targets in human tissues than in cell lines.

2.3 TF clusters differ across human tissues.

As shown in the previous section that TFs are driving the diversification of tissues, the next focus would be on the expression profiles of the TFs only. When only restricted to the TF expression profiles, we can detect a clear distinction between immortalized cell lines and healthy human tissues, as illustrated in Figure 2.2. As shown, the sub-transcriptome of 101 cell types are reduced in dimensionality using principal component analysis (PCA). As expected, some tissues that are functionally related are close to each other as they share similar transcriptome signatures. Such as duodenum, esophagus, small intestine, and colon, all come from gastrointestinal (GI) tract and aggregate in the lower part of the PCA plot. There are examples, however, of functionally adjacent tissues being distant, such as skeletal muscle being away from both smooth muscle and heart muscle. The contradiction is probably caused by the heterogeneity exhibited by the skeletal muscle tissue, with numerous subtypes and subpopulations [22], which are difficult to detect in bulk measurements.

Meanwhile, clusters of TFs show similarity when steering tissue specification, as seen in Figure 2.3. Some of these lineages are driven by the group-wise up-regulation fo TF cluster that stands out in a single tissue. For example, cerebral cortex possesses a TF gene group (OLIG1, OLIG2, ARNT2, SOX8, HEY1, MYT1I, LHX2, ZNF365) that are simultaneously up-regulated, with a combined expression level that is 5-folds of that of the second highest tissue. OLIG1/2, or oligodendrocyte transcription factor 1/2, are two basic helix–loop–helix transcription factors that specialize in facil-

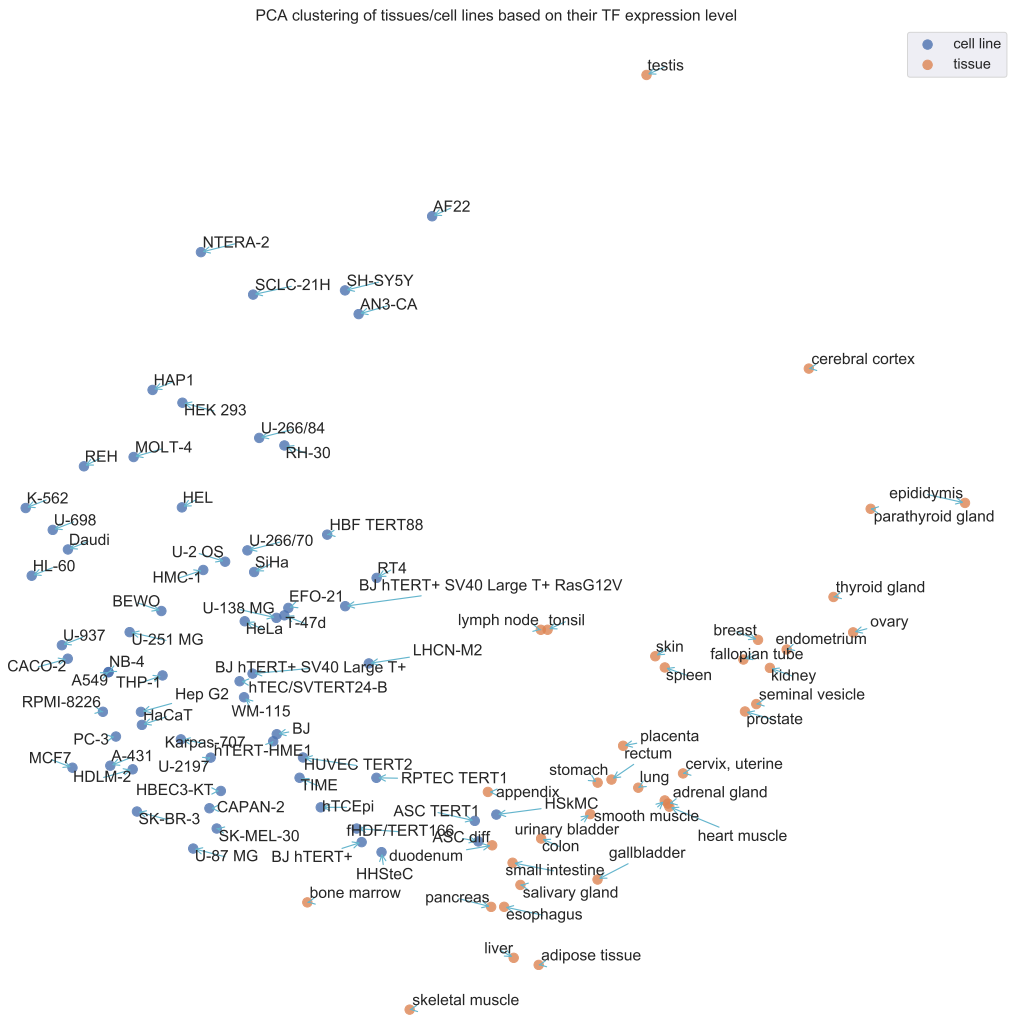


Figure 2.2: Dimensionality reduction for Tissues/Cell lines using PCA based on TF expression. The RNA-seq expression levels are from Human Protein Atlas [17] (www.proteinatlas.org), including 37 human tissues and 64 cell lines, with 1,595 TFs in total. The expression profiles are normalized by cell types before doing principle component analysis (PCA). Each point represents a cell, with its color indicating whether it is tissue or cell line. The clear separation between the two types outlined a fundamental difference between healthy tissues and immortalized cell lines. A closer look at the tissues reveals the aggregation of a collection from gastrointestinal (GI) tract: duodenum, esophagus, small intestine, and colon. In contrary, skeletal muscle is far away from both smooth muscle and heart muscle, as compared to other more different tissues, implicating that its heterogeneity complicates the true expression profile of skeletal muscle upon the bulk measurement of the transcriptome.

itating the formation of the central nervous system (CNS).[23] Apart from these two highly specialized TFs, there is also ETV1 present in this group, which is not only significantly up-regulated in the cerebral cortex, but also in salivary glands. The role of ETV1 in salivary gland is mainly related to epithelial-mesenchymal plasticity as salivary glands undergo branching during morphogenesis[24], whereas its role in the cerebral cortex is associated to regulation of neuronal maturation genes[25]. ETV1, as specialized as it is, does not seem to have a unified role across cell types, implying that its function is probably highly dependent on other co-expressed and co-functional TFs.

As for epididymis, a highly convoluted duct behind the testis, there are also highly specialized TFs responsible for its lineage (EMX2, HOXB3, HOXB7, HOXB9, HOXD8, ZNF189), each of which peaks in epididymis in expression level, giving a combined expression level of 3-folds of that of the second highest tissue (2.3c). Most TFs from this group are homeobox genes (HOX), with three from the B family. Homeobox genes are composed of a large family of TFs that are master regulators during embryogenesis, and orchestrate the limb development and organogenesis in pre- and post-natal life.[26,27] Not only their actual specific roles in the epididymis are mostly unknown, the three HOX genes, HOXB3/7/9, are also mysteriously simultaneously up-regulated in acute myeloid leukemia (AML), along with other HOX family genes.[28] Being a TF and oncogene at the same time, HOXB3/7/9 have proven to have a highly diverse role in maintaining the wellbeing of a human body, and the mechanisms are not easily unveiled behind such intricate network.

After seeing how two specialized TF groups are connected with certain tissues, we studied a family of TFs that are differentially co-upregulated in a selection of tissues. For example, the GATA family of genes got the family name from its shared binding specificity to the "GATA" sequence

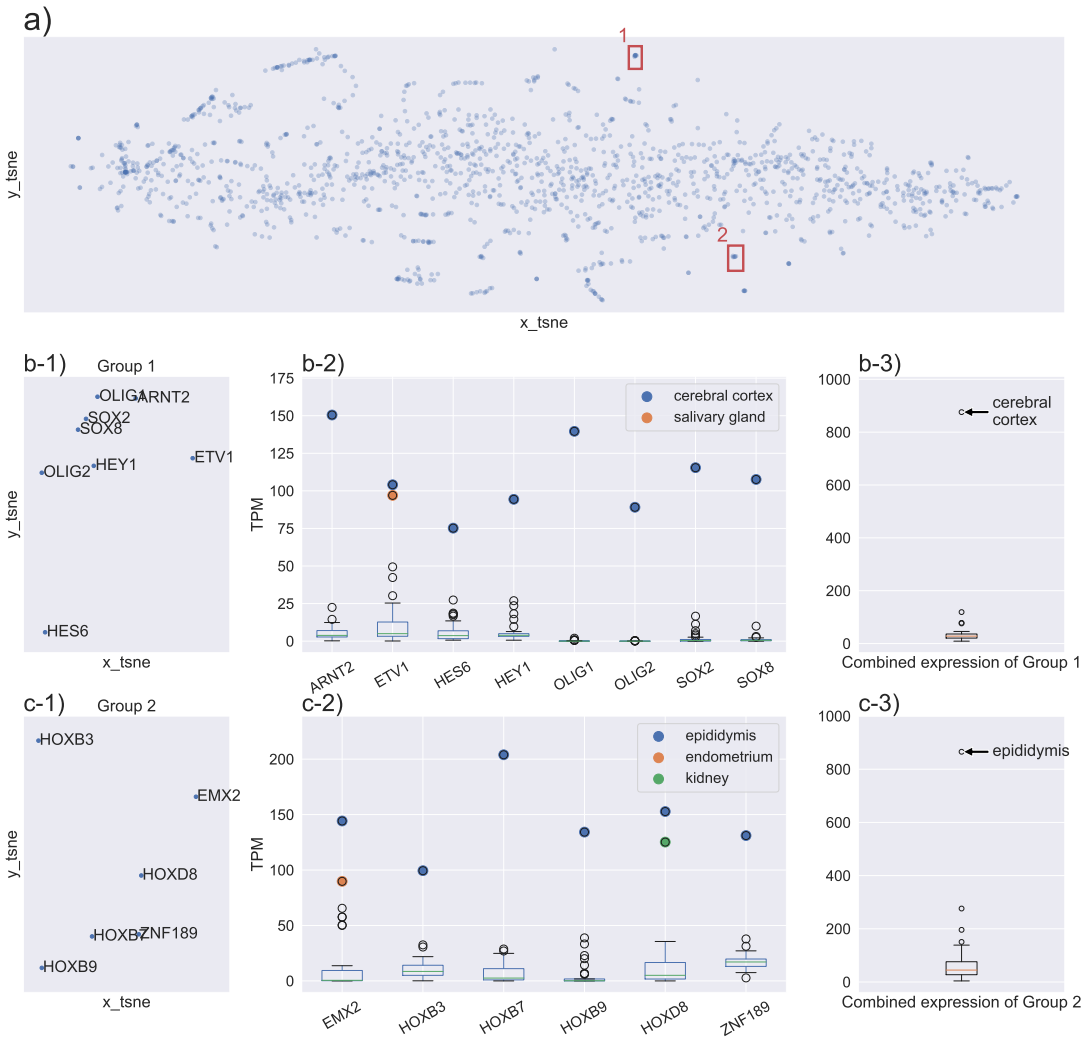


Figure 2.3: Dimensionality reduction for TF genes using t-SNE. a) The RNA-seq expression levels are from Human Protein Atlas [17] (www.proteinatlas.org), including 37 human tissues with 1,595 TFs in total. The expression profiles are normalized by cell types before doing t-Stochastic Neighbor Embedding(t-SNE). Each dot represents a TF gene. b-c) are two example groups of TFs that are clustered together because they are significantly up-regulated in one particular tissue, each of which is picked from regions of a) that are annotated with a red box. Of these two pairs of figures, (b/c)-1 show the t-SNE plot zoomed in on the groups, and (b/c)-2/3 show the distribution of the individual or combined expression of TFs from this group across all 37 human tissues, with the outlier tissue annotated. Group 1-2 are corresponding to cerebral cortex and epididymis, respectively.

and is required for the differentiation of mesoderm-, endoderm-, and ectoderm-derived tissues.[29,30] As shown in Figure 2.4, six members of the GATA family are coordinated across human tissues in different pairs. i.e., GATA4/6 show great similarity in expression patterns across human tissues, while being up-regulated at the same time in many tissues, such as heart muscle, ovary, and stomach, GATA4 is co-upregulated with a different partner, GATA5, in other tissues such as duodenum. These co-expression patterns are consistent with previous findings regarding the crucial role of GATA4/5/6 in organ development, demonstrated by mutations or knockout mice which usually resulted in severe defects in organ formation or even lethality at the early embryo stage.[30]. How GATA family members orchestrate the transcription regulation network in such diversified roles yet with highly conserved consensus binding sequence is still largely unknown.

In addition to TF groups that are co-upregulated in specific tissues, we also observed TF families that are universally expressed across human tissues, such as the activating transcription factor, ATF. ATFs do not have that much a disparity across tissues as compared to the GATA family, but rather exhibit a more uniform expression pattern, with each member slightly up- or down-regulated in a couple of different tissues. This uniformity in expression profiles is consistent with previous literature that each member of the ATF family is a multifaceted player in embryogenesis and organogenesis along with tissue homeostasis, with participation in various pathways.[31] Nevertheless, it is not possible to further dissect the specific roles of ATF family in different tissues just from the bulk measurements. Albeit exhibiting a similar expression level in the bone marrow and the muscle tissues, ATF1 targets different downstream genes in these two lineages. In vascular smooth muscle tissues, ATF1 induces the expression of NOX1, a catalytic subunit of NADPH oxidase involving in the mi-

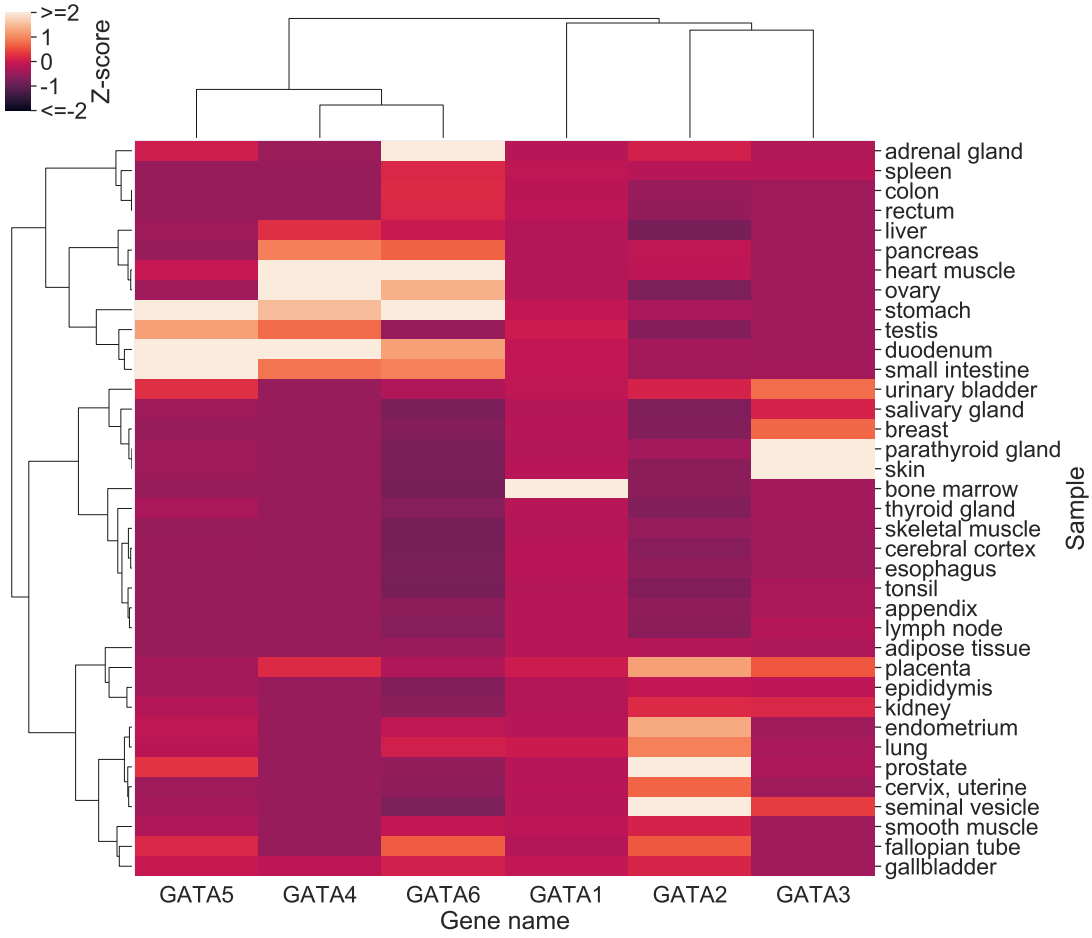


Figure 2.4: Differential Expression of GATA family across Tissues. The RNA-seq expression levels are from Human Protein Atlas [17] (www.proteinatlas.org) The expression profiles are first normalized across 37 human tissues, then the Z-scores are calculated for each TF gene. The dendrograms are clustered based on the Pearson correlation coefficient. Z-scores beyond 1.6 are clipped to show the difference. 37 human tissues are labeled on the y-axis.

tochondrial respiratory chain.[32] On the other hand, ATF1 phosphorylates the ataxia telangiectasia-mutated (ATM) protein kinase, which is one of the first responders to DNA double-strand breaks induced by IR.[33] The two different roles are not reflected in the similar expression levels of ATF1 in the two tissues.

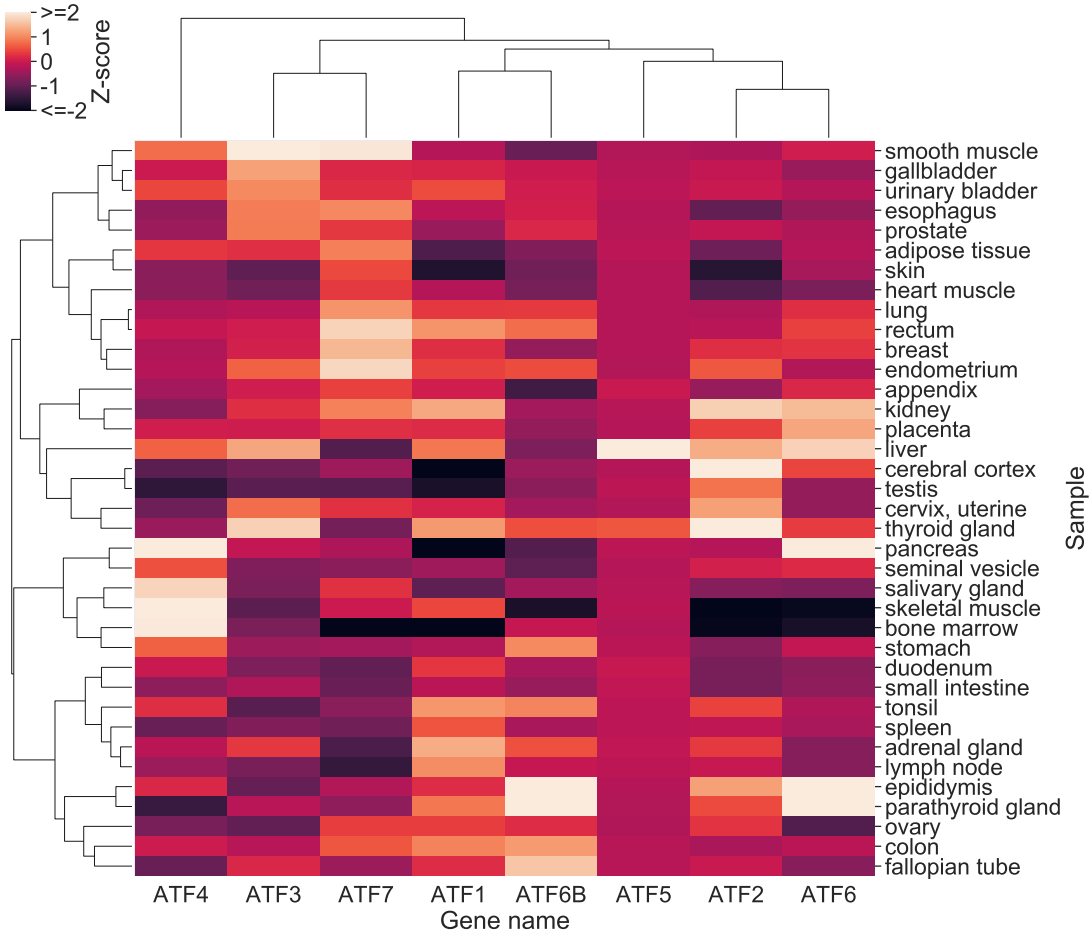


Figure 2.5: Differential Expression of ATF family across Tissues. The RNA-seq expression levels are from Human Protein Atlas[17] (www.proteinatlas.org) The expression profiles are first normalized across 37 human tissues, then the Z-scores are calculated for each TF gene. The dendrograms are clusterd based on Pearson correlation coefficient. Z-scores beyond 2.0 are clipped to show the difference. 37 human tissues are labeled on the y axis.

In summary, lists of TF genes that are up- or down-regulated in each tissue are summarized in

Table 2.1.

2.4 Conclusion: Transcription Factors as master mind

As demonstrated in this Chapter, transcription factors are indeed the key players in orchestrating the transcriptome to accommodate the dynamic nature of the cell population. The different gene:TF ratios in cell lines and human tissues underlies the fundamental differences in the TF regulation networks among these two major categories of cells, implying a much more intricate network structure for the dynamic tissues. Nevertheless, if we only look at the expression levels of TFs in each tissue, they can neither further dissect the heterogeneous population embedded within, nor shed light on the differential TF network the tissue exhibits.

Skeletal muscles are heterogeneous with its ubiquitous presence all over the human body, thus making the bulk measurement of transcriptome far off where it should be: close to smooth muscle and heart muscle, as compared to other functionally distant tissues. When it comes to extrapolating the functions of TFs based on their expression level, the puzzle gets more complicated. On the one hand, there are gene clusters that are expressed significantly higher in one specific tissue, implying their unique functions in it. On the other hand, these gene clusters are still expressed in the rest of tissues, despite in a much lower level, making the function inferred earlier not easily transferrable.

In addition to this, around half of the TFs are expressed uniformly across the tissues, such as ATF family that are known to possess a dynamic functional role across the tissues, making it even more challenging to compare and contrast their different statuses in various tissues. Different functions

Table 2.1: Differential TF expression across 37 human tissues. The RNA-seq expression levels are from Human Protein Atlas [17] (www.proteinatlas.org) The expression profiles are first normalized across 37 human tissues, then the Z-scores are calculated for each TF gene. Up-regulated genes are determined to be having z-score >5, whereas down-regulated genes are with z-score <-2.

Tissue Name	Up-regulated TF genes	Down-regulated TF genes
adipose tissue	SOX18	
adrenal gland	CREBL2/ DRGX/ NEUROD4/ PHOX2A/ PHOX2B/ TLX2/ ZNF331/ ZNF836	
appendix		
bone marrow	CEBPE/ E2F2/ GATA1/ GFI1/ GFI1B/ KLF1/ LTF/ NFE2/ RFX8/ ZBTB43/ ZNF394	ATF7/ EEA1/ JRKL/ SMAD4/ TFCP2/ THAP11/ ZNF146/ ZNF17/ ZNF181/ ZNF197/ ZNF2/ ZNF232/ ZNF234/ ZNF235/ ZNF30/ ZNF302/ ZNF322/ ZNF397/ ZNF449/ ZNF576/ ZNF691/ ZSCAN32
breast	ETV3L/ TRPS1	
cerebral cortex	ARNT2/ ATOH7/ BHLHE22/ CSRNP3/ DACH2/ DBX2/ DEAF1/ DLX1/ DLX2/ DPF1/ EGR4/ FERD3L/ FEZF2/ FOXG1/ GBX2/ HES6/ HEY1/ INSM2/ LHX2/ LHX3/ LHX5/ MYT1/ MYT1L/ NEUROD2/ NEUROD6/ NPAS3/ NPAS4/ NR2E1/ OLIG1/ OLIG2/ POU3F2/ POU3F4/ PRDM12/ SALL3/ SCRT1/ SCRT2/ SOX1/ SOX11/ SOX2/ SOX8/ ST18/ TBR1/ TFAP2D/ THRA/ VAX1/ VAX2/ ZIC1/ ZIC2/ ZIC4/ ZNF365/ ZNF536	
colon		
duodenum	NEUROG3/ PDX1	
endometrium		
epididymis	HOXB2/ HOXB3/ HOXB7/ HOXB8/ HOXB9/ HOXD3/ PREB/ TFAP2B/ ZMAT1/ ZNF189/ ZNF266/ ZNF705D/ ZNF776	
esophagus	OLIG3/ PITX1/ ZNF426	
fallopian tube	CCDC17/ FOXJ1/ SOX3/ ZNF474	
gallbladder		
heart muscle	NKX2-5/ TBX20/ TBX5	
kidney	EMX1/ HMX2/ LHX1/ SIM1/ UNCX	
liver	ARID3C/ ATF5/ NR1H4/ NR1I3/ PROX1/ ZGPAT	ZNF384
lung		
lymph node		
ovary	ARX/ LHX9	
pancreas	BHLHA15/ PTF1A/ RBPJL/ ZNF18/ ZNF98	MYSM1
parathyroid gland	DMRT2/ DMRT3/ GCM2/ MAFB/ PAX1/ PBX3/ PRDM4/ SIX3/ T/ TIGD6	
placenta	ARID3A/ DLX4/ FOXI3/ FOXO4/ GCM1/ LIN28B/ MSX2/ NFE2L3/ NFE4/ PLAGL1/ SOX14/ SP6/ ZFAT/ ZNF595	
prostate	NKX3-1/ RAX/ SP8/ ZNF613/ ZNF761	
rectum		
salivary gland	ASCL3/ FOXC1/ LMX1B/ SOX10/ ZNF124	
seminal vesicle	GLIS1	
skeletal muscle	EN1/ GBX1/ HOXC10/ HOXC9/ LBX1/ MAFA/ MEF2C/ MEF2D/ MYF5/ MYF6/ MYOD1/ MYOG/ NFE2L1/ PAX7/ PITX3/ RXRG/ SCX/ SNAI3/ TAL2/ TBX1/ TEAD4/ YBX3/ ZNF784/ ZNF865	BAZ2B/ MYNN/ TMF1/ TOPORS/ ZBED6/ ZBTB48/ ZFP69/ ZNF136/ ZNF317/ ZNF326/ ZNF557/ ZNF654/ ZNF678/ ZNF845
skin	CDX4/ DLX3/ FOXN1/ HOXC12/ HOXC13/ NHLH2/ OTX1/ POU2F3/ POU3F1/ RARG/ SOX15/ TFAP2E/ TP63/ ZNF385A	
small intestine		
smooth muscle		
spleen	FLI1/ SPIC/ TLX1	
stomach	BARX1/ FOXQ1/ NKX6-2/ NKX6-3/ ONECUT3/ SOX21/ ZSCAN4	
testis	AHCTF1/ AHRR/ ANHX/ ARID2/ BARHL2/ BHLHE23/ BNC1/ CBX2/ CCDC169-SOHLH2/ CPEB1/ CPXCR1/ CTCFL/ DBX1/ DMBX1/ DMRT1/ DMRTB1/ DMRTC2/ DOT1L/ DUX4/ ESX1/ FAM170A/ FEZF1/ FIGLA/ FOXN4/ FOXR1/ FOXR2/ GSC2/ HES3/ HSF5/ HSFX2/ HSFY1/ HSFY2/ KDM5B/ KLF17/ LIN28A/ NEUROG2/ NKX2-4/ NKX2-8/ NR6A1/ OTX2/ OVOL3/ POU4F2/ POU5F2/ PRDM9/ RFX2/ RFX3/ RFX4/ RHOXF1/ RHOXF2/ RHOXF2B/ SIX6/ SKOR2/ SOHLH1/ SOHLH2/ SOX30/ SOX5/ SP7/ SPZ1/ TBP/ TBPL1/ TBX22/ TCFL5/ TERB1/ TFDP2/ TFDP3/ TGIF2LX/ TGIF2LY/ TIGD4/ TPRX1/ YBX2/ YY2/ ZBED9/ ZBTB32/ ZBTB37/ ZFHX2/ ZFP91/ ZIC5/ ZNF165/ ZNF200/ ZNF233/ ZNF280B/ ZNF385C/ ZNF433/ ZNF473/ ZNF479/ ZNF487/ ZNF507/ ZNF534/ ZNF541/ ZNF546/ ZNF560/ ZNF574/ ZNF578/ ZNF645/ ZNF646/ ZNF677/ ZNF679/ ZNF683/ ZNF689/ ZNF705B/ ZNF709/ ZNF728/ ZNF729/ ZNF829/ ZNF99/ ZSCAN5A/ ZSCAN5B	
thyroid gland	FOXE1/ PAX8/ ZBED2/ ZBTB2/ ZMAT4/ ZNF571/ ZNF804B	
tonsil		
urinary bladder	HOXA1	

and target genes of ATF1 in the bone marrow and the muscle tissues are hidden beneath its similar expression levels observed in the two tissues.

As the bulk measurement of expression levels being inadequate to dissect the changing roles of TFs in promoting tissue diversity, The inadequacy exhibited by the bulk measurements of expression levels to dissect the changing roles of TFs to promote tissue diversity demonstrated the need for a technique and data that can reveal the dynamics embedded within. What do we want? Single-cell RNA-seq.

2.5 References

[1] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics* **10**, 252–263 (2009).

[2] Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650 – 665 (2018).

[3] Geertz, M., Shore, D. & Maerkl, S. J. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform **109**, 16540–16545 (2012).

[4] Rohs, R. *et al.* Origins of specificity in protein-dna recognition. *Annual review of biochemistry* **79**, 233–269 (2010).

[5] Boija, A. *et al.* Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell* **175**, 1842 – 1855.e16 (2018).

[6] Marcon, E. *et al.* Human-chromatin-related protein interactions identify a demethylase complex required for chromosome segregation. *Cell Reports* **8**, 297 – 310 (2014).

[7] Gebhardt, J. C. M. *et al.* Single-molecule imaging of transcription factor binding to dna in live mammalian cells. *Nature Methods* **10**, 421 – 426 (2013).

[8] Deuschle, U., Gentz, R. & Bujard, H. lac repressor blocks transcribing rna polymerase and terminates transcription. *Proceedings of the National Academy of Sciences* **83**, 4134–4137 (1986).

[9] Peng, Y. & Jahroudi, N. The nfy transcription factor functions as a repressor and activator of the von willebrand factor promoter. *Blood* **99**, 2408–2417 (2002).

[10] Zabidi, M. A. & Stark, A. Regulatory enhancer–core-promoter communication via transcription factors and cofactors. *Trends in Genetics* **32**, 801–814 (2016).

[11] Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* **5**, 276–287 (2004).

[12] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

[13] Hsieh, T.-H. S. *et al.* Resolving the 3d landscape of transcription-linked mammalian chromatin folding. *bioRxiv* (2019).

[14] Krietenstein, N. *et al.* Ultrastructural details of mammalian chromosome architecture. *bioRxiv* (2019).

[15] Goode, D. K. *et al.* Dynamic gene regulatory networks drive hematopoietic specification and differentiation. *Developmental cell* **36**, 572–587 (2016).

[16] Lambert, M., Jambon, S., Depauw, S. & David-Cordonnier, M.-H. Targeting transcription factors for cancer treatment. *Molecules (Basel, Switzerland)* **23**, 1479 (2018).

[17] Uhlén, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357** (2017).

[18] The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

[19] Weirauch, M. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

[20] Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347** (2015).

[21] Hynds, R. E., Vladimirou, E. & Janes, S. M. The secret lives of cancer cell lines. *Disease Models & Mechanisms* **11** (2018).

[22] Terry, E. E. *et al.* Transcriptional profiling reveals extraordinary diversity among skeletal muscle tissues. *eLife* **7**, e34613 (2018).

[23] Meijer, D. H. *et al.* Separated at birth?: The functional and molecular divergence of olig1 and olig2. *Nature Reviews Neuroscience* **13**, 819–831 (2012).

[24] Heeg, S. *et al.* Ets-transcription factor etv1 regulates stromal expansion and metastasis in pancreatic cancer. *Gastroenterology* **151**, 540–553.e14 (2016).

[25] Abe, H., Okazawa, M. & Nakanishi, S. The etv1/er81 transcription factor orchestrates activity-dependent gene regulation in the terminal maturation program of cerebellar granule cells. *Proceedings of the National Academy of Sciences* **108**, 12497–12502 (2011).

[26] Lappin, T. R. J., Grier, D. G., Thompson, A. & Halliday, H. L. Hox genes: seductive science, mysterious mechanisms. *The Ulster medical journal* **75**, 23–31 (2006).

[27] Bradaschia-Correa, V. *et al.* Hox gene expression determines cell fate of adult periosteal stem/progenitor cells. *Scientific Reports* **9**, 5043 (2019).

[28] Rice, K. L. & Licht, J. D. Hox deregulation in acute myeloid leukemia. *The Journal of clinical investigation* **117**, 865–868 (2007).

[29] Merika, M. & Orkin, S. H. Dna-binding specificity of gata family transcription factors. *Molecular and cellular biology* **13**, 3999–4010 (1993).

[30] Lentjes, M. H. F. M. *et al.* The emerging role of gata transcription factors in development and disease. *Expert reviews in molecular medicine* **18**, e3–e3 (2016).

[31] Hillary, R. F. & FitzGerald, U. A lifetime of stress: Atf6 in development and homeostasis. *Journal of biomedical science* **25**, 48–48 (2018).

[32] Katsuyama, M. *et al.* Essential role of atf-1 in induction of nox1, a catalytic subunit of nadph oxidase: involvement of mitochondrial respiratory chain. *The Biochemical journal* **386**, 255–261 (2005).

[33] Guo, Y. *et al.* mir-30a radiosensitizes non-small cell lung cancer by targeting atf1 that is involved in the phosphorylation of atm. *Oncology reports* **37**, 1980–1988 (2017).

To see a world in a grain of sand,

And a heaven in a wild flower,

Hold infinity in the palm of your hand,

And eternity in an hour.

William Blake

3

Advancement of scRNA-Seq technique

No two cells are identical. Tissues are convoluted with a highly heterogeneous cell population. Bulk measurements of transcriptome can only produce a universal gene interaction network shared among several cell populations. In order to dissect further into the dynamic states embedded within each tissue, transcriptome as high resolution as single cells is needed. Typically, such high-

resolution transcriptomes are used to observe sub-populations. [1–6] However, by providing an additional dimension to deciphering the gene expression profiles, an excellent single cell transcriptome will also reveal the differential fluctuation pattern in addition to the varied mean expression levels for each sub-populations which is unenlightening when interpreting the cell-type specific functions the gene is involved with.

3.1 MALBAC-DT: Highly sensitive technique enabling dynamic transcriptome analysis at single-cell level

In order to further examine the transcriptome variation within a cell population, it is necessary to employ single-cell RNA-seq (scRNA-Seq). It has become a powerful technique to reveal the heterogeneity embedded within the cell population. Nevertheless, significant advances are still necessary to reach the technique's full potential. Many methods have been developed for single cell amplification for transcriptome [1,2,7–11], but all suffer from various combinations of poor counting accuracy, low detection sensitivity, or low throughput. While these methods have been successful in cell typing [1–6], their ability to shed light on the relationship among the genes in each cell type identified is limited. To further our understanding of how genes interact with each other in producing complex cellular behaviors, a technique with high accuracy, sensitivity, and throughput is required. To meet these unique technical demands, we designed a novel single-cell mRNA amplification method called Multiple Annealing and Looping Based Amplification Cycles for Digital Transcriptomics (MALBAC-DT) (Figure 3.1).

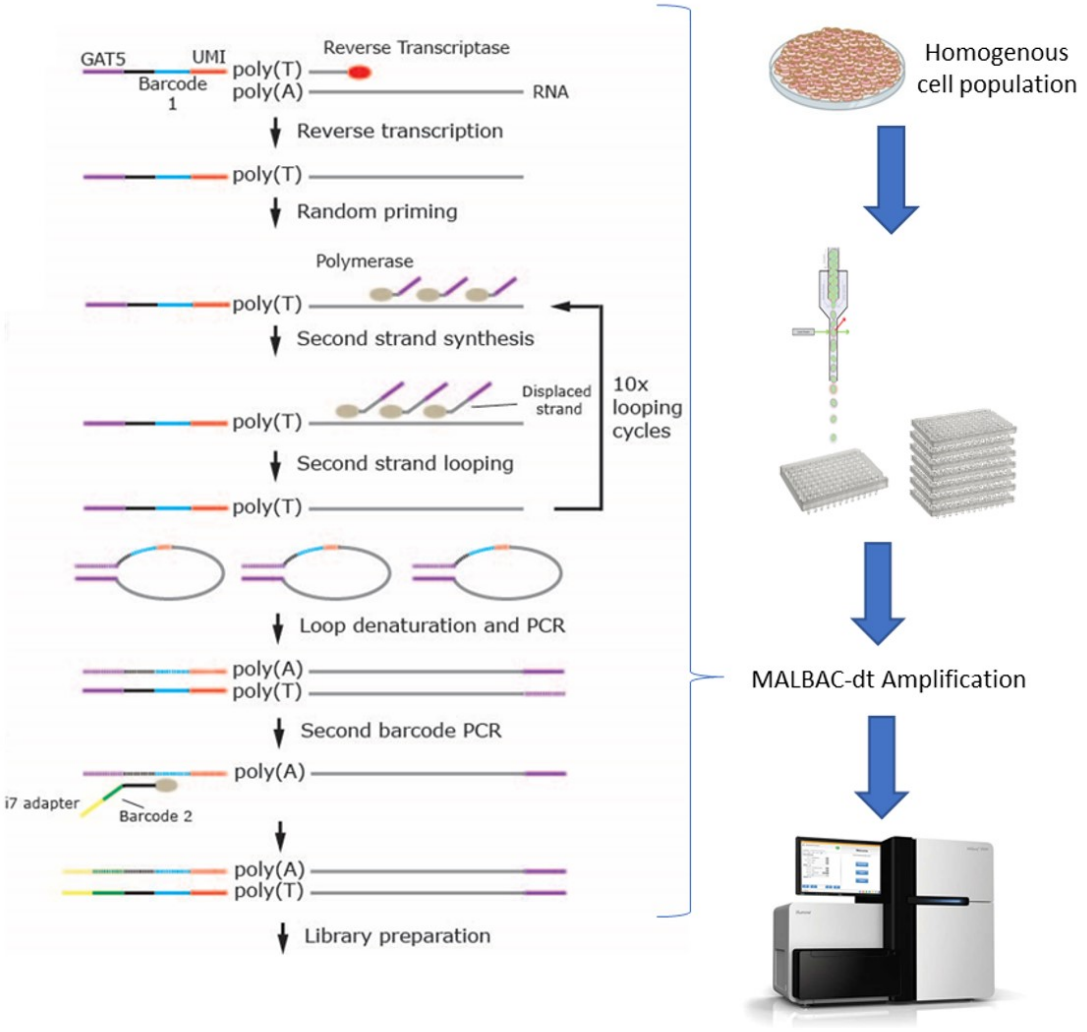


Figure 3.1: MALBAC-dt protocol and experimental workflow. A homogenous cell population is trypsinized and sorted into individual wells of 96-well plates by flow cytometry. Reverse transcription is carried out using a poly-T primer containing a cell-specific barcode and unique molecular identifier (UMI). First strand cDNA is amplified by random primers using MALBAC thermocycling to ensure linear amplification followed by additional cycles of exponential amplification by PCR. After amplification, samples are pooled together for library preparation and sequencing.

To demonstrate the ability of our method to generate unique insights into gene functions and interactions, we amplified and sequenced 5,000 cells from four different cell lines: GM12878, K562, HEK293T, and U2-OS. The selection represents a vast collection of human cells: GM12878 is a healthy B-lymphocyte cell line, K562 is a well studied chronic myeloid leukemia (CML) cell line, HEK293T is derived from human embryonic kidney, and U2-OS is human bone osteosarcoma epithelial cells. As shown in Figure 3.2, about 15,000 genes were observed in any of the four cell types, with half of them shared by all four populations. The highly shared genes are potentially essential for general cell functions, such as cell proliferation, cell cycle progression, and protein synthesis. Other than that, cell lines were so similar with each other to say any two shared more than the rest.

As expected for a homogeneous cell population, clustering of cells based on gene expression using t-stochastic neighbor embedding (t-SNE) did not reveal distinct subpopulations of cells for each cell type sequenced (Figure 3.3). On the other hand, for the same cell line, biological replicates only showed a minor batch effect that did not separate them from each other to be distinct subpopulations.

Meanwhile, inter-cell-line-wise, the two supposedly non-cancerous cell lines, GM12878 and HEK293T, did not show a clear distinction from the other two cancerous cell lines, suggesting that as being an immortalized cell line, cell lines are inherently 'cancerous'.

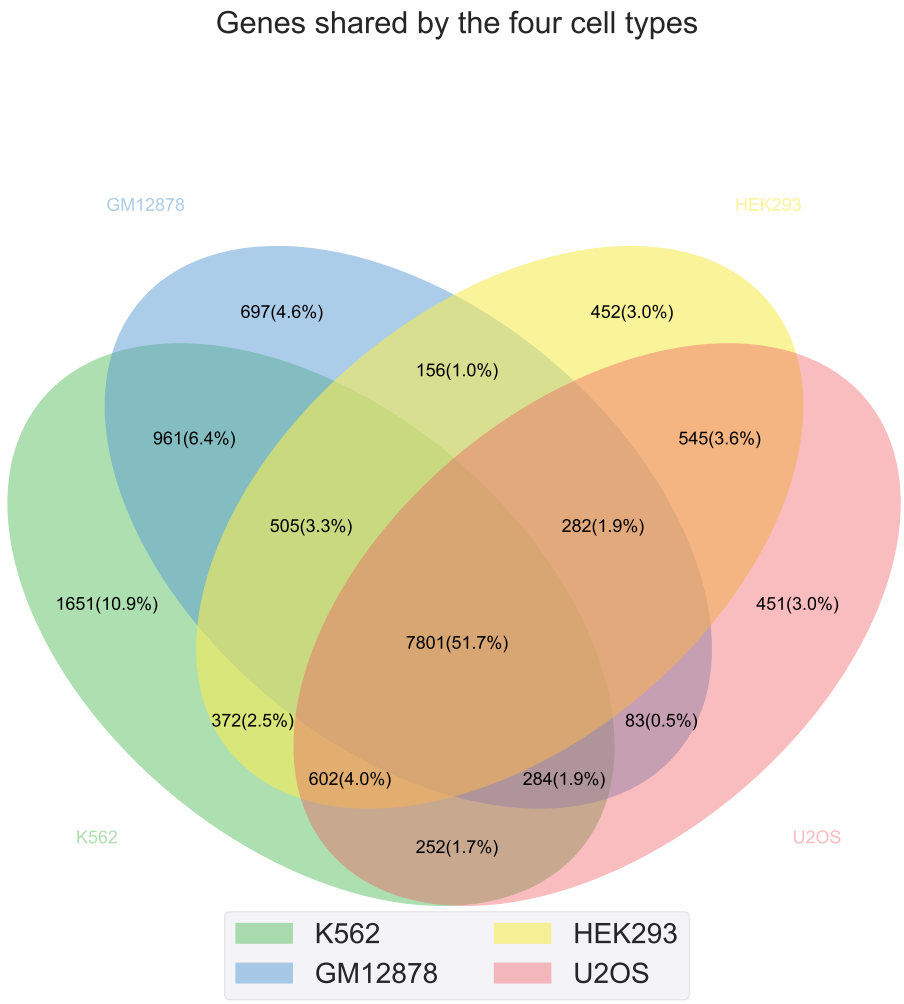


Figure 3.2: Genes shared by U2OS, HEK293T, GM12878 and K562 showed by Venn Diagram. The number of genes in each category is listed with the percentage of the total number of genes observed in any cell type in the bracket). The number of cells that passed quality filter and considered here is as follows: K562, 685 cells; GM12878, 2,863 cells; HEK293, 711 cells; U2OS, 741 cells. For each cell type, each gene is determined to be expressed by having it been observed in more than 10% of the cells. In total, 15,094 genes are expressed in any of the four cell types. About half of the genes are shared by all four cell types and thus potentially essential for general cell function.

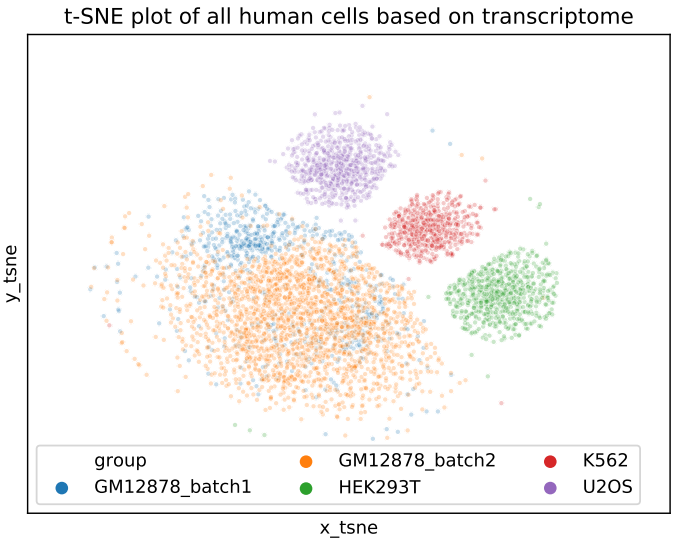


Figure 3.3: t-SNE plot of different human cells that were sequenced. Each dot is a cell, colored by its true identity. All counts were normalized to have the same UMI counts for each cell across all the cell types. GM12878 was divided into two batches in this figure, which were two biological replicates followed by the same experimental protocol. The figure implies a small batch effect, insignificant as compared to the difference among different cell types.

3.2 Enhanced correlation study yields co-regulation modules

While clustering of cells did not show any significant subpopulations, clustering of genes did reveal distinct sets of genes that displayed similar patterns of expression across cells within a cell population (Figure 3.4). Upon hierarchical clustering of the correlation matrix for every pair of genes within each cell line, we observed hundreds of correlated gene modules (CGMs), clusters of 10-200 genes highly correlated with each other. Most of the modules were enriched for a specific biological function. The detection of such correlated modules relies on the precise measurements of correlations between genes.

Before digging further in the correlation study, we want to investigate first the origin of such a correlation coefficient. We examined whether it could be explained by a simple model of regulation of gene expression (Figure 3.5). Due to the limited number of DNA molecules present in a single cell, usually in 2-4 copies, transcription is bursty, resulting in stochastic fluctuations of transcripts over time. When several genes share a common regulator protein, it is logically expected that the fluctuation of that regulator would be transduced to the targets, causing them to fluctuate synchronously. Therefore, when their expression is measured across any cells, the genes would appear as correlated. In this simple model proposed, each gene is transcribed and degraded at a constant rate. The protein is translated and degrade at a constant rate as well. The mathematical deduction of a theoretical correlation coefficient in such model, along with a Gillespie simulation of such process, proves that such fluctuation transduction from the regulator to its targets would indeed give the correlation coefficient of the same magnitude as observed in the study.

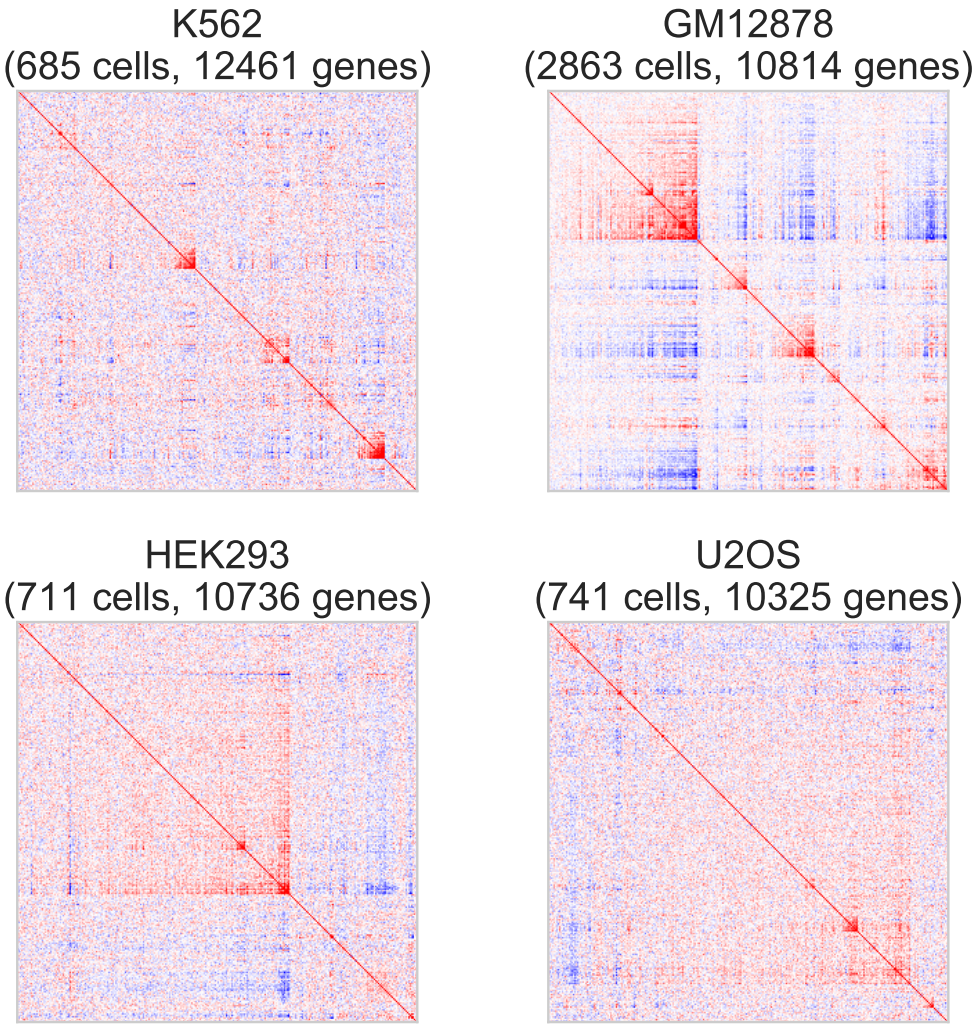


Figure 3.4: Correlation matrices for U2OS, HEK293T, GM12878 and K562. For each pair of genes in each cell line, the Pearson's correlation coefficient is computed and used to do the hierarchical clustering of genes. The number of cells and genes involved for each cell type is listed next to each name. Genes are ordered by hierarchical clustering to reveal various modules of highly correlated genes. The huge disparity among four cell types is observed.

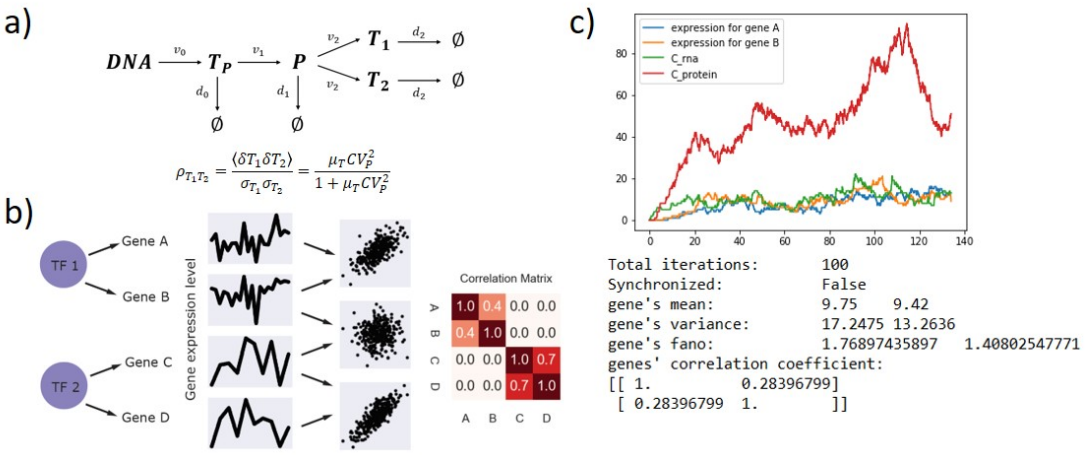


Figure 3.5: Co-regulation hypothesis for the rise of correlation between gene pairs. It is hypothesized that under the co-regulation of a common transcription factor, the expression of the downstream target genes would fluctuate in a synchronized pattern as transduced from the transcriptional fluctuation of the transcription factor itself. Through mathematical deduction in a) and a Gillespie simulation as illustrated in c), such fluctuation transduction would indeed be sufficient in giving the correlation coefficient of the same magnitude as observed in the study.

Now that we have a reasonable hypothesis regarding the rise of correlated gene pairs, it is crucial we make sure such correlation coefficients are reproducible. As shown in Figure 3.6, the correlation coefficients are reproducible among biological replicates, with a Spearman's correlation coefficient of 0.85. Meanwhile, the correlation coefficients are not preserved much from cell type to cell type, with a Spearman's coefficient as low as around 0.3. On the other hand, the mean expression levels are much more alike within and among various cell lines. Reproducible within the cell line, yet distinct among cell lines, correlation coefficients demonstrated its superiority by adding another dimension to contrast cell lines, providing the potential to further dissect into the different dynamics hidden within each cell line.

In other words, two cell lines might have a very similar mean expression level for the shared genes, or say similar transcriptome in general, thus making it harder to compare and contrast the different

networks of each cell line. Meanwhile, the correlation coefficient of each pair of genes makes the difference between cell lines much more significant. If further zoom in on this, this phenomenon arises from the fact that two genes might be expressed at around the same level in two different cell lines, whereas they are not co-expressed the same way, yielding a different correlation coefficient. This contradiction between co-expression and co-activation will be discussed with more details in a later section, 3.4.

Because correlational study can give rise of CGMs and such correlation is cell-type specific, the next step in the study is to investigate the functional validity of CGMs and their status in other cell lines. Here we focused on a CGM from U2OS cells, potentially to be the target module co-regulated by TP53. An shRNA knockdown on TP53 was done on U2OS cells, and differential expression against control cells was calculated for each gene from that TP53 target module. As shown in Figure 3.7, upon knockdown of TP53, the module was enriched for down-regulated genes, suggesting that the module indeed encapsulates a group of TP53 target genes. The same module was also enriched for high confident targets of TP53 curated by literature [12], also colliding with a higher degree of downregulation upon TP53 knockdown. The functionally verified p53 target module was also observed in GM12878, but not in K562 and HEK293T. The different activities of the TP53 module is consistent with the literature that both HEK293T and K562 lack p53 activity. [13,14]

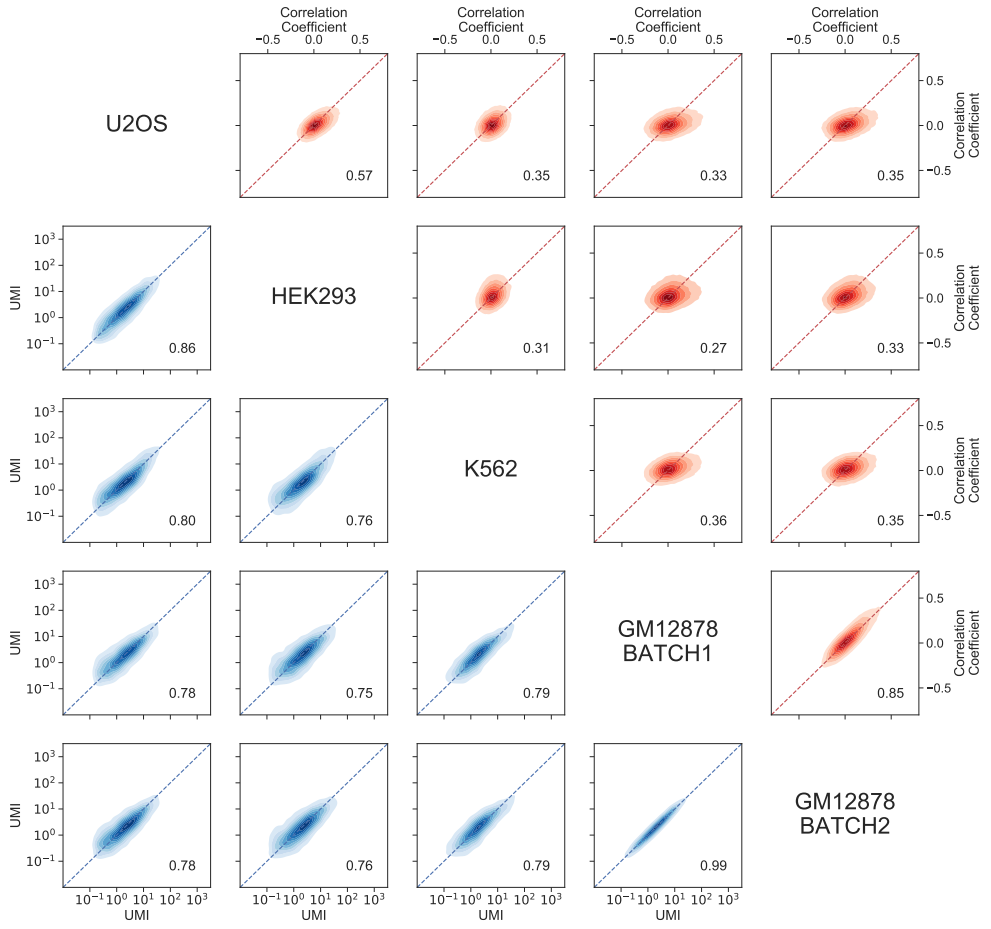


Figure 3.6: The correlational analysis is reproducible and is better at differentiate cell types. Bivariate kernel density estimation (KDE) plots are generated for each pair of the samples sequenced, with regard to the mean expression level of each gene (as in blue in the lower half of the matrix) and to the correlation coefficient of every gene pair (as in red in the upper half of the matrix). Only genes that are shared by all four cell types are shown, with the top 500 expressed in GM12878 are shown in the correlation coefficient comparison. The Spearman's coefficient is calculated for each pair-wise comparison between cell types and listed at the lower right corner. GM12878 is divided into two batches in this figure, which are two biological replicates followed by the same experimental protocol. Both the mean expression levels and correlation coefficients are well preserved from batch to batch for GM12878. On the other hand, the correlational analysis makes the biological replicates pair to stand out from the rest with a Spearman's coefficient as high as 0.85, as compared to that of the cross-cell-types pairs, ranging from 0.27 to 0.57. Meanwhile, this distinction is much more significant than that inferred from the mean expression comparison, with a Spearman's coefficient as high as 0.99 for the biological replicates pair, but also as high as 0.76 to 0.86 for the rest pairs.

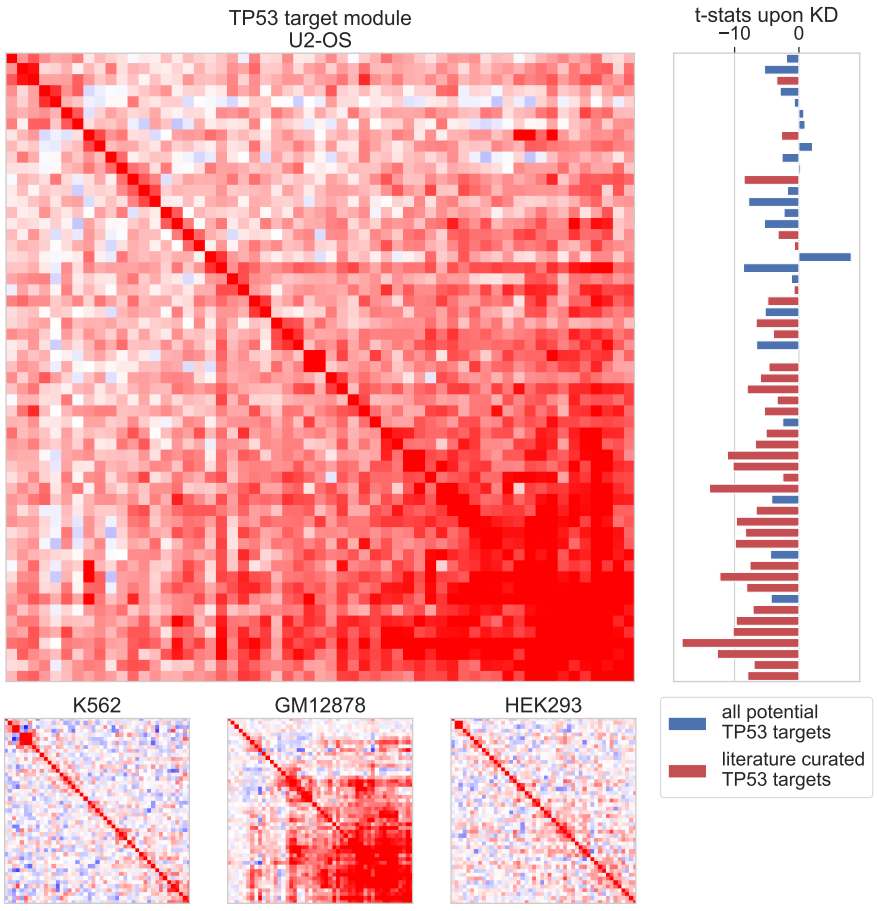


Figure 3.7: TP53 module across different cell lines and its verification *via* TP53 knockdown in U2OS. The module of potential target genes of TP53 inferred from U2OS is shown on the upper left, with the results of TP53 knockdown in the same cell line illustrated on the right as in the degree of differential expressions for each gene as measured in t-statistics. The literature curated targets of TP53 [12] are highlighted in red, which are enriched in this module and with a high degree of downregulation as expected. The covariance matrices of the same target genes for the other three cell lines are plotted to the bottom, showing the preservation of this module only in GM12878 while not in the other two, consistent with literature that those two lacking p53 activity [13,14].

3.3 Correction for cell-cycle effect using pseudotime inference

When interpreting CGMs, we also want to exclude the effect coming from cell-cycle. Cell-cycle is known to have pronounced effects on a wide range of genes, by down-/up-regulating them periodically as the cell goes through four phases: M (mitosis), Gap 1 (G1), S (synthesis), and Gap 2 (G2).[15] Such effect is more significant in scRNA-seq experiments since the cell population would have an extensive collection of cells at various phases of life, and genes connected to the same phase would hence be correlated. Thus, in order to focus on non-periodic pathways, such as developmental and signaling, the cell cycle effect on transcriptome needs to be removed. There have been methods invented before to remove cell-cycle effects in scRNA-seq[15]. However, they were not designed for a large steady state cell population as we have; therefore, a new method is described below.

In this new method to correct for cell-cycle effect, the first step is to infer a pseudotime for each cell, which describes how far along the cell is in its lifetime. The basic principle behind the pseudotime inferences relies on the assumption that the expression of cell-cycle genes follows a sinusoidal wave over time, with a different peak time for each of them.[16] Consequently, each different time point along the time axis would give a fluctuating expression profile of each cell-cycle gene, which is not in phase with each other. By fitting each cell with a pseudotime, we can fit the expression of its cell cycle genes to their corresponding sinusoidal wave, and thus reconstruct the pseudotime series for the cell population. The homogeneous cell population sequenced gives a comprehensive collection of cells at various time points, thus making it possible to fit the sinusoidal wave for all the cell cycle genes.

The actual expression of each cell-cycle gene was modeled as follows, a normal distribution centered around the level predicted by sinusoidal function, with variance aggregated from both stochastic expression variance and technical noise:

$$
\begin{aligned}
y_{g,c} &\sim \mathcal{N}(\mu_{g,c}, v_g^2 + v_{tech}^2) \\
\mu_{g,c} &= Amp_g * (\cos(t_c - T_{peak,g}) + 1)) + AmpShift_g
\end{aligned}
\tag{3.1}
$$

$y_{g,c}$: actual expression of gene g for cell c.

$\mu_{g,c}$: expected expression of g for c from sinusoidal function.

v_g^2: gene-specific variance from the stochastic expression for g.

v_tech^2: common technical noise.

$Amp_g, AmpShift_g$: amplitude of the sinusoidal function for g.

$T_{peak,g}$: The peak time of g, in the time scale of percentage into the cell-cycle. Retrieved from Cyclebase.org[16].

t_c: The pseudo-time of cell c.

The transcriptome was fitted against the described model, with a pseudo-time optimized for each cell to maximize the overall likelihood estimation (MLE). The MLE process was done using PyTorch[17]. In order to correct the covariance matrix for cell-cycle effect, cells were then ordered by the assigned pseudo-time, and the expression of each gene was corrected by subtracting the mean of the surrounding rolling window. The process of cell-cycle correction is illustrated in Figure 3.8 for two genes in U2OS, a typical cell-cycle gene, AURKA, and a non-periodic, MDM2. With the

pseudotime inference, the expression time trajectory of AURKA exhibits a sinusoidal waveform as expected, whereas that of MDM2 does not. After removing the cell cycle effect, AURKA becomes more like MDM2, with expression trajectory showing no significant time dependency. On the other hand, the process does not affect much on MDM2 expression pattern.

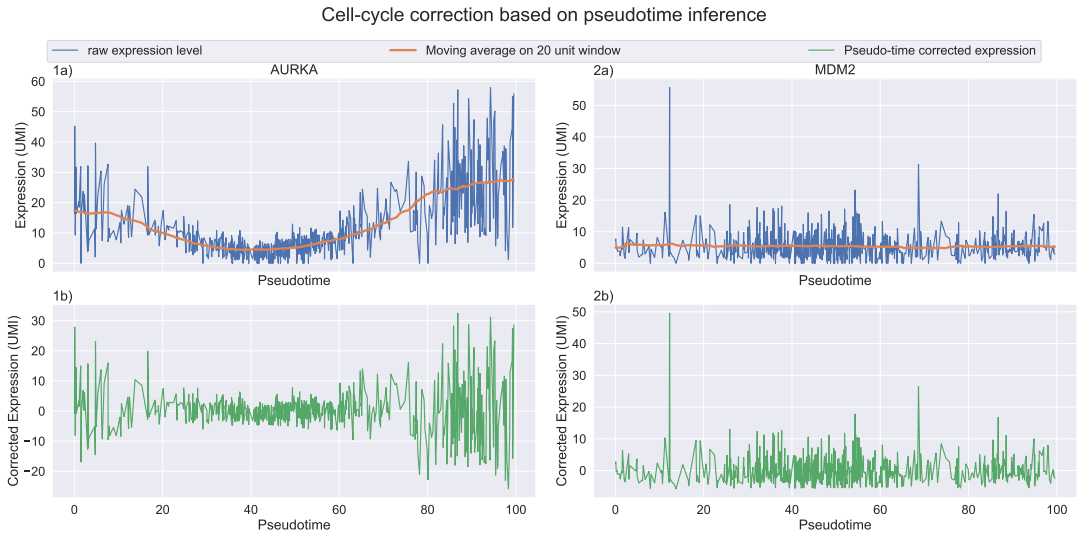


Figure 3.8: Cell cycle correction using pseudotime inference. The process of cell cycle correction is illustrated for U2OS. After the pseudotime is inferred for each cell by fitting all the cell cycle genes to their own sinusoidal curves, the cells are sorted by the pseudotime, and the raw expression level of each gene is corrected by subtracting a moving average across the time series (the orange line in the a)'s), to get the corrected expression, as illustrated in the green lines. Here, one typical cell-cycle gene AURKA along with a typical non-periodic gene MDM2 are shown together to show that the cell cycle correction only affects the cell cycle genes, and not so much effect on the rest.

With pseudotime inferred, now we can interrogate the effect of cell-cycle on CGMs. A side-by-side comparison of covariance matrices before and after cell-cycle correction shows no significant difference at the whole transcriptome level, except for cell-cycle related modules (Figure 3.9). TP53 target module, which is not directly related to the division cycle, is well preserved upon cell-cycle correction, while the other two M-/S- phase modules are slightly diminished. The most significant

change is the disappearance of the strong anti-correlation between the M- and S-phase modules, which are expected since they are two non-consecutive phases in the cycle. In summary, the cell-cycle effect is not the leading cause for the arising of CGMs, implying that the latters are connected to other pathways.

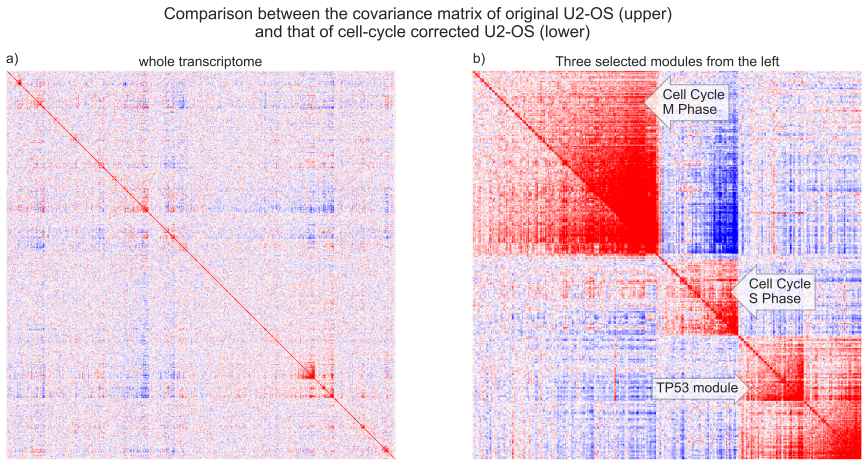


Figure 3.9: Cell cycle correction only affects the modules related to the cell cycle. The covariance matrices of the whole transcriptome before and after cell cycle correction are shown to demonstrate the preservation of the signatures after the correction. Enlarged in b) are three modules, two with annotated cell cycle phases M and S, and one that is not cell cycle related, a potential target module for TP53. With the cell cycle removed, the TP53 module remained the same, while the other two M-/S- phase modules are slightly diminished. The most significant change is the disappearance of the strong anti-correlation between the M- and S- phase modules, which are expected since they are two non-consecutive phases.

3.4 Co-expression vs. co-activation: inference of PPI from co-activation

When people refer to 'co-expression', it usually means how two or more genes are expressed at the same time. However, due to the limitation in previous techniques, the co-expression is more about the genes expressed in the same cell type, or population. Such inter-population co-expression pat-

tern is not enough to tell the whole story. If one dissects deeper into the population, two genes being co-expressed across different cell populations does not mean they are expressed at the same time within the population, or we refer to as 'co-activation' hereof. The term 'co-activation' focuses on whether the transcription of the genes are turned on at the same time, thus giving a positive correlation of expressions when they do. In other words, 'co-activation' focuses on a much smaller time scale as compared to 'co-expression', where the latter is an ensemble measurement: 'co-expression' means expressing in the same population, whereas 'co-activation' means expressing in the same cell.

For example, as shown in Figure 3.10, CENPE and MCM6 show similar expression pattern across different cell types but are strongly anti-correlated when dissected into a specific cell type. CENPE (Centrosome-associated protein E) is a kinesin-like motor protein that concentrates at G2 phase [18], whereas MCM6 (minichromosome maintenance complex component 6) is a DNA replication licensing factor and expression of it peaks at G1/S as induced by growth stimulation. Since the two genes are both necessary for cell-cycle progression, it is reasonable for the two to show similar co-expression pattern across tissues. However, the two genes are expressed and function at different phases of the cycle, and not co-activated at the same time point. This pair of genes demonstrate that co-expression across populations are not equal to genes being co-activated and turned on at the same time, and it further exemplifies the unique power of MALBAC-DT to reveal the gene interaction hidden beneath the bulk measurements.

Traditionally, people have been linking co-expression across tissues with protein-protein interaction (PPI). [19] The logic behind this is that if two genes are transcribed and translated at the same time, it would be more likely they are functionally connected. However, the timescale 'co-

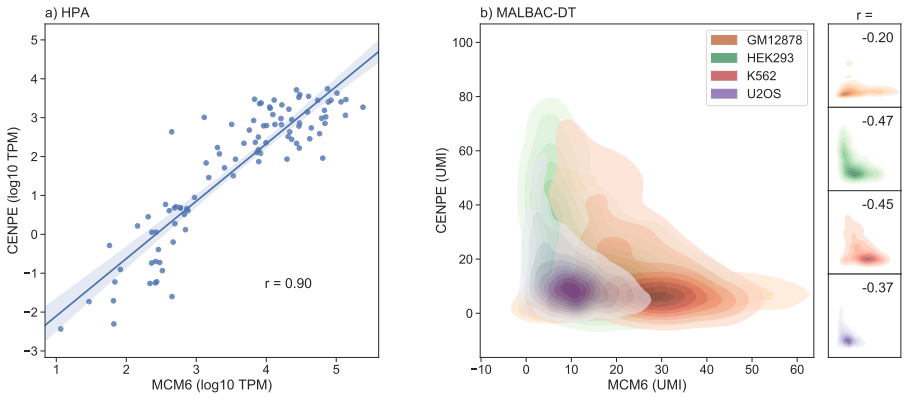


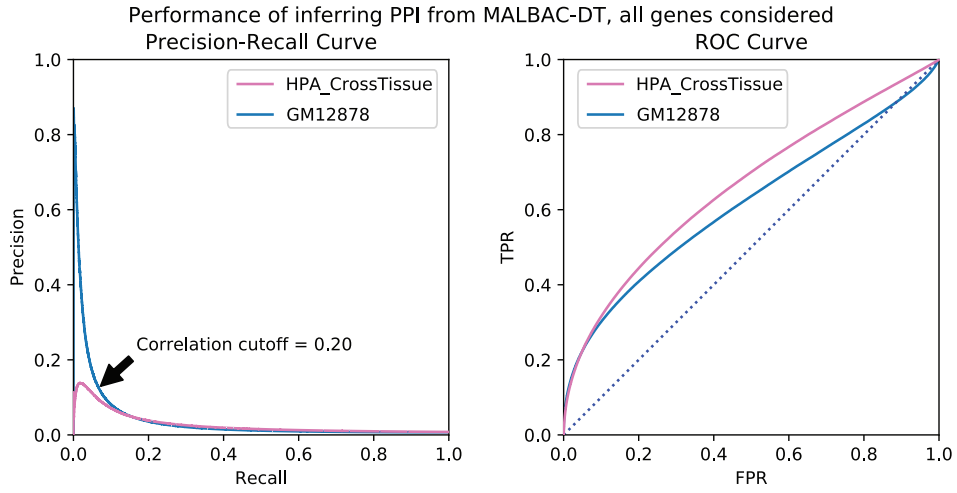
Figure 3.10: Co-expression across tissues does not necessarily give co-activation at the same time. Scatter plot of the expression level for the two genes in each tissue is generated in a) with data drawn from HPA, with a Pearson's correlation coefficient calculated and listed on the lower right corner. In b), Bivariate kernel density estimation (KDE) plots are generated for the relative distribution of the pair in each cell lines, with a corresponding Pearson's correlation coefficient noted accordingly. This pair of genes, CENPE and MCM6, showed a high positive correlation across different tissues. However, such co-expression pattern would not give a co-activation relationship between the two genes since they are not activated for transcription at the same, illustrated by a strong negative correlation coefficient when zooming in to the transcriptome fluctuations at steady state. This exemplifies the unique power of MALBAC-DT to reveal the gene interactions hidden beneath the bulk measurements.

expression' referred to is not always compatible with this logic. Just as we pointed out with the example pair of CENPE and MCM6, albeit with highly correlated co-expression patterns, the pair are not working together to the current knowledge. Therefore, 'co-expression' is fated to have a high false positive rate when used only itself to infer PPI.

On the other hand, co-activation, as supported by scRNA transcriptome with MALBAC-DT, provides another dimension of information when inferring PPI. To analytically compare the performance of inferring PPI from the two type of measurements, Precision-Recall Curve and Receiver Operating Characteristics (ROC) curves are generated, and the area under corresponding curves are calculated (Figure 3.11). Here, the golden truth of PPI edges is retrieved from STRING database v11 [19], with a combined score higher than 400 being regarded as true PPI. The shuffled truth is serving as negative control. Co-activation from the covariance matrix performs as good as two-fold better than co-expression from tissue panel as measured by AUPRC, area under precision-recall curve. When comparing to its peer 10x [11], MALBAC-DT yields a two-folds better AUPR in HEK293T. Not only is our data better than bulk measurements to infer more accurate PPI, but it is also better than its peers.

On the other hand, the data for GM12878 is much better than that of the other three cell lines, which can be explained by two reasons. First, GM12878 has three-folds more cells than the rest, thus more accurate correlation coefficients. Second, the cell culture for GM12878 is a single clonal amplification, making it a more homogeneous cell population as compared to the rest, suggesting that the optimal setup for the method to reveal gene interactions is a large steady-state population.

a)



b)

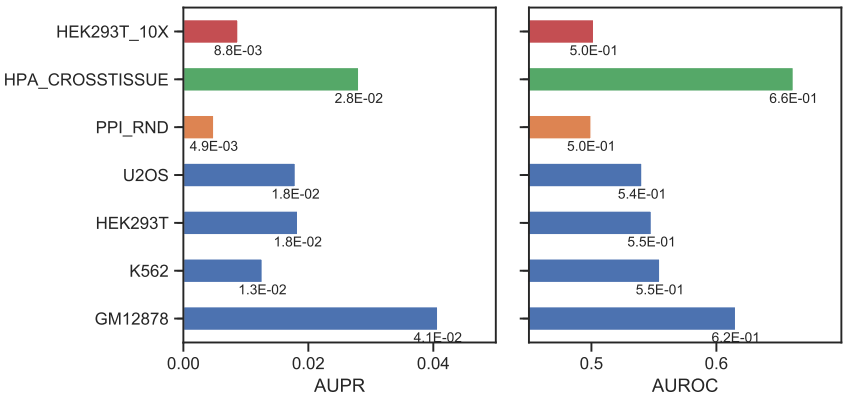


Figure 3.11: Performance of inferring Protein-Protein Interaction (PPI) from the covariance matrix. To evaluate the performance of inferring PPI from the covariance matrix from MALBAC-DT experiments, the Precision-Recall (PR) curve and Receiver Operating Characteristics (ROC) curves for GM12878 and that for the cross tissue co-expression from Human Protein Atlas (HPA) [20] are plotted in a), whereas the Area under PR Curve (AUPR) and Area under ROC curve (AUROC) for the rest of the cell types are listed in b). PPI_RND is the shuffled truth, serving as negative control. HEK293T_10x is retrieved from 10x Genomics [11]. The golden truth of PPI edges are retrieved from STRING database v11 [19], with a combined score greater than 400 being regarded as true PPI. All of the MALBAc-DT experiments along with HPA and 10x data showed an AUPR much better than the negative control, demonstrating their prediction power for PPI. With an AUPR of GM12878 as good as two-fold of HPA, the MALBAC-DT scRNA covariance matrix is superior in inferring PPI while providing cell type-specific information at the same time. GM12878, as single clonal cell culture, shows a much better AUPR as compared to the other three cell lines, suggesting that MALBAC-DT can reveal gene interactions better from a steady-state with a more homogeneous cell population. When comparing to its peer 10x, MALBAC-DT yields a two-fold better AUPR inferring PPI from HEK293T.

3.5 Conclusion: Unveiling the multilevel complex transcription regulation network

As pointed out at the beginning of this Chapter, scRNA-Seq development has been focused on cell-typing and subpopulation discovery in highly heterogeneous samples. In contrary, MALBAC-DT, innovated with its unique experiment setup and high accuracy, succeeded in revealing the more informational representation of gene interaction network by uncovering the pair-wise co-activation relationships among genes. Contrasted from the traditional terminology of 'co-expression', which describes two genes being expressed in the same tissue or cell types, 'co-activation' refers to the kinship of the two genes being expressed in the same cell at the same time. Two different levels of similar expression pattern give different information. As signified by a pair of cell-cycle genes that are co-expressed across tissues but strongly anti-co-activated in single cell level, co-activation proves to be a complementary side of information to co-expression, by expanding a steady state into a dynamic population.

The co-activation relationships were later proven to be a better indicator of protein-protein interaction (PPI), as compared to co-expression. This extends the knowledge one step further into unveiling the transcription regulation network by inferring protein interaction pairs that function combinatorially.

Nevertheless, one will get lost when looking at the whole transcriptome, which yields an interaction network as expansive as a galaxy (Figure 3.12). As a next step to further decipher this obscure system, we will start by looking at Transcription Factors (TF), the masters orchestrating the cell

functions, as outlined in the last Chapter.

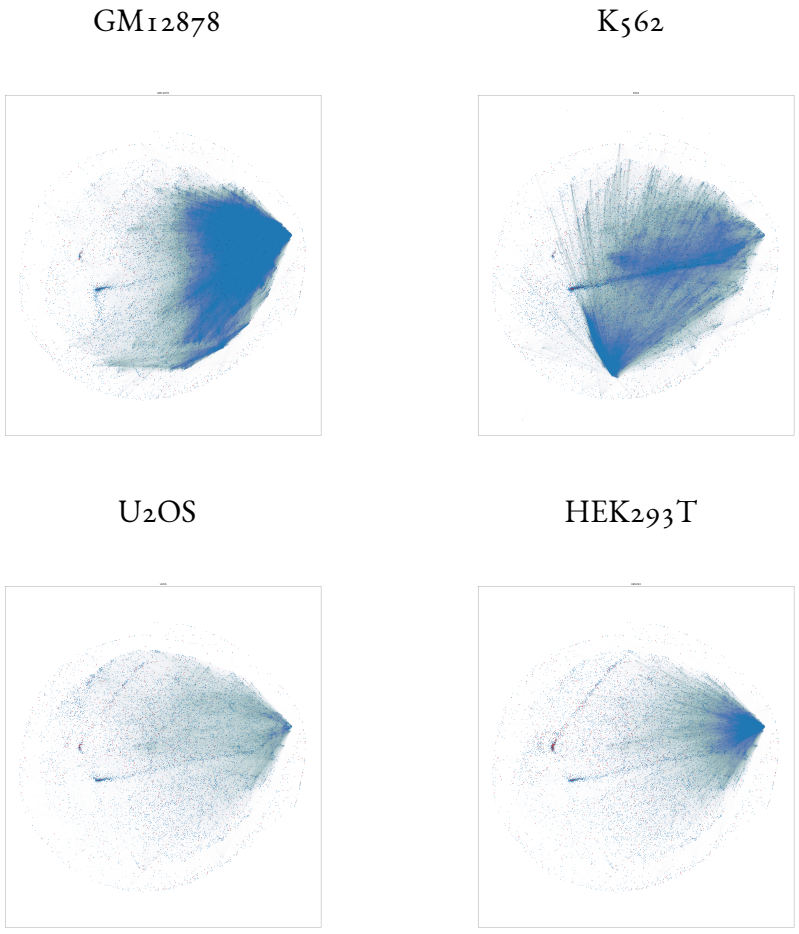


Figure 3.12: Human Protein-Protein Interaction (PPI) network inferred from MALBAC-DT. Each dot represents a gene, with expressed ones highlighted in blue and expressed TFs in red. For each cell line, the PPI network is drawn by affirming the edges from pairs with a correlation coefficient above a cell-type specific threshold (0.2 for GM12878 and K562, and 0.15 for U2OS and HEK293T). The complexity of the networks signifies the difficulty to unravel the regulation mechanism hidden.

3.6 References

[1] Klein, A. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

[2] Macosko, E. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

[3] Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236 – 240 (2013).

[4] Grün, D. *et al.* Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).

[5] Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science* **347**, 1138–1142 (2015).

[6] Usoskin, D. *et al.* Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature Neuroscience* **18**, 145–153 (2014).

[7] Tang, F. *et al.* mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).

[8] Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10**, 1096 – 1098 (2013).

[9] Chapman, A. R. *et al.* Single cell transcriptome amplification with malbac. *PLOS ONE* **10**, 1–12 (2015).

[10] Hashimshony, T. *et al.* Cel-seq2: sensitive highly-multiplexed single-cell rna-seq. *Genome Biology* **17**, 77 (2016).

[11] Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017).

[12] Fischer, M. Census and evaluation of p53 target genes. *Oncogene* **36**, 3943–3956 (2017). Review.

[13] Mattison, S. A., Blatch, G. L. & Edkins, A. L. Hop expression is regulated by p53 and ras and characteristic of a cancer gene signature. *Cell stress & chaperones* **22**, 213–223 (2017).

[14] Law, J., Ritke, M., Yalowich, J., Leder, G. & Ferrell, R. Mutational inactivation of the p53 gene in the human erythroid leukemic k562 cell line. *Leukemia Research* **17**, 1045–50 (1993).

[15] Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33**, 155–160 (2015).

[16] Santos, A., Wernersson, R. & Jensen, L. J. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Research* **43**, D1140–D1144 (2014).

[17] Paszke, A. *et al.* Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop* (2017).

[18] Mirzaa, G. M. *et al.* Mutations in cenpe define a novel kinetochore-centromeric mechanism for microcephalic primordial dwarfism. *Human genetics* **133**, 1023–1039 (2014).

[19] Szklarczyk, D. *et al.* The string database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research* **45**, D362–D368 (2017).

[20] Uhlén, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357** (2017).

Nothing in the world is single;

All things by a law divine

In one spirit meet and mingle.

Why not I with thine?—

Percy Bysshe Shelley

4

TF3C: Co-activated and Co-localized Combinatorial TF pairs.

Combination of Transcription Factors (TFs) are more than the sum of its parts.[1]

As discussed in the ending of Chapter 2, TFs exhibit a wide variability of functions across different

cell types and tissues. Is a single transcription factor enough to explain its changing roles? No. It needs partners. Nevertheless, it has been challenging to decode the cell-type specific TF partnerships from bulk measurements. Blessed with the innovated MALBAC-DT technique that captures the dynamic gene-gene interactions embedded within a steady state population as described in the last Chapter, we can further dissect into a cell population and get the cell-type specific TF pair cooperativity.

4.1 TF cooperativity is paramount to support diversified transcription regulation

4.1.1 Motivation to decipher TF cooperativity

The necessity of TF cooperativity stems from the contradiction that its binding specificity not being distinctive enough to support its diversified and dynamic roles. Although defined as a specific DNA binding protein, TF has a short consensus binding sequence or called motif, that ranges from 5 - 20 base pairs (bp) in length and is recurrent in the genome [2]. On one side, the wide distribution of the same motif sequence allows for TF to bind to a wide range of target genes when it needs. Nevertheless, TF seldom binds to all the matched motif sequences, notwithstanding them being open and accessible. [3] As stated by Wasserman and Sandelin as Futility Theorem, essentially all predicted transcription-factor binding sites that are generated with models for the binding of individual TFs will have no functional role. [4] In other words, a single TF can potentially bind to a wide pool of target sites, but the functional binding pattern is cell-type specific. This additional layer of cell-type

specific selectivity on top of TF's sequence preference points to a more complicated TF recruitment mechanism than just sequence scanning.

Besides the problem regarding single TFs, there is not enough specificity from sequence preference to differentiate TF from its family. TF families are TFs that share similar structures signatures, such as ETS family sharing a DNA binding domain called ETS-domain[5]. Therefore, TF shares similar binding sequence specificity with its homologs from the same family. At the same time, TFs from the same family can have distinct functions, and the similarity in binding specificity makes it an unsolved challenge to decipher the mechanism behind homologs having different functions.[6] Take GATA family for example. As mentioned in Section 2.3, despite sharing a highly conserved binding sequence of being as specific as "GATA", GATA family members target different downstream genes across the tissue panel.

The difficulty in reconciling a TF's different binding pattern in different cell types, and the challenge to explain the various roles held by TFs from the same family, implies the existence of TF cooperative pairs. With one or more partners, TF can have a combined selectivity over sequences, making it more specific when choosing and regulating target genes.[1]

In addition to increased sequence specificity, with different partners, TF can even have different types of regulations within the same cell population. As mentioned in Section 2.1, NFY can either activate or repress one of its target genes, VMF, when binding to the promoter on different consensus sequences depending on the co-factors involved. As extreme as this example goes, the same TF can either activate or suppress the same target gene just by binding to different sequences upstream of it and by partnering with other disparate TFs. Such combinatorial TF partnership further diver-

sifies TF's roles in transcription regulation, making it in better accordance with the fact that more than 10k genes need different regulations from a conserved pool of 1k TFs, as compared to treating individual TF independently.

4.1.2 Traditional ways to infer TF cooperative pair

As essential and universal as it is, TF cooperativity is difficult to experimentally capture *in vivo* due to TF's short residence time and the dynamic nature of such interactions.[3,7,8] Traditionally, the field has been tackling TF cooperativity from two indirect ways: protein-protein interaction itself, and the cooperative DNA binding.

Protein-protein interaction (PPI) is one of the significant origins of force enabling TFs to form complexes at the regulation sites that will later recruit co-factors, chromatin modifiers, and other proteins necessary to regulate the transcription.[9] Two diverging routes have been developed by the field to detect TF pair interaction, one through co-expression pattern, and the other through directly probing their physical capability of interacting with each other.

People need to cast a wide net before catching the big fish. In this scenario, co-expression pattern across the tissues is a net that has been employed by the field traditionally. Co-expression of two proteins are perceived as a pre-requisite for them to form complexes or to function together, thus making it a good indicator of protein-protein interactions.[10] As a matter of course, with the abundant transcriptome data becoming available for various human tissues, people have been able to get potential TF interacting pairs from their co-expression pattern.[11] Nevertheless, co-expressed in a cell population does not guarantee co-activation of the two proteins in a single cell, at the same time,

as discussed in Section 3.4. Therefore, such co-expression analysis is destined to give a lot of false positive TF pairs.

Therefore, the next step after co-expression analysis is to probe the potential pairs using in vitro experiment techniques such as consecutive affinity-purification systematic evolution of ligands by exponential enrichment (CAP-SELEX)[1] and the Mammalian Two-Hybrid (M2H) system[12]. M2H system directly targets the physical PPI between pairs by probing the capability of two proteins interacting with each other to form a complex. When two TFs can physically interact, they would bring the artificial TF proteins tagged to them to be close to each other at the transcription starting site, triggering the target gene expression. On the other hand, CAP-SELEX explores the DNA-mediated TF cooperativity by screening for TF pairs co-bound on DNA fragments from a random pool of DNA sequence. This technique can detect TF pairs formed with assistance from DNA-mediated allostery, which usually does not exhibits a direct physical interaction between the two proteins in the spotlight.[1,13] Covering both direct and indirect physical interactions, the two techniques represent the comprehensiveness of *in vitro* means to observe TF cooperativity.

Despite being successful in decreasing false positives brought on by co-expression analysis, the techniques described above are cursed with false negatives. Both techniques are *in vitro* protocols that interrogate TF pairs in an unnatural setting, depriving them of the necessary micro-environment to form a solid partnership, thus losing cell-type specific TF pairs. Besides, these experiments rely on protein editing on each TF, making it not scalable enough to cover all of the 1k TFs expressed in a tissue. Further, TFs are known to have a short residence time when bound to DNA, in the magnitude of seconds[3,7,8], making the dynamic interactions even harder to be captured by these *in vitro*

methods.

On the other side of the problem, PPI is not the whole story. In addition to probing for TF-TF interaction through protein interactions, one can also investigate by targetting the co-binding of the two DNA-binding proteins. Similarly, people have both theoretical and experimental tools to tackle this problem. Theoretically, one can scan through the genome using motif, the specific sequence preference profiles of the TFs being questioned, to predict their potential binding sites and test whether the two TFs are binding in proximal locations. However, as discussed in Section 4.1.1, motif analysis alone is insufficient to provide cell-type specificity. When combined with cell-type specific chromatin accessibility data, such as Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq)[14], one can refine the motif-predicted binding sites to open regions, which is about 1% of the human genome. The assumption behind this is that TFs can only bind to open regions, not the closed heterochromatin. This way, one can recover the cell-type specificity when predicting binding sites.[15] However, this still suffers from non-functional predicted sites and non-functional co-occurrence of motifs.

Co-bound can also be experimentally interrogated by chromatin immuno-precipitation followed by sequencing (ChIP-Seq) to get the binding profiles of both members of the TF pair at the question. By pulling down TF with the DNA bound using protein-specific antibody, ChIP-Seq can profile the precise genomic loci bound by the TF. Then by intersecting two ChIP-Seqs spectra of the TF pair, one can determine whether two TFs are co-bound in the same cell population. However, this technique has several inherent flaws.

First, the binding profiles have a high false positive rate and a lousy reproducibility due to in-

consistent antibody qualities.[3] The accuracy depends significantly on the quality of the antibody used, which is cursed with a signficant batch effect along with a non-specific protein recognition by pulling down homologs along with the target protein. The high false positives further contribute to the complexity when deconvoluting the TF co-binding patterns. Second, the ChIP-Seq experiment is limited by the number of antibodies available, making it less scalable when expanding to profiling the whole TF panel.

False positives from ChIP-Seq can be further corrected using sequential ChIP-Seq, referred to as ReChIP-Seq. By pulling down protein-DNA complex twice successively, the DNA pulled down will be more confidently to be announced as a protein pair bound targets.[16] However, genome-wide ReChIP-Seq suffers from its low efficiency, which is exponentially lower than single ChIP-Seq, making it hard to extend to TF pair profiling.[17]

As summarized, current techniques are all defective in inferring TF cooperative pairs. They are either not cell-type specific, or suffering from low accuracy. Mostly not high-throughput, they are even less cost-effective when trying to extrapolating the interaction among all the expressed TFs in a given cell population, especially if we want to profile various tissues, multiplying the cost by magnitudes.

Can we solve this?

4.2 Cell-type specific TF cooperative pair inferred from single-cell RNA measurements

Now with the new scRNA-Seq technique MALBAC-DT, innovated with high sensitivity and accuracy, we can extract the cell-type specific protein-protein interactions from the single cell transcriptomes, as outlined in Chapter 3. Therefore, we can easily use MALBAC-DT and the correlational studies to extrapolate the PPI among TFs in order to get the cooperative pairs that are co-activated in the same cells. Additionally, it is a high-throughput method that can profile thousands of TFs at the same time, making it easily scalable to cover the tissue panel.

4.2.1 Comparable/better than ChIP-Seq

When comparing to ChIP-Seqs, which is also a *in vivo* profiling of TFs in their natural environment, the co-activation studies demonstrate superiority. Highly co-activated TF pairs are enriched with proven PPI: when setting a cutoff at 0.2, 40% of TF pairs have PPI evidence in both K562 and GM12878, as shown in Figure 4.1. This is comparable to ChIP-Seq co-bound analysis, with highest co-binding pairs showing enriched interactions as well. However, while being comparable to ChIP-Seq in K562, the co-activation studies from MALBAC-DT are much better than ChIP-Seq in GM12878. This is a result of two reasons: a decreased accuracy of MALBAC-DT in K562 as compared to GM12878 due to heterogeneity as discussed in Section 3.4, and also an increased sample size of ChIP-Seqs in K562 as compared to GM12878. As an ENCODE cell line, K562 has more than 250 ChIP-Seqs for TFs available, with GM12878 only having more than one hundred

less. This points to another inherent flaw in using ChIP-Seq and co-bound analysis to infer cooperative TF pairs: it highly depends on the single experiments and is limited by the antibody. Therefore, MALBAC-DT is a more accurate, cost-effective, and high-throughput method to infer TF cooperative pairs as compared to ChIP-Seqs.

4.2.2 Compare and contrast to see if the inferred TF pairs across different cell types (GM12878 vs. K562 vs. HEK293T vs. U2OS)

With the interactions profiles of thousands TFs in four different cell lines at hand, we next set out to investigate the similarities and differences among the TF pairs. Surprisingly, although more than half of TFs are expressed in all four cell lines, little to none TF pairs are common, as demonstrated in Figure 4.2. This contradiction is well below by chance, with a p-value smaller than $1E-6$ from permutation test. Here, differential co-activation is a better representation of the source of differential expression. This striking disparity exemplifies the advantage of the co-activation analysis in dissecting a steady state, by adding another a whole new dimension when comparing and contrasting TF behaviors. By partnering with distinct entities, TFs direct the differential transcriptome among the cell lines, notwithstanding only one thousand of them being active in each cell line with more than half shared among the four cell lines.

The high disparity among four cell lines directs our attention to the only pair shared among four cell lines, PA2G4/YBX1, which has strong experiment evidence proving its protein-protein interaction [18]. However, the combinatorial function of the two are not easily extrapolatable from literature. PA2G4 (Proliferation-Associated 2G4) has a wide range of functions: it is involved in growth reg-

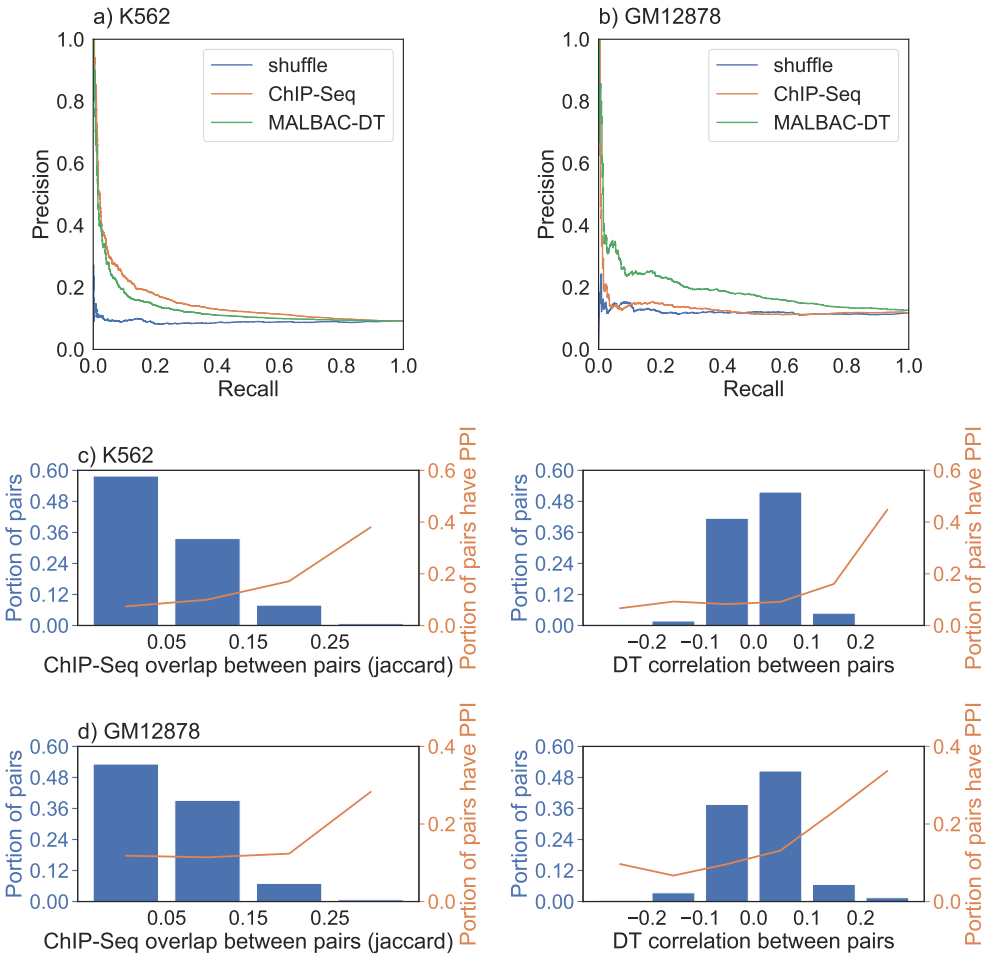


Figure 4.1: MALBAC-DT is better at inferring TF pairs than ChIP-Seq in GM12878, and comparable in K562. The performances of inferring PPI among TFs using co-binding from ChIP-Seq and using co-activation from MALBAC-DT are summarized as Precision-Recall (PR) Curve and Receiver Operating Characteristics (ROC) Curve side-by-side in GM12878 (a) and K562 (b). Co-activation is represented by correlation coefficients from single cell transcriptomes. Co-binding is measured by the overlap between the binding sites of each pair of TF (Jaccard index). All the TF ChIP-Seq data are retrieved from ENCODE project, with 135 TFs for GM12878, and 251 TFs for K562. In order to have a fair comparison, only TFs covered by ChIP-Seq data are included for the analysis, although co-activation covers all the expressed TFs. Here, the golden truth of PPI edges is retrieved from STRING database v11 [10], with a combined score greater than 400 regarded as true PPI. As a baseline, a shuffled truth is also included in the PR curve and ROC curve, with an average positive rate being 0.12. In a more intuitive representation of the same data, the percentage of pairs being PPI in each confidence level of co-binding/co-activation are demonstrated for K562 (c) and GM12878 (d). With an AUPR (area under PR curve) of 0.19, co-activation is much superior to co-binding (AUPR = 0.13) in inferring TF PPI in GM12878. On the other hand, co-activation is slightly worse than co-binding in K562 (AUPR = 0.13 vs. 0.15). This is because GM12878 is a more homogenized population as compared to K562, making the correlation coefficient more accurate and a better representation of the steady state. Overall, co-activation from MALBAC-DT is giving a PPI inference with 40% true positive rate for the highest confidence category, with correlation coefficient > 0.2.

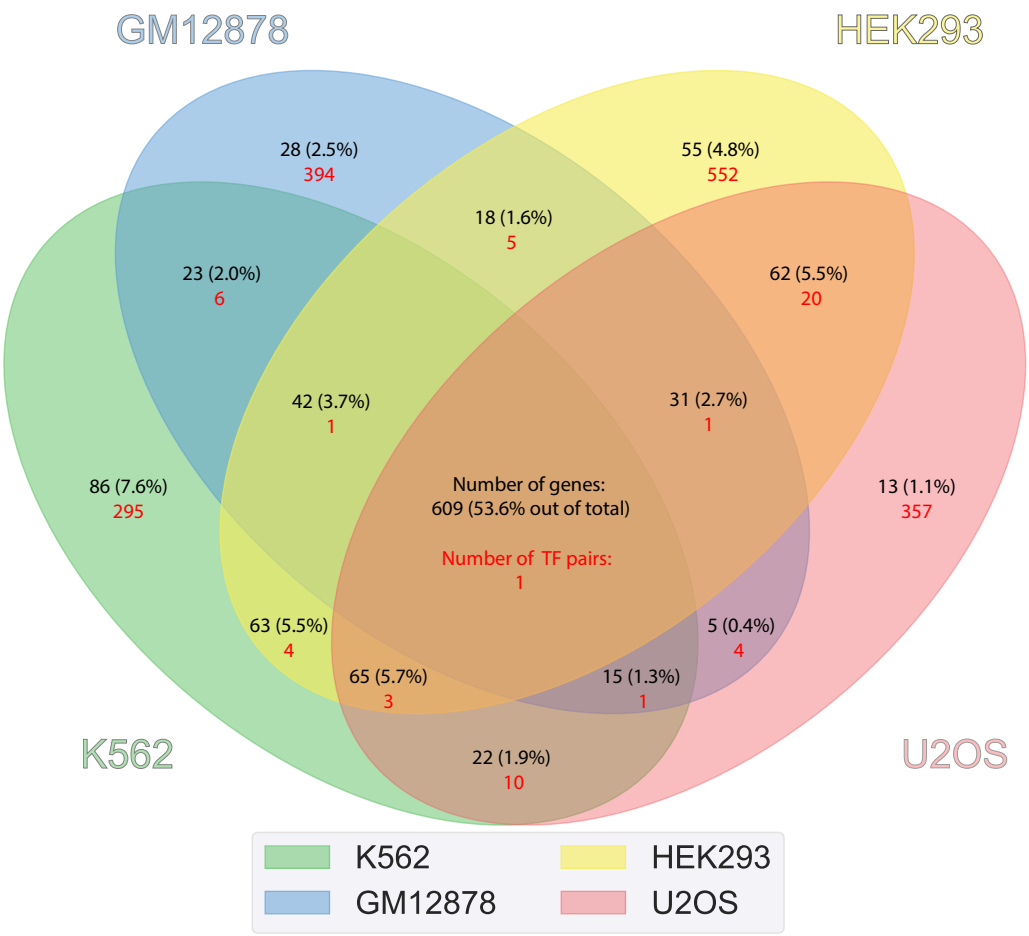


Figure 4.2: Different cell lines share little to none TF correlated pairs while sharing half of the expressed TF genes. Of all the TFs that are expressed in any cell lines in consideration, 53.6% are expressed in all four lineages. In contrary, little to none TF correlated pairs are shared, implying that despite that fact expression of a TF might be common, the function of it could be different by having different partners. The only shared pair among four cell lines is PA2G4/YBX1, which has strong experiment evidence proving its protein-protein interaction [18]. With unknown function as a cooperative pair yet shared by all four cell lines, ranging from lymphocyte to cancer cell lines, PA2G4/YBX1 serves as a great potential target for future experiments to unveil TF cooperation and regulation.

ulation, acts a corepressor of the androgen receptor (AR), and also is involved in the regulation of intermediate and late steps of rRNA processing.[19] On the other hand, YBX1 (Nuclease-sensitive element-binding protein 1) has a set of diversified roles in both mRNA splicing and regulating numerous genes, including enhancing expression of AR.[20] One common things the two shares is their ability to bind to RNA in addition to DNA. The two TFs not only do not share functions, but they also show contradicting roles in regulating AR. This is not surprising since even for the same single TF, it can be activator and repressor at the same time. As diversified in functions as the two TFs posses, the combinatorial relationship might be able to better annotate their roles in regulating transcription.

4.3 TF3C: Combination with ATAC-Seq to infer co-operative TF pairs

With viable candidates for TF cooperative pairs inferred, we can then combine with co-bind analysis to further filter out false positives and gauge possible regulation targets. Consolidated with chromatin accessibility data (ATAC-Seq) and motif, we set out to further insinuate the binding sites of TF pairs co-localized at open regions on genome.

4.3.1 TF pairs significantly reduce potential false positives of the motif and show cell-type specificity.

As a starter, it is evaluated whether the inference of TF pairs reduces the false positives from regular motif analysis. In average, for each 500 bp long ATAC-Seq open peak, there are 100 TFs showing matched motif sequences in that window, as illustrated in Figure 4.3. When restrict the co-bound

TFs to be co-activated at the same time, which is measured by MALBAC-DT correlation coefficients, number of TFs annotated to each open peak significantly decreases to about 10+ TF pairs in each open region. In this way, the combination of co-bound from motif profiles and co-activation from MALBAC-DT might decrease the high false positive motif sites as proposed by Futility Theorem, providing confident cooperative TF pairs colocalizing in the open regions.

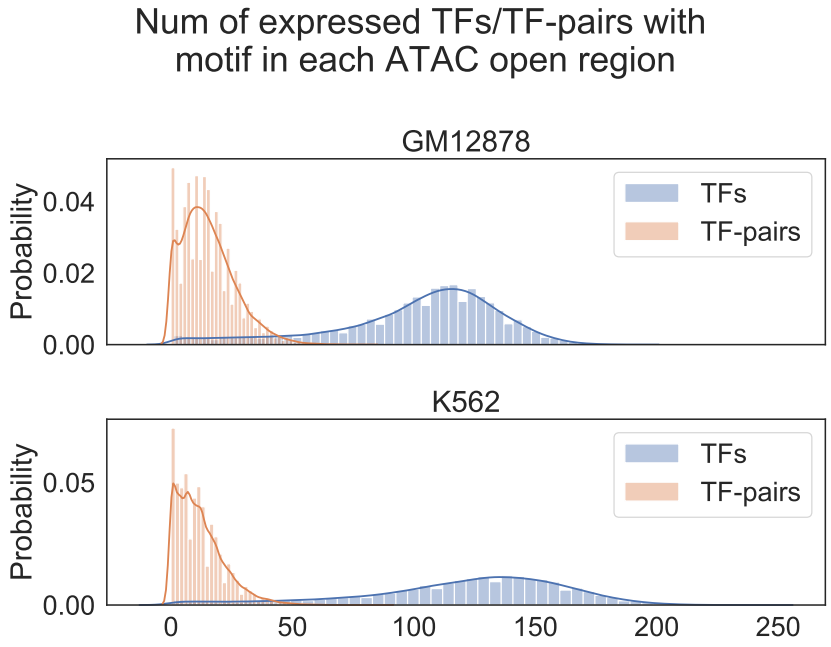


Figure 4.3: Inferring of correlated TF pairs significantly reduces potential false positives of motif scanning. Transcription factor binding sites in each ATAC-seq open region are predicted by scanning motif corresponding to the TF. Motif profiles are retrieved from Cis-BP Database [3,21]. ATAC-seq are retrieved from accession GSE65360. In average, each ATAC peak is annotated with 100 TFs if only using motif. When restrict the co-bound TFs to be co-activated at the same time, as measured by MALBAC-DT correlation coefficients, the number of TFs annotated to each peak significantly decreased to about 10+ TF pairs in each open region. The combination of co-bound from motif profiles and co-activation from MALBAC-DT might decrease the false positive motif sites as proposed by Futility Theorem, providing confident cooperative TF pairs.

In a more concrete example, we evaluate the open peak at promoter around Transcription Start-

ing Site (TSS) of Parkinsonism Associated Deglycase (PARK7), which encodes a protein involved in positive regulation of androgen receptor-dependent transcription, and also helps cells to fight oxidative stress and cell death.[22] It is highly expressed in both K562 and GM12878 (Figure 4.4b), with a two-fold difference between the two cell lines, indicating a differential regulation behind. A simple scanning for regulating TFs from motif shows about one hundred possible TFs bound to the promoter region in both cell lines, with about half of them being shared. This is probably because the two promoter peaks are nearly identical (Figure 4.6), thus similar genome sequence to scan for the motif, which yields highly resembled results. In order to further differentiate the two cell lines, the TF interactions are introduced into the analysis. Figure 4.4 demonstrates that only a dozen of TF and TF pairs are left upon filtering out the TFs that do not have co-activated pairs bound to the same region. Although half of the rest TFs are still in common, the TFs shared show a very distinctive interaction pattern. Such as the pair of HEY1 (Hes Related Family BHLH Transcription Factor With YRPW Motif 1) /EGR1 (Early Growth Response Protein 1), expressed in both cell lines, but only correlated in K562. Such distinction in correlational partnership is more in reconciliation with the two-fold differential expression of PARK7 observed in the two cell lines, as compared to only looking at single TF motif sites that are greatly shared among the two, and provides cell-type specific information more than motif analysis does.

Such cell-type specific combinatorial TF pairs revealed by colocalization and co-activation (TF3C) provides a whole new realm in decoding the transcription regulation. When expanding to genome-wide open regions, it would shed insight onto the whole transcriptome regulation.

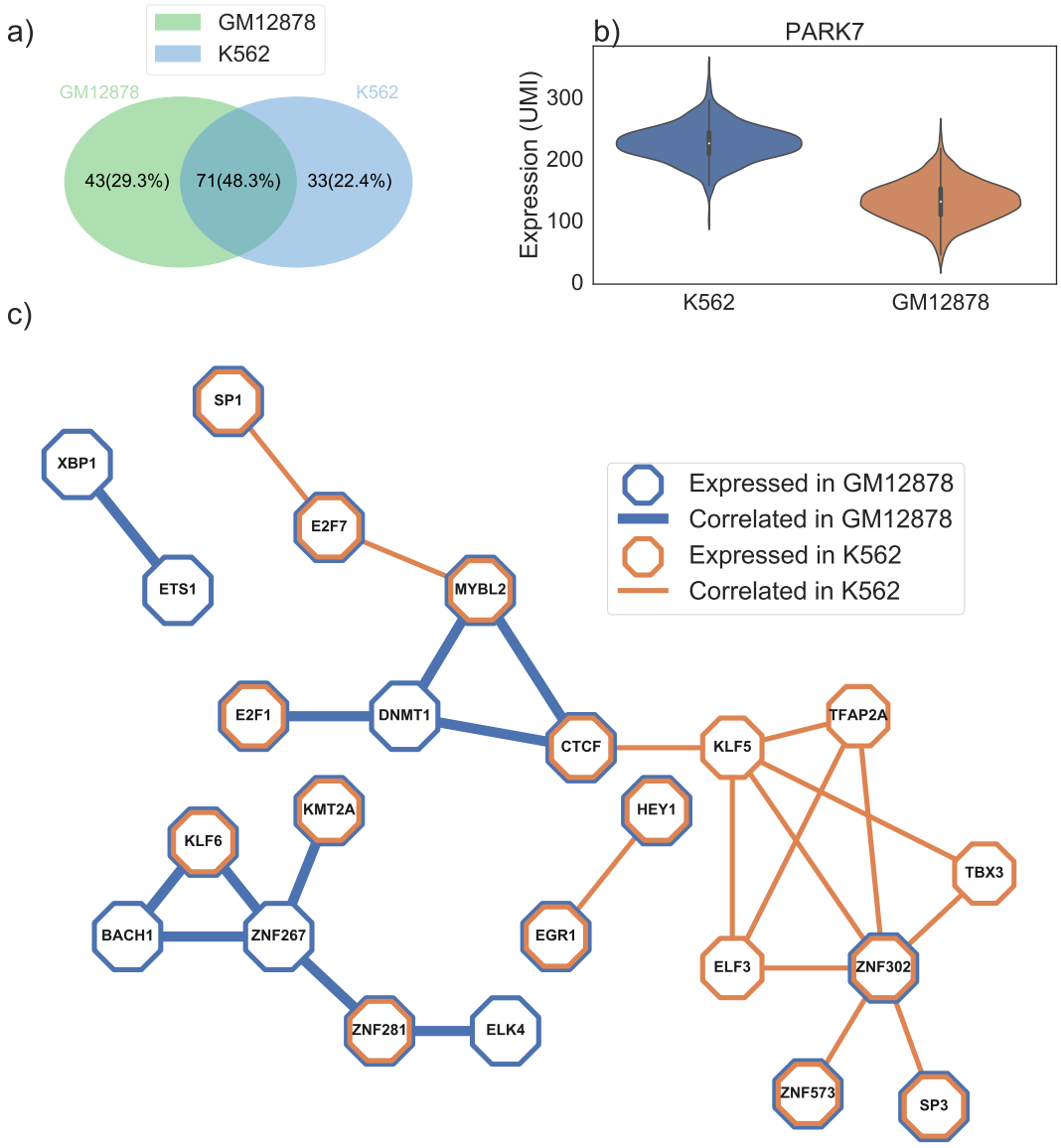


Figure 4.4: Differential TF partnership connected with differential PARK7 expression. a) Number of TFs with motif found in the ATAC peak annotated to be a promoter of PARK7 in GM12878 and K562. b) Violin plot showing differential expression of PARK7 in K562 and GM12878, which is not easily explainable by only looking at the differential TF motif sites, since half of them are shared between the two cell types. In c), the correlational interaction among TFs that have motif found in PARK7's promoter region is illustrated, showing a distinctive TF partnership consortium despite most of the TFs involved are expressed in both cell lines. Such as HEY1/EGR1, expressed in both cell lines, but only correlated in K562. Such distinction in correlational partnership is more in reconciliation with the two-folds differential expression of PARK7 observed, as compared to only looking at single TF motif sites.

4.3.2 Detour: better functional annotation of ATAC-Seq

Before moving on next to step the expand the search of co-localized co-operative TF pairs to genome-wide, we need to take a detour first to annotate ATAC-Seq peaks better. Currently, the annotation of ATAC-Seq peaks has been focusing on promoter/TTS peaks, which are believed to be directly related to the transcription of the corresponding target gene.[14] Promoter/TSS peaks comprise about 50% of the genome-wide collection of ATAC-Seq peaks, and the rest is poorly annotated. Efforts have been made to correlate ATAC-Seq peaks to enhancer regions.[23] However, due to the distal nature of enhancer regulation, it is challenging to connect enhancers to its corresponding target genes.

Here, the focus is cast to another side of the gene that has received much less attention, transcription termination site, TTS. As shown in Figure 4.5, the higher level one gene is transcribed, the more open its transcription start and termination sites are. It is expected to see open TSS connected with expressed genes, but surprisingly to see TTS having an open site nearby for highly expressed genes as well. With 70% of genes that have a mean expression level of higher than 50 UMI counts having both TSS and TTS open, the transcription regulation of highly expressed genes are clearly connected with both regions. An examination of gene sizes by the side shows that the genes with TSS and/or TTS open tends to be significantly shorter in linear size as compared to their peers with the corresponding sites not being open. Integrating the two observations, we hypothesize that the regulation of highly expressed genes is connected with the formation of a gene loop connecting TSS and TTS to stabilize the gene body to be open for transcription, and thus supporting the expression of high copy genes.

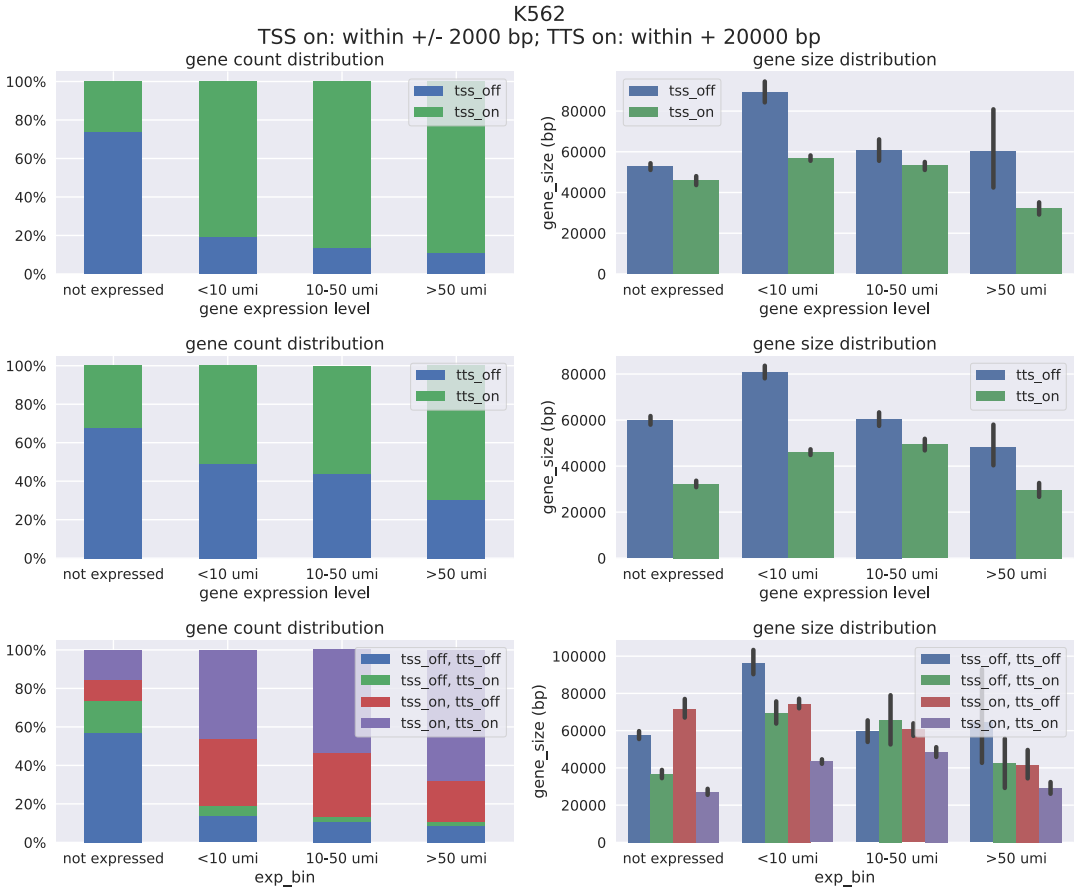


Figure 4.5: The higher expression level one gene exhibits, the more possible its transcription start and termination sites are both open. On the left, we have percentage of genes expressed in K562 at different levels that have open sites in Transcription Start Site (TSS), Transcription Termination Site (TTS), or both considered at the same time. To the right, we have gene size distribution for genes that have open sites in the TSS/TTS/both, categorizing by the expression level of the genes. The gene size is a linear genomic size between TSS and TTS. TSS 'on' means an ATAC-seq open peak is found within +/-2kb of TSS, whereas TTS 'on' means a peak within 20kb downstream of TTS. The figure demonstrates that the higher a gene expressed, the more possible both of its TTS and TSS would be on, with 70% of genes having >50 UMI counts having both sites open. Meanwhile, the gene that has open TTS/TSS tends to be shorter in size as compared to its peers. This observation allows for a hypothesis regarding the regulation of highly expressed genes, by suggesting the formation of a gene loop connecting TSS and TTS to stabilize the gene body to be open for transcription, and thus supporting the expression of high copy genes.

Unlike promoter or TSS that are highly similar in different cell types, the peak near TTS is cell-type specific. As shown in Figure 4.6, PARK7 has three open peaks in both GM12878 and K562, all with regulation activity proved by the overlay of H3K27ac and H3K4me1 peaks. Both cell lines exhibit the same promoter peak located at TSS of PARK7, in addition to a shared enhancer peak about 10kb upstream of the promoter. On the other hand, the two cell lines possess different active regulatory elements downstream of PARK7's TTS, with GM12878 having it 10kb downstream of TTS, and K562, 20kb downstream. The different active regulatory peaks near TTS of PARK7 in the two cell lines further provides an additional layer of reasoning in explaining the differential expression of PARK7 as illustrated in Figure 4.4b, besides the differential TF interacting pairs co-localized at the shared promoter site.

The annotation to a nearby TTS takes up about half of the total peaks. In K562, out of 50,000 ATAC-Seqs, 19,727(39%) peaks are annotated to be a promoter that is within 2kb of a TSS, and 23,105 (46%) peaks are annotated to be within 20kb of a TTS. Combined, 33,369 (66%) peaks are either near a TSS or TTS, suggesting that 19% of peaks have dual actions being close to a TSS and a TTS simultaneously. In the analysis, if a peak has already been annotated to the TSS of a gene, if the peak is also close to the same gene's TTS, the TTS connection will be excluded. So 19% of the dual action peaks are in connection of two or more different genes at the same time, revealing the multipurposeness of these open sites.

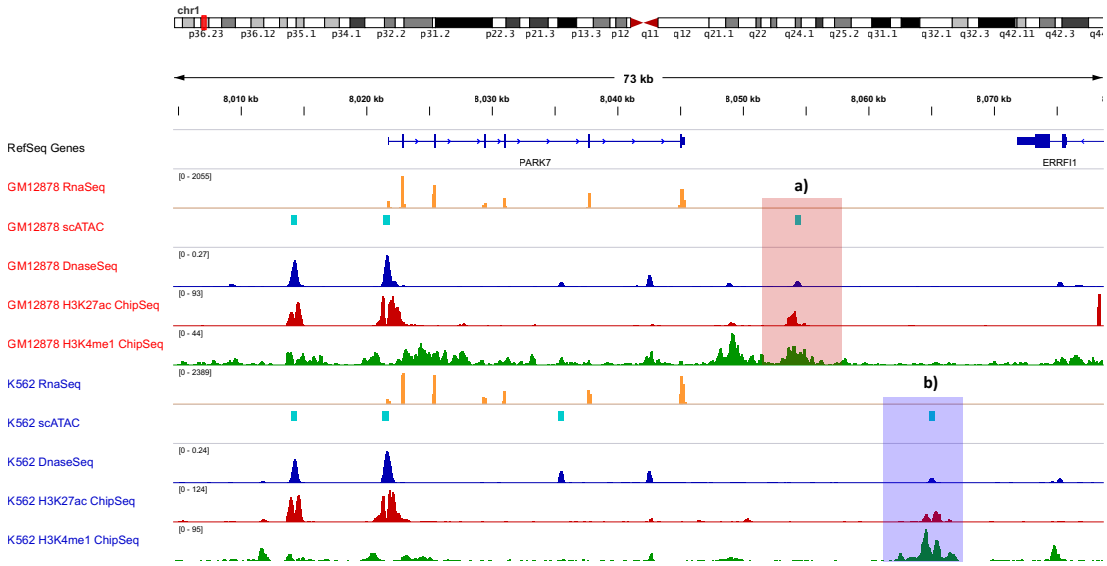


Figure 4.6: Cell-type specific open regions near the Transcription Termination Site (TTS) of PARK7. The open regions from ATAC-seq and Dnase-seq are shown around PARK7, along with binding profiles of H3K27ac and H3K4me1, the co-presence of which mark active promoters and enhancers. GM12878 (red) and K562 (blue) share the same promoter peak located at Transcription Starting Site (TSS) of PARK7, in addition to a shared enhancer peak about 10kb upstream of the promoter. On the other hand, the two cell lines possess a different active regulatory element downstream of PARK7's TTS, with GM12878 having one active peak 10kb downstream of TTS (box a), and K562, 20kb downstream (box b). The different enhancer peaks near TTS of PARK7 in the two cell lines further provides an additional dimension in explaining the differential expression of PARK7, as illustrated in Figure 4.4b.

4.3.3 Co-localized and co-activated combinatorial TF pairs (TF3C) shed light on differential regulation of PARK7

With co-localized and co-activated combinatorial TF pairs inferred for the promoter regions of PARK7 in Figure 4.4, we expand the search for TF3C to the enhancer and TTS regions of PARK7.

As revealed in Figure 4.7, TF3C are different when compared between GM12878 and K562 in all three regulatory elements: promoter, enhancer, and TTS. On the other hand, within K562, all three regulatory elements share a pair, CTCF/KLF5. Krüppel like factor 5 (KLF5) is a multifaceted transcription factor involved in cell growth, proliferation, and oncogenetic functions.[24] Likewise, CCCTC-binding factor (CTCF) is a multifunctional TF involved in various regulation pathways, also best known as one of the core architectural proteins that help establish a three-dimensional organization of the eukaryotic genome.[25,26] This lays the foundation for the hypothesis that such shared TF3C assists the contact between the regulatory elements, which is additionally affirmed by what we observed in GM12878. Slightly different from K562 where one pair is shared by all three regulatory elements, different TF3Cs are shared by different pair of genome elements. Enhancer and TTS peaks share the pair RUNX3/TCF7, with promoter sharing BACH1/ZNF267 with enhancer. Like K562 having CTCF involved in these three regions, GM12878 has RUNX3, that has presumably a role in maintaining 3D chromatin structure[27]. Thus, the shared TF3C and TFs are potentially responsible for a gene loop formed between enhancer, promoter, and TTS.

Recent literature with improved precision in profiling chromatin structure has proven the existence of such proposed gene loop between TSS and TTS, along with enriched contact between

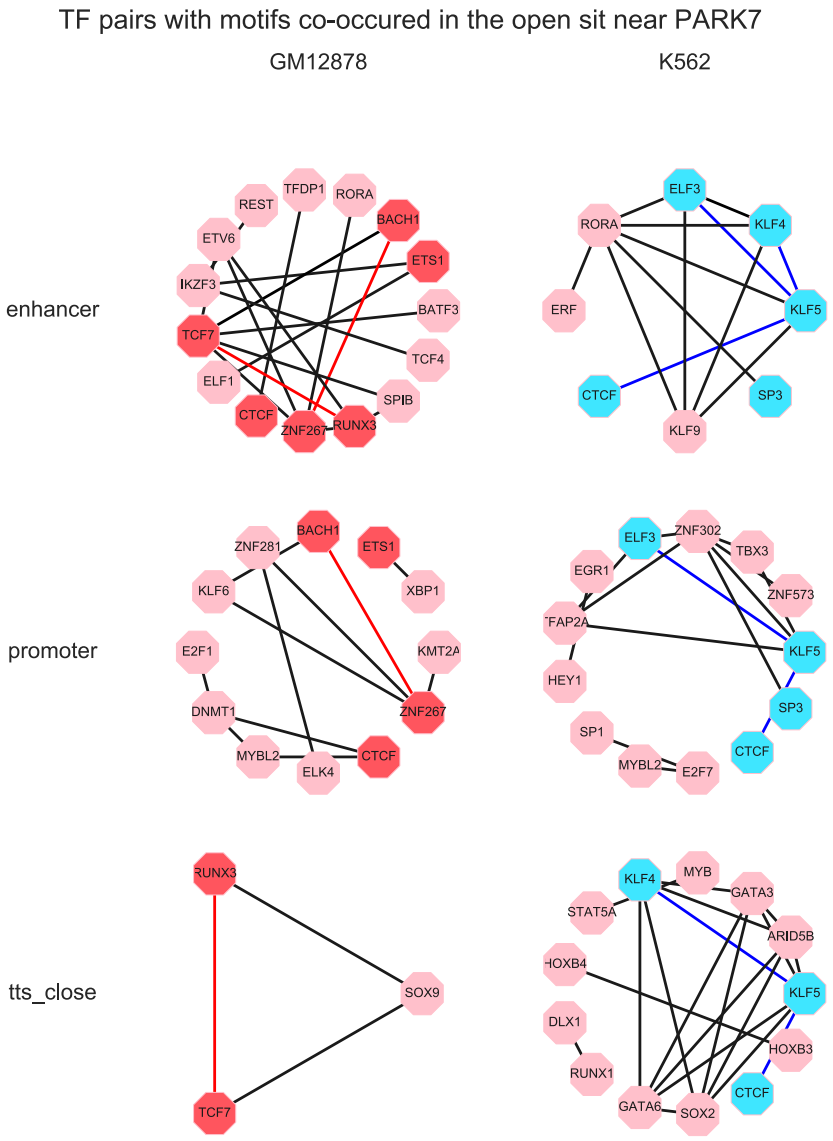


Figure 4.7: Differential TF3C located in the regulatory elements of PARK7 in GM12878 and K562. Co-located and co-activated combinatorial TF pairs (TF3C) annotated to PARK7's corresponding regulatory elements: promoter, enhancer, and the ATAC-peak downstream of TTS (tts_close, as illustrated in Figure 4.6). The annotation is based on motif co-occurrence of the correlated pair in each cell line. The figure demonstrates that TF3C are different for GM12878 and K562 in all three elements. On the other hand, within K562, all three regulatory elements share CTCF/KLF5. GM12878 is a slightly different story: enhancer and tts_close peak share the pair RUNX3/TCF7, with promoter sharing BACH1/ZNF267 with enhancer. CTCF and RUNX3 are believed to have roles in maintaining chromatin structure.[26,27] All these common TF3C and TFs (highlighted in red and blue for GM12878 and K562 respectively) are potentially responsible for a possible gene loop formed between enhancer, promoter, and the tts_close peak.

enhancer and promoter/TSS.[9] A proposed scheme of a local gene loop for PARK7 in GM12878 enlightened by the shared TF3Cs are drawn in Figure 4.8. As opposed to the promoter/enhancer/TTS being brought together by different TF3Cs in GM12878, the TF complex maintaining PARK7 genome structure in K562 would be dominated by CTCF and KLF5. Upon forming hetero-multimers, TF pairs stabilize the genome structure around PARK7, exposing the gene body to be constantly accessible by RNA Polymerase II in order for the gene to be transcribed in a high copy.

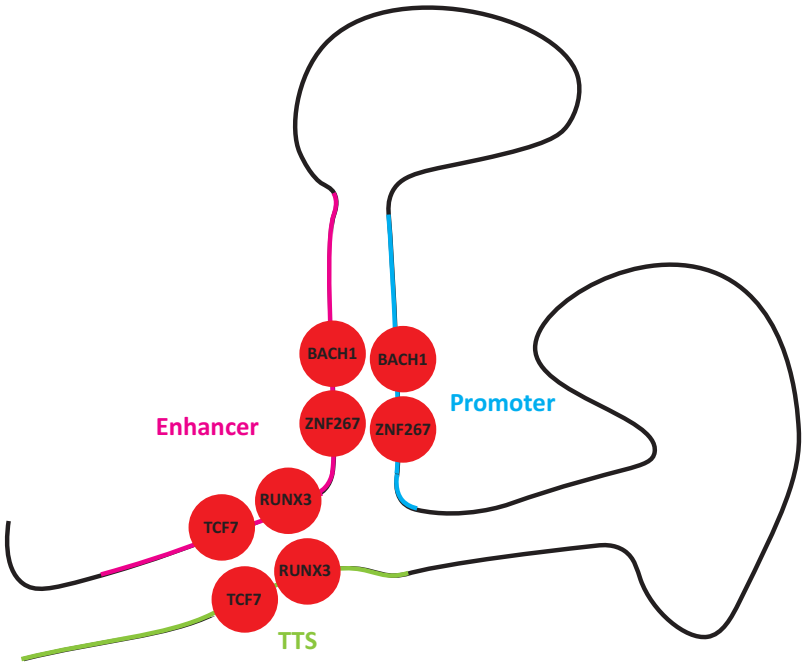


Figure 4.8: Proposed scheme of gene loop formation around PARK7 in GM12878 with assist from TF3C pairs. TF3C pairs are inferred based on co-activation from MALBAC-DT, and co-localization deduced from motif co-occurance in open regions. It is hypothesized that *via* the assist from shared TF3C, open regions from enhancer/promoter/TTS form a gene-loop together, stabilizing the transcription of the corresponding target gene in consideration, e.g. PARK7 in the example plotted.

One might wonder, can an open window of 500 bp be wide enough to accommodate multimers formed by a handful TFs. One of the best-studied TF complexes bound enhancer, called an

enhanceosome, is the one that regulates interferon-β (IFN-β). It is packed with eight different TFs on a 50bp genome window, with binding sites being contiguous to and even overlapped with each other.[28] In a recently published single-molecule nanoscopy experiment, in an enhancer bound with four different types of TFs, they counted more than a dozen molecules of each TF bound.[29]

These pieces of experiment evidence support the hypothesized scheme that a hetero-multimer is formed in an enhancer region, which is in proximity to promoter and TTS spatially. This further demonstrates the strength of TF3C in unveiling the TF regulation network by shedding light on differential TF hierarchy in the regulation of specific genes, which calls for a differential chromatin organization patterns to be accompanied with. The existence of multimers also reconciles with the short residence time of TFs. With a dozen of identical molecules nearby, even with quick turnaround time, TF complex can still maintain a stable structure in a highly dynamic fashion.

4.4 Construction of TF regulatory network from TF3C

After demonstrating the significance of TF3C in inferring TF regulation mechanism for a single gene PARK7, we expand the search to genome-wide. Chapter 3 showed that MALBAC-DT was able to give correlated gene modules (CGM), which are believed to be co-regulated by the same mechanism. Here, we explore whether that observation is in accordance with the new insight brought by TF3C.

In each cell line, there is a strongly correlated module enriched for protein synthesis functions, marked by YARS (Tyrosyl-TRNA Synthetase) and NARS (Asparaginyl-TRNA Synthetase), as il-

lustrated for K562 (Figure 4.9c) and GM12878 (Figure 4.10). We aim at decoding the co-regulating mechanism responsible for bringing these genes together.

By considering all the promoter/enhancer/TTS for each gene involved in the protein synthesis module, a bunch of TF3Cs is revealed to be enriched for each type of regulatory elements for the two cell lines considered. In K562, There are a couple of TF3Cs shared among the three types of genome regulatory elements. Such as the highly interconnected complex composed of ELF3/SOX2/ARID5B, which are shared between promoter and TTS. On the other hand, there are ELF3/SOX2/GATA6/GATA3 shared between enhancer and promoter. ELF3 (E74 Like ETS Transcription Factor 3) and SOX2 (SRY-Box 2) are both essential for maintaining self-renewal of cells and are known to aggregate in enhancers.[29,30] A similar scheme of the transcription factor complex formation in conjunction with gene looping can be proposed from these common TFs, as we did for the single gene PARK7 in Figure 4.8.

The shared TF3Cs and TFs are different in GM12878. No enriched TF3Cs are observed for enhancers in GM12878, with promoter and TTS sharing the pair MYBL2/DNMT1. MYBL2 (MYB Proto-Oncogene Like 2) is a crucial regulator of cell proliferation, cell survival, and differentiation involved in oncogenesis.[31] In contrast, MYBL2 is enriched by partnering with HOXB7 in enhancers of protein synthesis genes in K562. This further exemplifies that same TF can be involved in the same pathway while in varied roles with distinct partners.

In summary, we present with a genome-wide combinatorial regulation map, with top expressed genes exemplified in Table 4.1 and Table 4.2. The more TF3C's than average annotated to the highly expressed genes indicates that more factors are needed to accommodate the diversified roles a high copy gene holds.

4.5 Conclusion: Glimpse into combinatorial regulation map

As necessary and essential as it can be, cooperative TF pairs play a central role in transcription regulation. However, due to its dynamic nature, it has been challenging to capture. Now with the innovated single transcriptome technique, MALBAC-DT, we have demonstrated our ability to extract the cell-type specific TF-TF interactions embedded in a steady-state cell population. The highly diversified TF pairs among different cell lines signify how the common TFs can direct different cell fates.

In order to further dissect the functional roles of TF pairs in transcription regulation, the correlation studies were combined with chromatin accessibility data from ATAC-Seq and motif to infer co-localized and co-activated combinatorial TF pairs (TF3C). Such highly confident co-operative pairs shed light on the differential regulation and thus the differential expression of the same gene in different cell lines. Along the way to decode the regulation roles, we also discovered that highly expressed genes did not only have open promoters but also have open termination site, TTS, with a peak in proximity. Integrated with this knowledge, we proposed a regulation scheme by looping together enhancer, promoter, and TTS, as facilitated by the shared TF3C pairs. Such gene loops

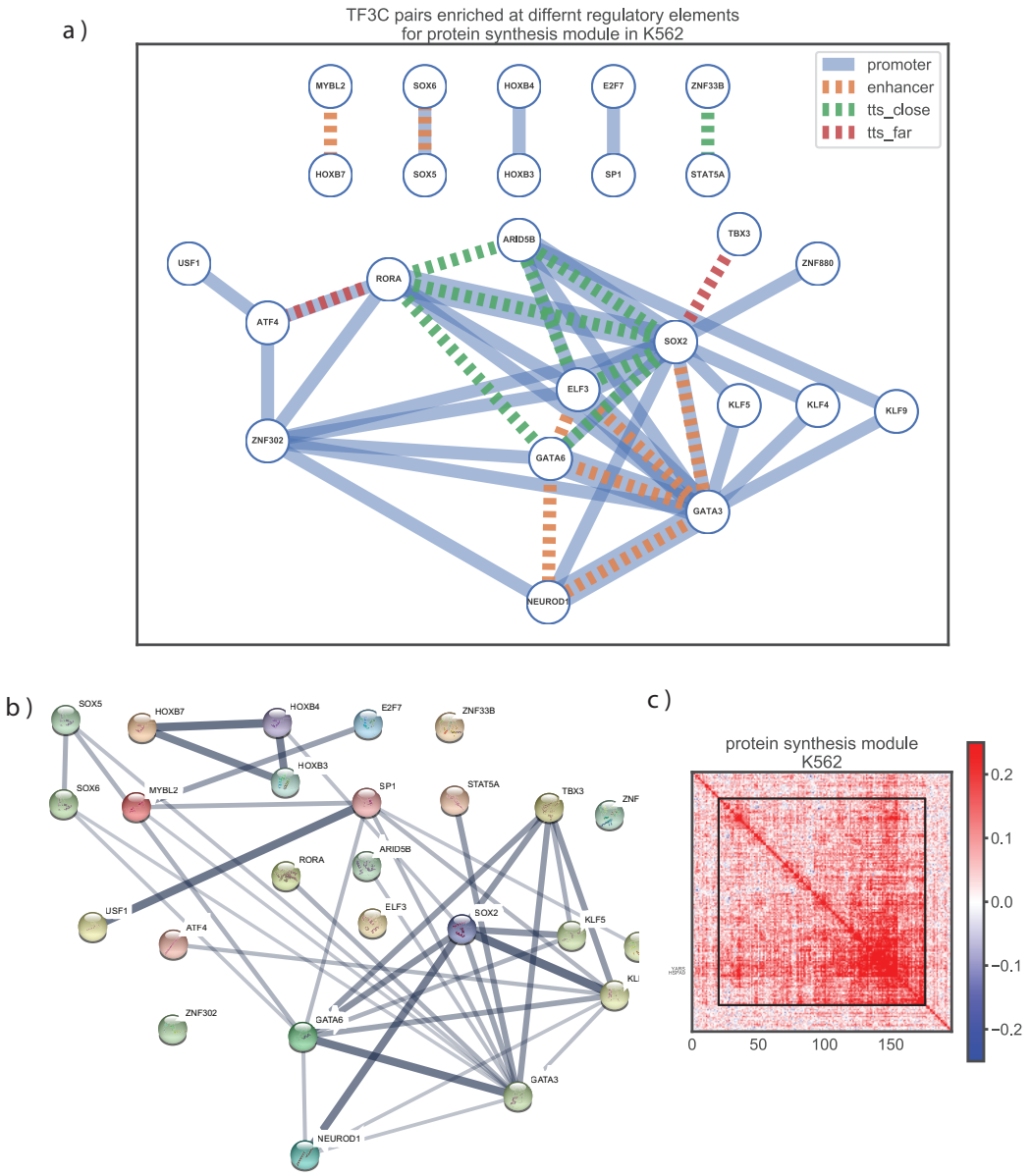


Figure 4.9: TF3C enriched for the regulatory elements regulating protein synthesis module in K562. a) TF3C pairs enriched at promoter (within 2kb of TSS), enhancer (within 20kb of TSS), or TTS_close and TTS_far (open peak within 20kb or 50kb of TTS, respectively) of genes from the protein synthesis module of K562 (c). Known protein-protein interactions (PPI) from STRING Database v11 [10] are illustrated in b, showing high resemblance to the interaction network demonstrated by MALBAC-DT in a). TF3C are shown to be shared between promoters and enhancers, and between promoters and TTS peaks, implying the possibility of them assisting the gene loop formation by bringing promoter/enhancer/TTS together.

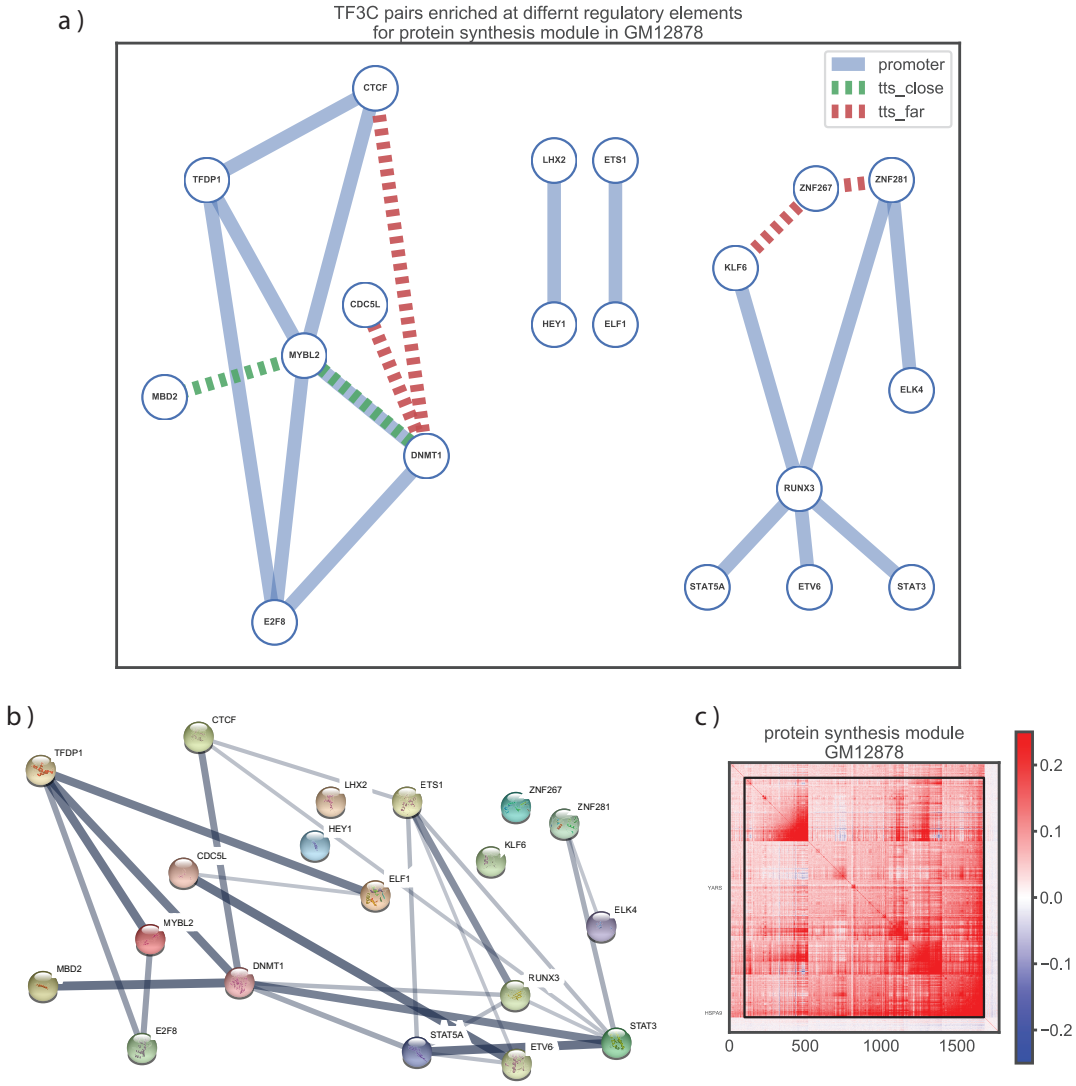


Figure 4.10: TF3C enriched for the regulatory elements regulating protein synthesis module in GM12878. a) TF3C pairs enriched at promoter, enhancer, or TTS_close and TTS_far (open peak within 20kb or 50kb of TTS, respectively) of genes from the protein synthesis module of GM12878 (c). Known protein-protein interactions (PPI) from STRING Database v11 [10] are illustrated in b, showing high resemblance to the interaction network demonstrated by MALBAC-DT in a). TF3C are shown to be shared between promoters and enhancers, and between promoters and TTS peaks, implying the possibility of them assisting the gene loop formation by bringing promoter/enhancer/TTS together.

Table 4.1: Combinatorial regulation map for the top 20 expressed genes in GM12878. Combinatorial co-activated Transcription Factor pairs co-localized at promoters (open peak within 2kb of TSS), enhancers (open peak within 20kb of TSS), and TTS_close (open peak within 20kb of TTS) of the top 3 expressed genes in GM12878.

Target Gene	Promoter	Enhancer	TTS_close
MALAT1	ZNF267|TCF7; IRF4|PRDM1; GTF3A|TFDP1; MYB|IKZF1; ZNF267|ZNF281; RELB|ETV6; BACH1|TCF7; RELB|STAT6; TCF7|BATF3; ETV6|TCF7; CTCF|TFDP1; MYBL2|TFDP1; IKZF3|IRF4; KLF6|BACH1; RELB|SPIB; RELB|ZNF267; STAT6|SPIB; SPIB|TCF7; GTF3A|MYBL2; IKZF3|PRDM1; IKZF3|REST; ETV6|ZNF267; MYBL2|CTCF; KLF6|TCF7; RELB|KLF6; RELB|TCF7; HEY1|BATF3; ZNF267|BACH1; STAT1|STAT2; KLF6|ZNF267; HEY1|TCF7	ETV6|NFKB1; ZNF267|NFKB1; ZNF267|ZNF281; XBP1|ETS1; HEY1|NFKB1; KLF6|NFKB1; KLF6|BACH1; BHLHE40|NFKB1; IKZF3|ETS1; BACH1|NFKB1; ETV6|ZNF267; IKZF3|REST; NFKB1|BATF3; ELF1|ETS1; ETV6|EBF1; NFKB1|EBF1; HEY1|BATF3; ZNF267|BACH1; HEY1|EBF1; KLF6|ZNF267; ATF6|ZNF281	IRF4|PRDM1; NR6A1|RELB; BACH1|TCF7; TCF7|LHX2; NR3C1|NFKB1; RELB|NFKB1; KLF6|SOX9; SOX9|SPIB; REL|NFKB1; NFKB1|EBF1; PRDM1|FOXO1; SOX9|TCF7; HIF1A|TCF7; RELB|NR3C1; KLF6|BACH1; NR3C1|EBF1; NR3C1|SOX9; POU2F2|SPIB; SMAD3|NR1D2; REL|TCF7; SPIB|TCF7; NR3C1|TCF7; POU2F2|BPTF; KLF6|TCF7; RELB|KLF6; SMAD3|IRF4; SOX9|LHX2; SMAD3|NR3C1; RELB|TCF7; NR6A1|NFKB1; DNMT1|CTCF; RELB|REL; DNMT1|E2F1; KLF6|NR3C1; STAT1|STAT2
ACTB	GTF3A|TFDP1; ETV6|NFKB1; MYBL2|DNMT1; REL|SOX9; ZNF281|RUNX3; ZNF267|RUNX3; MYB|IKZF1; NR3C1|NFKB1; ELF1|MEF2C; KLF6|SOX9; NFKB1|EBF1; MYBL2|TFDP1; KLF6|BACH1; SOX9|NFKB1; GTF3A|DNMT1; NR3C1|ZNF267; NR3C1|EBF1; SOX9|EBF1; POU2F2|SPIB; STAT6|SPIB; GTF3A|MYBL2; IKZF3|REST; NFKB1|BATF3; ELF1|ETS1; MYBL2|E2F8; ETV6|EBF1; ETV6|SOX9; SOX9|BACH1; KLF6|RUNX3; ZNF267|BACH1; SOX9|SPIB; ATF6|ZNF281; ZNF267|RORA; ZNF267|NFKB1; XBP1|ETS1; KLF6|ZNF267; SOX9|RUNX3; CTCF|TFDP1; KLF6|NFKB1; DNMT1|E2F8; RUNX3|SPIB; REL|SPIB; IRF8|NFAT5; DNMT1|TFDP1; IKZF3|ETS1; SMAD3|RORA; MYBL2|CTCF; TFDP1|E2F8; ZNF267|SOX9; DNMT1|CTCF; XBP1|PRDM1; ELK4|ZNF281; REL|RUNX3; DNMT1|E2F1; KLF6|NR3C1; ZNF267|ZNF281; ELF1|KLF2		GTF3A|TFDP1; IRF4|PRDM1; ETV6|NFKB1; ELF1|KLF2; ZNF281|RUNX3; RELB|ETV6; NR3C1|NFKB1; ZNF267|RUNX3; ZNF267|STAT5A; REL|NFKB1; NFKB1|EBF1; IKZF3|TCF4; SOX9|NFKB1; NR3C1|ZNF267; RELB|SPIB; ETV6|RUNX3; STAT5A|NFKB1; RELB|ZNF267; STAT6|SPIB; BACH1|NFKB1; ELF1|ETS1; SMAD3|IRF4; SMAD3|NR3C1; ETV6|EBF1; RELB|REL; NR3C1|STAT5A; RUNX3|NFKB1; ZNF267|RORA; NR6A1|RELB; ZNF267|NFKB1; RELB|RUNX3; RELB|NFKB1; KLF6|ZNF24; RELB|STAT6; SOX9|RUNX3; CTCF|TFDP1; RUNX3|SPIB; KLF6|NFKB1; PRDM1|FOXO1; RELB|NR3C1; IKZF3|IRF4; REL|SPIB; IKZF3|SMAD3; IKZF3|ETS1; NR3C1|SOX9; RELB|STAT5A; SMAD3|RORA; STAT5A|RUNX3; IKZF3|PRDM1; REL|STAT5A; TBX15|EOMES; RELB|KLF6; IRF8|REL; NR3C1|RUNX3; RUNX3|STAT3; SOX9|STAT5A; NR6A1|NFKB1; REL|RUNX3; ELK4|ZNF281; NR6A1|STAT5A; KLF6|NR3C1; ZNF267|ZNF281; KLF6|ZNF267; RUNX3|EBF1
HSP90AB1	GTF3A|TFDP1; IRF4|PRDM1; ZNF281|RUNX3; ZNF267|RUNX3; ZNF267|STAT5A; KLF6|SOX9; TP53|TFDP1; KLF6|BACH1; GTF3A|DNMT1; MEF2C|TCF4; STAT5A|NFKB1; NFKB1|BATF3; SMAD3|IRF4; SOX9|BACH1; KLF6|RUNX3; ZNF267|BACH1; ATF6|ZNF281; ZNF267|RORA; ZNF267|NFKB1; DNMT1|E2F1; XBP1|ETS1; PRDM1|FOXO1; CTCF|TFDP1; RUNX3|SPIB; KLF6|NFKB1; IKZF3|IRF4; REL|SPIB; DNMT1|TFDP1; IKZF3|MEF2C; SMAD3|NR1D2; IKZF3|PRDM1; MEF2C|KLF2; RUNX3|STAT3; ZNF267|SOX9; REL|RUNX3; DNMT1|CTCF; IKZF3|TCF4; MEF2C|PRDM1; XBP1|PRDM1; ZNF267|ZNF281; KLF6|ZNF267		ZNF281|RUNX3; HEY1|NFKB1; KLF6|SOX9; REL|NFKB1; NFKB1|EBF1; KLF6|BACH1; RELB|SPIB; POU2F2|SPIB; STAT6|SPIB; IKZF3|REST; ELF1|ETS1; NFKB1|BATF3; SMAD3|IRF4; MEF2C|STAT1; SOX9|BACH1; RELB|REL; ETV6|SOX9; ZNF267|BACH1; STAT1|STAT2; SOX9|SPIB; ZNF267|RORA; RELB|SOX9; KLF6|ZNF24; RELB|NFKB1; RELB|STAT6; KLF6|NFKB1; RUNX3|SPIB; IRF8|NFAT5; IKZF3|ETS1; IKZF3|MEF2C; IKZF3|PRDM1; IRF8|REL; NR3C1|RUNX3; RUNX3|STAT3; REL|RUNX3; XBP1|PRDM1; MEF2C|PRDM1; MYBL2|E2F8; CTCF|TFDP1; ELF1|KLF2; KLF6|ZNF267; RUNX3|EBF1; ETV6|NFKB1; IRF4|PRDM1; GTF3A|TFDP1; REL|SOX9; RELB|ETV6; NR3C1|NFKB1; MYB|IKZF1; ZNF267|RUNX3; ELF1|MEF2C; ZNF267|ZNF24; IKZF3|TCF4; TP53|TFDP1; SOX9|NFKB1; NR3C1|EBF1; NR3C1|BACH1; RELB|ZNF267; SOX9|EBF1; BACH1|NFKB1; SMAD3|NR3C1; ETV6|EBF1; KLF6|RUNX3; ATF6|ZNF281; RUNX3|NFKB1; ZNF267|NFKB1; XBP1|ETS1; PRDM1|FOXO1; IKZF3|IRF4; REL|SPIB; BHLHE40|NFKB1; DNMT1|TFDP1; SMAD3|RORA; SMAD3|NR1D2; MYBL2|CTCF; ETV6|ZNF267; TBX15|EOMES; RELB|KLF6; ZNF267|SOX9; DNMT1|CTCF; NR3C1|ETV6; ELK4|ZNF281; DNMT1|E2F1; KLF6|NR3C1; ZNF267|ZNF281

Table 4.2: Combinatorial regulation map for the top 20 expressed genes in GM12878. Combinatorial co-activated Transcription Factor pairs co-localized at promoters (open peak within 2kb of TSS), enhancers (open peak within 20kb of TSS), and TTS_close (open peak within 20kb of TTS) of the top 3 expressed genes in K562.

Target Gene	Promoter	Enhancer	TTS_close
RPL14	RBPJ|IKZF1; KLF5|TOPORS; KLF9|KLF4; KLF9|ELF3; KLF5|ELF3; KLF5|KLF4; ELF3|KLF4; KLF9|KLF5		ZNF302|KLF4; KLF4|NEUROD1; RORA|ELF3; ARID5B|ZNF302; KLF9|KLF4; KLF5|CTCF; ZNF302|NEUROD1; KLF5|ZNF333; KLF5|ZNF302; KLF9|ARID5B; KLF5|ARID5B; ARID5B|KLF4; STAT5A|STAT5B; GATA3|GATA6; SOX2|GATA6; ELF3|GATA6; ATF4|ZNF333; ZNF333|KLF4; KLF5|NEUROD1; SOX2|GATA3; ZNF880|SOX2; ARID5B|ZNF333; KLF9|ZNF333; KLF9|ZNF302; ELF3|GATA3; KLF5|KLF4; ATF4|RORA; ARID5B|NEUROD1; RORA|GATA6; ZNF302|SP3; MAFG|MYBL2; KLF9|NEUROD1; ZNF302|ZNF333; SOX2|ZNF302; ATF4|ZNF302; SOX2|ELF3; ZNF333|NEUROD1; RORA|GATA3; SOX2|RORA; KLF9|KLF5
MALAT1	RORA|ELF3; ZNF333|ELF3; TFAP2A|GATA3; KLF5|CTCF; KLF9|KLF4; KLF9|TFCP2L1; SOX2|ZNF333; RORA|ZNF333; GATA3|SP3; TFAP2A|KLF4; RBPJ|IKZF1; E2F7|SP1; TFCP2L1|KLF4; RORA|SP3; KLF4|GATA6; RORA|TFAP2A; RORA|NEUROD1; SOX2|GATA6; STAT5A|STAT5B; ELF3|GATA6; GATA3|GATA6; RORA|ZNF573; KLF5|TFCP2L1; ATF4|ZNF333; SOX2|NEUROD1; TFAP2A|ELF3; KLF9|GATA3; ZNF33B|STAT5A; SOX2|GATA3; ZNF880|SOX2; TFCP2L1|GATA3; KLF5|TFAP2A; ELF3|NEUROD1; GATA6|NEUROD1; TFCP2L1|GATA6; KLF9|TFAP2A; ELF3|GATA3; KLF5|KLF4; ATF4|RORA; TFAP2A|GATA6; KLF5|GATA3; GATA3|ZNF891; RORA|GATA6; KLF5|GATA6; ZNF333|GATA6; KLF5|TOPORS; MYB|STAT5A; TFCP2L1|TFAP2A; ZNF333|ZNF573; E2F7|MYBL2; GATA3|KLF4; KLF9|GATA6; SOX2|ELF3; ZNF333|NEUROD1; RORA|GATA3; SOX2|RORA; KLF9|KLF5	E2F7|SP1; TFAP2A|MTF2; BHLHE40|EGR1; HEY1|EGR1	RORA|ELF3; KLF4|NEUROD1; ZNF333|ELF3; TFAP2A|GATA3; KLF5|CTCF; MYBL2|HES6; KLF9|KLF4; SOX2|ZNF333; RORA|ZNF333; GATA3|SP3; TFAP2A|KLF4; ATF4|USF1; KLF5|ZNF333; GATA3|TBX3; YBX1|TFDP1; SOX2|KLF4; KLF9|ARID5B; SOX5|SOX6; KLF5|ARID5B; E2F7|SP1; ARID5B|KLF4; RORA|SP3; MYCN|ELF3; STAT5A|STAT5B; KLF9|SOX2; KLF5|ELF3; GATA3|GATA6; SOX2|GATA6; ELF3|GATA6; ARID5B|SOX2; RORA|ZNF573; ELF3|KLF4; MYCN|RORA; SOX2|NEUROD1; TFAP2A|ELF3; KLF5|NEUROD1; ARID5B|ELF3; KLF5|SOX2; KLF9|GATA3; SOX2|GATA3; ZNF880|SOX2; KLF5|TFAP2A; KLF9|ELF3; ELF3|NEUROD1; ARID5B|GATA3; GATA6|NEUROD1; KLF9|ZNF333; KLF9|TFAP2A; ELF3|GATA3; KLF5|KLF4; ATF4|RORA; KLF5|GATA3; ARID5B|NEUROD1; RORA|GATA6; ZNF333|GATA3; ZNF333|GATA6; MAFG|MYBL2; SOX2|TBX3; KLF9|NEUROD1; E2F7|MYBL2; GATA3|KLF4; SOX2|ELF3; RORA|GATA3; SOX2|RORA; KLF9|KLF5; GATA3|NEUROD1
HSP90AB1	ZNF302|KLF4; KLF4|NEUROD1; KLF9|KLF4; KLF5|CTCF; GATA3|SP3; ZNF121|ZNF302; GATA3|TBX3; KLF5|ZNF302; SOX2|KLF4; KLF9|RORA; RORA|KLF4; KLF5|RORA; KLF5|TBX3; E2F7|SP1; RORA|SP3; KLF9|SOX2; KLF4|TBX3; SOX2|NEUROD1; KLF5|SOX2; KLF9|GATA3; KLF5|NEUROD1; SOX2|GATA3; ZNF880|SOX2; KLF9|ZNF302; KLF5|KLF4; ATF4|RORA; KLF5|GATA3; ZNF302|SP3; KLF9|NEUROD1; SOX2|TBX3; GATA3|KLF4; RORA|GATA3; SOX2|RORA; KLF9|KLF5; GATA3|NEUROD1	RORA|ELF3; KLF4|NEUROD1; ZNF333|ELF3; TFAP2A|GATA3; KLF9|KLF4; RORA|ZNF333; GATA3|SP3; TFAP2A|KLF4; GATA3|TBX3; KLF9|RORA; RORA|KLF4; ZNF333|TBX3; RORA|SP3; KLF4|GATA6; RORA|TFAP2A; RORA|NEUROD1; GATA3|GATA6; STAT5A|STAT5B; ELF3|GATA6; RORA|ZNF573; KLF4|TBX3; ELF3|KLF4; ZNF333|KLF4; TFAP2A|ELF3; KLF9|GATA3; KLF9|ELF3; ELF3|NEUROD1; GATA6|NEUROD1; KLF9|ZNF333; KLF9|TFAP2A; ELF3|GATA3; TFAP2A|GATA6; RORA|GATA6; ZNF333|GATA3; ZNF333|GATA6; KLF9|NEUROD1; ZNF333|ZNF573; GATA3|KLF4; KLF9|GATA6; ZNF333|NEUROD1; RORA|GATA3; GATA3|NEUROD1	ZNF302|KLF4; KLF4|NEUROD1; ZNF333|ELF3; KLF9|KLF4; KLF5|CTCF; MYBL2|HES6; TFAP2A|GATA3; TFAP2A|KLF4; KLF5|ZNF333; KLF5|ZNF302; ZNF302|GATA3; MYCN|TFAP2A; RORA|KLF4; E2F7|DNMT1; RBPJ|IKZF1; KLF5|TBX3; RORA|SP3; STAT5A|STAT5B; MYCN|KLF5; SOX2|GATA6; KLF9|SOX2; KLF4|TBX3; ELF3|KLF4; ARID5B|GATA6; SOX2|NEUROD1; TFAP2A|ELF3; ARID5B|ELF3; ZNF33B|STAT5A; SOX2|GATA3; KLF9|ELF3; KLF9|ZNF333; KLF9|TFAP2A; KLF9|ZNF302; MYCN|ZNF333; ATF4|RORA; MYB|STAT5B; KLF9|NEUROD1; ZNF302|ZNF333; E2F7|MYBL2; KLF9|GATA6; RORA|GATA3; SOX2|RORA; ZNF302|TFAP2A; GATA3|NEUROD1; RORA|ELF3; ZNF302|GATA6; ARID5B|ZNF302; NR6A1|MYB; MYCN|KLF9; SOX2|ZNF333; GATA3|SP3; ATF4|USF1; ZNF121|ZNF302; GATA3|TBX3; SOX2|KLF4; KLF9|RORA; SOX5|SOX6; MYCN|ZNF302; KLF5|RORA; E2F7|SP1; RORA|TFAP2A; RORA|NEUROD1; GATA3|GATA6; KLF5|ELF3; ZNF302|ELF3; ELF3|GATA6; ARID5B|SOX2; ZNF333|KLF4; BHLHE40|EGR1; HEY1|EGR1; KLF5|NEUROD1; KLF5|SOX2; KLF9|GATA3; ZNF880|SOX2; KLF5|TFAP2A; ELF3|NEUROD1; ARID5B|GATA3; ELF3|GATA3; KLF5|KLF4; KLF5|GATA3; TFAP2A|GATA6; GATA3|ZNF891; RORA|GATA6; KLF5|GATA6; ZNF302|SP3; MYB|STAT5A; MAFG|MYBL2; KLF5|TOPORS; SOX2|TBX3; SOX2|ZNF302; ZNF333|ZNF573; GATA3|KLF4; SOX2|ELF3; KLF9|KLF5

might be indicative of how the gene body is stabilized to allow for high copy gene expression.

Expanding from a single gene, we profiled the differential regulation of a gene cluster enriched for protein synthesis, which was a general function shared in diversified ways by cells. Showing the differential TF3C as expected, we caught a glimpse of the genome-wide combinatorial regulation map (Table), setting us one step further in decoding the control of the human genome.

4.6 References

[1] Jolma, A. *et al.* Dna-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).

[2] Alcántara-Silva, R. *et al.* Pisma: A visual representation of motif distribution in dna sequences. *Bioinformatics and biology insights* **11**, 1177932217700907–1177932217700907 (2017).

[3] Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650 – 665 (2018).

[4] Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* **5**, 276–287 (2004).

[5] Sharrocks, A. D. The ets-domain transcription factor family. *Nature Reviews Molecular Cell Biology* **2**, 827–837 (2001).

[6] Rohs, R. *et al.* Origins of specificity in protein-dna recognition. *Annual review of biochemistry* **79**, 233–269 (2010).

[7] Marcon, E. *et al.* Human-chromatin-related protein interactions identify a demethylase complex required for chromosome segregation. *Cell Reports* **8**, 297 – 310 (2014).

[8] Gebhardt, J. C. M. *et al.* Single-molecule imaging of transcription factor binding to dna in live mammalian cells. *Nature Methods* **10**, 421 – 426 (2013).

[9] Hsieh, T.-H. S. *et al.* Resolving the 3d landscape of transcription-linked mammalian chromatin folding. *bioRxiv* (2019).

[10] Szklarczyk, D. *et al.* The string database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research* **45**, D362–D368 (2017).

[11] Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744 – 752 (2010).

[12] Suzuki, H. *et al.* Protein–protein interaction panel using mouse full-length cdnas. *Genome Research* **11**, 1758–1765 (2001).

[13] Kim, S. *et al.* Probing allostery through dna. *Science* **339**, 816–819 (2013).

[14] Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

[15] Pranzatelli, T. J. F., Michael, D. G. & Chiorini, J. A. Atac2grn: optimized atac-seq and dnase1-seq pipelines for rapid and accurate genome regulatory network inference. *BMC Genomics* **19**, 563 (2018).

[16] Ross-Innes, C. S. *et al.* Cooperative interaction between retinoic acid receptor-α and estrogen receptor in breast cancer. *Genes & Development* **24**, 171–182 (2010).

[17] Kinkley, S. *et al.* rechip-seq reveals widespread bivalency of h3k4me3 and h3k27me3 in cd4+ memory t cells. *Nature Communications* **7**, 12514 (2016).

[18] Kristensen, A. R., Gsponer, J. & Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nature Methods* **9**, 907–909 (2012).

[19] O'Leary, N. A. *et al.* Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733–D745 (2016).

[20] Yue, D., Wang, Y., Sun, Y., Niu, Y. & Chang, C. C1qbp regulates ybx1 to suppress the androgen receptor (ar)-enhanced rcc cell invasion. *Neoplasia (New York, N.Y.)* **19**, 135–144 (2017).

[21] Weirauch, M. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

[22] Stelzer, G. *et al.* The genecards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics* **54**, 1.30.1–1.30.33 (2016).

[23] Daugherty, A. C. *et al.* Chromatin accessibility dynamics reveal novel functional enhancers in c. elegans. *Genome research* **27**, 2096–2107 (2017).

[24] Diakiw, S. M., D'Andrea, R. J. & Brown, A. L. The double life of klf5: Opposing roles in regulation of gene-expression, cellular function, and transformation. *IUBMB Life* **65**, 999–1011.

[25] Kim, S., Yu, N.-K. & Kaang, B.-K. Ctcf as a multifunctional protein in genome regulation and gene expression. *Experimental &Amp; Molecular Medicine* **47**, e166 (2015).

[26] Tang, Z. *et al.* Ctcf-mediated human 3d genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).

[27] Ibn-Salem, J. & Andrade-Navarro, M. A. Computational chromosome conformation capture by correlation of chip-seq at ctcf motifs. *bioRxiv* (2019).

[28] Panne, D. The enhanceosome. *Current Opinion in Structural Biology* **18**, 236 – 242 (2008).

[29] Li, J. *et al.* Single-molecule nanoscopy elucidates rna polymerase ii transcription at single genes in live cells. *Cell* (2019).

[30] Novo, C. L. *et al.* Long-range enhancer interactions are prevalent in mouse embryonic stem cells and are reorganized upon pluripotent state transition. *Cell reports* **22**, 2615–2627 (2018).

[31] Musa, J., Aynaud, M.-M., Mirabeau, O., Delattre, O. & Grünewald, T. G. Mybl2 (b-myb): a central regulator of cell proliferation, cell survival and differentiation involved in tumorigenesis. *Cell Death &Amp; Disease* **8**, e2895 (2017).

5

Conclusion and Future

With our innovated high accuracy scRNA-Seq technique, MALBAC-DT, we successfully provide gene pair-wise co-activation relationship as a representation of the dynamic nature unique to each cell type's steady state. Together with motif analysis, we deciphered co-localized and co-activated combinatorial TF pairs (TF3C), the differential combination of which was proved to be in connection with differential expression of the regulated gene. In accompany, we proposed a general

scheme for gene transcription, where the regulatory elements of a gene are brought together to form a gene loop, bridged by the shared TF3Cs. Based on that, we present a genome-wide combinatorial regulation map unique to cell types, taking one step further in deciphering the human genome transcription.

Our TF3C and its map provide a comprehensive candidate pool to be experimented on. As a next step, the verification can be planned from two perspectives: first prove the existence of gene loop and interrogate it to uncover its potential dynamic sates, then to prove the TF partnership at such looping loci. As scalable as it is, our technique is going to deduce the TF3C and the map for numerous cell types and tissues in the near future.

With the map in hand, endless opportunities await.

...

A

Methods

A.1 Cell culture and handling

K562, GM12878, U2OS and HEK293T cell lines were obtained from ATCC and cultured at 37°C in RPMI-1640 medium with 10% Fetal Bovine Serum and 1% Penicillin-Streptomycin. To form single cell suspensions for flow sorting, culture medium was removed, cultures were rinsed with Dulbecco's phosphate-buffered saline (D-PBS), and incubated with 1mL of 0.25% trypsin for 5 minutes.

Detached cells in D-PBS were pelleted by centrifugation at 300g for 5 minutes and resuspended in D-PBS. Single cell suspensions were kept on ice until flow sorting.

A.2 MALBAC-DT protocol

Cells are flow sorted into 3uL of lysis buffer consisting of 1uL H_2O, 0.6uL 5x SSIV buffer, 0.15uL 10% ICA-630, 0.8uL 5M betaine, 0.05uL SUPERase In, 0.2uL 50uM RT-An primer, and 0.2uL 10mM dNTP mix. Plates are stored at -80°C until ready for amplification. Plates are kept on ice while pipetting and vortexed and briefly centrifuged after all pipetting steps.

To perform reverse transcription, plates are incubated at 72°C for 3 minutes, then 1uL of RT mix is added consisting of 0.264uL H_2O, 0.16uL 5x SSIV buffer, 0.2uL 100mM DTT, 0.152uL SUPERase In, 0.024uL 1M $MgSO_4$, and 0.2uL SuperScript IV. Plates are incubated for 10 minutes at 55°C.

Next, excess reverse transcription primers are degraded by exonuclease digestion. 1uL of exonuclease mix is added consisting of 0.1uL ExoI buffer, 0.1uL H_2O, 0.6uL ExoI, and 0.2uL 50uM RT-Bn primer. Plates are incubated for 30 minutes at 37°C and then 20 minutes at 80°C.

Amplification is performed by adding 24uL of amplification mix consisting of 18.64uL H_2O, 3uL ThermoPol buffer, 0.4uL 10mM dNTP mix, 0.16uL 100mM $MgSO_4$, 0.4uL 50uM GAT-7N, 0.4uL 50uM GAT-COM, and 1uL Deep Vent (exo-). The following thermocycle program is run:

Step	Temperature	Time
1	95	5:00
2	4	0:50
3	10	0:50
4	20	0:50
5	30	0:50
6	40	0:45
7	50	0:45
8	65	4:00
9	95	0:20
10	58	0:20
11	Goto 2	10x
12	95	1:00
13	95	0:20
14	58	0:30
15	72	3:00
16	Goto 13	17x
17	72	5:00
18	4	0:00

Finally, amplification is completed by adding 0.4uL 50uM Tru2-Gn-RT primers and running an additional 5 cycles of PCR steps 12-15. Amplified plates are stored at -20°C until library preparation. To prepare libraries for sequencing, 1uL from all wells are combined and purified using 0.8x Ampure beads. The Nextera library preparation kit is used to add Illumina adapters by tagmentation. During subsequent PCR steps, Ix-Tru2 primers are substituted for Nextera S5XX primers in order to select the 3' ends of transcripts containing cell barcodes and UMIs.

A.3 Sequence processing

Separate fastq files are generated for each cell based on the outer and inner barcode sequences. Barcodes not matching a cell exactly are discarded. Barcodes, adapter sequences, and UMIs are stripped from the reads, and reads and are aligned to the human GRCh38.p7 reference using STAR

2.5.2. For each gene, a list of UMIs is obtained for all reads mapping to that gene, excluding regions masked by RepeatMasker. To remove extraneous UMIs resulting from amplification or sequencing errors, UMIs for a particular gene are represented as nodes in a graph, with connections between UMIs differing at no more than 7 bases. Connected components are identified, and the consensus sequence within each component is determined. Consensus sequences matching the (HBDV)5 RT-An pattern and differing from the (VDBH)5 RT-Bn pattern at at least three bases are retained. To avoid potential cross-talk between wells, UMIs observed for the same gene in multiple cells are discarded.

After obtaining UMI counts for all genes and cells, cells for which more than 1% of transcripts are from ERCC spike-ins or contain fewer than 1000 total transcripts are discarded, as are genes which are observed in fewer than 10% of cells. Counts are normalized relative to the total number of transcripts in each cell prior to computing the correlation matrix. Hierarchical clustering is performed using the SciPy function scipy.cluster.hierarchy.linkage using method "average," and with a distance metric of $1 - abs(\rho_{ij})$, where ρ_{ij} is the correlation between genes i and j.

A.4 Cell cycle correction

Pseudo-time inference and cell-cycle correction Pseudo-time was inferred for each cell by assuming that the expression of cell-cycle genes followed a sinusoidal function along the time trajectory. The actual expression of each cell-cycle gene was further modeled as follows, a normal distribution centered around the level predicted by sinusoidal function, with variance aggregated from both

stochastic expression variance and technical noise.

$$
\begin{aligned}
y_{g,c} &\sim \mathcal{N}(\mu_{g,c}, v_g^2 + v_{tech}^2) \\
\mu_{g,c} &= Amp_g * (\cos(t_c - T_{peak,g}) + 1)) + AmpShift_g
\end{aligned}
\tag{A.1}
$$

$y_{g,c}$: actual expression of gene g for cell c, $\mu_{g,c}$: expected expression of g for c from sinusoidal function.

v_g^2: gene specific variance from stochastic expression for g, v_{tech}^2: common technical noise.

$Amp_g, AmpShift_g$: amplitude of the sinusoidal function for g.

$T_{peak,g}$: The peak time of g, in the time scale of percentage into the cell-cycle. Retrieved from Cyclebase.org [1].

t_c: The pseudo-time of cell c.

The transcriptome was fitted against the described model, with a pseudo-time optimized for each cell to maximize the overall likelihood estimation. The MLE process was done using PyTorch [2].

In order to correct the covariance matrix for cell-cycle effect, cells were then ordered by the assigned pseudo-time, and the expression of each gene was corrected by subtracting the mean of the surrounding rolling window.

A.5 Motif scanning to predict binding sites

Motif profiles were retrieved from Cis-BP Database [3,4] as PWMs. Open regions were extracted from ATAC-Seq peaks for GM12878 and K562 (GSE65360 [5]). Then FIMO from MEME-suite [6] was used to scan for the potential binding sites of each TF.

A.6 References

[1] Santos, A., Wernersson, R. & Jensen, L. J. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Research* **43**, D1140–D1144 (2014).

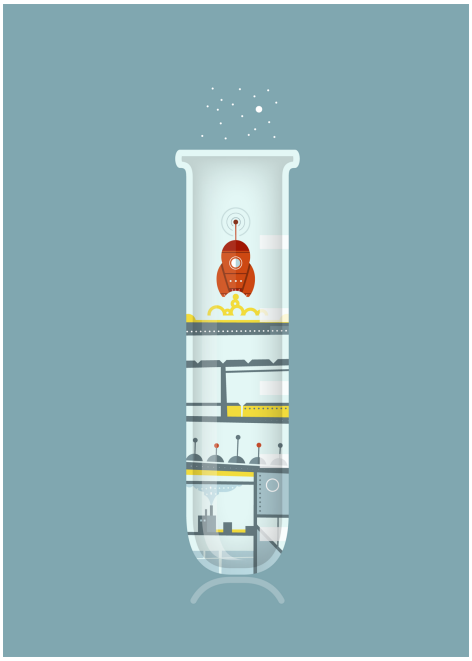
[2] Paszke, A. *et al.* Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop* (2017).

[3] Weirauch, M. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

[4] Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650 – 665 (2018).

[5] Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

[6] Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).



This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.