



Norwegian University
of Life Sciences

Master's Thesis 2020 60 ECTS

Faculty of Chemistry, Biotechnology and Food Science

Uncovering Key Transcription Factors in Breast Cancer Subtypes Using Matrix Factorization

Solveig Margrete Knoph Klokkerud

Bioinformatics and Applied Statistics (M.Sc.) - Bioinformatics

Acknowledgments

The following work was carried out for the Computational Biology & Gene Regulation group at Centre for Molecular Medicine Norway (NCMM), as part of the Master program in Bioinformatics and Applied Statistics at the University of Life Sciences (NMBU).

First, I would like to thank my main supervisor at NCMM, Anthony Mathelier, for guiding me through this process. Your enthusiasm for what could be achieved with my project has been a great inspiration, and you have encouraged me to work hard. Also, thank you for enduring me as I fumbled through the beginning of this steep learning curve - and for answering all my questions along the way. Thank you to my main supervisor at NMBU, Hilde Vinje, for helping me throughout this thesis. Your kind words of encouragement, your devotion to proofreading my drafts - I could not have done this without you.

Also, thank you to my co-supervisor at NCMM, Jaime Castro Mondragón, for all the help. Your tips and tricks in the world of programming have helped me become a better scientist. I would also like to thank the rest of the group at NCMM, for the discussions and help along the way. I especially want to thank Roza Berhanu Lemma - you have been my go-to for biology related questions, and you have always devoted your time to discuss the problems I have met. You have not only become a great inspiration, but also a friend.

I would also like to express my gratitude to my mum, dad and sister: Thank you for all your love and support, I am incredibly lucky to have you as a family. Last, but not least, I want to thank Preben for being there for me every step of the way. Your calm presence and belief in me has kept me going through the hard times.

Ås, June 2020

Solveig Klokkerud

Abstract

Breast cancer is the most common cancer type in women, and response to treatment varies immensely between subtypes. As of today, patients with Basal-like breast cancer lacks targeted treatment, which leads to poor prognosis for this group. Also other subtypes could benefit from a more targeted treatment. The molecular characteristics of each subtype remains an active area of research, and transcription factors that drive the subtypes need to be investigated in order to provide potential targets for more effective treatments. The molecular characteristics of each breast cancer subtype were inferred from ATAC-seq and RNA-seq data from 70 breast cancer patients, using two different matrix factorization methods. The first analysis used non-negative matrix factorization (NMF) on two separate data sets: One for ATAC-seq data, and one for RNA-seq data. The samples were clustered into five groups, based on molecular patterns shared within the groups, for both data sets. The DNA regions that were specifically open for each group were investigated for enriched transcription factor binding sites. The same was done for the promoter regions of the genes that were highly expressed in each group. The Basal-like subtype achieved the most successful clustering, and transcription factors likely to drive this subtype were uncovered. Also transcription factors responsible for driving a collective group of estrogen positive (ER+) subtypes were uncovered. The second analysis used Multi-Omics Factor Analysis (MOFA) to integrate the ATAC-seq and RNA-seq data in one combined analysis. The main purpose of this analysis was to support the findings of the first analysis, and possibly improve the clustering. The integration of multi-omics data resulted in two clusters, separating the Basal-like subtype from the rest of the subtypes. The clustering was not improved. However, some of the key transcription factors found for each group supported the results of the NMF analysis.

Sammendrag

Brystkreft er den krefttypen som rammer flest kvinner, og effekten pasienter har av behandling er svært avhengig av subtype. Fortsatt mangler pasienter med Basal brystkreft behandlingsalternativer som er målrettet mot denne subtypen, og prognosen er derfor dårlig for disse pasientene. Også pasienter med andre subtyper kunne ha dratt nytte av mer målrettet behandling. De molekylære egenskapene som kjennetegner hver subtype er et felt det forskes mye på, og transkripsjonsfaktorer som kan være viktige for hver av disse subtypene må undersøkes som potensielle mål for behandling. De molekylære egenskapene som kjennetegner de ulike subtypene ble funnet fra RNA-seq og ATAC-seq data fra 70 brystkreftpasienter, ved bruk av to ulike matrisefaktoriseringsteknikker. Den første analysen brukte ikke-negativ matrisefaktorisering (NMF) på to ulike datasett: Ett for ATAC-seq data, og ett for RNA-seq data. Prøvene ble gruppert i fem grupper, basert på de molekylære mønstrene som var felles for hver gruppe, for hvert datasett. DNA-regionene som var spesifikt åpne for hver gruppe ble undersøkt for å finne transkripsjonsfaktorbindingssetene som opptrådte oftest for hver gruppe. Det samme ble gjort for promoter-regionene til genene som var høyest uttrykt i hver gruppe. Den beste separasjonen ble oppnådd for den Basale subtypen, og for denne gruppen ble det funnet en rekke transkripsjonsfaktorer som trolig er viktige. Det ble også funnet transkripsjonsfaktorer som kan være viktige i subtyper som er drevet av østrogenreseptorer (ER+). Den andre analysen brukte «multi-omics» faktoranalyse (MOFA) for å integrere ATAC-seq og RNA-seq data i en kombinert analyse. Hovedmålet med denne analysen var å understøtte funnene fra den første analysen, og å forbedre grupperingene om mulig. Integreringen av «multi-omics» data resulterte i to grupper, som separerte den Basale subtypen fra resten av subtypene. Grupperingene ble ikke forbedret. Likevel kunne noen av transkripsjonsfaktorene som ble funnet for hver gruppe brukes til å støtte opp om resultatene fra NMF-analysen.

Contents

Acknowledgments	ii
Abstract	iv
Sammendrag	vi
Abbreviations	ix
1 Introduction	2
1.1 Gene regulation in breast cancer	5
1.1.1 Transcription factors	7
1.1.2 Chromatin	8
1.2 Investigating gene regulation	9
1.2.1 Measuring gene expression	10
1.2.2 The open chromatin landscape and TFs	11
1.2.3 Interpreting big data by dimensionality reduction	14
1.2.4 Predict transcription factor drivers from regulatory regions	18
2 Aim of thesis	20
3 Materials and methods	22
3.1 Data	22
3.2 NMF analysis	24
3.2.1 Non-negative matrix factorization (NMF)	24
3.2.2 Feature selection	25
3.2.3 Investigating genomic regions	27
3.2.4 Gene ontology enrichment analysis	28
3.2.5 Transcription factor binding site enrichment	28
3.3 Multi-omics analysis	29
3.3.1 Data preprocessing and normalization	29
3.3.2 Feature selection and signature analyses	30
4 Results	31
4.1 NMF	31
4.1.1 Clustering of samples into a priori subtypes	31
4.1.2 Clustering of features reveals the activity of each pattern .	36
4.1.3 Connecting the samples with pattern-specific features . .	39
4.1.4 Gene signatures and gene set enrichment analysis	41
4.1.5 Open regions and their chromosomal location	43
4.1.6 Key transcription factors	45

4.2	MOFA	50
4.2.1	Normalization	50
4.2.2	Sample clustering	51
4.2.3	Gene ontology enrichment analysis	52
4.2.4	UniBind TF enrichment	54
5	Discussion	56
5.1	NMF analysis	56
5.1.1	Subtype clustering	56
5.1.2	Gene and peak signatures	57
5.1.3	Gene ontology enrichment analysis	58
5.1.4	Chromosome distribution	58
5.1.5	Subtype-specific transcription factors	59
5.2	MOFA analysis	62
5.2.1	Subtype clustering	62
5.2.2	Gene ontology enrichment analysis	62
5.2.3	Subtype-specific transcription factors	63
5.3	Method discussion	63
6	Conclusion and future perspective	66
	Bibliography	67
	Attachments	82

Abbreviations

ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
bp	Base pairs
DNA	Deoxyribonucleic acid
ER	Estrogen receptor
HER2	Human epidermal growth factor receptor 2
ICA	Independent Component Analysis
IHC	Immunohistochemistry markers
MF	Matrix factorization
MOFA	Multi-Omics Factor Analysis
NMF	Non-Negative Matrix Factorization
PCA	Principal Component Analysis
PR	Progesterone receptor
RNA	Ribonucleic acid
RNAPII	RNA polymerase II
RNA-seq	RNA sequencing (next generation sequencing)
TF	Transcription factor
TFBS	Transcription factor binding site
TSS	Transcription start site

Chapter 1

Introduction

Cancer is a disease that occurs when normal cells turn into high replicating tumor cells. It can happen anywhere in the body and can be caused by multiple factors, for example genetic predisposition, viral infections, radiation or contaminants such as tobacco smoke. Age, diet and physical activity are highly contributing risk factors (WHO, 2018). The various types of cancer are responsible for nearly 10 million deaths globally each year, and is the second leading cause of death among the global population (WHO, 2018). As of today, cancer is one of the most challenging diseases to combat, as it varies immensely between patients. Cancers in different tissue types are driven by different factors, and so are cancers within the same tissue type, giving rise to multiple subtypes of each cancer type (Song et al., 2015) (Figure 1.1).

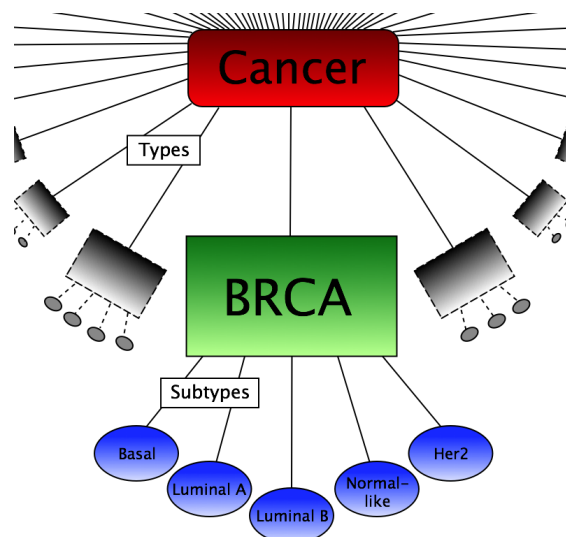


Figure 1.1: There are multiple cancer types, each of which have multiple subtypes. Here shown for breast cancer (BRCA), with five molecular subtypes: Basal-like, Luminal A, Luminal B, Normal-like and Her2.

Breast cancer is the most commonly diagnosed cancer type for women in developed countries (Bray et al., 2018). Approximately 8% of Norwegian women will get breast cancer at one point in their life (Kreftforeningen, 2020), and the number is estimated to be 12% for American women (Waks & Winer, 2019). Up until 2011, it was also the leading cause of cancer related deaths for women in developed countries (Jemal et al., 2011; Torre et al., 2015). In some countries, including Norway, breast cancer has now been surpassed by lung cancer, likely due to improved treatments (Bray et al., 2018; DeSantis et al., 2017). Current treatments include surgery, chemotherapy, radiation and hormone therapy (Waks & Winer, 2019). However, the clinical outcome of these treatments varies immensely from patient to patient, depending on subtype (Waks & Winer, 2019).

Breast cancer subtypes

Breast cancer can be subtyped in a number of ways, depending on which underlying characteristics that are in focus (Dai et al., 2016; Stingl & Caldas, 2007; Sun et al., 2014). According to the literature reviewed for this thesis, the three most common methods are:

1) Histological/morphological methods

Classification based on location or tissue in the breast where cancer cells are present, or morphological features from microscope examination. These subtypes are often referred to as breast cancer types in order to avoid confusion with other subtypes (Canadian Breast Cancer Network, 2020). Examples include ductal carcinoma in situ and metastatic (stage IV) breast cancer (Malhotra et al., 2010; Ivshina et al., 2006).

2) Receptor status

Method that uses immunohistochemistry markers (IHC) to describe different subtypes. These subtypes are normally created from a combination of estrogen receptor (ER) status, progesterone receptor (PR) status and human epidermal growth factor receptor 2 (HER2) status. Each receptor status can also be used independently, by for example dividing breast cancer into ER+ and ER- (Dai et al., 2016).

3) Molecular subtype

Classification of subtypes based on molecular profiles. The most common methods are i) integrative clustering and ii) intrinsic clustering (Russnes et al., 2017). Integrative clustering is based on a combination

of copy number drivers and gene expression, while intrinsic subtypes are based on gene expression alone. Molecular subtypes based on intrinsic clustering normally include Normal-like, Basal-like, Her2, Luminal A and Luminal B (Russnes et al., 2017). Prediction analysis of microarray-50 (PAM50) (Parker et al., 2009) has gained popularity as a robust way of classifying breast cancer into these five molecular subtypes based on the expression of 50 genes (Nielsen et al., 2010; Sabatier et al., 2014).

These methods for classifying subtypes are based on different criteria, which means that tumors that are grouped together using one criteria may not be grouped together using another. However, tumors with the same molecular subtype usually have the same receptor status (see Table 1.1).

Table 1.1: Relationship between molecular subtype and receptor status. Normal-like, Luminal A and Luminal B tumors are normally hormone receptor positive, while Her2 and Basal-like tumors are almost exclusively hormone receptor negative. Her2 and some Luminal B tumors are enriched for HER2 (HER2+) (Nguyen et al., 2008; Breastcancer.org, 2020). Some individual tumors can have an atypical profile, but this table is based on the characteristics of most tumors. *ER+ and/or PR+

Molecular subtype	ER status	PR status	HER2 status
Luminal A	ER+/-*	PR+/-	HER2-
Luminal B	ER+/-*	PR+/-	HER2+/-
Normal-like	ER+/-*	PR+/-	HER2-
Her2	ER-	PR-	HER2+
Basal-like	ER-	PR-	HER2-

Separating between estrogen receptor positive (ER+) and estrogen receptor negative (ER-) breast cancer has often been a main focus in clinical settings, as it has a great impact on current treatment. ER+ tumors (Luminal A, Luminal B and Normal-like) account for 75-80% of the breast cancer cases (Cui et al., 2005; Hart et al., 2015), and because ER+ tumors are enriched for estrogen receptors (ER), patients with this subtype are likely to respond to hormone therapy. About 65 % of these ER+ tumors are also positive for progesterone receptor (PR+), and the combination of both these receptors increases success rate of hormone treatment and survival further (Cui et al., 2005). On the other hand, patients with ER-breast cancer (Basal-like and Her2) normally also lack the progesterone receptor, and are much less likely to respond to hormone treatment (Itoh et al., 2014). However, Her2 and some Luminal B tumors usually respond to treatment that targets HER2 receptors (Arteaga et al., 2012), while the treatment of Basal-

like cancer is limited to chemotherapy (Waks & Winer, 2019). This lack of targeted treatment leads to poor prognosis for patients with Basal-like breast cancer compared to the other subtypes (Anders & Carey, 2008; Dai et al., 2016) (Figure 1.2).

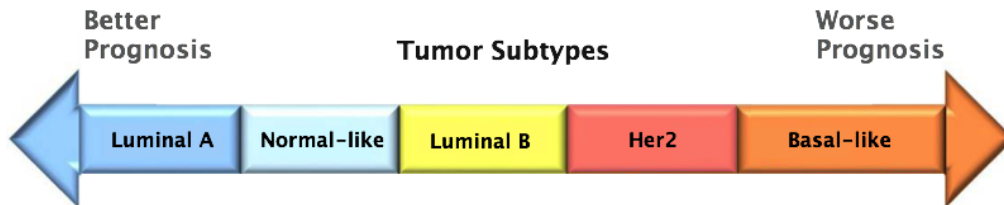


Figure 1.2: The five molecular subtypes (Basal-like, Luminal A, Luminal B, Normal-like and Her2) have different prognosis, based on response to current treatment. Hormone receptor positive subtypes (ER+/PR+) have the best prognosis, with Luminal A being the least deadly, followed by Normal-like and Luminal B. The hormone receptor negative subtypes have worse prognosis, with Basal-like being the most deadly (Wang et al., 2011; Breastcancer.org, 2020). Adapted from Dai et al. (2017).

In order to investigate the possibility for more successful treatments, especially for patients with Basal-like tumors, it is important to know the molecular subtypes and what characterize them. Molecular classifiers such as PAM50 has shown that each subtype can be characterized by a common gene expression pattern, called a gene signature (Cantini et al., 2017). Although these types of gene signatures have good prognostic value, they are intended for classification, and therefore contain the minimum number of genes needed to classify a sample (Nielsen et al., 2010). In order to explain more of the characteristics that define each subtype, larger gene signatures can be defined by using unsupervised learning methods on a full set of gene expression data. Molecular subtypes such as PAM50 can be used to validate the clusters, thus, combining a priori knowledge of subtypes with larger, data-driven gene signatures. However, to understand exactly why the subtypes exhibit different gene expression patterns, we have to understand the mechanisms behind. One of the most important mechanisms is gene regulation.

1.1 Gene regulation in breast cancer

Gene regulation is a set of mechanisms that increase or decrease the expression of genes, and previous research has shown that it plays an important part in the development and progression of breast cancer (Emmert-Streib et al., 2014; Hua

et al., 2008). Expression of genes are regulated at different stages: Transcription, translation and post-translation (Chen & Rajewsky, 2007; Kulis et al., 2013), as shown in Figure 1.3. Transcription of genes is largely regulated by a cooperation between transcription factors and chromatin structure (Bonifer & Cockerill, 2011). The regions where this regulation takes place are found in non-coding DNA, in regions previously called "junk DNA" (Nowak, 1994). Changes in gene regulation at transcriptional level will impact the later stages, and is therefore considered to be the most critical control point of gene regulation (Delgado & León, 2006).

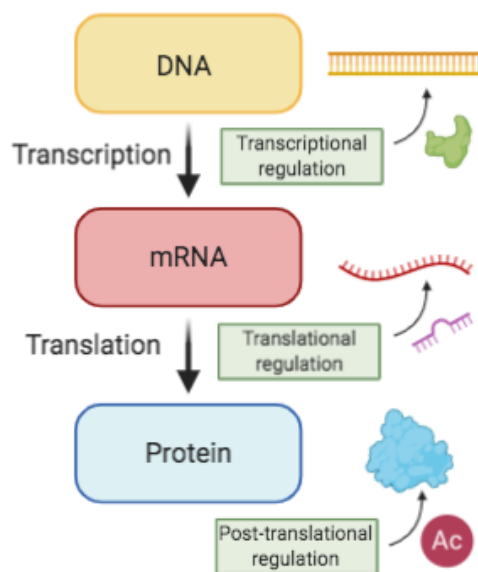


Figure 1.3: Regulation of gene expression is a multi-level process, that takes part in all steps of the central dogma. Transcriptional regulation is the first step, and affects the downstream processes of translation and post-translation. One example of regulation is shown for each step, starting from top: 1) TF binding, 2) microRNA binding and 3) acetylation. However, gene regulation is a complex process, and there are multiple other factors at play. Created with BioRender.

To better understand gene regulation, imagine a greenhouse where the goal is to get a tree to thrive. There are thousands of buttons with different functions: Some are responsible for watering, some are responsible for light, while others provide different kinds of nutrients to the roots. We have multiple janitors, each of them responsible for pushing one or more buttons. In order to do so, they need to have access to the buttons. Sometimes they are told they need to turn the buttons slightly up or down, so the tree gets exactly what it needs for normal function and growth. There are also some buttons that are supposed to stay turned off. These buttons are blocked, and the janitors are unable to push them. Now, imagine that this tree is a cell in our body. The buttons are genes, the

janitors are transcription factors and whether these janitors have access to the buttons or not correspond to open or closed chromatin. In normal cells, this system works: Chromatin opens where it is supposed to, and transcription factors only turn on the genes that should be on. However, this could change.

Let us say that one day, a lightning strikes and damages the system. Some of the buttons are unblocked by mistake, and the janitors push the buttons that are supposed to stay off. At the same time, some of the buttons that are supposed to be pushed are blocked, and the janitors responsible for pushing these buttons are unable to do so. Suddenly the tree starts growing uncontrollably; the trunk bulges into a thick structure and the branches start growing in every direction. This scenario is essentially what happens when normal cells turn into cancer cells. Here, the lightning strike represents an external cause that creates mutations in regions of the DNA related to growth. This is the case in about 90-95% of cancer cases, including breast cancer (Anand et al., 2008; Mehrgou & Akouchekian, 2016). These mutations initiate a cascade reaction where multiple genes that are involved in **promoting** cell division and growth are turned on, while genes involved in **suppressing** cell division and growth are turned off (Hua et al., 2008; Cox & Goding, 1991). The result is uncontrolled growth and tumor formation.

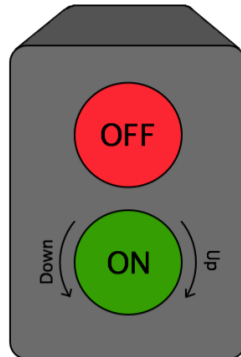


Figure 1.4: Gene regulation can be represented as the process of turning buttons, or genes, on/off or up/down.

1.1.1 Transcription factors

The transcription factors (TFs) - or "janitors" - are, together with chromatin, crucial for ensuring that the correct genes are expressed at the correct time in the correct cells in order for our body to function normally. There are roughly 1,600 different human TFs (Lambert et al., 2018), and previous research has estimated each TF to have as many as 10,000-300,000 copies within a single cell (Biggin,

2011; Simicevic et al., 2013). TFs can bind to different regions in the genome, but often bind the promoter region, which is located near and upstream of the transcription start site (TSS) of the gene they regulate (Tompa et al., 2005). TFs can also bind to regions located far away from the gene they regulate, as chromatin arrangements can allow them to be close in three-dimensional space (Marsman & Horsfield, 2012). These distant binding sites are located in regions called enhancers or silencers (Delgado & León, 2006). TFs that bind to promoter and enhancer regions are normally responsible for increased expression of the gene they regulate. On the other hand, TFs that bind to silencers can block RNA polymerase II (RNAPII) from binding and turn off all expression of a gene (Maston et al., 2006; Delgado & León, 2006).

Previous studies have shown that TFs have an important role in the development and progression of breast cancer (Shepherd et al., 2016). The most active TFs vary between breast cancer subtypes, resulting in different transcriptional profiles (Zhu et al., 2020).

There is a general agreement that TFs such as FOXA1, GATA3 and ER α (estrogen receptor alpha) are important drivers in ER+ subtypes like Luminal A/B and Normal-like (Theodorou et al., 2013). However, there is still a lot to discover about the TFs that potentially separate these subtypes.

On the other hand, there seems to be a lack of established consensus regarding the TFs that drive ER- tumors like Her2 and Basal-like. Some studies have suggested SOX2 as a possible driver of Basal-like tumors (Rodriguez-Pinilla et al., 2007; Chen et al., 2008), in addition to other SOX TFs, like SOX10 (Cimino-Mathews et al., 2013), SOX4 (Zhang et al., 2012) and SOX11 (Shepherd et al., 2016). TEAD4 (Wang et al., 2015; Adélaïde et al., 2007; Zhu et al., 2020), STAT3 (Zhu et al., 2020), CEBPB (Willis et al., 2015) and MYC (Xu et al., 2010) have also been suggested as potential Basal-like drivers. However, there seems to be large variations between different studies, depending on the data type and methods used.

For the Her2 subtype, Yin Yang 1 (YY1) has received attention as a likely TF driver (Begon et al., 2005; Powe et al., 2009). YY1 has been proposed to cooperate with TFs in the AP-2 (activator protein 2) transcription factor family (Woodfield et al., 2010; Powe et al., 2009). Although various AP-2 TFs appear to be enriched in Her2 tumors (Begon et al., 2005; Turner et al., 1998), other studies have found that they cooperate with ER α in ER+ luminal tumors (Cyr et al., 2015). TFAP2C (transcription factor AP-2 γ) is a member of the AP-2 family, and though this specific TF is often associated with ER+ tumors

(Woodfield et al., 2010), it has long been suggested to have important functions in different subtypes (Turner et al., 1998; Gee et al., 2009; Woodfield et al., 2010).

The TFs that drive the various breast cancer subtypes depend on differences in chromatin accessibility landscape, as TF binding is restricted to open chromatin regions. This gives possibilities for searching open regions for transcription factor binding sites (TFBSs).

1.1.2 Chromatin

The human genome contains roughly 3 billion base pairs (bp) of DNA (National Human Genome Research Institute (NIH), 2020), and this DNA is divided between 23 pairs of chromosomes. In order to fit these massive amounts of DNA in each cell, the DNA in each chromosome is tightly packed at several levels. The inner level of this packaging consists of DNA wrapped around proteins called histones, and this structure is called chromatin (Figure 1.5). Chromatin can either be tightly packed, in which case it is called closed chromatin, or loosely packed, which is called open chromatin. The reason why not all chromatin is tightly packed, is because the open and closed state of chromatin is one of the main contributors to gene regulation (Buenrostro et al., 2013).

If we think of the example above, the chromatin state decides whether the "janitors" are able to push the buttons or not. In other words, if a TF cannot reach its binding site due to a closed chromatin state, it is unable to regulate the gene. Similarly, if the TF has access, the genes they regulate will either be transcribed or repressed, depending on which region the binding site is located in. The process of opening closed chromatin and vice versa is called chromatin remodeling, and is driven by many different mechanisms. These include binding of pioneer TFs (Zaret & Carroll, 2011) and post-translational modifications of the histones (Delgado & León, 2006; Phillips & Shaw, 2008). In addition, complexes such as the CTCF/cohesin complex play a part in organizing the 3D structure of chromatin, which also affects transcription (Song & Kim, 2017). In order for a gene to be transcribed beyond basal levels, the chromatin in the promoter and enhancer regions of that gene needs to be accessible. The open regions gives RNAPII and TFs direct accessibility to the DNA, and thereby allow binding and subsequently transcription (Buenrostro et al., 2013).

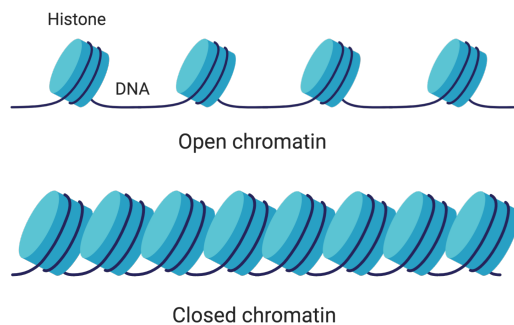


Figure 1.5: Chromatin is DNA wrapped around histone proteins, and can be open or closed. Created with BioRender.

1.2 Investigating gene regulation

There are multiple techniques available to investigate gene expression, open chromatin and TFs, in addition to methods for analyzing the data these techniques generate. Together, these techniques and methods can be used to understand gene regulation under different conditions, such as breast cancer.

1.2.1 Measuring gene expression

For many years, microarray was the leading method for measuring gene expression. RNA sequencing (RNA-seq) arose as a contender after the sequencing of the first human genome, and proved to be a more sensitive method able to detect genes expressed at very low and high level (Zhao et al., 2014). Following the commercialization of sequencing, the price of an RNA-seq experiment was drastically reduced, and is currently cheaper than microarray (Rao et al., 2019; Lachmann et al., 2018). As a result, RNA-seq is now the leading technique for measuring gene expression (Lachmann et al., 2018). RNA-seq uses next-generation sequencing to study different parts of the transcriptome. The transcriptome refers to all transcribed RNA in a given sample at a given time (Wang et al., 2009). The technique is commonly used to measure gene expression (Li et al., 2010), by adapting the sequencing library to sequence mRNA only (Wang et al., 2009). A brief overview of this procedure is described in Figure 1.6 (Wang et al., 2009). RNA-seq has often been used to perform gene set enrichment analysis and define gene signatures (Rapaport et al., 2013; Ackermann et al., 2016). Another usage is to search for TFBSs in the promoter of the most highly expressed genes. If combined with techniques that search for TFBSs in all open regions (including

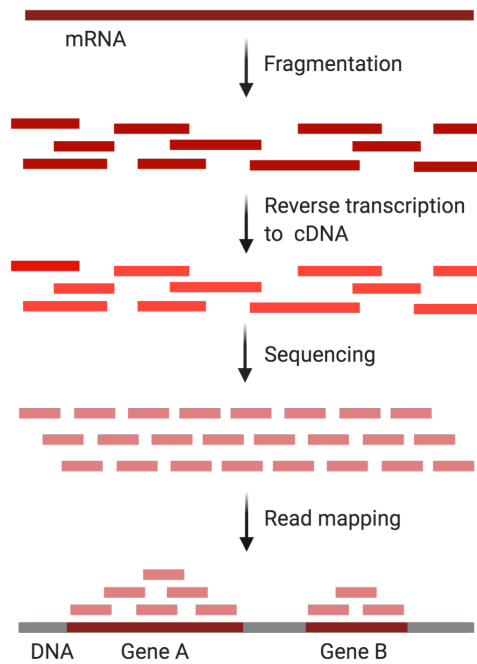


Figure 1.6: The first step of RNA-seq is normally fragmentation, as mRNA fragments are too big to be sequenced directly. After fragmentation, mRNA is converted to cDNA. However, the order of these two steps depends on the procedure. Regardless of order, the resulting cDNA fragments are then sequenced and mapped to the genome. As each mRNA molecule codes for a gene, the fragments will map to the positions of the corresponding gene. Generally, the genes that are highly expressed will have more reads mapped to it. However, larger genes will have larger mRNAs, which again will produce more reads. Correcting this bias by normalization is a crucial step before analyzing the data (Li et al., 2015). Created with BioRender.

enhancers/silencers), potential overlaps will represent a robust set of TFs located in promoter regions, that possibly regulate the most highly expressed genes.

1.2.2 The open chromatin landscape and TFs

Open chromatin has previously been captured by sequencing techniques such as DNase-seq (Song & Crawford, 2010), FAIRE-seq (Davie et al., 2015) and MNase-seq (Schones et al., 2008) (Figure 1.7), in order to gain information about transcriptionally active regions and the TFs that bind there. However, these methods require lots of cells and are expensive and time consuming (Buenrostro et al., 2015; Tsompana & Buck, 2014). ChIP-Seq (Landt et al., 2012) is another technique that has gained popularity in search of TFBSs, and uses antibodies to extract DNA bound to TFs of interest. Although ChIP-Seq has proven to be a successful technique with high resolution, it requires antibodies to extract specific TFs, making it time consuming and expensive (Park, 2009; Buenrostro et al., 2015).

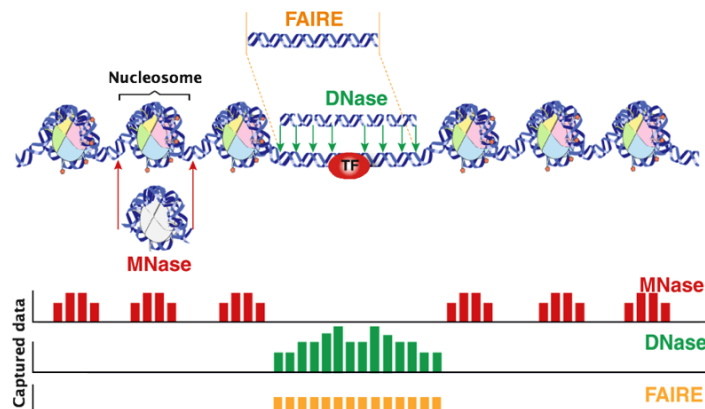


Figure 1.7: Different methods for capturing chromatin accessibility. MNase-seq finds accessible DNA indirectly, by probing closed regions (nucleosomic DNA). DNase-seq and FAIRE-seq capture accessible DNA directly. Adapted from Tsompana & Buck (2014).

A relatively new technique that has gained popularity lately due to low cost, low cell requirement and high speed is Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2015; Bajic et al., 2018). This technique has high sensitivity, and has been successfully used to find TFBSs in various studies (Davie et al., 2015; Corces et al., 2018). Also, the low cell requirement has also made it popular in single cell studies (Yan et al., 2020; Erbe et al., 2020).

ATAC-seq uses a protein called Tn5 transposase to extract DNA in accessible

chromatin regions, and sequences the DNA using next generation sequencing (Buenrostro et al., 2015). Tn5 proteins will only bind to regions that are loosely packed, for the simple reason that there is no physical space for them to bind to the DNA hidden within the tightly packed closed chromatin (Sun et al., 2019). After sequencing the accessible regions captured by Tn5, these regions will show up as peaks, and can be used to capture the accessibility landscape. More details on the technique is described in Figure 1.8.

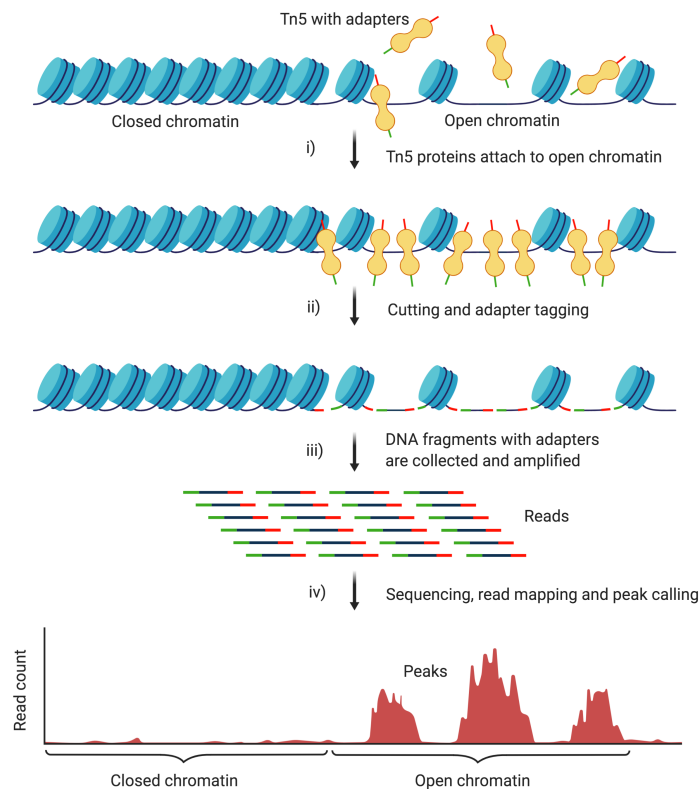


Figure 1.8: The hyperactive Tn5 proteins attach to accessible regions in the chromatin (i), where they cut the DNA and insert adapter sequences to the fragment ends (ii). The tagged DNA is then amplified by PCR and prepared for sequencing (iii). After preparations, the reads are sequenced using next generation sequencing. The resulting sequences are mapped to the genome, giving the coverage for each position. When plotted, this will reveal peaks, meaning regions in the genome with more overlapping reads than the background. Peak calling is applied in order to separate peaks that arise from truly open regions from background noise. Larger peaks correspond to more open regions. Created with BioRender.

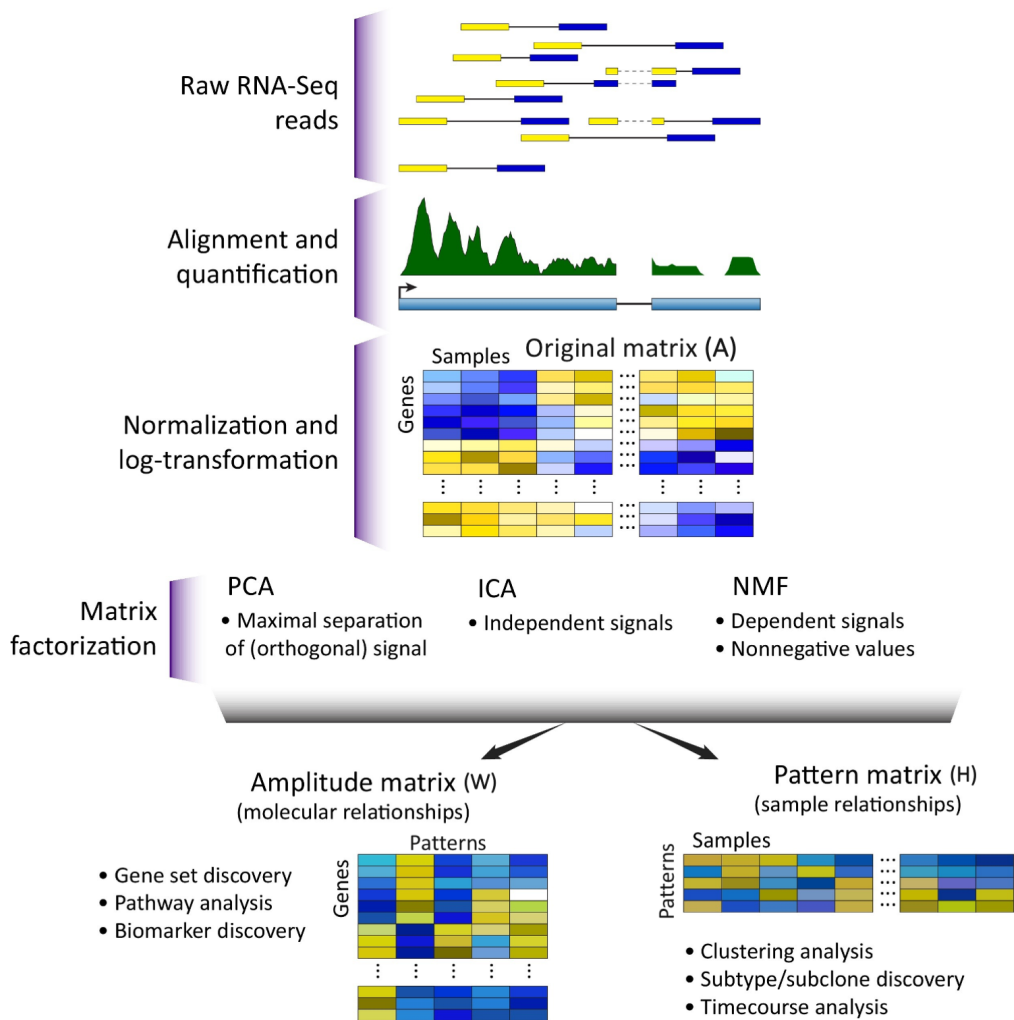
1.2.3 Interpreting big data by dimensionality reduction

High-throughput based sequencing techniques such as RNA-seq and ATAC-seq generate a vast amount of data, and in order to gain useful insight into the characteristics of each subtype, it is normally necessary to do some type of dimension reduction. Matrix factorization (MF) is a popular, unsupervised way of reducing dimensions (Gaujoux & Seoighe, 2010), and simultaneously cluster samples based on common features.

The most popular MF methods include Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Non-negative Matrix Factorization (NMF) (Stein-O'Brien et al., 2018). Common for all of these methods, is that they factorize one big matrix A into two smaller matrices: the amplitude matrix (W), which explains the relations between features (e.g. genes or peaks of open chromatin), and the pattern matrix (H), which explains the relations between samples (Stein-O'Brien et al., 2018). Each of these methods try to reduce the dimensions by using a low number of components, called patterns, while still preserving the original information (Figure 1.9).

In PCA, the patterns are called principal components and aim to maximize variation in the data. The first component describes most of the variation, and each component has to be orthogonal to the previous one. As a consequence, the first two or three principal components usually describe most of the variation, and meaningful patterns may be mixed (Stein-O'Brien et al., 2018). ICA and NMF are better suited for clustering, as all patterns have relatively equal variation and capture "co-variation" that is particular for different clusters of samples. Both require the rank (total number of patterns) to be chosen beforehand, and this remains one of the biggest challenges with these technique, especially if no a priori knowledge about classification exists. Both ICA and NMF have been used successfully to derive biologically meaningful gene signatures (Teschendorff et al., 2007; Brunet et al., 2004), but the non-negativity constraint of NMF makes the interpretation of the patterns much more intuitive than for ICA (Stein-O'Brien et al., 2018). NMF also has no restrictions on orthogonality or independence - each feature can contribute to multiple patterns, to a different degree. This dependency is valuable when explaining complex, biological data, as it allows the features (e.g. genes) to contribute to multiple patterns. These patterns may represent pathways or co-variation that is important for certain groups (Gaujoux & Seoighe, 2010).

Other MF methods, such as Multi-Omics Factor Analysis (MOFA) (Argelaguet et al., 2018), enables integration of data from multiple "omics" for the same set of



Trends in Genetics

Figure 1.9: Matrix factorization methods factorize the observed count matrix (A) into two smaller matrices: The amplitude matrix (W) and the pattern matrix (H). In NMF, both the original matrix A and the factorized matrices W and H contain non-negative values only. This is not a requirement for PCA and ICA (Stein-O'Brien et al., 2018). W and H can be used for multiple purposes, including gene/peak set discovery and clustering of subtypes, respectively. Here, we see part of the RNA-seq process and how the resulting data can be used in MF. The number of patterns equals the number of dimensions ("rank"). Adapted from Stein-O'Brien et al. (2018).

samples. Omics is merely an abbreviation for biological fields ending with -omics, such as transcriptomics (including RNA-seq) or epigenomics (including ATAC-seq) (Vailati-Riboni et al., 2017). MOFA works in a similar way as PCA, but instead of factorizing one big matrix into two smaller matrices, MOFA factorizes multiple matrices from multiple omics into one pattern matrix and multiple amplitude matrices - each describing the features from a omics (Figure 1.10). In that way, it clusters samples based on the combined signal from multiple data types. The advantage of this technique is that it can be used to highlight biological processes that are affected on multiple "omics levels", presuming that the molecular patterns of the omics data are highly connected (Vailati-Riboni et al., 2017). However, unlike NMF, MOFA has no non-negativity constraint, which can complicate the interpretation of the analysis.

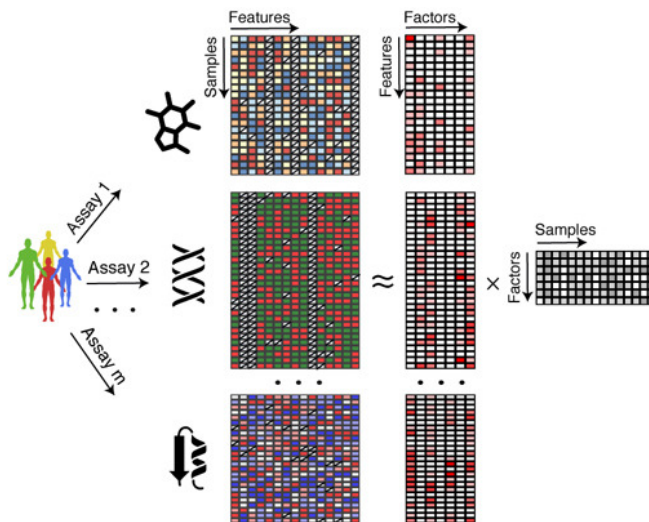


Figure 1.10: MOFA can handle input of multiple omics data sets for the same group of patients. The omics data can be count-based, continuous or binary. Adapted from (Vailati-Riboni et al., 2017).

NMF has previously been used for text mining, image processing and sound source separation (Shahnaz et al., 2006; Gillis, 2014; Virtanen, 2007)). In later years it has gained popularity in the field of bioinformatics, especially for positive, count-based data such as RNA-seq data and ATAC-seq data (Devarajan, 2008; Stein-O'Brien et al., 2018; Erbe et al., 2020).

The goal of NMF is to reduce dimensions, while still getting as close as possible to representing the original matrix. This is performed through the factorization of the original matrix A into W and H so that:

$$A \approx WH \text{ (Lee \& Seung, 2001)}$$

However, finding the perfect combination of W and H , where the difference to the original matrix is at a global minimum, is simply too time consuming - especially for larger data sets. Instead, the algorithm iterates until a local minimum is reached (Lee & Seung, 2001). Because the algorithm settles for a local minimum, there will always be some differences between the original matrix and the one recreated by combining W and H , especially when running on larger data sets.

In the initiation of NMF, random weights are assigned to H and W . These are adjusted after every iteration, when the algorithm measures how well the weights preserved the observed data in A . For each element of A , the corresponding vector of W and H must be multiplied together using a linear combination of the patterns. The values of W and H are then adjusted up or down, depending on whether the product attained is higher or lower than the element of the original matrix. In order to better understand how this works, an example is presented in Figure 1.11. In the example, the number of values have been reduced from 20 to 18, which is a small reduction. However, if we had a data set with 50,000 rows and 150 columns, the number of values would be reduced from 7,500,000 to 100,300 if we set the rank to 2. In other words, NMF effectively reduces dimensions in larger data sets.

In addition to dimension reduction, NMF can be used directly for clustering samples and features through the patterns it creates after factorization. This is called biclustering (Kim & Park, 2007). For the pattern matrix (H), these patterns will represent the different co-variation of the features, and cluster assignment can be done based on which pattern contributes most to each sample. An example of cluster assignment is shown in Figure 1.12. The amplitude matrix (W) describes the contribution of each feature to the different patterns, and features that contribute strongly to the same pattern can be clustered together. The pattern matrix and the amplitude matrix are strongly connected: The features that cluster to Pattern 1 are the features that are most important in the samples that cluster to Pattern 1. By using this simple, inherent clustering, it is possible to get a direct link between a cluster of features and a cluster of samples, which is very useful in the search for features that characterize a group of samples. For use in analyses that involve prior knowledge of subtypes (e.g. PAM50), each cluster should preferably represent one subtype. In that way, when some characteristics are learned about a pattern or a cluster, this knowledge can be directly transferred to a subtype. Thus, it is possible to

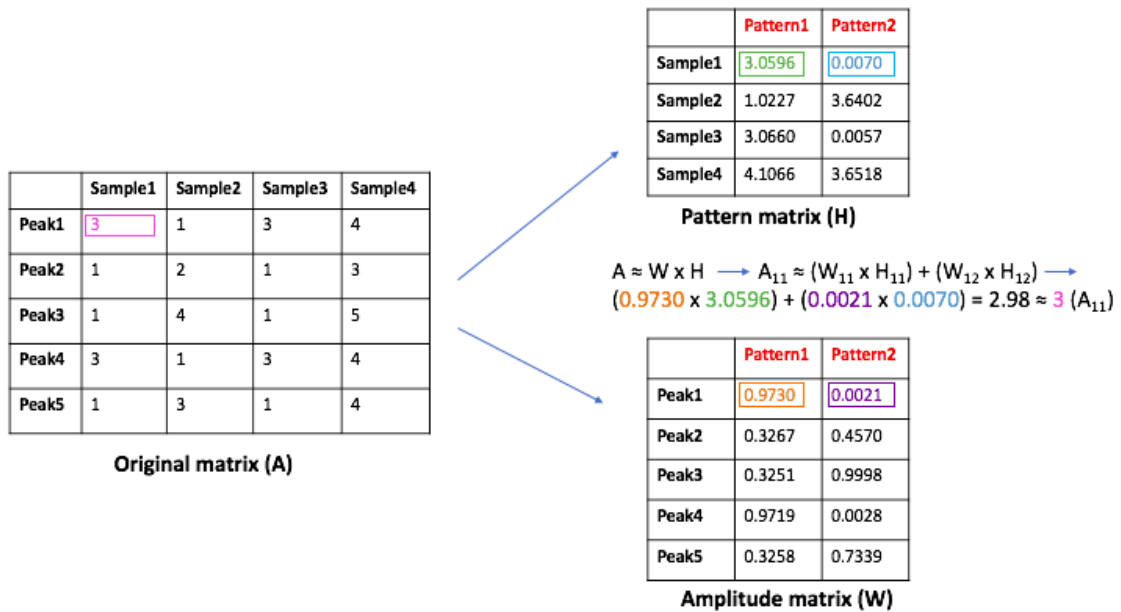


Figure 1.11: As an example, we create a 4x5 count matrix (A) with peaks in rows and samples in columns. In order to estimate the first value for A after factorization, we combine the first row in W and H. First, we multiply the element for [Pattern 1, Peak 1] in W with the element for [Pattern 1, Sample 1] in H, and repeat the procedure for Pattern 2. The two values are then added together. This gives the value 2.98, which means that the algorithm came close to representing the real value 3. If the local minimum had not yet been reached, NMF would do more iterations and adjust the colored weights slightly up in order to get closer to 3.

combine unsupervised and supervised learning to learn important features of pre-defined subtypes, without making assumptions about an expected outcome. When NMF is used for data types with features that can be connected to genomic regions, these features can be searched for enriched TFBSs in order to find out which TFs are active in a set of regions.

	Pattern1	Pattern2
Sample1	3.0596	0.0070
Sample2	1.0227	3.6402
Sample3	3.0660	0.0057
Sample4	4.1066	3.6518

Pattern matrix (H)

Figure 1.12: An example of cluster assignment using the pattern matrix. Sample 1, 3 and 4 will be assigned to the Pattern 1 cluster, while Sample 2 will be assigned to the Pattern 2 cluster. Sample 4 has high values for both Pattern 1 and Pattern 2, which means that this sample shares similar features with both groups.

1.2.4 Predict transcription factor drivers from regulatory regions

There are multiple tools available for finding enriched TFBSs in genomic regions, including MEME Suite (Bailey et al., 2009), Enrichr (Kuleshov et al., 2016), UniBind Enrichment Analysis (UniBind, 2020) and HOMER (Heinz et al., 2010). Some tools, like HOMER, scan the genomic regions for sequences that match a set of pre-defined motifs for regulatory regions, such as TFBSs. Since most TFs can bind to multiple, similar sequences, these motifs are variable sequences. The motifs can be represented as position weight matrices, based on alignment of known binding sites in the cell type of question (Wasserman & Sandelin, 2004; EMBL-EBI, 2020; Ren et al., 2016). PWMs are often visualized as sequence logos, where the size of the letters indicate their relative frequency (Figure 1.13). The binding site of each TF can also have multiple motifs, if computed from different sets of TFBSs, from different cell types or conditions. Other tools, like UniBind (UniBind, 2020), use a combination of motifs and known genomic positions of TFBSs. These known TFBS regions can be intersected with the regions provided as input. The TFBSs in the UniBind database have been located by ChIP-Seq experiments, which is an advantage - it has been experimentally shown that a certain TF binds in certain regions, and it therefore avoids only relying on motifs

that might be similar for different TFs within the same family.

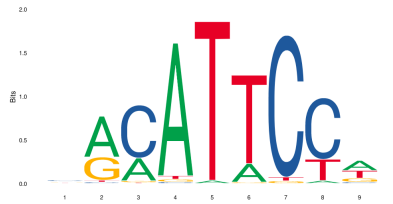


Figure 1.13: An example of a motif, showing the binding site of the TEAD4 transcription factor. Retrieved from the JASPAR database (Fornes et al., 2020).

All these techniques have previously been used in various combinations to explore gene regulation in breast cancer. For example, RNA-seq has been used together with techniques like ChIP-Seq in order to infer gene regulatory networks in cancer (Angelini & Costa, 2014). However, there is still a lot to explore about the unique characteristics of each breast cancer subtype. Combining relatively recent techniques in the field, like ATAC-seq, RNA-seq and NMF, can uncover new information and help us get a better understanding of gene regulation in different subtypes of breast cancer. Finding the TFs that drive each subtype can provide potential targets in the search for new treatments.

Chapter 2

Aim of thesis

The main goal of my thesis is to find out which transcription factors drive the different subtypes of breast cancer, and possibly associate these with subtype-specific, highly expressed genes. This is done in order to understand the molecular mechanisms behind the different gene regulatory profiles, and provide research that can be used for potential treatments.

To achieve this goal, three subgoals were formed:

- 1) Use NMF on RNA-seq and ATAC-seq data from the same patients to derive subtype-specific gene and peak signatures
- 2) Search for enriched transcription factor binding sites within the regions of the most subtype-specific features
- 3) Explore information gained by combining RNA-seq and ATAC-seq data in a multi-omics experiment

An outline of the methods and data flow for subgoal 1) and 2) is showed in Figure 2.1.

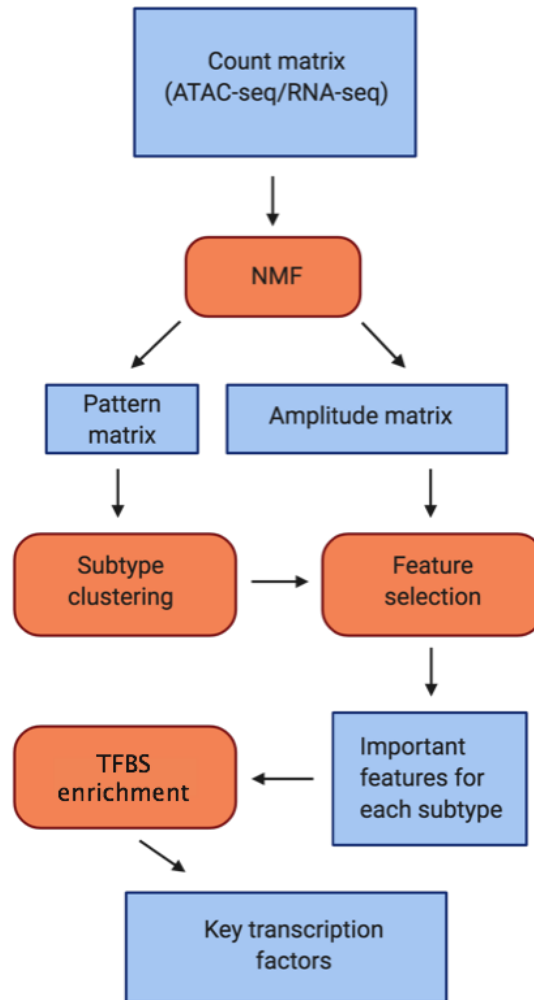


Figure 2.1: Overview of methods and data flow. The initial input is a normalized count matrix, which is factorized into two smaller matrices through NMF. The pattern matrix is used for clustering, and top features are selected from the amplitude matrix for each cluster. The top features are then searched for enriched transcription factor binding sites in each cluster. Blue/sharp edge describes data, while red/rounded edge describes method. Created with BioRender.

Chapter 3

Materials and methods

The materials and methods in this chapter were used to achieve the goal of uncovering the key transcription factors in different breast cancer subtypes. All data analysis in R was conducted with RStudio version 3.6.0 (RStudio Team, 2015), and all plots were created with ggplot2 (Wickham, 2016) unless specified otherwise.

3.1 Data

ATAC-seq and RNA-seq data from a matching cohort of 70 breast cancer patients was analyzed in order to explore characteristics of different subtypes at a molecular level.

ATAC-seq data

The ATAC-seq data used in this thesis is based on BRCA-US data produced by the cancer genome atlas (TCGA) (Weinstein et al., 2013) and has been preprocessed and normalized by Corces et al. (2018). The normalized count matrix was retrieved from Supplemental Data Files and available through the GDC database (National Cancer Institute, 2020). Corces et al. (2018) had defined a subset of peaks that were specific for breast cancer, and also used a set width for all peaks (501 bp) in order to decrease bias in motif analyses, which was an advantage for the purpose of this master thesis. ATAC-seq data was available for 74 donors with one sample and one or two technical replicates per sample, resulting in a total of 141 samples/replicates. The data was subsetted further to only keep samples with corresponding RNA-seq data. After subsetting, 70 unique samples remained. When including technical replicates for these samples, the numbers added up to 134 samples (samples + replicates). Log2 transformation from former normalization had to be removed before further processing, as the

data contained negative values. After preprocessing, the result was a matrix with 49,748 peaks (rows) and 134 samples (columns).

RNA-seq data

In order to create a corresponding RNA-seq count matrix, normalized mRNA-seq data for the BRCA-US project was downloaded from the International Cancer Genome Consortium (ICGC) data portal (International Cancer Genome Consortium (ICGC), 2020). The data for the BRCA-US project was generated by TCGA (Weinstein et al., 2013). Before the data could be used for downstream analysis, it needed to go through multiple preprocessing steps. First, the donor information was used to match the ICGC ID with the ICGC ID of donors with available ATAC-seq data. The 70 common IDs were then converted to donor (DO) IDs, and the file containing the donor IDs was used to search the RNA-seq file using the Unix command *grep*. This resulted in a subset that only contained data for donors that also had available ATAC-seq data. The genes with unknown gene ID were filtered out, along with the SLC35E2 gene. SLC35E2 appeared twice in the data for all samples, with highly variable expression within the same samples. There was no way to determine if one of them was due to a sequencing error, and if so, which. As a result, both duplicates of the gene were removed. The data was then reshaped into a matrix with 20,500 genes and 85 samples from the 70 common donors. A last column filtering was done to only keep samples that also had ATAC-seq data available, and this resulted in a matrix with 20,500 genes and 72 samples (columns), where 70 of the samples were unique and 2 were biological replicates of one of the samples. The biological replicates were kept to validate the clustering, as they should preferably cluster together. Last, the genes that contained 0 for all samples or all samples but one, were filtered out. These did not contribute to separating the data, and would cause trouble for the next steps. The final count matrix contained 19,766 genes (rows) and 72 samples (columns).

Metadata

The metadata containing ER status and PAM50 subtypes were retrieved for all samples. The PAM50 subtypes were extracted using the `TCGA_MolecularSubtype` function from the `TCGAbiolinks` R package, while information about ER status was available as supplement through Corces et al. (2018). All samples had subtype information, except for the two biological replicates of a Basal-like sample in the RNA-seq data. These were imputed as Basal-like, as they came from the

same sample and were expected to be the same subtype.

3.2 NMF analysis

The NMF analysis section describes the methods and data flow shown in Figure 2.1.

3.2.1 Non-negative matrix factorization (NMF)

NMF was performed on the ATAC-seq and RNA-seq data in order to reduce dimensionality. There were multiple tools available for NMF, and for R there were two packages commonly used for biological data: NMF (Gaujoux & Seoighe, 2010) and CoGAPS (Fertig et al., 2010). CoGAPS uses Bayesian inference to reduce local optima, and has been shown to be successful for biological count data such as single-cell ATAC-seq, single-cell RNA-seq and microarrays (Erbe et al., 2020; Fertig et al., 2010). The NMF package uses multiplicative update rules from Lee & Seung (2001), which makes it more prone to uncertainty, but also speeds up the algorithm considerably (Sherman et al., 2019). Both packages were tried for this thesis, but the NMF package was chosen based on the big difference in computational speed combined with indistinguishable differences in result.

All runs of NMF were performed with the `nmf` function, with default parameters. The default algorithm was 'brunet', based on Kullback-Leibler divergence from Brunet et al. (2004). The choice of rank was guided by calculating the cophenetic correlation coefficient, which is a measure of cluster stability (Brunet et al., 2004). The cophenetic correlation coefficient was calculated using the `nmfEstimateRank` function from the NMF package, with the number of runs per rank set to 3 and otherwise default parameters, corresponding to the parameters of the `nmf` function. A common seed was used for all runs. Because the estimation was time consuming for large data sets, the cophenetic correlation coefficient was compared for a limited number of ranks. To guide this selection, we used four different ranks that would be likely to form meaningful clusters in our data: 2, 3, 4 and 5. These ranks were chosen because the subtype information available contained two different ER based subtypes (ER+/ER-) and five different PAM50 based subtypes (Luminal A, Luminal B, Normal-like, Her2 and Basal-like).

NMF was then performed using the `nmf` function on the ATAC-seq and RNA-seq data, with ranks ranging from 2 to 5. Multiple runs were performed in order to see how the choice of rank affected the concordance with prior subtypes. The dimensions of the pattern matrices were then reduced further, using the `umap` function from the `uwot` R package. This was done in order to project

multiple ranks in a two-dimensional space (McInnes et al., 2018). Then, the new pattern matrices were plotted with colors corresponding to subtype, and shape corresponding to cluster. The cluster assignment was made on the basis of which pattern was the strongest in each sample, as described in Figure 1.12 in the Introduction. UMAP was only used for visualization purposes, and had no impact on the clustering.

3.2.2 Feature selection

A subset of genes and peaks that contributed most to each pattern had to be defined, in a process called feature selection. The first step of feature selection was feature scoring.

Feature scoring

The features of the amplitude matrices were scored using the `featureScore` function from the NMF package with method 'Kim' (Kim & Park, 2007). The Kim method scores each feature based on how pattern-specific it is, but also how important it is. If the feature (peak or gene) contributes almost solely to one pattern, it will receive a higher score than features that contribute evenly to multiple patterns. Higher value for a pattern-specific feature means higher importance, and also contributes to the score. A feature score is calculated for each pattern, and the highest score is kept for each feature (See example in Figure 3.1).

	Pattern1	Pattern2	Pattern3	Pattern4	
Peak1	11.48	12.71	17.61	50.00	→ Score = 0.14
Peak2	261.37	243.45	272.27	3.05	→ Score = 0.19
Peak3	10.58	7.63	3.79	203.39	→ Score = 0.70

Amplitude matrix (W)

Figure 3.1: An example of feature scoring, shown with the highest feature score for each peak. The first peak contributes more to Pattern 4 than the other patterns, but receives a low feature score due to lack of importance. The second peak are important for the first three patterns, but receives a low score due to lack of pattern-specificity. The third peak receives a high score, due to a combination of high pattern-specificity and importance.

The Kim scores always range between 0 and 1, and can be calculated as follows:

$$\text{Gene_score}(i) = 1 + \frac{1}{\log_2(k)} \sum_{r=1}^k p(i, r) \log_2(p(i, r)),$$

where k is the total number of patterns (rank) and $p(i,r)$ is the probability that the i -th feature contributes to a pattern r . $p(i,r)$ for each pattern Ω is calculated as follows:

$$p(i,\Omega) = \frac{W(i,\Omega)}{\sum_{r=1}^k W(i,r)},$$

where $W(i,\Omega)$ is the weight of the i -th feature for a given pattern, and $\sum_{r=1}^k W(i,r)$ is the sum of the weights for all patterns of the i -th feature.

Feature extraction

The `featureScore` function is built into the `extractFeatures` function, which was also available from the NMF package. Here, the scores are calculated and extracted in one step. The Kim method uses the scores to choose the most pattern-specific features based on a certain threshold. In order to pass this threshold, scores need to be greater than:

$$\hat{\mu} + 3\hat{\sigma},$$

where $\hat{\mu}$ is the median of the scores, and $\hat{\sigma}$ is the median absolute deviation of the scores (Kim & Park, 2007). In addition, the maximum values in the corresponding rows of the amplitude matrix (W) has to be larger than the median value of all weights in W .

The `extractFeatures` function was performed on the amplitude matrices from both data sets, with the Kim method for feature scoring and extraction, and the rest of the parameters set to default. This resulted in one list for each data set, with list elements corresponding to the indices of top contributing features for each of the patterns. In addition to the Kim feature selection, a new feature selection method was created and tried on the data, extracting the n most pattern-specific peaks based on pattern-specificity and importance. In addition, it took into account the score distribution of other peaks for each pattern before selection. However, the method yielded similar results as the Kim feature selection, which so far is a more reliable method. Therefore, only the Kim method was used for further processing.

The pattern-specific features (peaks and genes) obtained from the ATAC-seq and RNA-seq data using Kim were our defined gene/peak signatures for each pattern.

3.2.3 Investigating genomic regions

In order to investigate whether the gene signatures were likely to be associated with the peak signature, the chromosomal regions where the peaks and genes reside were investigated. The percentage of peaks and genes coming from each chromosome was plotted, in order to see if the transcriptionally active regions from the ATAC-seq data resided in the same chromosomes as the highly transcribed genes. First, the location of the genes were extracted. All gene locations were downloaded in bed format from the UCSC Table Browser (UCSC, 2020). Next, the table was read into R, and each of the five gene signatures were matched by gene names in order to extract the genomic regions. The percentage of genes residing in each chromosome was then plotted into pie charts, using the `pie` function from the base R package. The regions of the peaks were already available through (Corces et al., 2018), and pie charts showing chromosome distribution were created for the peaks as well. The UCSC genome browser (Kent et al., 2002) was then used to investigate specific regions, for subtypes that were dominated by activity on a specific chromosome.

3.2.4 Gene ontology enrichment analysis

The gene signatures could also be used to perform a gene ontology enrichment analysis, in order to find enriched processes and functions. This was tried using Gene Ontology enRiChment anaLysis and visuaLizAtion tool (GORilla) (Eden et al., 2009). However, no significant hits were found, except for one cluster. The search was therefore extended to include a wide range of ontology and gene signature databases. This was done using HOMER `findMotifs.pl`, while simultaneously searching for enriched TFBS motifs in the promoter regions. The parameters are described in the next part.

3.2.5 Transcription factor binding site enrichment

Using the top features for each cluster in both data sets, we wanted to see if there was an enrichment of any particular TFBSs located in the top peaks of the ATAC-seq data, and in the promoter regions of the gene sets derived from the RNA-seq data. The first TFBS enrichment analysis was performed with HOMER, on the regions defined as informative features for each pattern.

HOMER

HOMER had two tools available for motif enrichment: `findMotifsGenome.pl` for search in genomic regions, and `findMotifs.pl` for search in promoters of genes in a gene list (Heinz et al., 2010). HOMER returns motifs that are enriched with $p < 0.05$, compared to the background.

First, we tried to uncover the TFBSs hidden within the peaks from the ATAC-seq data. `findMotifsGenome.pl` was performed with the genomic regions of the peaks and the hg38 genome as input. The size parameter was set to 200, which meant that 200 bp on each side of the peak center were searched for motifs. The background was generated automatically by HOMER, using similar GC content as in the input peaks. In addition, we used the `-mask` option in order to reduce bias from repeated regions in the genome. The rest of the parameters were set to default.

Second, `findMotifs.pl` was used to search for TFBS motifs in the promoters of the pattern-specific genes derived from the RNA-seq data. The only mandatory parameter, except for a gene list, was the organism the promoter regions would be extracted from, which was set to 'human'. Otherwise, the tool was run with default parameters.

UniBind

The second TFBS enrichment analysis was done using UniBind Enrichment Analysis (UniBind, 2020; Gheorghe et al., 2019). UniBind finds overlaps between sets of regions and sets of TFBSs, and the enrichment is calculated using the LOLA R package (Sheffield & Bock, 2016). TFBS set enrichment was performed on the ATAC-seq peaks with two different settings: 1) Background consisting of all top peaks for all the different patterns and 2) No provided background. All peaks were kept at their original length (501 bp) for this TFBS enrichment analysis.

3.3 Multi-omics analysis

The ATAC-seq and RNA-seq data was combined in a multi-omics analysis, using a matrix factorization tool called MOFA (Argelaguet et al., 2018). This was done in order to compare with the NMF analysis, where each data set was analyzed separately.

3.3.1 Data preprocessing and normalization

Technical/biological replicates were removed from each data set, so that each ATAC-seq sample had exactly one matching RNA-seq sample, and so on. This left 70 samples, one per patient.

Two normalization methods were tried on both data sets. The first method was to calculate counts per million using the `cpm` function from the `edgeR` R package (Robinson et al., 2010). A pseudocount of 5 was added to the data, before taking the `log2`. For the second method, the `estimateSizeFactors` function from the `DESeq2` R package (Love et al., 2014) was used to estimate and normalize for size factors, before adding a pseudocount of 5 and taking `log2` of the values. The `poscount` method was used for estimating size factors, because it calculates a modified geometric mean that is better suited for handling multiple rows with zero counts than the default median ratio. The two methods were then compared. Multi-omics analysis was performed with the `MOFA2` R package (Argelaguet et al., 2018). Similarly to NMF and ICA, it required the choosing of a rank beforehand. Different ranks were tried in order to best describe the data. The normalized data was used to create a MOFA model, which contained a pattern matrix and two amplitude matrices: One for the ATAC-seq peaks, and one for the RNA-seq genes. The pattern matrix was then used to cluster the samples according to the strongest factor (similar to pattern in NMF).

3.3.2 Feature selection and signature analyses

Feature selection was performed for each factor in each data type (ATAC-seq and RNA-seq), using `extractFeatures` with method `Kim`, the same feature selection method described in Figure 3.2.2. As `extractFeatures` was unable to handle the negative weights resulting from MOFA, a "pseudocount" of +1 was added to all values before selecting the most pattern-specific genes and peaks.

The resulting gene and peak signatures for each factor were used for two purposes: 1) Do a gene ontology enrichment analysis for enriched biological processes, by loading the gene list in `GORilla` (Eden et al., 2009), and 2) Find enriched TFBS sets from the peak signatures using `UniBind` Enrichment analysis (`UniBind`, 2020; Gheorghe et al., 2019). The latter was tried both with a background consisting of the top peaks, and with no background.

Chapter 4

Results

The results presented here show how the different methods were able to provide insight into the gene regulatory profiles of each breast cancer subtype. The different parts will fulfill the subgoals, which together will achieve the main goal: To find out which transcription factors drive the different subtypes of cancer.

4.1 NMF

4.1.1 Clustering of samples into a priori subtypes

In order to find a peak/gene signature for each subtype, NMF was performed on ATAC-seq and RNA-seq data for breast cancer patients. To achieve the best possible clustering of subtypes, the rank was chosen by combining unsupervised and supervised methods. For the purpose of this study, we would preferably opt for a rank that was close to the number of subtypes, so that each cluster could possibly represent one subtype. However, if the data naturally clustered into another number of groups, the clusters should not be forced to match prior subtypes that were based on other criteria. To investigate the most stable and inherent clustering, the cophenetic correlation coefficient was calculated as suggested by Brunet et al. (2004). The results of the cophenetic correlation calculations show that rank 5 achieves the most stable clustering for the ATAC-seq data, as the clusters vary less between each run than for the other ranks (Figure 4.1). The RNA-seq data has the highest stability of clusters for rank 4 and 5, which means that the data can naturally be divided into 4 or 5 clusters. In general, however, the clusters are more stable for the ATAC-seq data, as the cophenetic correlation coefficients are closer to 1 (Figure 4.1).

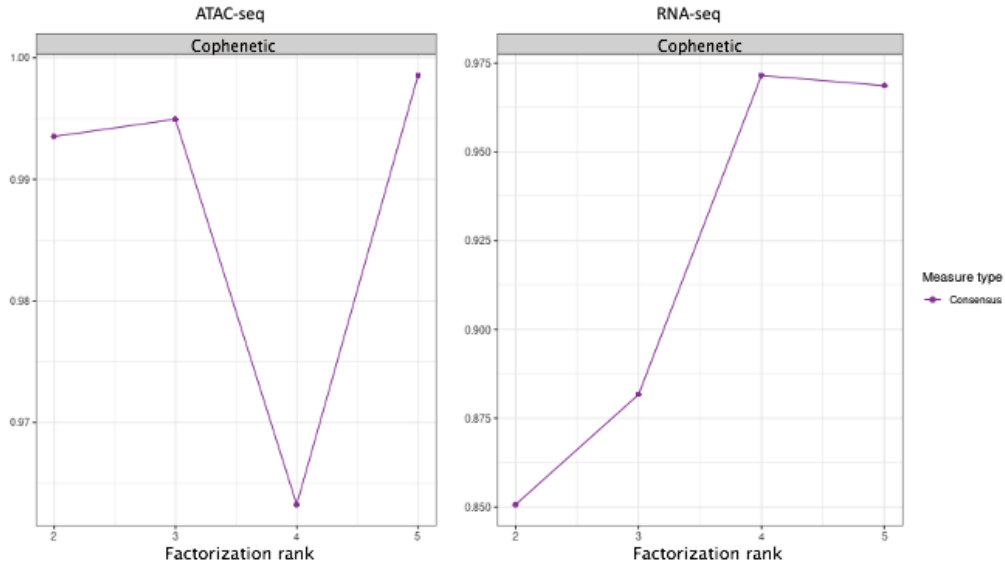


Figure 4.1: Cophenetic correlation coefficient of ATAC-seq data (left) and RNA-seq data (right). Each point in the graph is the result of 3 runs of NMF with "brunet" method, and represent cluster stability for each rank. The cophenetic correlation coefficient is always between 0 and 1, where 1 represents a perfect reproduction of clusters throughout the runs.

Afterwards, the samples were clustered through the pattern matrices from the different ranks, in order to show compatibility with a priori subtypes (PAM50/ER status). For the ATAC-seq data, the clustering of samples shows that the most accurate clusters (shapes) according to prior subtypes (color) is achieved with rank 5 and PAM50 subtypes (Figure 4.2). With rank 5, the Basal-like samples cluster alone (right), and so do most Her2 samples (top). Two of the clusters are dominated by Luminal A samples (middle/bottom), while a third (left) contains a more even mix of Luminal A and Luminal B samples. However, none of the ranks are able to truly separate the Luminal A, Luminal B and Normal-like subtypes. For the RNA-seq data, the clustering of the samples shows that most clusters are slightly less compatible with prior subtypes (Figure 4.3), compared with the ATAC-seq data. However, both rank 4 and rank 5 were able to separate most Basal-like and Her2 samples from other subtypes. These ranks also have the most stable clusters according to the cophenetic correlation coefficient. The best separation of Basal-like samples is achieved with rank 5. Here, only three samples with different subtype share their cluster, versus seven for rank 4 and thirteen for rank 3. As a result, the pattern matrix and amplitude matrix gained from running NMF with rank 5 were used for further processing. There were also two Basal-like samples that formed their own cluster. These were the samples

that were imputed as Basal-like, as they were biological replicates of a Basal-like sample. However, they appear to have a distinct pattern that separates them from other Basal-like samples.

There is an overall similarity between the distribution of subtypes in the clusters for both data sets. However, that does not mean that the exact samples necessarily cluster together between the data sets, which makes it harder to create a link between the clusters in some of the subtypes in the ATAC-seq and RNA-seq data. This is especially the case for the Luminal A samples, which are separated between multiple clusters. The list of samples and their cluster assignments is listed in Table S5 in Attachments.

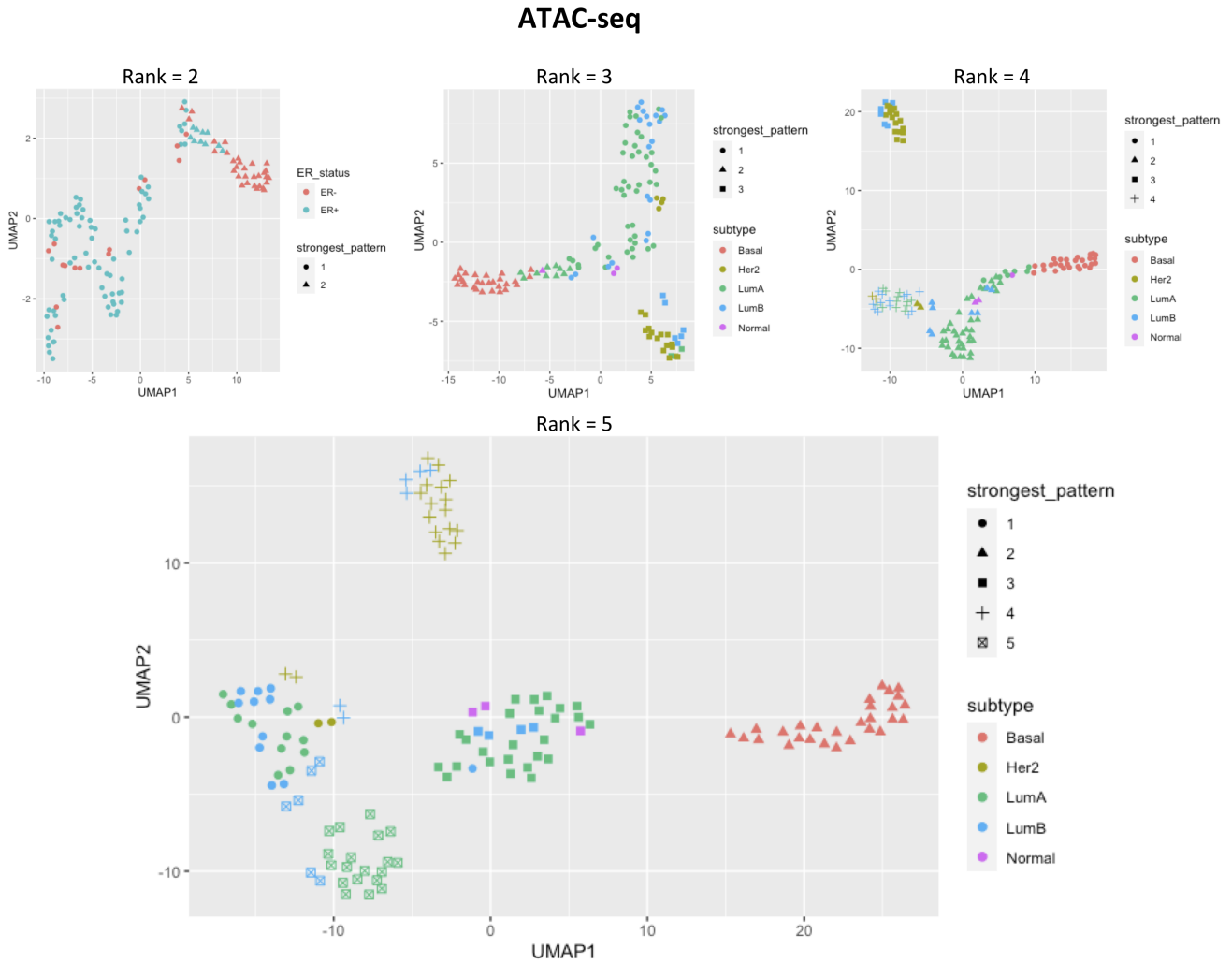


Figure 4.2: ATAC-seq pattern matrices for 4 runs of NMF. The dimensions have been further reduced using UMAP, and the axes correspond to each UMAP factor. Each plot contains 134 samples/technical replicates from 70 samples. The cluster assignment varies from each run of NMF, and their exact names (strongest pattern) are therefore not comparable between ranks. Subtype names have been abbreviated for the plot. Basal = Basal-like, Her2 = Her2, LumA = Luminal A, LumB = Luminal B and Normal = Normal-like.

RNA-seq

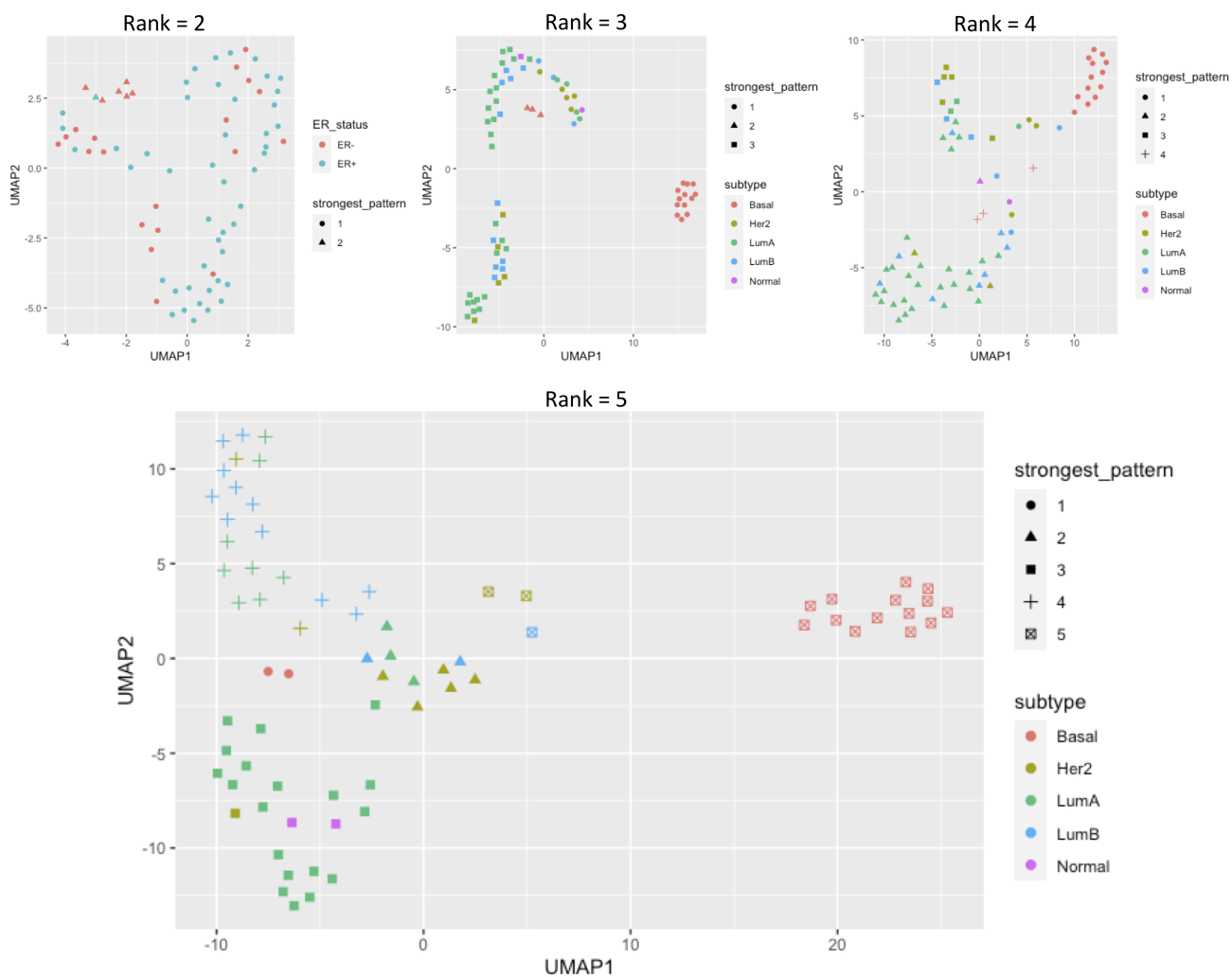


Figure 4.3: RNA-seq pattern matrices for 4 runs of NMF. The dimensions have been further reduced using UMAP, and the axes correspond to each UMAP factor. The plot contains 72 samples/biological replicates from 70 samples. The cluster assignment varies from each run of NMF, and their exact names (strongest pattern) are therefore not comparable between ranks. Subtype names have been abbreviated for the plot. Basal = Basal-like, Her2 = Her2, LumA = Luminal A, LumB = Luminal B and Normal = Normal-like.

NMF creates patterns in a random order, meaning that Pattern 1 in the ATAC-seq data does not necessarily correspond to Pattern 1 in the RNA-seq data, and so on. In order to keep track throughout the analysis, names that reflect the dominating subtype were created for each of the five clusters. The exception was for Luminal A, which dominated multiple clusters. Therefore, the cluster containing all the Normal-like samples and the majority of the Luminal A samples was named as a combination of these, for both data sets. The same was done for the cluster that contains most of the Luminal B samples, but also a large number of Luminal A samples in both data sets. The distribution of subtypes in each cluster and the given cluster name is shown in Table 4.1 and 4.2 for the ATAC-seq and RNA-seq data, respectively.

Table 4.1: Distribution of subtypes belonging to each cluster for the ATAC-seq data with rank 5. The values correspond to number of samples (labeled with a priori PAM50 subtypes) belonging to each cluster.

Cluster	Luminal B	Basal	Normal	Her2	LumA	Cluster name
Pattern 1	11	0	0	2	12	LumA/B
Pattern 2	0	28	0	0	0	Basal
Pattern 3	4	0	3	0	25	LumA/Normal
Pattern 4	6	0	0	18	0	Her2
Pattern 5	6	0	0	0	19	LumA

Table 4.2: Distribution of subtypes belonging to each cluster for the RNA-seq data with rank 5. The values correspond to number of samples (labeled with prior PAM50 subtypes) belonging to each cluster.

Cluster	Luminal B	Basal	Normal	Her2	LumA	Cluster name
Pattern 1	0	2	0	0	0	Basal outlier
Pattern 2	2	0	0	5	3	Her2mix
Pattern 3	0	0	2	1	19	LumA/Normal
Pattern 4	11	0	0	2	8	LumA/B
Pattern 5	1	14	0	2	0	Basal

4.1.2 Clustering of features reveals the activity of each pattern

After clustering the samples into patterns, we needed to cluster the features to find out which peaks and genes contributed most to each pattern. The heatmaps in Figure 4.4 show the amplitude matrix of the ATAC-seq data (a) and RNA-seq data (b), with the contribution of all features (peaks/genes) to the different patterns. The rows are scaled so that each row sums up to 1,

and ordered by which pattern the features contribute most to (the strongest pattern). Similarly to the clustering of samples, the features can be clustered by the strongest pattern. For the ATAC-seq data (Figure 4.4(a)), Pattern 2 has 14,160 regions contributing mostly to this pattern. We can roughly say that this pattern is defined by a large number of accessible regions that are mainly specific to this pattern. Overall, the clear, red squares indicate that there is little overlap between the clusters, meaning that a lot of the peaks are specific for one pattern. For the RNA-seq data (Figure 4.4(b)), Pattern 4 and 3 have the highest number of genes contributing to their clusters (7087 and 6530, respectively). However, their squares are less defined and overlap somewhat with each other, indicating that some of these genes are important for both patterns. Other patterns, like Pattern 1, has a small number of genes contributing to it. However, the red color of the square indicates that these genes are highly specific for this pattern.

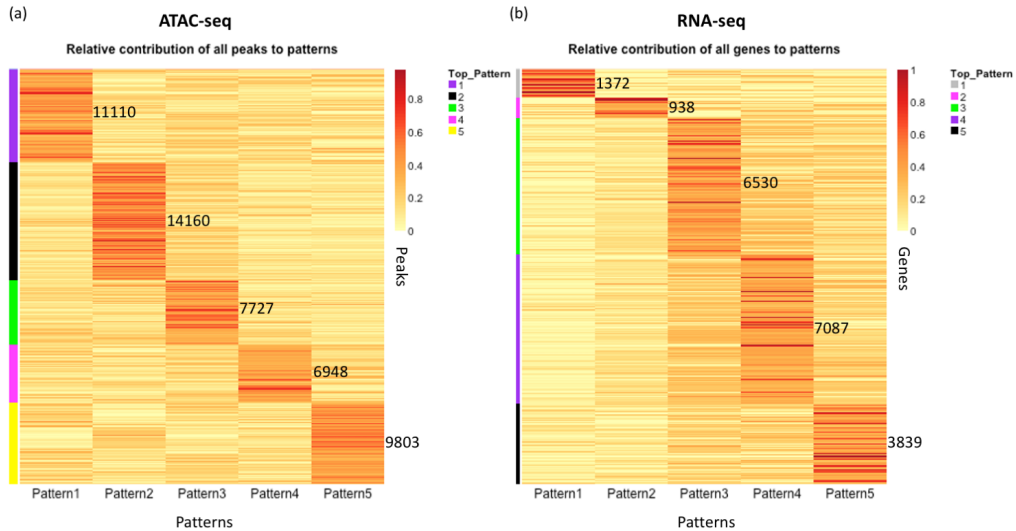


Figure 4.4: Heatmaps showing the contribution of features (rows) to patterns (column). The number of features contributing to each pattern is shown on the right side of each cluster of features. The rows are scaled so that each row sums up to 1, and ordered by which pattern the features contribute most to. The color bar on the left represents the cluster, which is assigned by the strongest pattern. (a) Heatmap of the ATAC-seq amplitude matrix showing the contribution of each of the 49,748 peaks to each pattern. (b) Heatmap of the RNA-seq amplitude matrix showing the contribution of each of the 19,766 genes to each pattern.

4.1.3 Connecting the samples with pattern-specific features

The feature selection resulted in a set of peaks and genes that were highly important and specific for each subtype. The plots in Figure 4.5 and 4.6 show heatmaps of the amplitude matrix containing the top features (a), and the pattern

matrix containing all samples (b). The rows in Figure 4.5(a) and 4.6(a) are scaled so that each row sums up to 1, and ordered by which pattern the features contribute most to (strongest pattern). The number of pattern-specific features has been added to the plots, to the right of each peak/gene cluster. By looking at both heatmaps for each of the data sets, we can see how much each pattern-specific peak/gene contributes to each pattern, and how strong those patterns are in the samples. For the ATAC-seq data in Figure 4.5, we can see that Pattern 2 contains the largest amount of pattern-specific peaks (904 peaks). By looking at the corresponding sample clusters, we see that Pattern 2 is strongest for Basal-like samples. In other words, Basal-like samples have a lot of regions that are very accessible in this subtype, but not in other subtypes. For the RNA-seq data, the number of highly expressed, pattern-specific genes seems to be more equally distributed between the patterns (Figure 4.6), however, Pattern 3 and Pattern 5 contain the largest amount of pattern-specific peaks (214 and 202 peaks, respectively). From the sample heatmap to the right, we can see that Pattern 3 corresponds to mostly Luminal A/Normal-like samples, while Pattern 5 corresponds to mostly Basal-like samples.

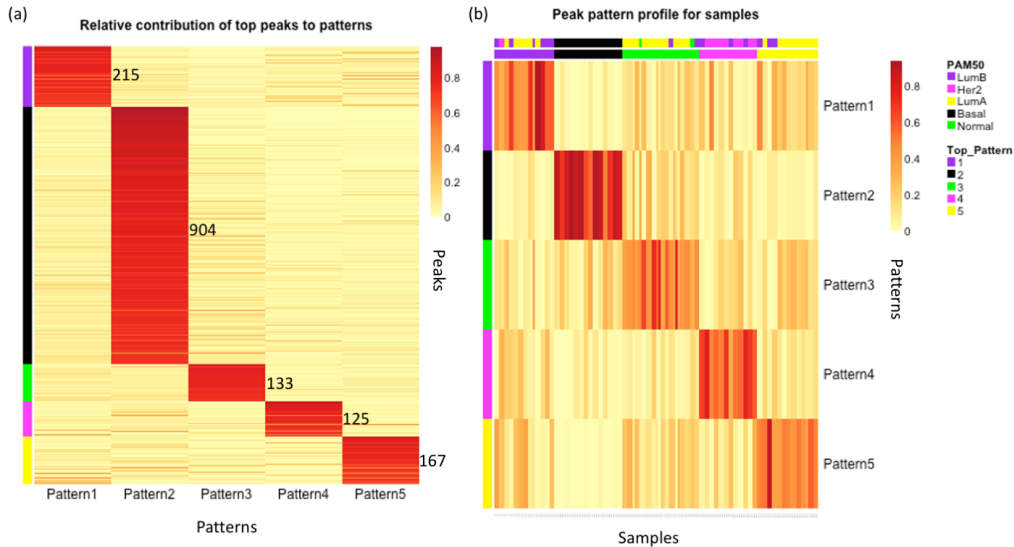


Figure 4.5: Heatmaps of ATAC-seq data after NMF. The color bars on the left (top pattern) are colored by the cluster names that represent each pattern. Purple = LumA/B (Pattern 1), black = Basal (Pattern 2), green = LumA/Normal (Pattern 3), pink = Her2 (Pattern 4) and Yellow = LumA (Pattern 5). (a) Contribution of peaks to each pattern. The peaks (rows) represent the 1544 most pattern-specific peaks out of 49,748 peaks. (b) Peak pattern profile of all 134 samples (samples + technical replicates), showing how strong each pattern is in each of the samples. The color bar on top shows the PAM50 subtype for each sample, while the lower bar matches the side bar and shows the strongest pattern.

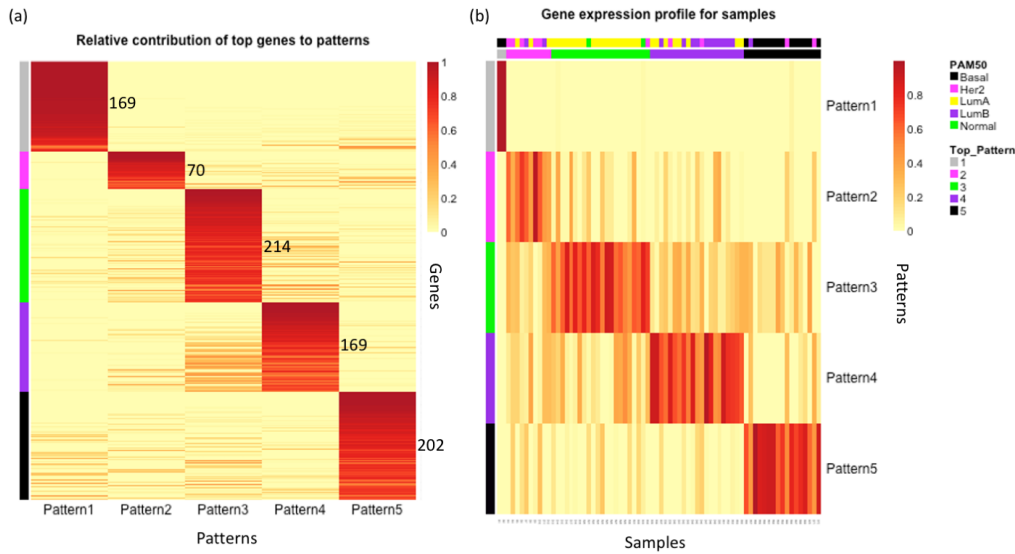


Figure 4.6: Heatmaps of RNA-seq data after NMF. The color bars on the left (top pattern) are colored by the cluster names that represent each pattern. Grey = Basal outlier (Pattern 1), pink = Her2mix (Pattern 2), green = LumA/Normal (Pattern 3), Purple = LumA/B (Pattern 4) and black = Basal (Pattern 5). (a) Contribution of genes to each pattern. The genes (rows) represent the 824 most pattern-specific genes out of 19,766 genes. (b) Gene expression profile of all 72 samples (samples + biological replicates), showing how strong each pattern is in each of the samples. The color bar on top shows the PAM50 subtype for each sample, while the lower bar matches the side bar, and shows the strongest pattern.

4.1.4 Gene signatures and gene set enrichment analysis

The set of pattern-specific genes that make up our gene signatures could be compared to known, subtype-specific genes directly. The top 5 most highly expressed genes for each cluster is shown in Table 4.3, together with some hits from further down the list. These were included due to previous findings, which will be discussed further in the Discussion. The full pattern-specific gene list is available in Attachments, in Table S1.

Table 4.3: The top 5 genes from the gene signatures, for each pattern. The cluster names are used as header, without the pattern number, in order to save space. Basal outliers = Pattern 1, Her2mix = Pattern 2, LumA/Normal = Pattern 3, LumA/B = Pattern 4 and Basal = Pattern 5.

Basal outliers	Her2mix	LumA/Normal	LumA/B	Basal
SNORD9	LACRT	CSN2	CPB1	CARD18
SNORA47	SULT1C3	C10orf96	CYP2A7	SERPINB3
SNORD67	UGT2B28	IRS4	UCN3	SPRR2D
SCARNA4	TBX10	ARHGAP36	CGA	MAGEA4
SNORD8	MUCL1	PROL1	CYP2A6	KRT79
-	-	-	SERPINA6	EGFR
-	-	-	BEX1	SOX8
-	-	-	AGTR1	SOX10

However, looking at each gene separately is time consuming, and the collection of a number of genes can provide more biological meaning. Therefore, the gene sets were used to perform gene set enrichment analyses. Due to a low number of pattern-specific genes per pattern, only one of the patterns were enriched for biological processes using GOrilla. This was Pattern 1 (the Basal outliers), which was most enriched for RNA metabolic process ($P = 6.34E-8$).

By widening the search to include gene signature databases, using HOMER, these genes could be used to check for enrichment in pre-existing gene signatures. The top enriched gene signature for the different subtypes can be seen in Table 4.4. Pattern 1 - Basal outlier describes two distinct Basal-like samples, and the genes are enriched for a gene signature related to the RNA polymerase I promoter opening pathway, which plays a role in the regulation and synthesis of rRNA (Reactome, 2020). This is similar to what was found in the GOrilla results for this pattern. Pattern 2 - Her2mix contains mostly Her2 samples, and our gene list for this pattern is enriched for genes previously found to define the Her2 subtype. The same goes for Pattern 3 - LumA/Normal, which is dominated by Luminal A, and is enriched for Luminal A-typical genes. Pattern 4 - LumA/B, which consists of slightly more Luminal B than Luminal A samples, is enriched for Luminal B-like genes. Similarly, Pattern 5 - Basal, which consists of Basal-like samples only, is enriched for a predefined Basal-like gene signature. All of these gene signatures are stored in the Molecular Signatures Database (MSigDB), which is part of a cooperation between UC San Diego and BROAD Institute (GSEA, 2020). The results for Pattern 2-5 show that our gene signatures share similarities

Table 4.4: The most enriched ontology term for the 5 different patterns defined in this thesis. The top result for each pattern was found in the MSigDB, and most corresponded to gene signatures previously defined for these subtypes. Cluster names are derived from Table 4.2. ERBB2, which is the name of the gene that encodes HER2, is an alternative naming of the Her2 subtype. The table is adapted from HOMER Gene Ontology Enrichment results.

Pattern	Cluster name	Term	P-value
Pattern 1	Basal outliers	REACTOME_RNA_POL_I_PROMOTER_OPENING	3.392e-43
Pattern 2	Her2mix	SMID_BREAST_CANCER_ERBB2_UP	9.745e-16
Pattern 3	LumA/Normal	SMID_BREAST_CANCER_LUMINAL_A_UP	2.053e-53
Pattern 4	LumA/B	SMID_BREAST_CANCER_LUMINAL_B_UP	2.687e-58
Pattern 5	Basal	SMID_BREAST_CANCER_BASAL_UP	2.226e-77

with the gene signatures found by Smid et al. (2008) using microarray data, for the subtypes that dominate our clusters.

4.1.5 Open regions and their chromosomal location

In order to get an overview of the active regions for each subtype, the chromosome regions were plotted for the pattern-specific peaks. This was also done for the most pattern-specific genes, in order to compare the regions (Figure 4.7). From the pie charts, we can see that some of the top peaks (left) are largely accessible at certain chromosomes, while the distribution is more even for the most highly expressed set of genes (right). When comparing the similar clusters in the ATAC-seq and RNA-seq data (row-wise), there seems to be little overlap between the chromosome distribution of the top peaks and the top genes. For example, the peaks from the LumA/B cluster of the ATAC-seq data are mostly located in chromosome 8, while this is not the case for the genes from the LumA/B cluster of the RNA-seq data, and so on.

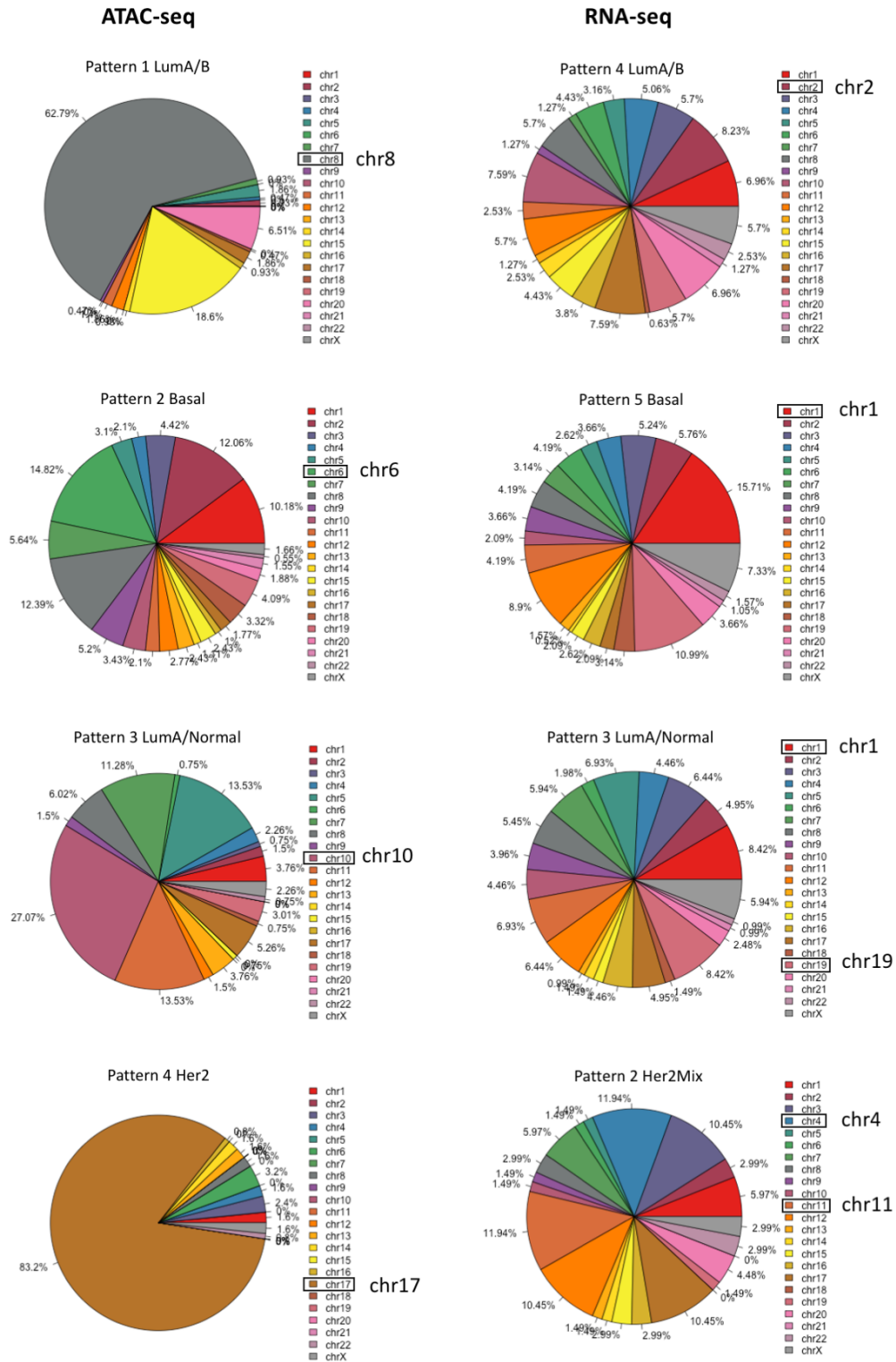


Figure 4.7: Pie charts showing the percentage of each chromosome for the active regions of the ATAC-seq data (left) and RNA-seq data (right), for each cluster. The most enriched chromosome is shown with a square, and in the cases where two chromosomes have the same enrichment, both are shown.

Due to the very high accessibility levels on chromosome 17 for the Her2 cluster, the regions of these pattern-specific peaks were investigated closer using the UCSC genome browser. The top peaks spanned two major regions, which were investigated separately. The first region contained the ERBB2 gene, which codes for the human epidermal growth factor 2 (HER2). The gene is situated on chromosome 17, from base 39,700,080 to 39,728,662. This region also contained miR-4728, a microRNA known to be encoded by ERBB2, located within the introns of the ERBB2 gene. Nine out of the 125 top peaks for the Her2 cluster were situated within the region of the ERBB2/HER2 gene (+/- 500 bp), and three of these peaks were on the top four most accessible peaks in the Her2 cluster, confirming that there is a lot of transcriptional activity related to the ERBB2 gene for the Her2 cluster, which contains 18 Her2 samples and 6 Luminal B samples.

The second active area was also present in chromosome 17, from base 39,461,486 to 39,534,544. The region contained the CDK12 gene, which codes for Cyclin-dependent kinase 12 (CDK12), a protein that regulates transcriptional and post-transcriptional processes (Lui et al., 2018). Eleven of the 125 top peaks were found within the region of this gene (+/-500 bp), and six of these were found in the top 25

4.1.6 Key transcription factors

In order to reveal the TFs that bind to regions within the subtype-specific peaks and promoter of genes, HOMER and UniBind were used for TFBS enrichment.

ATAC-seq TFBS enrichment results

The UniBind and HOMER TFBS enrichment for the most pattern-specific ATAC-seq peaks are shown for Pattern 1 (LumA/B) and Pattern 2 (Basal) in Figure 4.8 and 4.9, respectively. The results for the rest of the patterns are shown in Attachments (Figure 4.8-6.3), together with the UniBind results without background (Figure 6.4-6.8).

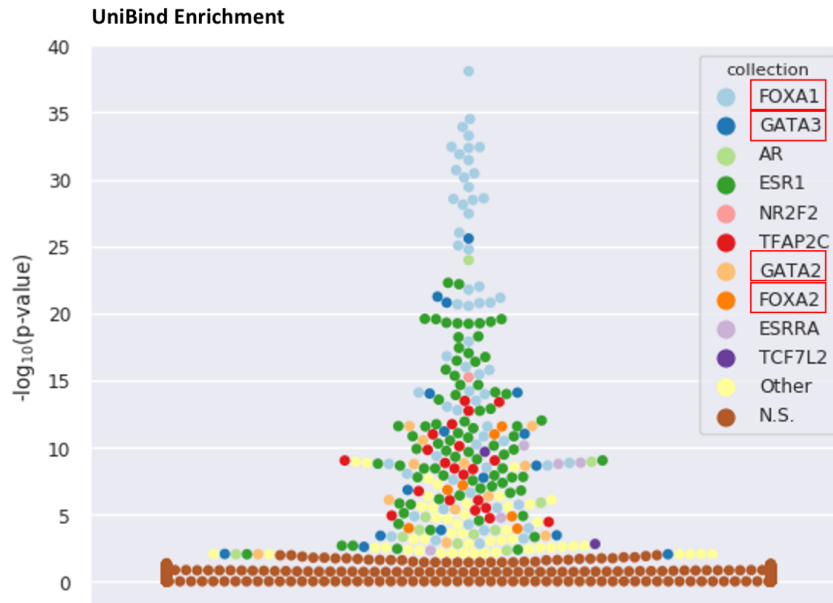
Each dot in the UniBind swarm plots represents the enrichment of a TFBS set, and the sets are derived from different cell types with different treatments. The colors represent the TF that binds to the TFBS set. The location of the TFBSs for a certain TF vary between cell types and condition. This is because the TFBSs found are the ones currently bound by a TF (using ChIP-Seq), and the regulatory state affecting the binding varies. In some cell types and conditions, the TFBS set for a given TF overlaps a lot with the input set of regions, while

other sets for the same TF overlap less. This explains the different enrichment of the TFBS sets representing the same TF. The names of the top 10 TFs with enriched TFBS sets are shown to the right. The HOMER table shows the most enriched TFBS motifs for the most pattern-specific regions of the ATAC-seq data. The name of the TF is shown first (bold), followed by the name of the TF family in parenthesis. Some of the TFs have multiple motifs, derived from different sets of TFBSs.

The result of both enrichment tools for Pattern 1 (LumA/B) is shown in Figure 4.8. The UniBind results (top) describe a "typical" ER+ TF profile according to the literature, with FOXA1, GATA3 and ER α (ESR1) highly enriched. The HOMER results are mainly dominated by motifs from the Forkhead family, including FOXA1 and FOXA2 (bottom). Two different GATA motifs are also enriched.

The result for Pattern 2 (Basal) is shown in Figure 4.9. The UniBind results (top) show that the TFBS sets for TFAP2C are highly enriched compared to those representing other TFs. The other results include two members of the SOX family (SOX10 and SOX2), TEAD4 and GRHL2, among others. The HOMER results show an enrichment for motifs corresponding to the binding site of the SOX family, including SOX10 and SOX2.

Pattern 1: LumA/B TFBS Enrichment Results (ATAC-seq)

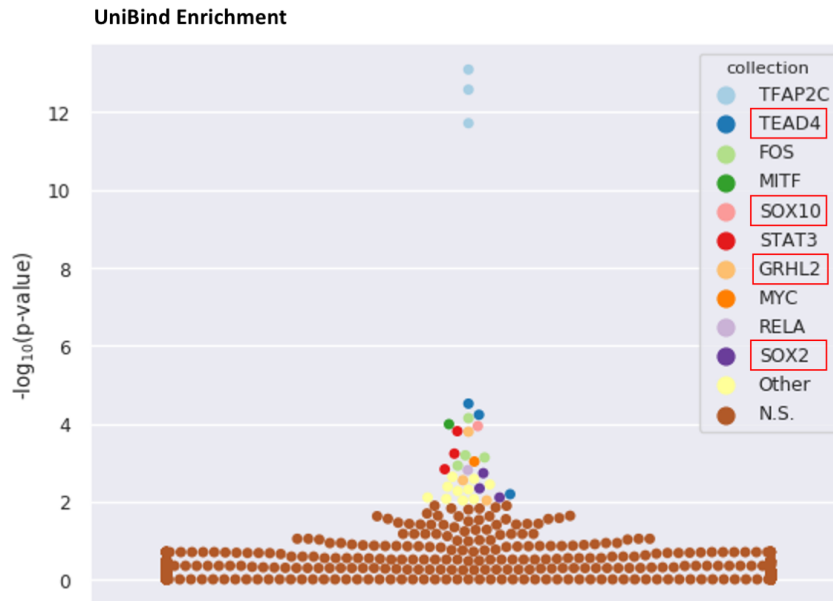


Homer Known Motif Enrichment

	Motif	Name	P-value	No. of target sequences with motif
1		FOXA2(Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer	1e-35	89
2		FOXA1(Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-30	104
3		FOXA1(Forkhead)/MCF7-FOXA1-ChIP-Seq(GSE26831)/Homer	1e-28	93
4		Fox:Ebox(Forkhead,bHLH)/Panc1-Foxa2-ChIP-Seq(GSE47459)/Homer	1e-25	82
5		Foxo1(Forkhead)/RAW-Foxo1-ChIP-Seq(Fan et al.)/Homer	1e-13	93
6		FOXP1(Forkhead)/H9-FOXP1-ChIP-Seq(GSE31006)/Homer	1e-13	41
7		PHA-4(Forkhead)/cElegans-Embryos-PHA4-ChIP-Seq(modEncode)/Homer	1e-11	132
8		ERE(NR),IR3/MCF7-Era-ChIP-Seq(Unpublished)/Homer	1e-9	24
9		GATA(Zf),IR3/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer	1e-7	19
10		Gata4(Zf)/Heart-Gata4-ChIP-Seq(GSE35151)/Homer	1e-6	50

Figure 4.8: TFBS enrichment for the Pattern 1 clusters of peaks, which are most important in Luminal B and Luminal A samples. The UniBind swarm plot (top) represents the top 10 TFs with enriched TFBS sets, and the HOMER table (bottom) shows the top 10 enriched TFBS motifs. The overlap between the two methods are marked with a red square.

Pattern 2: Basal TFBS Enrichment Results (ATAC-seq)



Homer Known Motif Enrichment

	Motif	Name	P-value	No. of target sequences with motif
1		Sox3(HMG)/NPC-Sox3-ChIP-Seq(GSE33059)/Homer	1e-62	385
2		Sox10(HMG)/SciaticNerve-Sox3-ChIP-Seq(GSE35132)/Homer	1e-53	355
3		NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq(Unpublished)/Homer	1e-51	372
4		Sox6(HMG)/Myotubes-Sox6-ChIP-Seq(GSE32627)/Homer	1e-50	337
5		Sox4(HMG)/proB-Sox4-ChIP-Seq(GSE50066)/Homer	1e-42	219
6		Sox2(HMG)/mES-Sox2-ChIP-Seq(GSE11431)/Homer	1e-38	211
7		GRHL2(CP2)/HBE-GRHL2-ChIP-Seq(GSE46194)/Homer	1e-32	116
8		NF1:FOXA1(CTF,Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-20	34
9		AP-2alpha(AP2)/Hela-AP2alpha-ChIP-Seq(GSE31477)/Homer	1e-15	134
10		TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	1e-14	159

Figure 4.9: TFBS enrichment for the Pattern 2 clusters of peaks, which are most important in Basal-like samples. The UniBind swarm plot (top) represents the top 10 TFs with enriched TFBS sets, and the HOMER table (bottom) shows the top 10 enriched TFBS motifs. The overlap between the two methods are marked with a red square.















RNA-seq TFBS enrichment results

A TFBS enrichment analysis was also performed for the RNA-seq data, using HOMER. This analysis was performed on the promoter regions of the genes in our gene signatures, and would therefore not capture TFs binding to enhancer regions. The results for the LumA/B and Basal cluster are shown in Table 4.5 and 4.6, respectively. The name of the TF is shown first (bold), followed by the name of the TF family in parenthesis. The motifs found correspond to variable binding sites of TFs that are present in the promoter regions of our gene set. The motifs found for the LumA/B cluster mostly belong to the Forkhead TF family. These include FOXA1 and FOXA2, which were also found from the ATAC-seq data for the corresponding cluster. For the Basal cluster, we find a lot of motifs for TFs in the bZIP family. In addition, we find GHRL2 and OCT4-SOX2-TCF-NANOG. The latter is the enrichment of a motif representing multiple TFs that bind together. GHRL2 and SOX2 were also found to be enriched for the ATAC-seq data. The results for the rest of the clusters are shown in Table S2-S4 in Attachments.

Table 4.5: Enriched TFBS motifs for Pattern 4 (LumA/B) from the RNA-seq data. The table is ordered by P-value.

Rank	Motif	Name	P-value	No. of target sequences with motif
1		FOXA1 (Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-4	32
2		Foxa2 (Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer	1e-4	26
3		FOXA1 (Forkhead)/MCF7-FOXA1-ChIP-Seq(GSE26831)/Homer	1e-4	27
4		FOXP1 (Forkhead)/H9-FOXP1-ChIP-Seq(GSE31006)/Homer	1e-4	17
5		VDR (NR), DR3/GM10855-VDR+vitD-ChIP-Seq(GSE22484)/Homer	1e-2	11
6		TATA-Box (TBP)/Promoter/Homer	1e-2	38
7		Fox:Ebox (Forkhead,bHLH)/Panc1-Foxa2-ChIP-Seq(GSE47459)/Homer	1e-2	24
8		Foxo1 (Forkhead)/RAW-Foxo1-ChIP-Seq(Fan et al.)/Homer	1e-2	48

Table 4.6: Enriched TFBS motifs for Pattern 5 (Basal) from the RNA-seq data. The table is ordered by P-value.

Rank	Motif	Name	P-value	No. of target sequences with motif
1		Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1e-5	34
2		BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	1e-4	32
3		AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1e-4	35
4		TATA-Box(TBP)/Promoter/Homer	1e-4	55
5		Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	1e-4	23
6		Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	1e-3	28
7		Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	1e-3	18
8		GRHL2(CP2)/HBE-GRHL2-ChIP-Seq(GSE46194)/Homer	1e-3	18
9		HOXD13(Homeobox)/Chicken-Hoxd13-ChIP-Seq(GSE38910)/Homer	1e-3	33
10		NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq(Unpublished)/Homer	1e-3	74
11		CEBP(bZIP)/ThioMac-CEBPb-ChIP-Seq(GSE21512)/Homer	1e-2	26
12		OCT4-SOX2-TCF-NANOG(POU,Homeobox,HMG)/mES-Oct4-ChIP-Seq(GSE11431)/Homer	1e-2	8
13		Pit1(Homeobox)/GCrat-Pit1-ChIP-Seq(GSE58009)/Homer	1e-2	28
14		NFkB-p65(RHD)/GM12787-p65-ChIP-Seq(GSE19485)/Homer	1e-2	29

4.2 MOFA

ATAC-seq and RNA-seq data was combined in a multi-omics analysis in order to see if this would improve the clustering. In addition, this would give a more direct link between the different clusters of the two data sets, meaning that we could do a gene ontology (GO) enrichment analysis on the most pattern-specific genes, that could possibly be connected with the TFs binding to the most pattern-specific peaks.

4.2.1 Normalization

Because the two data sets used for NMF had been normalized in different ways by different research groups, the raw data had to be normalized again, in a more similar manner. This was done using different methods from different packages, and the methods were compared by creating boxplots of the count data after each normalization (Figure 4.10). The boxplots show that the best normalization method for these data sets was to normalize by size factor (DESeq2).

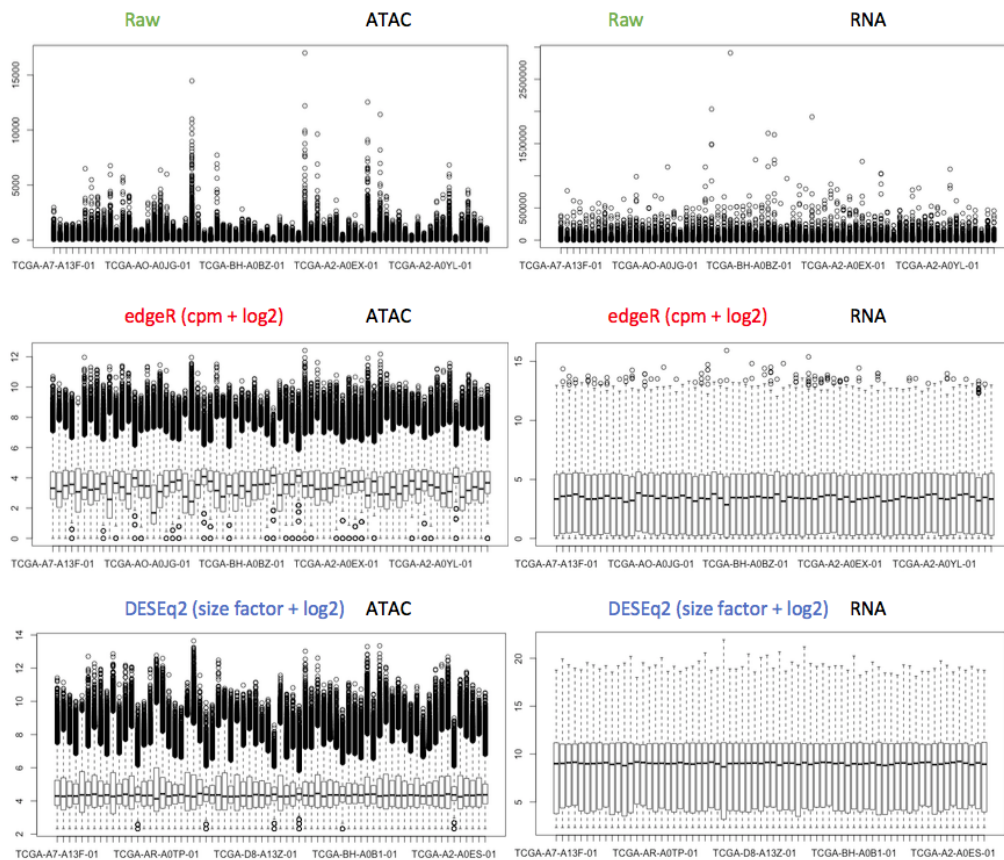


Figure 4.10: Boxplots showing different normalization methods for our data. ATAC-seq data (left) and RNA-seq data (right) is shown with raw data (top), edgeR (middle) and DESeq2 (bottom).

4.2.2 Sample clustering

The normalized data was used to create a MOFA model. Rank 2 yielded the best results according to prior subtypes, and was therefore used to create a MOFA model. Also, using more than two factors caused the remaining factors to explain less and less variation. The model was used to cluster the samples according to the strongest factor (similar to pattern in NMF). The clusters were plotted using UMAP, and different subtypes were attempted as labels, including PAM50 subtypes, ER⁺/ER⁻ and Basal/Non-basal. The latter was derived from the PAM50 subtypes, and the non-basal group contained all subtypes except for Basal-like: Luminal A, Luminal B, Normal-like and Her. The Basal/Non-basal labels were used due to low concordance with the other subtypes (ER status/PAM50). The sample clusters derived from the multi-omics analysis using MOFA is shown in Figure 4.11. The subtype labels have been modified to Basal (Basal-like) and Non-basal (Luminal A, Luminal B, Normal-like and Her), as

these two groups explained most of the variation. All the Basal-like samples clustered in one group, but six of the Non-basal samples also clustered in the Basal group. However, none of the Basal-like samples were present in the Non-basal cluster. The cluster assignment is shown with shapes, and the groups are shown with color.

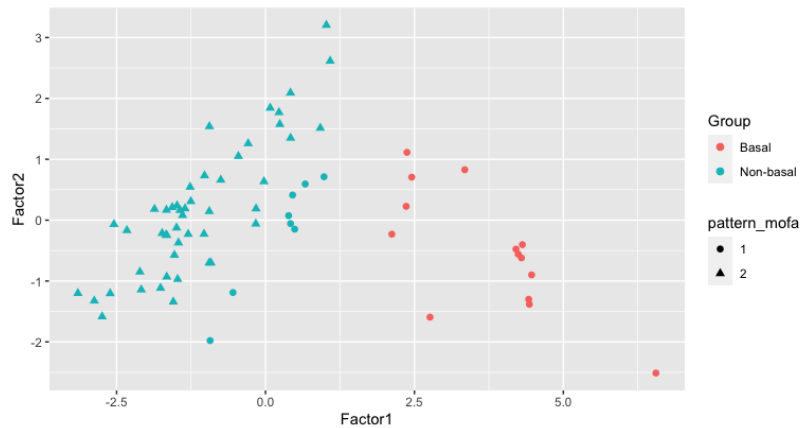


Figure 4.11: Clustering of samples based on the two factors created by MOFA.

4.2.3 Gene ontology enrichment analysis

In order to find out if the genes defining the Basal and Non-basal groups were enriched for any particular processes, a gene ontology enrichment analysis was performed on each factor from the MOFA analysis (Figure 4.12). The gene set corresponding to the Basal group (Factor 1) is enriched for processes involving mitosis (cell division), and especially for the organization of different cell components during mitosis. The gene set corresponding to the Non-basal group (Factor 2) is enriched for processes involving cell differentiation in different tissues. It is also enriched for different hormone receptor pathways, including positive regulation of the estrogen receptor signaling pathway.

Factor 1: Basal				
GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
GO:1902850	microtubule cytoskeleton organization involved in mitosis	8.47E-6	7.25E-2	4.20 (2082.27,257,14)
GO:0051383	kinetochore organization	8.45E-5	3.62E-1	11.35 (2082.7,131,5)
GO:1903047	mitotic cell cycle process	1.3E-4	3.71E-1	1.73 (2082.140,455,53)
GO:0007052	mitotic spindle organization	2.2E-4	4.7E-1	4.03 (2082.23,247,11)
GO:0008608	attachment of spindle microtubules to kinetochore	2.59E-4	4.44E-1	9.93 (2082.8,131,5)
GO:0006564	L-serine biosynthetic process	2.83E-4	4.04E-1	115.67 (2082.4,9,2)
GO:0090307	mitotic spindle assembly	3.37E-4	4.13E-1	6.32 (2082.8,247,6)
GO:0007051	spindle organization	3.45E-4	3.69E-1	3.27 (2082.35,255,14)
GO:0022402	cell cycle process	4.59E-4	4.37E-1	1.56 (2082.200,455,68)
GO:0006563	L-serine metabolic process	5.99E-4	5.12E-1	92.53 (2082.5,9,2)
GO:0030322	stabilization of membrane potential	6.22E-4	4.84E-1	5.63 (2082.5,370,5)
GO:0006865	amino acid transport	6.23E-4	4.44E-1	5.22 (2082.21,152,8)
GO:0048255	mRNA stabilization	7.77E-4	5.12E-1	7.29 (2082.7,204,5)
GO:0051315	attachment of mitotic spindle microtubules to kinetochore	8.74E-4	5.34E-1	10.60 (2082.6,131,4)

Factor 2: Non-basal (LumA, LumB, Normal and Her2)				
GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
GO:0002067	glandular epithelial cell differentiation	1.39E-7	1.25E-3	242.79 (2266.4,7,3)
GO:0060479	lung cell differentiation	4.34E-7	1.96E-3	194.23 (2266.5,7,3)
GO:0060487	lung epithelial cell differentiation	4.34E-7	1.31E-3	194.23 (2266.5,7,3)
GO:0033145	positive regulation of intracellular steroid hormone receptor signaling pathway	1.05E-6	2.38E-3	133.29 (2266.3,17,3)
GO:0033148	positive regulation of intracellular estrogen receptor signaling pathway	1.05E-6	1.9E-3	133.29 (2266.3,17,3)
GO:0002065	columnar/cuboidal epithelial cell differentiation	1.33E-6	2E-3	161.86 (2266.6,7,3)
GO:0030855	epithelial cell differentiation	1.59E-6	2.06E-3	22.85 (2266.35,17,6)
GO:0033146	regulation of intracellular estrogen receptor signaling pathway	2.95E-6	3.33E-3	53.32 (2266.10,17,4)
GO:0060480	lung goblet cell differentiation	1.64E-5	1.64E-2	323.71 (2266.2,7,2)
GO:0060749	mammary gland alveolus development	3.03E-5	2.74E-2	66.65 (2266.6,17,3)
GO:0060740	prostate gland epithelium morphogenesis	3.15E-5	2.58E-2	66.65 (2266.6,17,3)
GO:0002070	epithelial cell maturation	3.48E-5	2.62E-2	37.15 (2266.3,61,3)
GO:0061140	lung secretory cell differentiation	4.07E-5	2.83E-2	215.81 (2266.3,7,2)
GO:0006338	chromatin remodeling	4.65E-5	3E-2	75.53 (2266.15,6,3)

Figure 4.12: GOrilla gene ontology enrichment results showing enriched processes for the genes that were specific for the Basal (top) and Non-basal (bottom) groups. All 14 significantly enriched processes for the Basal group is shown, while the Non-basal group were enriched for 38 processes, and only the top 15 is shown.

4.2.4 UniBind TF enrichment

In order to see if the Basal and Non-basal groups achieved a similar transcriptional profile as in the NMF analysis, enriched TFBS sets were found using UniBind. The analysis was performed both with and without background. However, only the results using no background is presented. The top 10 TFs with enriched TFBS sets for the Non-basal group is shown in Figure 4.13, and include FOXA1, ER α and GATA2/3, among others. The top 10 TFs with enriched TFBS sets for the Basal group include GHRL2, TEAD4, FOXA1 and GATA2 (Figure 4.14).

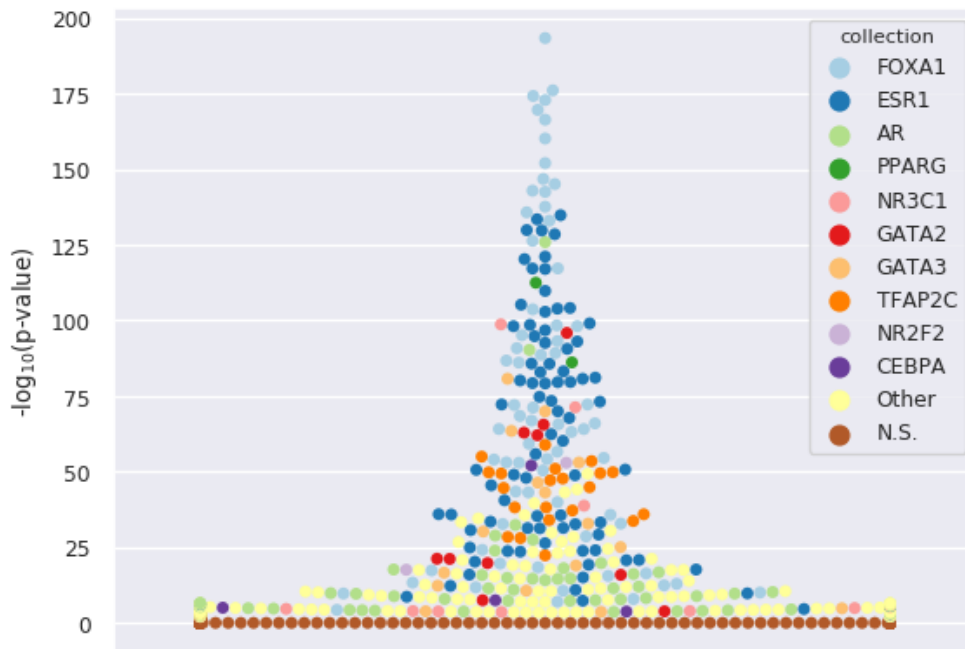


Figure 4.13: UniBind enrichment for 657 factor-specific genes corresponding to the Non-basal group.

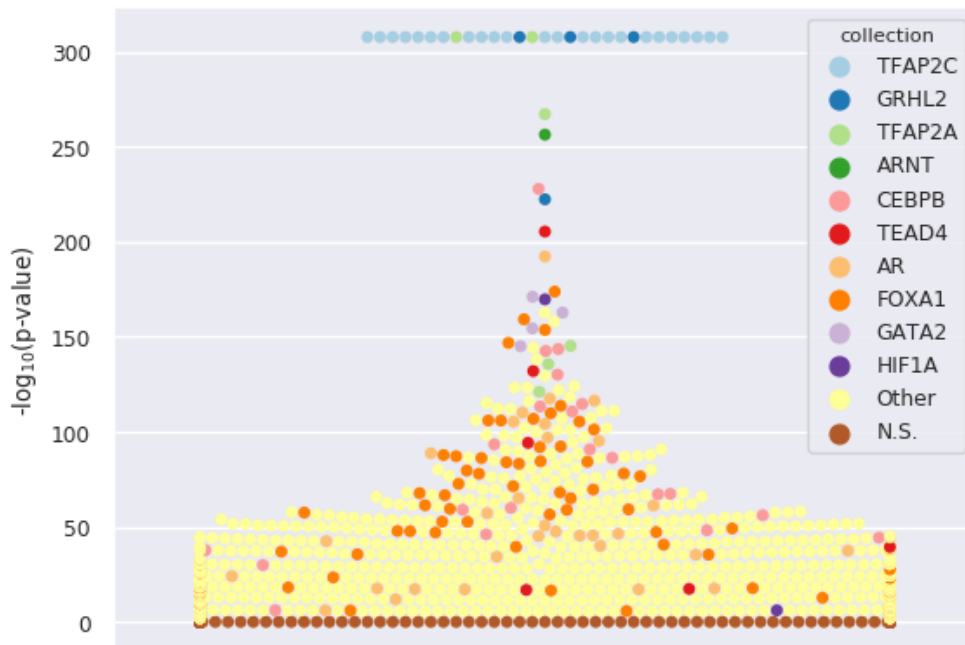


Figure 4.14: UniBind enrichment for 5117 factor-specific genes corresponding to the Basal group.

Chapter 5

Discussion

For this study, two different matrix factorization methods were used in order to find the key transcription factors (TFs) that drive each subtype: NMF and MOFA. In this chapter, we will discuss the results gained from each study, in addition to aspects of the methods that may have affected the results.

5.1 NMF analysis

The aim of the non-negative matrix factorization (NMF) analysis was to divide the samples into clusters that reflected the prior subtypes, and derive a set of peaks and genes that defined each cluster. The clustering of samples showed to be largely related to the PAM50 subtypes for both data sets, especially for the Basal-like subtype. The gene signatures derived for each subtype were consistent with gene signatures from previous studies.

5.1.1 Subtype clustering

The clustering of samples shows that when divided into five clusters, the data has the biggest overlap with PAM50 subtypes, for both the ATAC-seq and RNA-seq data (Figure 4.2 and 4.3, respectively). Because the Basal-like subtype was the most interesting due to the lack of targeted treatment, the focus was on preserving this group in a separate cluster. This was especially successful for the ATAC-seq data, where all Basal-like samples clustered alone (Figure 4.2). Thus, this cluster can be directly interpreted as the Basal-like subtype, and the characteristics found are important for that subtype alone. The credibility of the clusters is supported by the cophenetic correlation coefficient, which indicates stable clusters for rank 5 (Figure 4.1). The concordance with PAM50 subtypes was somewhat expected, as PAM50 is based on gene expression data. The PAM50 subtypes were also expected to make most sense for the RNA-seq data, as they

were based on the same data type. Therefore, it was surprising that the ATAC-seq data achieved the best concordance with these subtypes, particularly for the Basal-like and Her2 subtypes (Figure 4.2). However, the ATAC-seq data had almost twice as many samples due to the number of technical replicates, which may have impacted the results. In addition, the clustering of the RNA-seq data might have been impacted by the two Basal outliers. These share a distinct pattern that separate them from the other samples, which may be the result of contamination in the prior RNA-seq process. Their gene expression profiles could also be the result of technical errors during sequencing. In that case, these replicates should perhaps have been removed in order to see if this improved the clustering. However, if the gene expression pattern is not due to errors, these samples could have biological characteristics that would explain some of the diversity found within the same tumor, and should not be excluded. The exact reason for their distinct pattern was hard to know without analyzing the data further, and since the Basal-like pattern clustered well regardless, the replicates were kept for further analysis.

While PAM50 worked well for separating clusters, the use of ER status with rank 2 gained mixed clusters, which was not surprising considering that the two ER-subtypes (Basal-like and Her2) behave widely differently. As stated by Perou et al. (2000): "The clinical designation of 'oestrogen receptor negative' breast carcinoma encompasses at least two biologically distinct subtypes of tumours (basal-like and ErB-B2 [HER2] positive), which may need to be treated as distinct diseases". This was also indicated in the instability of the clusters, especially for the RNA-seq data. Using these clusters for further analysis would have lead to mixed signals when deriving the characteristics from each cluster. Although the Basal-like subtype - which was the main focus for this thesis - clustered well using PAM50, the overall clustering could potentially have been improved by including more samples from other subtypes than Luminal A. These are however the only samples with available ATAC-seq and RNA-seq data from the US-BRCA project, so far.

5.1.2 Gene and peak signatures

From Figure 4.3, we can see some of the most pattern-specific genes resulting from NMF. Some of these genes have previously been described in other breast cancer related studies, for example EGFR, SOX8 and SOX10, which were found in the Basal gene signature. EGFR is a co-activator that has been investigated as a potential target for Basal-like breast cancer (Siziopikou & Cobleigh, 2007), and is highly expressed in the majority of Basal-like tumors (Cleator et al., 2007).

The SOX genes code for a family of TFs related to development (Kamachi & Kondoh, 2013), and various members of this family have previously been found to be overexpressed in Basal-like tumors (Zhang et al., 2012; Liu et al., 2018), including SOX8 (Tang et al., 2019) and SOX10 (Cimino-Mathews et al., 2013). In addition, some of the genes from the LumA/B cluster, namely SERPINA6, BEX1 and AGTR1, have been proposed as markers for poor response to chemotherapy for patients with HER2 negative tumors (De Ronde et al., 2013). These genes were not found in the LumA/Normal cluster, which would make sense as patients with Luminal B tumors have a worse prognosis than most patients with Luminal A or Normal-like tumors (Figure 1.2).

5.1.3 Gene ontology enrichment analysis

The gene signatures were also used to search for enriched ontology terms and gene signatures. Although the GOrilla gene ontology enrichment analysis did not result in any enriched biological processes or functions for the NMF analysis, HOMER discovered an enrichment for previously defined gene signatures (Table 4.4). The gene signature for the Basal outliers was not enriched for any type of breast cancer signatures, but for RNAPII promoter opening. However, the gene signatures for the rest of the clusters were enriched for the subtype that dominated the clusters, which was a confirmation that these were truly active, subtype-specific genes. This indicates that NMF was successful at defining meaningful gene signatures, and that the algorithm itself preserved a lot of the original information.

5.1.4 Chromosome distribution

The pie charts showing the percentage of peaks/genes residing at each chromosome were created in order to visually compare between the corresponding clusters of ATAC-seq and RNA-seq data (Figure 4.7). The pie charts should preferably show similar chromosome distribution if the TFs that bind to open regions in a subtype regulate the most highly expressed genes for the same subtype. This assumption was made because most TFBSs are found on the same chromosome as the genes they regulate (van Arensbergen et al., 2014). However, no such connection was apparent. There could be several reasons for this. First, some of the samples that cluster together in the ATAC-seq data, cluster together with other samples in the RNA-seq data, causing mixed signals (Table S5). Second, some of the open regions could possibly be bound by TFs at silencer regions. As we only look at highly expressed genes, the effect of potential silencer regions

would not be uncovered.

Although the regions did not overlap between the data sets, the disproportionate distribution of chromosome locations in the most pattern-specific ATAC-seq peaks were highly interesting. The Pattern 1 (LumA/B) and Pattern 4 (Her2) clusters were the most extreme cases, showing great accessibility in chromosome 8 and chromosome 17, respectively. The simultaneous accessibility through open chromatin could suggest that multiple regions work together, in a potential transcription regulatory network. For the Her2 cluster, which contains 18 Her2 samples and 6 Luminal B samples, multiple peaks were located in close proximity to each other. These regions were investigated through the UCSC genome browser, uncovering the ERBB2 (encodes HER2) and CDK12 genes. The high accessibility surrounding the ERBB2 gene in this cluster is not surprising, considering that Her2 and some Luminal B tumors are known to overexpress ERBB2 (Hugh et al., 2009). The overexpression of ERBB2 and other genes in the surrounding regions of chromosome 17 has previously been linked to DNA amplification (Perou et al., 2000), which might explain the major activity in these regions. CDK12 is known to be commonly co-amplified with ERBB2 in breast cancer, although its exact function and relation to ERBB2 remains largely unknown (Tien et al., 2017). The fact that the top peaks for the Her2 cluster were found in these regions, indicated that also the peak signatures derived from NMF made sense. Because the peak signatures succeeds in describing the differences in chromatin accessibility between subtypes, at least for the Basal-like and Her2 subtypes, these peak signatures could potentially be stored and used for classification purposes.

5.1.5 Subtype-specific transcription factors

For the ATAC-seq data, TFBS enrichment was performed using both UniBind and HOMER. HOMER was also used to find enriched TFBS motifs from the RNA-seq gene signatures. The latter only involved enrichment in promoter regions, which means that the TFBS enrichment for the ATAC-seq data gave a fuller picture by involving possible enhancer regions as well. In addition, the cluster assignment differed slightly between the data sets (Figure S5). Therefore, the results from both data sets were not expected to fully overlap.

LumA/B

The results of the TFBS enrichment analyses for the ATAC-seq data show that FOXA1, FOXA2 and GATA(3/2) are enriched for the LumA/B cluster, for both TFBS enrichment tools (Figure 4.8). The roles of FOXA1 and GATA3 in

Luminal breast cancer have been widely documented in multiple studies, while FOXA2 and GATA2 have received less attention. However, some studies have shown that FOXA2 acts to prevent metastasis in breast cancer (Zhang et al., 2015). In addition, multiple sets of TFBSs for ER α (showed as "ESR1" in the plot) were found enriched using UniBind. ER α is known to be a main driver of ER+ subtypes, and was therefore expected to be enriched. It is unclear why this TF was not enriched using HOMER for the ATAC-seq data. The TFBS enrichment for the LumA/B cluster from the RNA-seq data showed an enrichment for FOXA1 and FOXA2, showing that these TFs could possibly upregulate some of the most highly expressed genes in the LumA/B peak signature through the promoter regions (Figure 4.5).

Basal

For the Basal cluster of the ATAC-seq data, which only contained Basal-like samples, SOX10, SOX2, TEAD4 and GHRL2 were found enriched, regardless of TFBS enrichment tool (Figure 4.9). All of these TFs have previously been found to be enriched in the Basal-like (or triple negative) subtype (Cimino-Mathews et al., 2013; Rodriguez-Pinilla et al., 2007; Wang et al., 2015). In addition, STAT3 and MYC, which have also been proposed as potential Basal drivers (Zhu et al., 2020; Xu et al., 2010), were found enriched using UniBind, but not HOMER. The TFBS enrichment for the Basal cluster from the RNA-seq data showed an enrichment for GHRL2 and OCT4-SOX2-TCF-NANOG (Figure 4.6). As these overlaps with some of the TFs found in the ATAC-seq data, this strengthens the hypothesis that SOX TFs and GHRL2 are important drivers of Basal-like breast cancer. Also, because the clustering of Basal-like samples was strong also for the RNA-seq data, it is likely that these TFs are involved in regulating some of the genes in the Basal gene signature.

The results of the following clusters are shown in Attachments.

LumA/Normal

The LumA/Normal cluster differed from the LumA/B cluster, as the peaks that characterized this cluster were enriched for CEBP and members of the STAT family, including STAT5 (Figure 6.1). STAT5 has previously been associated with good prognosis in ER/PR+ breast cancers (Barash, 2012), indicating that this group might represent the group with the best prognosis. These TFs were however not found in the promoter regions from the RNA-seq data, which contained very few enriched TFBSs (Figure ??). Previous research has shown that both

CEBP and STAT bind to enhancer regions (Ramji & Foka, 2002; Vahedi et al., 2012), which explains why these were not found in the promoter region of the highly expressed genes.

Her2

The results for the Her2 cluster showed no overlap between HOMER and UniBind (Figure 6.2). However, using UniBind, TFAP2C and YY1 were found to be enriched. These TFs are both previously found to be important for the Her2 subtype (Begon et al., 2005; Powe et al., 2009; Woodfield et al., 2010). Also here, UniBind is more consistent with previous research. The TFs found enriched in the promoter regions of the top genes for the Her2mix cluster (Figure ??) did not correspond with the TFs found for the Her2 cluster from the ATAC-seq data. Here, FOXA1 and FOXA2, among others, were found enriched. These are normally associated with ER+ subtypes, such as Luminal A and Luminal B. The enrichment seen for these TFs are likely because the Her2mix cluster contains a mix of five Her2 samples, three Luminal A samples and 2 Luminal B samples, and some of the highly specific genes for this group might be related to the Luminal subtypes.

LumA

The LumA cluster from the ATAC-seq data contained a large number of Luminal A samples, and some Luminal B samples. Although the locations of the most open regions differ from the LumA/B cluster, they involve a lot of the same TFs (Figure 6.3). These include FOXA1, FOXA2 and GATA3/GATA2. The overlap indicates that the tumors making up the samples of this cluster might behave in a similar matter as those from the LumA/B cluster. This cluster had no corresponding cluster in the RNA-seq data, as the RNA-seq data only contained two Luminal clusters, instead of three. Therefore, most of the samples in this group were distributed between the LumA/B and LumA/Normal RNA-seq clusters. This complicated the results of the analysis for the ER+ subtypes (Luminal A, Luminal B and Normal-like) considerably, especially when making connections between the data sets. Understanding which TFs of the ATAC-seq cluster regulates which genes of the RNA-seq clusters is therefore a hard connection to make from this analysis. However, the TFs inferred from the ATAC-seq data are largely supported by literature, and give valuable information on its own.

5.2 MOFA analysis

The aim of the MOFA analysis was to explore the possible advantages of integrating the RNA-seq and ATAC-seq data in one analysis. MOFA was used to perform a gene set enrichment analysis, and TFBS enrichment was performed also here in order to possibly produce a more robust set of key TFs.

5.2.1 Subtype clustering

The MOFA analysis was performed with two factors (rank 2), in order to separate the Basal-like samples from the other subtypes. Compared to the clustering performed with NMF and rank 5, the separation of Basal-like samples was poor (Figure 4.11). Some of the Non-basal samples were clustered with the Basal group, which was not ideal. Nevertheless, the resulting pattern matrix and the two amplitude matrices were used for further analysis, as the samples that dominated each cluster would hopefully provide a strong signal.

5.2.2 Gene ontology enrichment analysis

The gene ontology enrichment showed that the genes selected as factor-specific for the Basal group (Figure 4.12 (top)) were largely involved in processes related to cell cycle and cell division (mitosis), including the organization of cell division related compartments, such as the cytoskeleton and spindle. The Basal-like subtype is an aggressive subtype, and enhanced cell division is expected. A study by Yang et al. (2019) also found that genes specific for Basal-like breast cancer were associated with pathways involving the cell cycle.

The gene ontology enrichment for the Non-basal group showed that the genes were largely involved in processes that involved cell differentiation (Figure 4.12 (bottom)). Higher degree of cell differentiation is associated with less aggressive tumors and better prognosis, which is in line with the prognosis of the majority of the subtypes residing in this cluster. GATA3, which was found for the Luminal subtypes in almost all TFBS enrichment analysis in this study, has previously found to be an important regulator of Luminal cell differentiation (Asselin-Labat et al., 2007). Therefore, it is likely that GATA3 regulate some of the genes that are enriched for this biological process. In addition to cell differentiation, the genes for this group were enriched for positive regulation of estrogen receptor (ER) signaling pathway. The enrichment for this pathway was expected, as most of the samples in this group are ER+, except for some Her2 samples. If we had excluded the Her2 samples before the analysis, it is possible that the most

factor-specific genes would be more enriched for this pathway.

5.2.3 Subtype-specific transcription factors

TFBS enrichment was performed for the MOFA peaks using UniBind, with no background. The choice of using no background was made based on the fact that we only had two groups, and most of the top peaks were more open in the Basal group (5117 out of 5774). The results of using no background were also more in line with previous studies, including the NMF analysis performed in this study. The TFs found to be most enriched for the Non-basal group include FOXA1, ER α , GATA2 and GATA3 (Figure 4.13). These were repeatedly found in the Luminal clusters for the NMF analysis, which makes sense as this group is mostly dominated by Luminal samples.

The most enriched TFs for the Basal group include GHRL2, TEAD4, FOXA1 and GATA2 (4.14). Seen in the context of the results from the NMF analysis, this appears to be a mixed signal of Basal-like and Luminal TFs. The enrichment of FOXA1 and GATA2 is likely due to the fact that this group also contains some Luminal samples. However, the presence of GHRL2 and TEAD4 strengthen the chance of these TFs being key drivers of the Basal-like subtype.

5.3 Method discussion

This section will discuss the different tools and parameters used in both analyses, and potential improvements that can be made.

NMF method

As previously mentioned, NMF has some drawbacks. First, the choice of rank will affect the clusters and the value of the information gained. We attempted to overcome potential bias by also looking at the stability of the clusters, independent of prior knowledge of groupings. Another drawback is the initialization of NMF: As it starts with random numbers for the pattern matrix and amplitude matrix, the path it takes before it reaches a local minimum vary from run to run, and the results are therefore not reproducible (Pehkonen et al., 2005). Methods that aim for a more deterministic solution have been suggested, and others are still in development (Wild et al., 2004; Sauwen et al., 2016; Janecek & Tan, 2011; Gong & Nandi, 2013). Although the lack of reproducibility using a random initialization remains a problem, the results gained from this study proves that it can be a useful tool for exploratory studies.

The probability of reaching a bad local minimum could have been reduced by running the algorithm multiple times and choosing the run with the lowest error, but this was discovered too late. Instead, multiple runs were performed with visual inspection, in order to make sure that the TFs uncovered were the same, which they were. Also, the overlap of the peak and gene signatures with previous studies suggested that no bad local minimum was reached. Nevertheless, this is a possible source of error that should be accounted for when interpreting the results.

MOFA method

MOFA is a new method, first published in 2018 (Argelaguet et al., 2018). Therefore, the advantages and disadvantages of this method are still being established. However, one disadvantage is the lack of a non-negativity constraint, which makes it less intuitive when interpreting the data, compared to NMF. Another disadvantage that has been mentioned is when the multiple data sets being used do not have a direct, linear relationship (Peng et al., 2020). This could have impacted the results of this study, as some of the open regions might have been bound by TFs at silencers in some samples, and at enhancers in others. The openness of the peaks would thereby be similar, but the effect on the genes being expressed would be different. In this study, the clustering of samples were not improved compared to NMF, indicating that there could potentially be some inconsistencies between the data sets. However, it could also be the underlying characteristics of the algorithm itself, which is hard to tell. In order to make full use of the MOFA analysis, the clustering of these subtypes should have been improved. Regardless, MOFA was a useful comparison to NMF for finding a robust set of TFs for the Basal-like and Luminal subtypes.

TFBS enrichment methods

The two TFBS enrichment methods used in this study gained some differences in results. In general, it appears that the most enriched TFs found using UniBind are more consistent with previous research. Often, the top HOMER hits involved multiple TFs from the same family, which is likely due to the fact that TFs within the same family share similar motifs. In addition, similar motifs derived from different cell types were presented as different hits, thereby pushing other results further down the list. Enrichment for different cell types is better represented in the UniBind plot, as only the top TFBS set is used to define the degree of

enrichment for each TF. In general, UniBind is a newer method that combines the use of motifs with known TFBSs, which makes it more reliable.

There were also some parameters that differed between the methods, which may have impacted the results. First, the choice of size for the ATAC-seq data differed. In UniBind, the full peaks (501 bp) were analyzed, while only 401 bp of each peak were analyzed in HOMER. For ChIP-Seq data, this would not matter, as the TFs are usually found to bind within 50 bp from the peak center (Bailey, 2011). However, ATAC-seq extracts regions in a different way, and TFBSs might potentially be shifted a bit further from the peak center. To ensure that this did not impact the results, HOMER was also tried with size 2000, which yielded highly similar results. Thus, it is likely that the TFBSs are well covered with size 200, and that this parameter had little effect on the differences in results between the methods. Second, the backgrounds differed between the methods. For UniBind, a background consisting of the top peaks were used for the NMF analysis. For the analyses performed with the no background option, a background consisting of all TFBS sets stored in the UniBind database was used. HOMER, on the other hand, creates random backgrounds that match the GC content of the input sequences, when no customized background is provided. The use of random sequences in HOMER versus real regions known to be active in UniBind could have impacted the results. In order to possibly improve the HOMER analysis, a customized background could have been provided.

Chapter 6

Conclusion and future perspective

In this study, two different matrix factorization methods were used in order to uncover the TFs that drive each subtype of breast cancer.

The first aim was to use NMF on RNA-seq and ATAC-seq data, in order to define gene and peak signatures for each subtype. Five different gene and peak signatures were defined for each cluster of samples, and these were further explored and validated. The clusters did not correspond directly to a subtype, with the exception of the Basal-like subtype. The clustering could possibly have been improved by using more samples, if ATAC-seq data had been available for these. An increase in available data is expected as the ATAC-seq technique improves.

The second aim was to find enriched TFBSs within the peaks and promoter regions of the signatures derived from NMF. The TFs that were found to drive the Basal-like subtype include members of the SOX family (specifically SOX2 and SOX10), GRHL2 and TEAD4, all of which have previously been suggested as potential drivers in different studies. In addition, MYC and STAT3 are possible candidates. The Luminal subtypes are largely driven by FOXA1, ER α and GATA3, as found in previous studies. They were also found to be enriched for FOXA2 and GATA2, which have gotten less attention than their family members (FOXA1 and GATA3, respectively), for their potential roles in Luminal breast cancers.

The third aim was to explore information gained by combining RNA-seq and ATAC-seq data in a multi-omics experiment. A gene ontology enrichment analysis was performed, which revealed that the Basal-like breast cancer is enriched for processes involving cell division. Although the results of the MOFA analysis was impacted by the poor clustering, the TFBS enrichment analysis supported some of the results found for the NMF analysis, suggesting that GRHL2 and TEAD4

are key drivers of the Basal-like subtype.

The unsupervised matrix factorization methods used in this study have shown great potential for learning characteristics of different groups, and should be applied to other cancer types in order to potentially discover new subtypes and their molecular characteristics. The computational process involving these different tools should also be made available for public use, possibly as an R package. The key TFs found for each breast cancer subtype throughout this study, especially for the Basal-like subtype, should be investigated as potential targets for new treatments.

Bibliography

- Ackermann, A. M., Wang, Z., Schug, J., Naji, A., & Kaestner, K. H. (2016). Integration of atac-seq and rna-seq identifies human alpha cell and beta cell signature genes. *Molecular metabolism*, 5(3), 233–244.
- Adélaïde, J., Finetti, P., Bekhouche, I., Repellini, L., Geneix, J., Sircoulomb, F., Charafe-Jauffret, E., Cervera, N., Desplans, J., Parzy, D., et al. (2007). Integrated profiling of basal and luminal breast cancers. *Cancer research*, 67(24), 11565–11575.
- Anand, P., Kunnumakara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., Sung, B., & Aggarwal, B. B. (2008). Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical research*, 25(9), 2097–2116.
- Anders, C., & Carey, L. A. (2008). Understanding and treating triple-negative breast cancer. *Oncology (Williston Park, NY)*, 22(11), 1233.
- Angelini, C., & Costa, V. (2014). Understanding gene regulatory mechanisms by integrating chip-seq and rna-seq data: statistical solutions to biological problems. *Frontiers in cell and developmental biology*, 2, 51.
- Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6).
- Arteaga, C. L., Sliwkowski, M. X., Osborne, C. K., Perez, E. A., Puglisi, F., & Gianni, L. (2012). Treatment of her2-positive breast cancer: current status and future perspectives. *Nature reviews Clinical oncology*, 9(1), 16–32.
- Asselin-Labat, M.-L., Sutherland, K. D., Barker, H., Thomas, R., Shackleton, M., Forrest, N. C., Hartley, L., Robb, L., Grosveld, F. G., van der Wees, J.,

- et al. (2007). Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. *Nature cell biology*, *9*(2), 201–209.
- Bailey, T. L. (2011). Dreme: motif discovery in transcription factor chip-seq data. *Bioinformatics*, *27*(12), 1653–1659.
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). Meme suite: tools for motif discovery and searching. *Nucleic acids research*, *37*(suppl_2), W202–W208.
- Bajic, M., Maher, K. A., & Deal, R. B. (2018). Identification of open chromatin regions in plant genomes using atac-seq. In *Plant Chromatin Dynamics*, (pp. 183–201). Springer.
- Barash, I. (2012). Stat5 in breast cancer: potential oncogenic activity coincides with positive prognosis for the disease. *Carcinogenesis*, *33*(12), 2320–2325.
- Begon, D. Y., Delacroix, L., Vernimmen, D., Jackers, P., & Winkler, R. (2005). Yin yang 1 cooperates with activator protein 2 to stimulate erbb2 gene expression in mammary cancer cells. *Journal of Biological Chemistry*, *280*(26), 24428–24434.
- Biggin, M. D. (2011). Animal transcription networks as highly connected, quantitative continua. *Developmental cell*, *21*(4), 611–626.
- Bonifer, C., & Cockerill, P. N. (2011). *Chromatin mechanisms regulating gene expression in health and disease*. Springer.
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6), 394–424.
- Breastcancer.org (2020). Molecular subtypes of breast cancer. Last accessed 2020-05-24.
 URL <https://www.breastcancer.org/symptoms/types/molecular-subtypes>
- Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, *101*(12), 4164–4169.

- Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, *10*(12), 1213.
- Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). Atac-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, *109*(1), 21–29.
- Canadian Breast Cancer Network (2020). Types & sub-types. Last accessed 2020-05-24.
URL <https://www.cbcn.ca/en/types-and-subtypes>
- Cantini, L., Calzone, L., Martignetti, L., Rydenfelt, M., Blüthgen, N., Barillot, E., & Zinovyev, A. (2017). Classification of gene signatures for their information value and functional redundancy. *NPJ systems biology and applications*, *4*(1), 1–11.
- Chen, K., & Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and micrnas. *Nature Reviews Genetics*, *8*(2), 93–103.
- Chen, Y., Shi, L., Zhang, L., Li, R., Liang, J., Yu, W., Sun, L., Yang, X., Wang, Y., Zhang, Y., et al. (2008). The molecular mechanism governing the oncogenic potential of sox2 in breast cancer. *Journal of Biological Chemistry*, *283*(26), 17969–17978.
- Cimino-Mathews, A., Subhawong, A. P., Elwood, H., Warzecha, H. N., Sharma, R., Park, B. H., Taube, J. M., Illei, P. B., & Argani, P. (2013). Neural crest transcription factor sox10 is preferentially expressed in triple-negative and metaplastic breast carcinomas. *Human pathology*, *44*(6), 959–965.
- Cleator, S., Heller, W., & Coombes, R. C. (2007). Triple-negative breast cancer: therapeutic options. *The lancet oncology*, *8*(3), 235–244.
- Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., Silva, T. C., Groeneveld, C., Wong, C. K., Cho, S. W., Satpathy, A. T., Mumbach, M. R., Hoadley, K. A., Robertson, A. G., Sheffield, N. C., Felau, I., Castro, M. A. A., Berman, B. P., Staudt, L. M., Zenklusen, J. C., Laird, P. W., Curtis, C., Greenleaf, W. J., & Chang, H. Y. (2018). The chromatin accessibility landscape of primary human cancers. *Science*, *362*(6413).
- Cox, P., & Goding, C. (1991). Transcription and cancer. *British journal of cancer*, *63*(5), 651–662.

- Cui, X., Schiff, R., Arpino, G., Osborne, C. K., & Lee, A. V. (2005). Biology of progesterone receptor loss in breast cancer and its implications for endocrine therapy. *Journal of clinical oncology*, *23*(30), 7721–7735.
- Cyr, A. R., Kulak, M. V., Park, J. M., Bogachek, M. V., Spanheimer, P. M., Woodfield, G. W., White-Baer, L. S., O'Malley, Y. Q., Sugg, S. L., Olivier, A. K., et al. (2015). Tfp2c governs the luminal epithelial phenotype in mammary development and carcinogenesis. *Oncogene*, *34*(4), 436–444.
- Dai, X., Cheng, H., Bai, Z., & Li, J. (2017). Breast cancer cell line classification and its relevance with breast tumor subtyping. *Journal of Cancer*, *8*(16), 3131.
- Dai, X., Xiang, L., Li, T., & Bai, Z. (2016). Cancer hallmarks, biomarkers and breast cancer molecular subtypes. *Journal of Cancer*, *7*(10), 1281.
- Davie, K., Jacobs, J., Atkins, M., Potier, D., Christiaens, V., Halder, G., & Aerts, S. (2015). Discovery of transcription factors and regulatory regions driving in vivo tumor development by atac-seq and faire-seq open chromatin profiling. *PLoS genetics*, *11*(2).
- De Ronde, J. J., Lips, E. H., Mulder, L., Vincent, A. D., Wesseling, J., Nieuwland, M., Kerkhoven, R., Peeters, M.-J. T. V., Sonke, G. S., Rodenhuis, S., et al. (2013). Serpina6, bex1, agtr1, slc26a3, and laptm4b are markers of resistance to neoadjuvant chemotherapy in her2-negative breast cancer. *Breast cancer research and treatment*, *137*(1), 213–223.
- Delgado, M. D., & León, J. (2006). Gene expression regulation and cancer. *Clinical and Translational Oncology*, *8*(11), 780–787.
- DeSantis, C. E., Ma, J., Goding Sauer, A., Newman, L. A., & Jemal, A. (2017). Breast cancer statistics, 2017, racial disparity in mortality by state. *CA: a cancer journal for clinicians*, *67*(6), 439–448.
- Devarajan, K. (2008). Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS computational biology*, *4*(7).
- Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, *10*(1), 48.
- EMBL-EBI (2020). Variants in transcription factor binding motifs. Last accessed 2020-06-11.
- URL <https://www.ebi.ac.uk/training/online/course/>

human-genetic-variation-i-introduction-2019/
what-genetic-variation/variants-transcription

- Emmert-Streib, F., de Matos Simoes, R., Mullan, P., Haibe-Kains, B., & Dehmer, M. (2014). The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Frontiers in genetics*, 5, 15.
- Erbe, R., Kessler, M., Favorov, A., Easwaran, H., Gaykalova, D., & Fertig, E. (2020). Matrix factorization and transfer learning uncover regulatory biology across multiple single-cell atac-seq data sets. *Nucleic acids research*.
- Fertig, E. J., Ding, J., Favorov, A. V., Parmigiani, G., & Ochs, M. F. (2010). Cogaps: an r/c++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics*, 26(21), 2792–2793.
- Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). Jaspar 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 48(D1), D87–D92.
- Gaujoux, R., & Seoighe, C. (2010). A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11(1), 367.
- Gee, J. M. W., Eloranta, J., Ibbitt, J., Robertson, J., Ellis, I., Williams, T., Nicholson, R. I., & Hurst, H. (2009). Overexpression of tfap2c in invasive breast cancer correlates with a poorer response to anti-hormone therapy and reduced patient survival. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 217(1), 32–41.
- Gheorghe, M., Sandve, G. K., Khan, A., Cheneby, J., Ballester, B., & Mathelier, A. (2019). A map of direct tf-dna interactions in the human genome. *Nucleic acids research*, 47(4), e21–e21.
- Gillis, N. (2014). The why and how of nonnegative matrix factorization. *Regularization, optimization, kernels, and support vector machines*, 12(257), 257–291.
- Gong, L., & Nandi, A. K. (2013). An enhanced initialization method for non-negative matrix factorization. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, (pp. 1–6). IEEE.
- GSEA (2020). Msigdb: Molecular signatures database. Last accessed 2020-05-31. URL <https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>

- Hart, C. D., Migliaccio, I., Malorni, L., Guarducci, C., Biganzoli, L., & Di Leo, A. (2015). Challenges in the management of advanced, er-positive, her2-negative breast cancer. *Nature reviews Clinical oncology*, *12*(9), 541.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, *38*(4), 576–589.
- Hua, S., Kallen, C. B., Dhar, R., Baquero, M. T., Mason, C. E., Russell, B. A., Shah, P. K., Liu, J., Khramtsov, A., Tretiakova, M. S., et al. (2008). Genomic analysis of estrogen cascade reveals histone variant h2a. z associated with breast cancer progression. *Molecular systems biology*, *4*(1).
- Hugh, J., Hanson, J., Cheang, M. C. U., Nielsen, T. O., Perou, C. M., Dumontet, C., Reed, J., Krajewska, M., Treilleux, I., Rupin, M., et al. (2009). Breast cancer subtypes and response to docetaxel in node-positive breast cancer: use of an immunohistochemical definition in the bcirg 001 trial. *Journal of clinical oncology*, *27*(8), 1168.
- International Cancer Genome Consortium (ICGC) (2020). Icgc data portal. Last accessed 2020-06-01.
URL <https://dcc.icgc.org/>
- Itoh, M., Iwamoto, T., Matsuoka, J., Nogami, T., Motoki, T., Shien, T., Taira, N., Niikura, N., Hayashi, N., Ohtani, S., et al. (2014). Estrogen receptor (er) mrna expression and molecular subtype distribution in er-negative/progesterone receptor-positive breast cancers. *Breast cancer research and treatment*, *143*(2), 403–409.
- Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., et al. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*, *66*(21), 10292–10301.
- Janecek, A., & Tan, Y. (2011). Using population based algorithms for initializing nonnegative matrix factorization. In *International Conference in Swarm Intelligence*, (pp. 307–316). Springer.
- Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: a cancer journal for clinicians*, *61*(2), 69–90.

- Kamachi, Y., & Kondoh, H. (2013). Sox proteins: regulators of cell fate specification and differentiation. *Development*, *140*(20), 4129–4144.
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at ucsc. *Genome research*, *12*(6), 996–1006.
- Kim, H., & Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, *23*(12), 1495–1502.
- Kreftforeningen (2020). Om brystkreft [about breast cancer]. Last accessed 2020-05-26.
URL <https://kreftforeningen.no/rosasloyfe/om-brystkreft/>
- Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, *44*(W1), W90–W97.
- Kulis, M., Queirós, A. C., Beekman, R., & Martín-Subero, J. I. (2013). Intragenic dna methylation in transcriptional regulation, normal differentiation and cancer. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1829*(11), 1161–1174.
- Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., Silverstein, M. C., & Ma’ayan, A. (2018). Massive mining of publicly available rna-seq data from human and mouse. *Nature communications*, *9*(1), 1–10.
- Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The human transcription factors. *Cell*, *172*(4), 650–665.
- Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., et al. (2012). Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, *22*(9), 1813–1831.
- Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, (pp. 556–562).

- Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*(4), 493–500.
- Li, P., Piao, Y., Shon, H. S., & Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data. *BMC bioinformatics*, *16*(1), 347.
- Liu, P., Tang, H., Song, C., Wang, J., Chen, B., Huang, X., Pei, X., & Liu, L. (2018). Sox2 promotes cell proliferation and metastasis in triple negative breast cancer. *Frontiers in pharmacology*, *9*, 942.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, *15*(12), 550.
- Lui, G. Y., Grandori, C., & Kemp, C. J. (2018). Cdk12: an emerging therapeutic target for cancer. *Journal of clinical pathology*, *71*(11), 957–962.
- Malhotra, G. K., Zhao, X., Band, H., & Band, V. (2010). Histological, molecular and functional subtypes of breast cancers. *Cancer biology & therapy*, *10*(10), 955–960.
- Marsman, J., & Horsfield, J. A. (2012). Long distance relationships: enhancer-promoter communication and dynamic gene transcription. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1819*(11-12), 1217–1227.
- Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, *7*, 29–59.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Mehrgou, A., & Akouchejian, M. (2016). The importance of brca1 and brca2 genes mutations in breast cancer development. *Medical journal of the Islamic Republic of Iran*, *30*, 369.
- National Cancer Institute (2020). The chromatin accessibility landscape of primary human cancers: Supplemental data files. Last accessed 2020-06-01. URL <https://gdc.cancer.gov/about-data/publications/ATACseq-AWG>

- National Human Genome Research Institute (NIH) (2020). Genome. Last accessed 2020-04-13.
 URL <https://www.genome.gov/genetics-glossary/Genome>
- Nguyen, P. L., Taghian, A. G., Katz, M. S., Niemierko, A., Abi Raad, R. F., Boon, W. L., Bellon, J. R., Wong, J. S., Smith, B. L., & Harris, J. R. (2008). Breast cancer subtype approximated by estrogen receptor, progesterone receptor, and her-2 is associated with local and distant recurrence after breast-conserving therapy. *Journal of clinical oncology*, *26*(14), 2373–2378.
- Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S. R., Snider, J., Stijleman, I. J., Reed, J., et al. (2010). A comparison of pam50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clinical cancer research*, *16*(21), 5222–5232.
- Nowak, R. (1994). Mining treasures from 'junk dna.' (includes related glossary). *Science*, *263*(5147), 608–611.
- Park, P. J. (2009). Chip-seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, *10*(10), 669–680.
- Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, *27*(8), 1160.
- Pehkonen, P., Wong, G., & Törönen, P. (2005). Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC bioinformatics*, *6*(1), 162.
- Peng, A., Mao, X., Zhong, J., Fan, S., & Hu, Y. (2020). Single-cell multi-omics and its prospective application in cancer biology. *Proteomics*, (p. 1900271).
- Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *nature*, *406*(6797), 747–752.
- Phillips, T., & Shaw, K. (2008). Chromatin remodeling in eukaryotes. *Nature Education*, *1*(1), 209.
- Powe, D. G., Akhtar, G., Habashy, H. O., Abdel-Fatah, T., Rakha, E. A., Green, A. R., & Ellis, I. O. (2009). Investigating ap-2 and yy1 protein expression as

- a cause of high her2 gene transcription in breast cancers with discordant her2 gene amplification. *Breast Cancer Research*, 11(6), R90.
- Ramji, D. P., & Foka, P. (2002). Ccaat/enhancer-binding proteins: structure, function and regulation. *Biochemical Journal*, 365(3), 561–575.
- Rao, M. S., Van Vleet, T. R., Ciurlionis, R., Buck, W. R., Mittelstadt, S. W., Blomme, E. A., & Liguori, M. J. (2019). Comparison of rna-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. *Frontiers in genetics*, 9, 636.
- Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., & Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome biology*, 14(9), 3158.
- Reactome (2020). Rna polymerase i promoter opening. Last accessed 2020-05-31. URL <https://reactome.org/content/detail/R-HSA-73728>
- Ren, C., Chen, H., Yang, B., Liu, F., Ouyang, Z., Bo, X., & Shu, W. (2016). iform: incorporating find occurrence of regulatory motifs. *PloS one*, 11(12).
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140.
- Rodriguez-Pinilla, S. M., Sarrío, D., Moreno-Bueno, G., Rodríguez-Gil, Y., Martínez, M. A., Hernández, L., Hardisson, D., Reis-Filho, J. S., & Palacios, J. (2007). Sox2: a possible driver of the basal-like phenotype in sporadic breast cancer. *Modern pathology*, 20(4), 474–481.
- RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.
- Russnes, H. G., Lingjaerde, O. C., Børresen-Dale, A.-L., & Caldas, C. (2017). Breast cancer molecular stratification: from intrinsic subtypes to integrative clusters. *The American journal of pathology*, 187(10), 2152–2162.
- Sabatier, R., Goncalves, A., & Bertucci, F. (2014). Personalized medicine: present and future of breast cancer management. *Critical reviews in oncology/hematology*, 91(3), 223–233.

- Sauwen, N., Acou, M., Bharath, H., Sima, D., Veraart, J., Maes, F., Himmelreich, U., Achten, E., Van Huffel, S., & Biomedical, M. (2016). Initializing nonnegative matrix factorization using the successive projection algorithm for multi-parametric medical image segmentation.
- Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., & Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell*, *132*(5), 887–898.
- Shahnaz, F., Berry, M. W., Pauca, V. P., & Plemmons, R. J. (2006). Document clustering using nonnegative matrix factorization. *Information Processing & Management*, *42*(2), 373–386.
- Sheffield, N. C., & Bock, C. (2016). Lola: enrichment analysis for genomic region sets and regulatory elements in r and bioconductor. *Bioinformatics*, *32*(4), 587–589.
- Shepherd, J. H., Uray, I. P., Mazumdar, A., Tsimelzon, A., Savage, M., Hilsenbeck, S. G., & Brown, P. H. (2016). The sox11 transcription factor is a critical regulator of basal-like breast cancer growth, invasion, and basal-like gene expression. *Oncotarget*, *7*(11), 13106.
- Sherman, T., Gao, T., & Fertig, E. (2019). Cogaps 3: Bayesian non-negative matrix factorization for single-cell analysis with asynchronous updates and sparse data structures. *BioRxiv*, (p. 699041).
- Simicevic, J., Schmid, A. W., Gilardoni, P. A., Zoller, B., Raghav, S. K., Krier, I., Gubelmann, C., Lisacek, F., Naef, F., Moniatte, M., et al. (2013). Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. *Nature methods*, *10*(6), 570.
- Siziopikou, K. P., & Cobleigh, M. (2007). The basal subtype of breast carcinomas may represent the group of breast tumors that could benefit from egfr-targeted therapies. *The Breast*, *16*(1), 104–107.
- Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A. M., Yu, J., Klijn, J. G., Foekens, J. A., & Martens, J. W. (2008). Subtypes of breast cancer show preferential site of relapse. *Cancer research*, *68*(9), 3108–3114.
- Song, L., & Crawford, G. E. (2010). Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, *2010*(2), pdb-prot5384.

- Song, Q., Merajver, S. D., & Li, J. Z. (2015). Cancer classification in the genomic era: five contemporary problems. *Human genomics*, *9*(1), 27.
- Song, S.-H., & Kim, T.-Y. (2017). Ctf, cohesin, and chromatin in human cancer. *Genomics & informatics*, *15*(4), 114.
- Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., Goff, L. A., Li, Y., Ngom, A., Ochs, M. F., Xu, Y., & Fertig, E. J. (2018). Enter the matrix: Factorization uncovers knowledge from omics. *Trends in Genetics*, *34*(10), 790–805.
- Stingl, J., & Caldas, C. (2007). Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nature Reviews Cancer*, *7*(10), 791–799.
- Sun, E., Zhou, Q., Liu, K., Wei, W., Wang, C., Liu, X., Lu, C., & Ma, D. (2014). Screening mirnas related to different subtypes of breast cancer with mirnas microarray. *Eur Rev Med Pharmacol Sci*, *18*(19), 2783–2788.
- Sun, Y., Miao, N., & Sun, T. (2019). Detect accessible chromatin using atac-sequencing, from principle to applications. *Hereditas*, *156*(1), 1–9.
- Tang, H., Chen, B., Liu, P., Xie, X., He, R., Zhang, L., Huang, X., Xiao, X., & Xie, X. (2019). Sox8 acts as a prognostic factor and mediator to regulate the progression of triple-negative breast cancer. *Carcinogenesis*, *40*(10), 1278–1287.
- Teschendorff, A. E., Journée, M., Absil, P. A., Sepulchre, R., & Caldas, C. (2007). Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS computational biology*, *3*(8).
- Theodorou, V., Stark, R., Menon, S., & Carroll, J. S. (2013). Gata3 acts upstream of foxa1 in mediating esr1 binding by shaping enhancer accessibility. *Genome research*, *23*(1), 12–22.
- Tien, J. F., Mazloomian, A., Cheng, S.-W. G., Hughes, C. S., Chow, C. C., Canapi, L. T., Oloumi, A., Trigo-Gonzalez, G., Bashashati, A., Xu, J., et al. (2017). Cdk12 regulates alternative last exon mrna splicing and promotes breast cancer cell invasion. *Nucleic acids research*, *45*(11), 6698–6716.
- Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, *23*(1), 137–144.

- Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, *65*(2), 87–108.
- Tsompana, M., & Buck, M. J. (2014). Chromatin accessibility: a window into the genome. *Epigenetics & chromatin*, *7*(1), 33.
- Turner, B. C., Zhang, J., Gumbs, A. A., Maher, M. G., Kaplan, L., Carter, D., Glazer, P. M., Hurst, H. C., Haffty, B. G., & Williams, T. (1998). Expression of ap-2 transcription factors in human breast cancer correlates with the regulation of multiple growth factor signalling pathways. *Cancer research*, *58*(23), 5466–5472.
- UCSC (2020). Table browser. Last accessed 2020-06-13.
URL https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=797931347_STeclhFb3q7inyPbS0eTAXYXQJaA&clade=mammal&org=Human&db=hg38&hgta_group=genes&hgta_track=refSeqComposite&hgta_table=0&hgta_regionType=genome&position=chr1%3A11%2C102%2C837-11%2C267%2C747&hgta_outputType=bed&hgta_outFileName=ncbi_refseq_hg38.txt
- UniBind (2020). Unibind enrichment analysis. Last accessed 2020-06-15.
URL <https://unibind.uio.no/enrichment/>
- Vahedi, G., Takahashi, H., Nakayamada, S., Sun, H.-w., Sartorelli, V., Kanno, Y., & O’shea, J. J. (2012). Stats shape the active enhancer landscape of t cell populations. *Cell*, *151*(5), 981–993.
- Vailati-Riboni, M., Palombo, V., & Loor, J. J. (2017). What are omics sciences? In *Periparturient Diseases of Dairy Cows*, (pp. 1–7). Springer.
- van Arensbergen, J., van Steensel, B., & Bussemaker, H. J. (2014). In search of the determinants of enhancer–promoter interaction specificity. *Trends in cell biology*, *24*(11), 695–702.
- Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, *15*(3), 1066–1074.
- Waks, A. G., & Winer, E. P. (2019). Breast cancer treatment: a review. *Jama*, *321*(3), 288–300.
- Wang, C., Nie, Z., Zhou, Z., Zhang, H., Liu, R., Wu, J., Qin, J., Ma, Y., Chen, L., Li, S., et al. (2015). The interplay between tead4 and klf5 promotes breast

- cancer partially through inhibiting the transcription of p27kip1. *Oncotarget*, *6*(19), 17685.
- Wang, Y., Yin, Q., Yu, Q., Zhang, J., Liu, Z., Wang, S., Lv, S., & Niu, Y. (2011). A retrospective study of breast cancer subtypes: the risk of relapse and the relations with treatments. *Breast cancer research and treatment*, *130*(2), 489.
- Wang, Z., Gerstein, M., & Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57–63.
- Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, *5*(4), 276–287.
- Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, *45*(10), 1113.
- WHO (2018). Cancer. Last accessed 2020-02-21.
URL <https://www.who.int/news-room/fact-sheets/detail/cancer>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
URL <https://ggplot2.tidyverse.org>
- Wild, S., Curry, J., & Dougherty, A. (2004). Improving non-negative matrix factorizations through structured initialization. *Pattern recognition*, *37*(11), 2217–2232.
- Willis, S., De, P., Dey, N., Long, B., Young, B., Sparano, J. A., Wang, V., Davidson, N. E., & Leyland-Jones, B. R. (2015). Enriched transcription factor signatures in triple negative breast cancer indicates possible targeted therapies with existing drugs. *Meta gene*, *4*, 129–141.
- Woodfield, G. W., Chen, Y., Bair, T. B., Domann, F. E., & Weigel, R. J. (2010). Identification of primary gene targets of tfap2c in hormone responsive breast carcinoma cells. *Genes, Chromosomes and Cancer*, *49*(10), 948–962.
- Xu, J., Chen, Y., & Olopade, O. I. (2010). Myc and breast cancer. *Genes & cancer*, *1*(6), 629–640.
- Yan, F., Powell, D. R., Curtis, D. J., & Wong, N. C. (2020). From reads to insight: a hitchhiker’s guide to atac-seq data analysis. *Genome biology*, *21*(1), 22.

- Yang, K., Gao, J., & Luo, M. (2019). Identification of key pathways and hub genes in basal-like breast cancer using bioinformatics analysis. *OncoTargets and therapy*, *12*, 1319.
- Zaret, K. S., & Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes & development*, *25*(21), 2227–2241.
- Zhang, J., Liang, Q., Lei, Y., Yao, M., Li, L., Gao, X., Feng, J., Zhang, Y., Gao, H., Liu, D.-X., et al. (2012). Sox4 induces epithelial-mesenchymal transition and contributes to breast cancer progression. *Cancer research*, *72*(17), 4597–4608.
- Zhang, Z., Yang, C., Gao, W., Chen, T., Qian, T., Hu, J., & Tan, Y. (2015). Foxa2 attenuates the epithelial to mesenchymal transition by regulating the transcription of e-cadherin and zeb2 in human breast cancer. *Cancer letters*, *361*(2), 240–250.
- Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, *9*(1).
- Zhu, Q., Tekpli, X., Troyanskaya, O. G., & Kristensen, V. N. (2020). Subtype-specific transcriptional regulators in breast tumors subjected to genetic and epigenetic alterations. *Bioinformatics*, *36*(4), 994–999.

Attachments

Table S1: Gene signatures for each pattern, after feature selection. Each gene in the table has been chosen because it is pattern-specific. The cluster names are used as header, without the pattern number, in order to save space. Basal outliers = Pattern 1, Her2mix = Pattern 2, LumA/Normal = Pattern 3, LumA/B = Pattern 4 and Basal = Pattern 5.

Basal outliers	Her2mix	LumA/Normal	LumA/B	Basal
SNORD9	LACRT	CSN2	CPB1	CARD18
SNORA47	SULT1C3	C10orf96	CYP2A7	SERPINB3
SNORD67	UGT2B28	IRS4	UCN3	SPRR2D
SCARNA4	TBX10	ARHGAP36	CGA	MAGEA4
SNORD8	MUCL1	PROL1	CYP2A6	KRT79
SNORA28	SCGB2A2	SMR3B	GAGE12D	CYP2F1
SNORA66	LST-3TM12	OLFM3	KCNJ3	SERPINB4
SCARNA3	TGM4	ART4	MUC2	NCAN
SCARNA22	AICDA	CBLN2	UGT2B4	MAGEA9B
SNORA13	KRT12	SHISA2	TRH	CTAG1B
SCARNA23	CDH10	LALBA	CHGA	PAGE2
SCARNA11	KRT20	MYBPC1	CHGB	SPINK6
SNORA51	DCD	CIDEA	CPLX2	SPRR2E
SNORA41	FGA	FXYD1	CT45A1	GABBR2
SNORA58	SCGB2A1	CYP4F22	ASCL1	C4BPA
SCARNA14	HPD	SCARA5	SERPINA6	C1orf105
SNAR-A3	SCGB1D2	TUSC5	PAGE5	SERPINB13
RPPH1	FGG	LOC389033	SLIT1	CTAG2
SNORA71D	MYL1	HBA1	PHGR1	SPRR1B
SCARNA10	FGB	CIDEC	MYT1	LY6D
SNORA1	ERVFRDE1	PI16	CST5	LGALS7B
SNORA68	UGT2B11	MYL7	SGCG	KRTDAP
SNORA54	DLK1	GPD1	VTN	SLC1A6

Table S1 continued from previous page

Basal outliers	Her2mix	LumA/Normal	LumA/B	Basal
SNORD89	FAM177B	FOSB	BEX1	C4orf7
RNU6ATAC	WFDC6	GDF9	SEZ6L	CAPNS2
RMRP	APOD	ZNF385D	GRIA2	TRPM8
SNORA80	GLRA3	PLIN4	PCDH19	SPIB
SNORA38	ALB	EGR3	SYT13	PRSS33
SNORA50	TAT	ADH1B	NPY6R	SPRR2A
SCARNA1	FGL1	CXCL13	PCDH10	KRT83
HIST1H1A	CLEC4D	CCL14	RIMS4	KLK6
SNORA44	SPINK8	DARC	CYP2B7P1	IVL
SNORA48	AKR1B15	NDP	TPH1	AMTN
SNORA71A	ACE2	AQP7	ALDOB	LGALS7
SNORA42	CST4	LEP	HAO2	ECEL1
SCARNA18	XAGE1D	C2orf40	FAM5B	KRT6A
SNORA14A	C6orf223	PGLYRP2	MSMB	SERPINB7
SNORA12	PIP	TNN	PVALB	GFRA3
SNORA53	ALOX15B	RERGL	ANO2	KRT1
RNU4ATAC	PPP1R14D	HRASLS5	PCSK1	WIF1
SNORA16A	ACY3	SCGB3A1	SOX2	MAGEA10
HIST1H4L	HPGD	THBS4	TRPA1	BPI
SCARNA7	LRRC31	MFAP4	IL20	S100A8
SCARNA8	GABRB3	FABP4	PRSS1	SPINK5
SNORA37	ATP13A4	CRISPLD1	PDZRN4	C1QL4
SCARNA5	ATP13A5	C7	AGTR1	MLC1
SNORA63	CXCL17	SLC26A3	GDF15	SOX8
SNORA49	STAC2	ADAMTS16	TMPRSS6	ORM2
SNORA75	GLYATL2	PLAC9	CPA6	MSMP
RNU11	HMGCS2	CILP	GNG13	CASP14
TMEM14E	OLFM4	COL14A1	RTBDN	VGLL1
SNORA40	MUC20	PNMA2	TRY6	MIA
HIST1H3I	SLC26A4	ADIPOQ	AFP	AMY1A
SNORA14B	ABCC11	HBB	FAM25A	PRPH
SNORA9	TARP	MUM1L1	CACNA1H	NPB
SNORA78	SNPH	ANKRD43	LOC145837	CRABP1
SNORA81	PNLIPRP3	EGR1	NKAIN1	CCL20
SCARNA2	MUC6	PGM5	UPF0639	FZD9
SNORD17	PAX7	PYDC1	AMBP	CAPN6

Table S1 continued from previous page

Basal outliers	Her2mix	LumA/Normal	LumA/B	Basal
SNORA84	FAIM2	OGN	SLC5A8	PI3
SNORA23	TCAP	STK32B	CRISP3	CLDN10
HIST1H4C	CYP4Z2P	FAM155A	INHA	GJB6
SNORD10	GJB1	S100G	HSPB8	BCL2A1
SNORA57	PPP1R1B	TNXB	B4GALNT2	MAPK4
SNORA24	ST6GALNAC1	ABCA8	SERPINA5	LOC100124692
SNORA7B	PSCA	GPIHBP1	AKR7A3	MSLN
HIST1H1B	SERHL2	FGFR2	EEF1A2	MUC13
SNORA52	SLC25A18	FOS	SYT1	S100A7
SNORA20	LRRC26	NGFR	DNAJC12	SMOC1
SNORA74A	NUDT8	SSTR2	KCNK15	POMC
SNORA77		HSPB6	TSPAN8	S100A7A
SNORA69		NEK10	FAM196A	FGFBP1
SNORD97		PLIN1	BCAS1	MMP12
TERC		CDC20B	GDPD3	RAET1L
SNORA74B		AK5	ABP1	SLC26A9
SNORA27		MMRN1	FOXJ1	GTSF1
SNORA26		FCER1A	RGS22	KRT16
HIST1H2BI		COL17A1	BMPR1B	COL11A2
HIST1H4B		MGP	IFI27	S100A9
SNORA8		TIMP4	ROBO2	FBN3
HIST2H2AB		KCNIP2	C15orf59	PRR4
SNORA71C		SDPR	KRT13	KRT6C
SNORA70		ABCC8	MAT1A	FGFBP2
HIST1H3F		CTSG	ARC	SBSN
HIST1H2AB		ADAM33	ISG15	SIX3
SNORA32		DES	GFRA1	S100B
SNORA22		ZBTB16	COX6C	ROPN1
SNORD15B		PGR	KCNF1	KLK5
SNORA25		MMP23B	FLJ45983	C8orf46
SNORA61		SORCS2	ADCY1	CXorf61
SNORD94		ELN	CA12	ACTG2
SNORA18		ST8SIA6	NBPF4	CA9
SCARNA6		VSTM2A	ELOVL2	RARRES1
SNORA64		LRRC17	AFF3	COL9A3
SNORA10		PTN	IFI6	GJB3

Table S1 continued from previous page

Basal outliers	Her2mix	LumA/Normal	LumA/B	Basal
PRO0611		ABCC13	ACTL8	S100A2
SNORA3		SCN4B	DIO1	EPHB1
HIST1H4A		HBA2	TFF1	PKP1
SNORA72		TPSAB1	SEMA3B	TMSB15A
SNORA79		PNMT	COL2A1	KLK8
SNORA2B		SLC40A1	NEURL	MARCO
HIST1H4D		CCDC8	C10orf82	MT1H
HIST1H1E		MYH11	CAPSL	CRLF1
SNORA38B		CHRD1	CST9	CALML3
SCARNA12		WISP2	IGSF1	KRT81
SNORA19		BTG2	INSM1	VCAM1
SNORA71B		HSPB7	KIF5C	KIF1A
SNORA34		GALNTL1	CST2	DNER
HIST1H3B		NPY1R	TMEM150C	GAL
SNORA6		PTGER3	CXCL14	FAM3D
SNORA5A		SCUBE2	TNNT1	IGF2
SCARNA9		CYP4Z1	TUBA3E	PTGS2
SNORA15		GPR162	TFF3	SYT8
SNORA46		CADM3	SPAG6	KRT6B
SNORA4		GRP	RAMP1	EPHX3
HIST2H3C		TFAP2B	C20orf85	CXCL1
HIST1H3A		NOVA1	WNK4	GPR64
SNORA31		PAMR1	CHAD	KRT14
SNORA11D		PTHLH	SLC16A6	SCRG1
SNORA62		TNNT3	SERPINI1	PPP1R14C
SCARNA20		GRPR	CLSTN2	CHI3L2
SNORA45		MS4A2	DOK7	GSDMC
SCARNA21		STC2	IGFALS	SOX10
SNORA11B		OXTR	LINGO1	LEMD1
HIST1H2AJ		PDE8B	CACNG4	GABRE
HIST1H4E		COL4A6	CEACAM6	UBD
SNORA5C		SAA2	GNMT	IDO1
WDR74		TPSB2	C6orf141	DEFB1
HIST1H2AL		WNT5A	HSH2D	IRX1
HIST1H2AH		SLC19A3	KCNE4	SLC6A14
SNORA55		MYL2	FSIP1	CHI3L1

Table S1 continued from previous page

Basal outliers	Her2mix	LumA/Normal	LumA/B	Basal
CDR1		MATN3	TUBA3D	SPRR1A
SNORA65		RDH5	PADI3	C2orf82
SCARNA27		CPA3	LRP2	GJB5
SCARNA16		PODN	TPRXL	TTYH1
SNORA2A		AOX1	GATA3	A2ML1
SNORA21		CKM	ADRA2C	SOSTDC1
SNORA76		CITED1	DNALI1	LGR6
SNORA11		ITGA10	SLC6A4	APCDD1L
HIST1H1D		NCAM2	CEACAM5	KRT5
HIST1H2BL		C20orf103	C16orf89	TUBB2B
MALAT1		NME5	ESR1	BCL2L14
SNORD15A		PI15	CORO6	FAM150B
HIST2H2AC		PDK4	MPP2	IL20RB
ABCA13		SELP	CNTD2	TRIM29
SNORA36A		FST	FAM134B	UCA1
DSG1		GP2	RBM24	CDKN2A
SNHG7		CLEC3B	MAGEA1	XDH
SCARNA17		PLEKHA4	RGS11	GPR87
HIST4H4		CLDN11	DACH1	MAGEA2
HIST2H3D		PLA2G2A	LIN7A	HRCT1
HIST1H3J		MEOX1	PRAME	GDF5
HIST1H2AD		CLDN5	PCSK6	EN1
PCDH11X		LRRN1	CCDC48	GABRP
HIST1H2BF		MT1M	NAT2	KRT17
SNORA59B		LAMA2	L1CAM	SLC15A1
SNORD22		NPY5R	SLC9A3R1	PCP4
SNORA11E		EGR2	SLC7A2	ORM1
SCGB3A2		SSC5D	C19orf21	CD1A
ART3		NNAT	SYBU	CRTAC1
HORMAD1		RAI2	LONRF2	COL22A1
GPR12		INMT	STARD10	EGFR
ZNF460		SUSD3	REPS2	FERMT1
C6orf15		FAM189A2	GREB1	AQP5
ANKRD36BP1		SYT9	ANKRD29	RHCG
C11orf90		SEMA6D	MKX	KCNK5
CCDC144A		SFRP4	LYPD6B	PCOLCE2

Table S1 continued from previous page

Basal outliers	Her2mix	LumA/Normal	LumA/B	Basal
DSG3		NTRK2	THSD4	LAMB3
SPDYE8P		SLC30A8	NAT1	NCCRP1
		LRRC4C		RGMA
		CNTN1		ROPN1B
		KLK4		C6orf176
		ANKRD30A		MAGEA6
		LTC4S		RBP4
		OMD		ANXA8L2
		AMPH		FABP7
		RANBP3L		CPA4
		GALNTL2		CWH43
		TMEM100		EPHB6
		TP63		SELE
		LYPD6		DKK1
		ASPN		YJEFN3
		CFD		CRYAB
		NRK		sep.03
		PLIN5		LOC84740
		LPL		ELF5
		UPK1A		ANKRD35
		LRRC48		FAM83A
		ITIH5		DLX5
		NELL2		IGF2BP2
		IQCA1		HLA-DOB
		CYP4F8		SFRP1
		FAM38B		KRT4
		TSPAN7		LBP
		LOC642587		DSC3
		ADCY5		PRSS12
		MAPT		MFI2
		CNN1		CLCA2
		ZNF423		CSAG3
		ERBB4		TMCC2
		TMEM26		C1orf186
		ZFP36		NRTN
		CA2		

Table S1 continued from previous page

Basal outliers	Her2mix	LumA/Normal	LumA/B	Basal
				TGFBR3
				HOXA7
				KCNJ11
				ATP1A4
				GRB14
				CYR61
				SHROOM1
				CPXM1
				NEFH
				NTN4
				ISM1

Full HOMER results for the RNA-seq data

The HOMER results in Table S2-S4 show all enriched TFBS motifs for the gene signatures of Pattern 1 (Basal outliers), Pattern 2 (Her2mix) and Pattern 3 (LumA/Normal).

Table S2: All enriched TFBS motifs for Pattern 1 - Basal outliers from the RNA-seq data.

Rank	Motif	Name	P-value	No. of target sequences with motif
1		E2F(E2F)/Hela-CellCycle-Expression/Homer	1e-3	9
2		E2F7(E2F)/Hela-E2F7-ChIP-Seq(GSE32673)/Homer	1e-3	11
3		RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq(GSE29180)/Homer	1e-2	36
4		NFY(CCAAT)/Promoter/Homer	1e-2	44
5		HNF6(Homeobox)/Liver-Hnf6-ChIP-Seq(ERP000394)/Homer	1e-2	24
6		Oct2(POU,Homeobox)/Bcell-Oct2-ChIP-Seq(GSE21512)/Homer	1e-2	17
7		RUNX2(Runt)/PCa-RUNX2-ChIP-Seq(GSE33889)/Homer	1e-2	29
8		Nanog(Homeobox)/mES-Nanog-ChIP-Seq(GSE11724)/Homer	1e-2	108
9		TATA-Box(TBP)/Promoter/Homer	1e-2	50

Table S3: All enriched TFBS motifs for Pattern 2 - Her2mix from the RNA-seq data.

Rank	Motif	Name	P-value	No. of target sequences with motif
1		NF1.FOXA1(CTF,Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-2	4
2		FOXA1(Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-2	21
3		Esrrb(NR)/mES-Esrrb-ChIP-Seq(GSE11431)/Homer	1e-2	12
4		Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer	1e-2	15
5		Nr5a2(NR)/Pancreas-LRH1-ChIP-Seq(GSE34295)/Homer	1e-2	12
6		Nr5a2(NR)/mES-Nr5a2-ChIP-Seq(GSE19019)/Homer	1e-2	10
7		Atf4(bZIP)/MEF-Atf4-ChIP-Seq(GSE35681)/Homer	1e-2	8

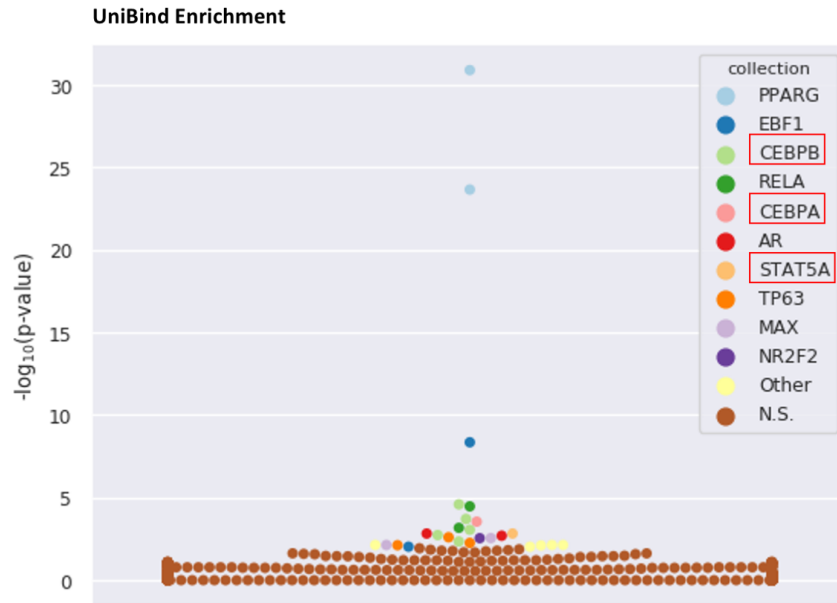
Table S4: All enriched TFBS motifs for Pattern 3 - LumA/Normal from the RNA-seq data.

Rank	Motif	Name	P-value	No. of target sequences with motif
1		CArG(MADS)/PUER-Srf-ChIP-Seq(Sullivan et al.)/Homer	1e-3	19
2		TATA-Box(TBP)/Promoter/Homer	1e-2	50
3		HOXD13(Homeobox)/Chicken-Hoxd13-ChIP-Seq(GSE38910)/Homer	1e-2	30
4		NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq(Unpublished)/Homer	1e-2	70
5		PRDM1(Zf)/Hela-PRDM1-ChIP-Seq(GSE31477)/Homer	1e-2	23

UniBind with background and HOMER supplementary results

Figure 6.1-6.3 show the HOMER and UniBind results for Pattern 3, 4 and 5.

Pattern 3: LumA/Normal TFBS Enrichment Results (ATAC-seq)

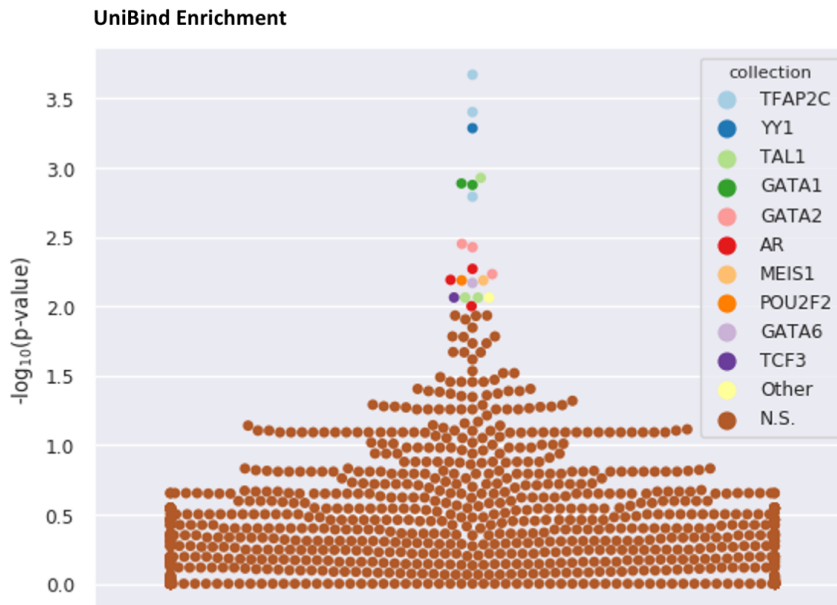


Homer Known Motif Enrichment

	Motif	Name	P-value	No. of target sequences with motif
1		CEBP:AP1(bZIP)/ThioMac-CEBPb-ChIP-Seq(GSE21512)/Homer	1e-5	27
2		Stat3+il21(Stat)/CD4-Stat3-ChIP-Seq(GSE19198)/Homer	1e-5	24
3		Atf4(bZIP)/MEF-Atf4-ChIP-Seq(GSE35681)/Homer	1e-5	15
4		Stat3(Stat)/mES-Stat3-ChIP-Seq(GSE11431)/Homer	1e-5	19
5		TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	1e-4	28
6		STAT4(Stat)/CD4-Stat4-ChIP-Seq(GSE22104)/Homer	1e-4	28
7		STAT1(Stat)/HelaS3-STAT1-ChIP-Seq(GSE12782)/Homer	1e-4	13
8		CEBP(bZIP)/ThioMac-CEBPb-ChIP-Seq(GSE21512)/Homer	1e-4	24
9		GRE(NR),IR3/A549-GR-ChIP-Seq(GSE32465)/Homer	1e-4	8
10		STAT5(Stat)/mCD4+-Stat5-ChIP-Seq(GSE12346)/Homer	1e-3	14

Figure 6.1: Enriched TFBSs for the Pattern 3 cluster of peaks, which are most important in Normal-like and most Luminal A samples. Top: Top 10 TFs with enriched TFBS sets. Bottom: Top 10 enriched TFBS motifs with corresponding TF name in bold.

Pattern 4: Her2 TFBS Enrichment Results (ATAC-seq)

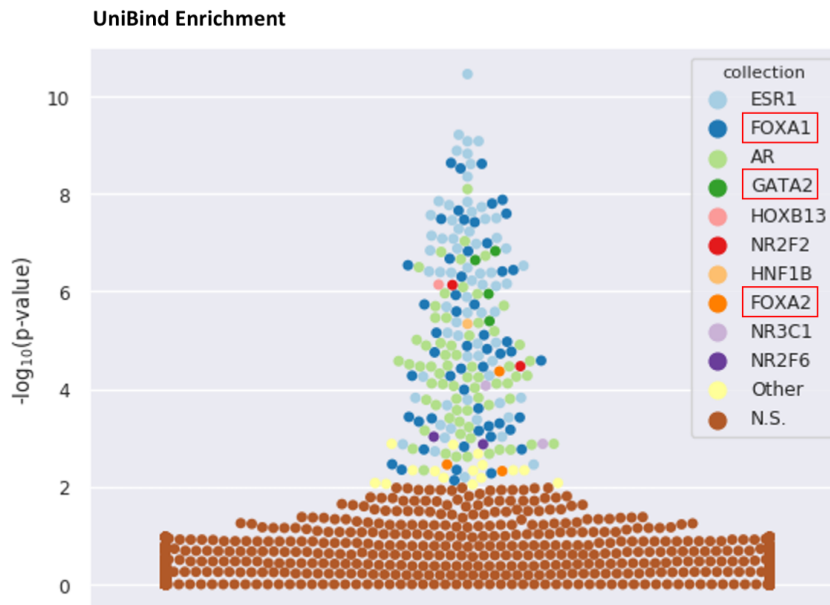


Homer Known Motif Enrichment

	Motif	Name	P-value	No. of target sequences with motif
1		FOXA1 (Forkhead)/MCF7-FOXA1-ChIP-Seq(GSE26831)/Homer	1e-5	27
2		FOXA1 (Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-5	30
3		AP-2gamma (AP2)/MCF7-TFAP2C-ChIP-Seq(GSE21234)/Homer	1e-4	29
4		Fox:Ebox (Forkhead,bHLH)/Panc1-Foxa2-ChIP-Seq(GSE47459)/Homer	1e-3	24
5		Six1 (Homeobox)/Myoblast-Six1-ChIP-Chip(GSE20150)/Homer	1e-3	9
6		Tlx? (NR)/NPC-H3K4me1-ChIP-Seq(GSE16256)/Homer	1e-3	15
7		AP-2alpha (AP2)/Hela-AP2alpha-ChIP-Seq(GSE31477)/Homer	1e-2	22
8		Foxa2 (Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer	1e-2	18
9		PHA-4 (Forkhead)/cElegans-Embryos-PHA4-ChIP-Seq(modEncode)/Homer	1e-2	52
10		NF1 (CTF)/LNCAP-NF1-ChIP-Seq(Unpublished)/Homer	1e-2	13

Figure 6.2: Enriched TFBSs for the Pattern 4 cluster of peaks, which are most important in Her2 samples. Top: Top 10 TFs with enriched TFBS sets. Bottom: Top 10 enriched TFBS motifs with corresponding TF name in bold.

Pattern 5: LumA TFBS Enrichment Results (ATAC-seq)



Homer Known Motif Enrichment

	Motif	Name	P-value	No. of target sequences with motif
1		FOXA1 (Forkhead)/MCF7-FOXA1-ChIP-Seq(GSE26831)/Homer	1e-21	75
2		FOXA1 (Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-19	79
3		Foxa2 (Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer	1e-12	51
4		ERE(NR),IR3/MCF7-ERa-ChIP-Seq(Unpublished)/Homer	1e-12	22
5		PHA-4 (Forkhead)/cElegans-Embryos-PHA4-ChIP-Seq(modEncode)/Homer	1e-11	112
6		Fox:Ebox (Forkhead,bHLH)/Panc1-Foxa2-ChIP-Seq(GSE47459)/Homer	1e-11	52
7		NF1:FOXA1 (CTF,Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-4	7
8		GATA (Zf),IR4/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer	1e-4	8
9		AMYB (HTH)/Testes-AMYB-ChIP-Seq(GSE44588)/Homer	1e-3	45
10		MYB (HTH)/ERMVB-Myb-ChIPSeq(GSE22095)/Homer	1e-3	48

Figure 6.3: Enriched TFBSs for the Pattern 5 cluster of peaks, which are most important in Luminal A and some Luminal B samples. Top: Top 10 TFs with enriched TFBS sets. Bottom: Top 10 enriched TFBS motifs with corresponding TF name in bold.

UniBind with no background

Figure 6.4-6.8 show the output of the UniBind Enrichment analysis with no background, for each of the patterns in the ATAC-seq data.

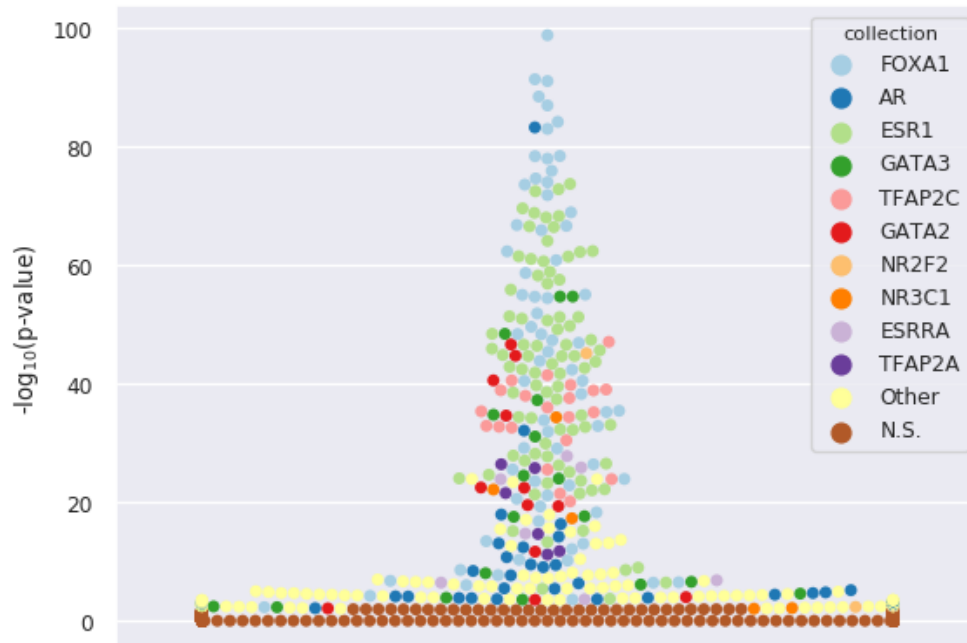


Figure 6.4: UniBind Enrichment without background for Pattern 1: LumA/B from the ATAC-seq data.

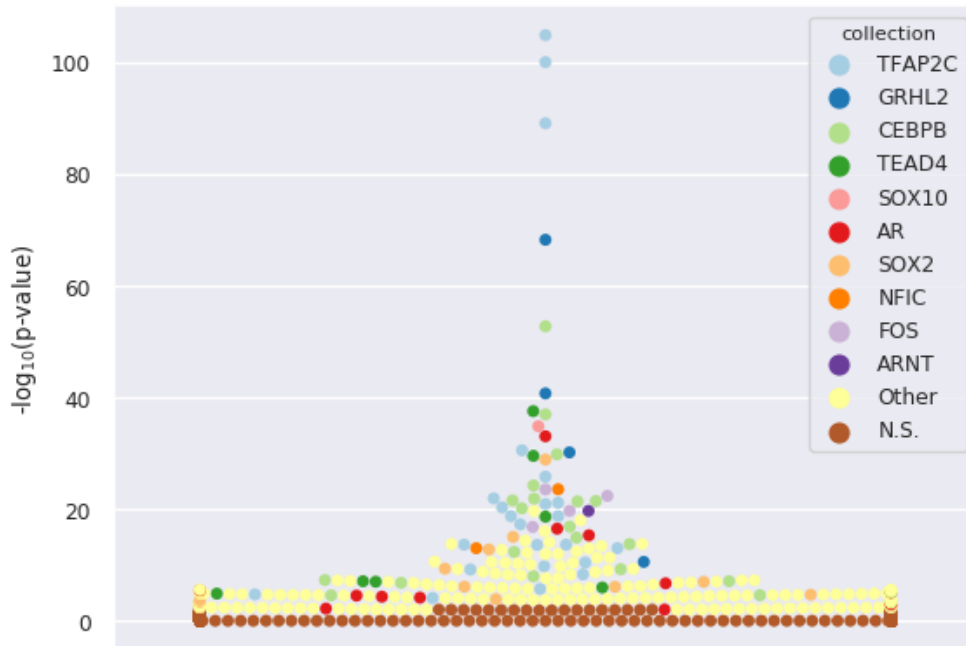


Figure 6.5: UniBind Enrichment without background for Pattern 2: Basal from the ATAC-seq data.

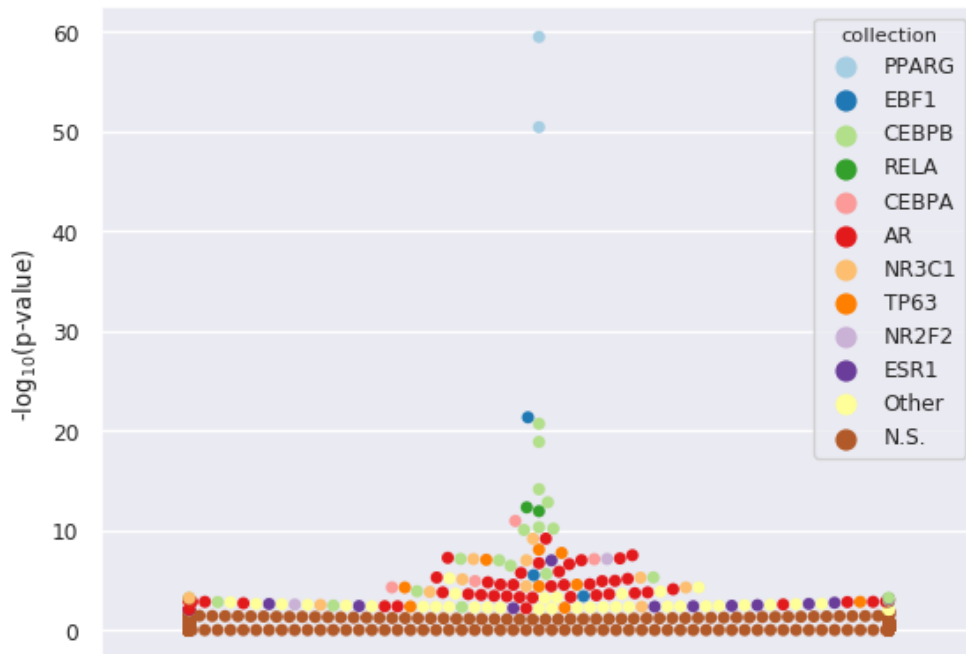


Figure 6.6: UniBind Enrichment without background for Pattern 3: LumA/Normal from the ATAC-seq data.

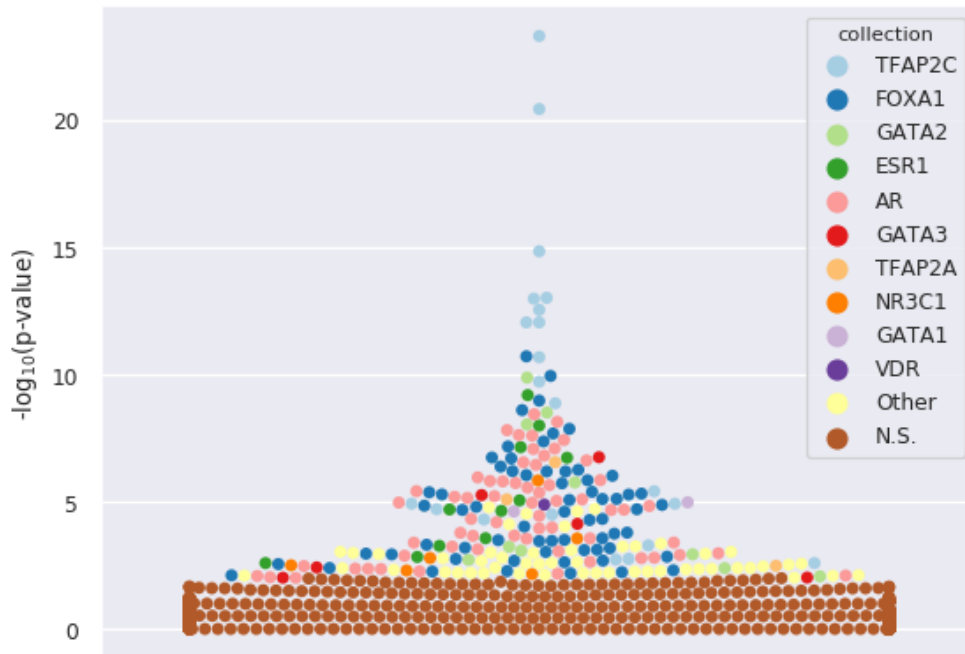


Figure 6.7: UniBind Enrichment without background or Pattern 4: Her2 from the ATAC-seq data.

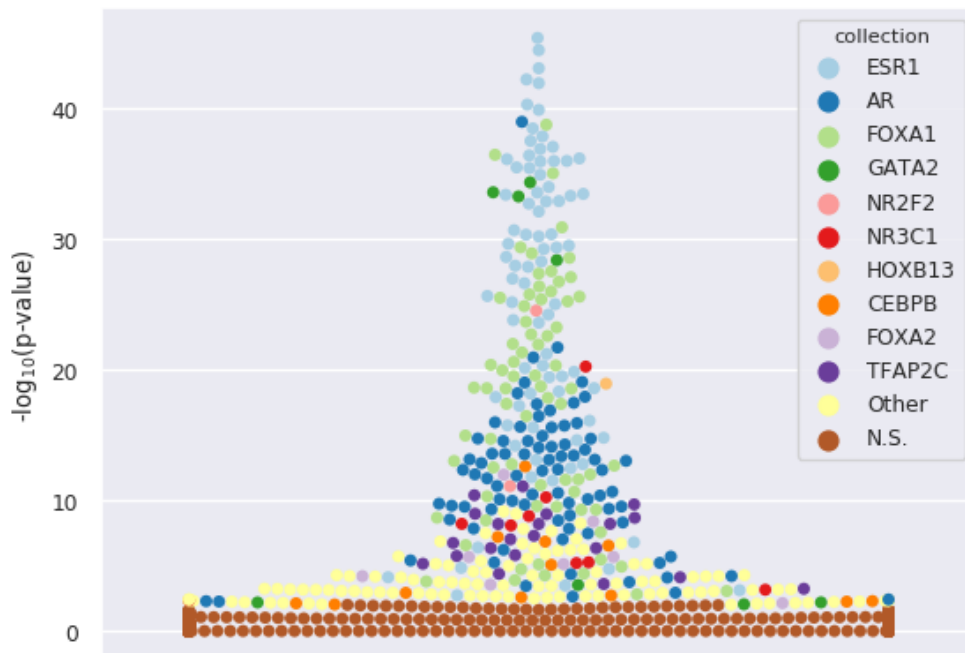


Figure 6.8: UniBind Enrichment without background for Pattern 5: LumA from the ATAC-seq data.

Cluster assignment

The cluster assignment of the samples are shown in Table S5. The table shows cluster names, together with the "true" PAM50 subtype.

Table S5: Comparison of cluster assignments for the ATAC-seq data and RNA-seq data.

	sample	clusterATAC	clusterRNA	truePAM50
1	TCGA-BH-A0E0-01	Basal	Basal	Basal
2	TCGA-A7-A13E-01	Basal	Basal	Basal
3	TCGA-A2-A0SX-01	Basal	Basal	Basal
4	TCGA-A2-A0YJ-01	Basal	Basal	Basal
5	TCGA-A2-A4RX-01	Basal	Basal	Basal
6	TCGA-AO-A124-01	Basal	Basal	Basal
7	TCGA-AO-A12F-01	Basal	Basal	Basal
8	TCGA-AR-A0TP-01	Basal	Basal	Basal
9	TCGA-C8-A12K-01	Basal	Basal	Basal
10	TCGA-C8-A12V-01	Basal	Basal	Basal
11	TCGA-S3-AA0Z-01	Basal	Basal	Basal
12	TCGA-AR-A0U0-01	Basal	Basal	Basal
13	TCGA-AR-A0U4-01	Basal	Basal	Basal
14	TCGA-BH-A0DL-01	Basal	Basal	Basal
15	TCGA-C8-A12Q-01	Her2	Her2mix	Her2
16	TCGA-A2-A0CX-01	Her2	Her2mix	Her2
17	TCGA-A8-A08J-01	Her2	LumAB	Her2
18	TCGA-A8-A094-01	Her2	Her2mix	Her2
19	TCGA-AO-A12D-01	Her2	Her2mix	Her2
20	TCGA-C8-A137-01	Her2	LumAB	Her2
21	TCGA-BH-A0EE-01	Her2	Basal	Her2
22	TCGA-BH-A1EV-01	Her2	Her2mix	Her2
23	TCGA-D8-A13Z-01	Her2	Basal	Her2
24	TCGA-C8-A12T-01	LumAB	LumANormal	Her2
25	TCGA-A2-A0ES-01	LumANormal	LumANormal	LumA
26	TCGA-A7-A0D9-01	LumAB	LumAB	LumA
27	TCGA-AO-A0J5-01	LumANormal	LumANormal	LumA
28	TCGA-AO-A0JG-01	LumANormal	LumAB	LumA
29	TCGA-BH-A0B1-01	LumAB	LumANormal	LumA
30	TCGA-BH-A0B5-01	LumA	LumANormal	LumA
31	TCGA-C8-A12O-01	LumAB	Her2mix	LumA

Table S5 continued from previous page

	sample	clusterATAC	clusterRNA	truePAM50
32	TCGA-A2-A0T4-01	LumANormal	LumANormal	LumA
33	TCGA-A2-A0T5-01	LumA	LumANormal	LumA
34	TCGA-AO-A03L-01	LumAB	LumAB	LumA
35	TCGA-A2-A0ET-01	LumANormal	LumAB	LumA
36	TCGA-A2-A0EV-01	LumANormal	Her2mix	LumA
37	TCGA-A2-A0T6-01	LumANormal	LumANormal	LumA
38	TCGA-A2-A0YC-01	LumA	LumAB	LumA
39	TCGA-A2-A0YD-01	LumA	LumANormal	LumA
40	TCGA-A2-A0YL-01	LumANormal	LumANormal	LumA
41	TCGA-AO-A0J8-01	LumA	LumANormal	LumA
42	TCGA-BH-A0BA-01	LumANormal	Her2mix	LumA
43	TCGA-BH-A0C1-01	LumANormal	LumANormal	LumA
44	TCGA-BH-A0DP-01	LumANormal	LumANormal	LumA
45	TCGA-BH-A0HP-01	LumANormal	LumANormal	LumA
46	TCGA-C8-A12Y-01	LumA	LumAB	LumA
47	TCGA-C8-A133-01	LumA	LumAB	LumA
48	TCGA-A2-A0EX-01	LumAB	LumANormal	LumA
49	TCGA-A2-A0T7-01	LumAB	LumANormal	LumA
50	TCGA-A7-A0CH-01	LumA	LumANormal	LumA
51	TCGA-AQ-A04L-01	LumANormal	LumANormal	LumA
52	TCGA-BH-A0DV-01	LumA	LumANormal	LumA
53	TCGA-A2-A0YF-01	LumA	LumAB	LumA
54	TCGA-A2-A0EW-01	LumANormal	LumANormal	LumA
55	TCGA-A2-A0EY-01	Her2	Her2mix	LumB
56	TCGA-A2-A0YH-01	LumAB	LumAB	LumB
57	TCGA-BH-A0BZ-01	LumAB	LumAB	LumB
58	TCGA-A2-A0SV-01	LumA	LumAB	LumB
59	TCGA-A2-A0YG-01	LumANormal	LumAB	LumB
60	TCGA-AO-A03N-01	LumANormal	Her2mix	LumB
61	TCGA-AO-A0JM-01	Her2	LumAB	LumB
62	TCGA-A2-A0SW-01	Her2	LumAB	LumB
63	TCGA-A7-A13F-01	LumA	LumAB	LumB
64	TCGA-C8-A12M-01	LumAB	LumAB	LumB
65	TCGA-C8-A130-01	LumA	Basal	LumB
66	TCGA-A2-A0YT-01	LumAB	LumAB	LumB
67	TCGA-AR-A0TV-01	LumAB	LumAB	LumB

Table S5 continued from previous page

	sample	clusterATAC	clusterRNA	truePAM50
68	TCGA-C8-A12U-01	LumAB	LumAB	LumB
69	TCGA-A2-A0YK-01	LumANormal	LumANormal	Normal
70	TCGA-AO-A0JB-01	LumANormal	LumANormal	Normal

Annotation of peaks

The annotation of peaks are shown in Table S6.

Table S6: Distribution of peak region annotations for the top peaks of the ATAC-seq data. Annotation data is retrieved from [\cite{corces2018chromatin}](#).

perc3UTR	perc5UTR	percDistal	percExon	percIntron	percPromoter
1.554	0.194	42.681	2.008	50.259	3.303



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitenskapelige universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway