Transcription factor-mediated gene regulation of chromosome 1q in breast cancer

by

Ferzan Betül Berber

A Dissertation Submitted to the
Graduate School of Sciences and Engineering

in Partial Fulfillment of the Requirements for

the Degree of

Master of Science

in

Molecular Biology and Genetics

August 12, 2022

Transcription factor-mediated gene regulation of chromosome 1q in breast cancer

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Ferzan Betül Berber

and have found that it is complete and satisfactory in all respects

and that any and all revisions required by the final

examining committee have been made.

Committee Members:

Asst. Prof. Gözde KORKMAZ (Advisor)

Asst. Prof. Hamzah SYED

Prof. Dr. Elif Damla ARISAN

Date: 12.08.2022

*I dedicate this thesis to the **initiative** of letting*
fear, doubt and disbelief go:
***freeing the mind**...*
*And the people who have contributed to making **it** happen,*
for me, and for themselves.

ABSTRACT

Transcription factor-mediated gene regulation of chromosome 1q in breast cancer

Ferzan Betül Berber

Master of Science in Molecular Biology and Genetics

August 12, 2022

Breast cancer is the most commonly diagnosed cancer among women and the fifth cause of cancer deaths in the world. Approximately 80% of breast cancer patients are estrogen receptor alpha (ERα) positive, which is a sequence-specific transcription factor that interacts with DNA either directly by estrogen response elements (EREs) or recruited by other transcription factors (TFs). Forkhead Box A1 (FOXA1) and GATA Binding Protein 3 (GATA3) are two of the major factors that regulate ERα binding and their abnormal expression is associated specifically with luminal A subtype breast cancer. Copy Number Variation (CNV) is a structural variant in the human genome, and chromosome 1 is shown to be one of the most altered regions of breast cancer cells. Consequently, more than 70% of breast tumors display copy number gain in the 1q arm.

In this study, the aim is to investigate the association of ERα-GATA3-FOXA1 cistrome with copy number amplifications of chromosome 1q and, finally, gene regulation. For this purpose, chromatin immunoprecipitation followed by sequencing (ChIP-seq) focused analysis of ERα, FOXA1, and GATA3 binding sites, RNA-seq data for the gene expression in MCF7, T47D, and ZR75-1 cell lines, and copy number data collected from Cancer Cell Line Encyclopedia are used. After obtaining the direct and indirect binding sites of targeted TFs, along with the overlapping binding sites among the cell lines, chromosome 1q cytobands with distinguished binding patterns are determined. Custom-made CRISPR/Cas9 libraries are generated in order to target those genomic regions in distinct ways. ChIP-seq and RNA-seq data association focused on specific cytobands revealed upregulated genes that might be promising candidates for further studies.

ÖZETÇE

Meme kanserinde kromozom 1q bölgesinin transkripsiyon faktörü aracılı gen regülasyonu

Ferzan Betül Berber
Moleküler Biyoloji ve Genetik, Yüksek Lisans
12 Ağustos 2022

Meme kanseri, kadınlarda en sık görülen kanser türü olmakla birlikte dünyada kansere bağlı ölümlerin beşinci nedenidir. Meme kanserlerinin yaklaşık %80'i, DNA ile doğrudan etkileşime östrojen yanıt elemanları (ERE) veya başka transkripsiyon faktörleri (TF) sayesinde giren diziye özgü transkripsiyon faktörü olan östrojen reseptör α (ERα) pozitiftir. Forkhead Box A1 (FOXA1) ve GATA Binding Protein 3 (GATA3), ERα bağlanmasını düzenleyen başlıca transkripsiyon faktörlerinden ikisidir ve bunların anormal ekspresyonu, özellikle luminal A moleküler alt tipi meme kanseri ile ilişkilidir. Kopya Sayısı Varyantı (CNV), insan genomunda meydana gelen yapısal bir varyanttır ve meme kanseri hücrelerinde en çok değişen bölgelerden birinin 1. kromozom olduğu daha önce gösterilmiştir. Buna bağlı olarak, meme kanserlerinin %70'inden fazlası 1. kromozomun q kolunda kopya sayısı farklılığı (artışı) gösterir.

Bu çalışmada, ERα-GATA3-FOXA1 cistromunun kromozom 1q kopya sayısı artışı ile ilişkisi ve buna bağlı olabilecek muhtemel gen regülasyonu değişikliklerinin araştırılması amaçlanmıştır. Bu doğrultuda, kromatin immünopresipitasyon dizileme (ChIP-seq) analizi ile belirlenen ERα, FOXA1 ve GATA3 bağlanma bölgeleri, RNA dizileme (RNA-seq) analizi ile elde edilen MCF7, T47D ve ZR75-1 hücre hatlarındaki gen ekspresyonu verileri ve Cancer Cell Line Encyclopedia (CCLE) ' dan elde edilen kopya sayısı varyant verileri kullanılmıştır. Hedeflenen TF'lerin doğrudan ve dolaylı bağlanma bölgelerinin yanı sıra hücre hatları arasındaki örtüşen bağlanma bölgelerinin elde edilmesinden sonra, kromozom 1q da belirgin bağlanma gösteren sitobandlar belirlenmiştir. Bu genomik bölgelerin fonksiyonlarını farklı bağlamlarda hedeflemek için CRISPR/Cas9 kütüphaneleri oluşturulmuştur. Belirli sitobandlara odaklanan ChIP-seq ve RNA-seq verisi arasındaki ilişkisinin araştırılması, daha ileri çalışmalar için umut verici adaylar olabilecek yukarı regüle edilmiş genleri ortaya çıkarmıştır.

AKNOWLEDGEMENTS

Firstly, I would like to express my gratitude to my advisor, Assistant Professor Gözde Korkmaz, for giving me the space to built my abilities in the field of bioinformatics, made me able to understand what I am dealing with with her patient explainings, as well as guiding me to pursue scientific thinking. She is a role model for me with her endless scientific curiosity, her patience and respect to her students with the willingness of teaching. It would not be possible to complete my master's thesis without her leading.

I also wish to express my thanks to Asst. Prof. Hamzah Syed and Prof. Dr. Elif Damla Arısan for accepting to be in my thesis defense jury.

I would like to thank my co-advisor, Emrah Nikerel and project advisor, Hamzah Syed, for sharing their knowledge. They guided me to see a flare while swimming in the immense sea of bioinformatics. I also thank Ufuk Yılmaz who had been the KUACC System Manager during most of my master's degree. I would not be able to manage the difficulties of working in another reality, a supercomputer, without his help. He was always kind and constructive towards my questions.

I met a few people that I always felt their warmness and friendship. After long COVID-19 quarantine times spent at home, they helped me to reorient myself to socialize in real life. I am specifically thankful to them for the scientific and philosophical conversations.

I would also like to thank my lab mates, Gamze Turan, Elif Yapıcı, Hande Özkan, Elif Kansız, Merve Kayış, Elif Güzar, and Dr. Gülfem Öztürk for their support during my master's and thesis writing period. We shared a lot of warm memories together.

Undoubtedly, all the achievements in my life would not be possible without the unconditional love and belief of my family. I am more than grateful for their encouragement not only during the most desperate times of my master's during the quarantine times but through my whole life. My mother, my brothers, my whole family, and especially NB, their presence made me come through all the obstacles.

Last but not the least, I am blessed to have Yunus in my life. He is not only the best scientific partner but also the ultimate source of happiness, tranquility, knowledge, kindness and patience. We have been walking the unique life roads of ours together for a while and it is a journey full of joy. He is Dominique to my Roark, Japanese culture to my cyberpunk mind.

TABLE OF CONTENTS

List of Tables

List of Figures

ABBREVIATIONS

ChIP-seq	Chromatin Immunoprecipitation Sequencing
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
GATA3	GATA Binding Protein 3
GEO	Gene Expression Omnibus
TF	Transcription Factor
GATA3	Single Nucleotide Variants
SNVs	Copy Number Variation
RNA-seq	RNA sequencinq
NHEJ	Non-Homologous End Joining
HDR	Homology-Directed Repair
PAM	Protospacer Adjacent Motif
TCGA	The Cancer Genome Atlas
NGS	Next Generation Sequencing
sgRNA	Single Guide RNA
FOXA1	Forkhead Box A1

Chapter 1:

INTRODUCTION

1.1 Breast Cancer

Breast cancer is the most common cancer type among women [1] and the fifth leading cause of cancer mortality [2]. In 2020, female breast cancer surpassed lung cancer as being the most commonly diagnosed cancer type, with an estimated 2.3 million new cases (11.7%) among both genders globally [2].

As one of the solid tumor types, breast cancer has a dynamic and heterogeneous nature both in an inter- and intra-tumoral manner. These cell populations evolve over time and may undergo tumorigenesis upon the emergence of mutagenesis and structural rearrangements [3–6]. Due to this heterogeneity, the classification of breast cancer is crucial to achieve successful clinical outcomes. There are several distinct classifications relying on different characteristics of breast cancer. According to its different molecular characteristics and gene expression profiles, breast cancer is divided into four main molecular (also called intrinsic) subtypes; luminal A, luminal B, triple negative, and HER2 enriched [7] . Among these subtypes, luminal breast cancer represents the most heterogeneous profile in regards to gene expression, mutation spectrum, copy number changes, and patient outcomes [5].

1.1.1 Genomic Alterations in Breast Cancer

Genomic aberrations are the key source of the neoplastic transformation and involve somatic mutations (substitutions, insertions, or deletions of small or large fragments of DNA), genomic amplification, and structural rearrangements (recurrent chromosomal translocation sites and frequent gains or losses involving both chromosomal arms [8,9]. As a result of those genomic aberrations, tumor formation and development, metastasis, and drug resistance are observed [10]. Somatic mutations in several key genes, such as *TP53, PIK3CA, CDH1, AKT1, GATA3, MAP2K7, MYC,* duplications such as *ERBB2,* and deletions in *PTEN or MAP2K4 [11–13]* are the most significant genomic

aberrations of the breast-derived tumor samples, and their frequency is changing according to the molecular subtype [12].

1.1.2 Somatic Mutations in Breast Cancer

Both endogenous and exogenous factors can cause DNA damage which is repaired through DNA repair mechanisms [14]. Any errors in the DNA replication mechanisms cause mutations, and cancer emerges as a consequence of the accumulation of somatic mutations [15].

A whole genome sequencing study conducted on 560 breast cancer patient data pinpointed driver mutations in 93 protein-coding genes. A controversial idea as one of the driver mutations is evident to trigger breast cancer [11]. Also, a network analysis using The Cancer Genome Atlas, which is a project that provides the characterization of different cancer types including breast cancer (TCGA; http://cancergenome.nih.gov/), on the gene expression and somatic mutation of 825 patients and indicated the presence of mutations in *TP53*, *PIK3CA*, and *GATA3* in more than 10% of the patients [5]. Another important driver of breast cancer, *ESR1*, is highly mutated upon endocrine therapy and distant organ metastasis [16]. In another study, a driver mutation in 40 genes was determined, and a link among the copy number alteration, gene expression, and also clinical data was established. They associated chromosome 11q13-14 with the aggressiveness of the tumor. As a result of the study, it is once again stated that genome-based classification is a prerequisite to designing and developing better therapeutic strategies [12].

1.1.3 Copy Number Variations in Breast Cancer

CNV (Copy Number Variation) is a structural variant in the human genome and is defined as “a segment of DNA that is 1 kb or larger and is present at a variable copy number in comparison with a reference genome.” [17]. CNVs encompass deletions, insertions, duplications, and complex multi-site variants [18], while these changes constitute 4.8% to 9.5% of the total variability in the human genome [19]. CNVs vary structurally and can be seen either as simple forms such as tandem duplications, or they can contain complex gains or losses of homologous sequences at multiple sites in the genome [20]. The formation of CNVs is based on different mechanisms that have been

implicated in genomic rearrangements [21]; nonallelic homologous recombination (NAHR); nonhomologous end-joining (NHEJ) [22], and retrotransposition [23–26]. NHEJ is the primary repair mechanism for DNA double-strand breaks in prokaryotes and eukaryotes [27], whereas NAHR is a homologous recombination occurring between similar DNA fragments [28,29]. Retrotransposition is involved in the integration of cDNA encoded in the cell to a new genomic site [30]. Fork stalling and template switching (FoSTeS) is a replication-based mechanism that has been suggested to be responsible for the observed complex genomic rearrangements that cannot be readily explained by NAHR, NHEJ, or retrotransposition [31]. CNVs in cells can interrupt the function of tumor suppressor genes and activate malignant cells. Tumor suppressor genes are responsible for the control of crucial pathways related to cancer, such as cell growth, proliferation and apoptosis [32,33]. Mutated tumor suppressor genes and its complete loss of function give rise to abnormal cell growth, tumor formation and development[34,35]. The most known example is observed in the *BRCA1* gene, which has a role in DNA damage repair and acts as a tumor suppressor. Inherited germ-line mutations, including CNVs [36,37] in *BRCA1*, is the main cause of genetic and hereditary breast and ovarian cancers [38,39].

Comparative genomic hybridization (CGH) was used for detecting the gains and losses in chromosomal regions as a conventional method [40–43]. However, copy number alterations and copy number variations in cancer still need to be studied in more depth. In solid tumor cells, it is well known that a vast amount of chromosomal rearrangements, such as amplifications and translocations, accumulate in the genome [44]. CGH studies on solid tumors show numerous copy-number abnormalities occurring on different chromosomes and in different regions. In the breast cancer genome, chr1, chr8q, chr8p, chr10p, and chr5q regions represent frequent copy number losses and gains [5,43]. Studies on breast cancer cell lines reveal high-level gene amplifications in those regions often distributed across many different chromosomes from their origin chromosomes [45]. These localizations, as well as amplicon structures, give rise to the formation of novel fusion genes, gene dysregulation, gene inactivation, and novel tumor-specific oncogene formation [46].

1.1.4 Copy Number Changes of Chromosome 1q

In breast cancer development, copy number alterations are known to be acquired frequently [47,48]. Chromosome 1 (chr1) is shown to be one of the most altered regions of breast cancer [49–51] as well as some other human malignancies such as multiple myeloma [52] and neuroblastoma [53]. Studies show that more than 70% of breast tumors, among all the subtypes, display copy number gain in q arm of chr1, while p arm majorly consists of copy number loss except the 1p31–p32 regions (gains occur in these regions) [49,51,54–59].

Recurrent copy number alterations profile of chr1q in breast cancer and the known oncogenes that are mapped to especially chr1q, such as *SHC1* and *NCSTN* [60] altogether, suggest the presence of important genes having an impact on breast cancer in chromosome 1q [61].

1.2 Transcription Factors

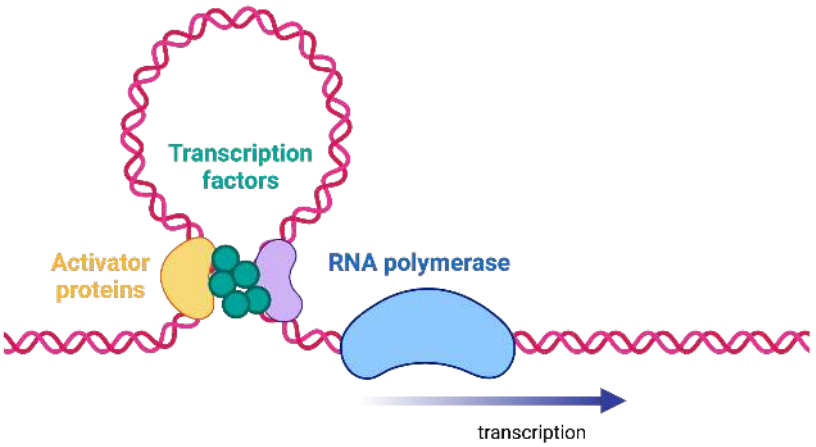


Figure 1.1: Activation of gene expression with the transcription factors. TFs regulate gene expression via binding to specific DNA regions (binding sites) and enhance the activity of RNA polymerase with the cooperation of other activator proteins. (*Created with Biorender)*

Regulation of gene expression depends on the transcriptional regulation and mRNA decay mechanisms [62,63]. Regulation of the transcription and, more importantly, frequency of the activation/repression depends on the factors that interact

with the specific elements located in the gene promoters and the regulatory regions [64]. Genes in eukaryotes have a particular combination of positive and negative regulatory *cis*-elements that are uniquely arranged in number, type, and spatial array, and they form binding sites for sequence-specific transcription factors [64,65]. Transcription factors are a specific protein group that plays an essential role in the maintaining of the unique gene expression profile specific to cell type [66]. Transcriptional regulation through transcription factors is dependent on the high-order chromatin architecture and is mediated by their interaction with regulatory elements such as promoter and enhancer region [67]. Mammalian cells harbor more than 1000 transcription factors that are selectively binding to short DNA motifs [68–72]. Many of these transcription factors (TFs) bind as homo- or heterodimers to multiple DNA motifs; thus, individual binding sites of related TFs can be located in distinct directions and lengths relative to each other [73–75]. For example, ERα forms a homodimer upon interaction with its ligand, estrogen, and binds to the target DNA sequence. Any alteration observed in the transcription factor activity results in an aberrant gene expression profile [76]. Sequencing studies have shown that mutations that occur in the ligand binding domain of ERα play a role in the development of hormone therapy resistance in approximately 30% of metastatic ER-positive tumors [77–79]. As transcription factors have a critical role in the regulation of gene expression, failure during their synthesis or activation [80] causes repression of tumor suppressor genes or over-expression of proto-oncogenes resulting in their conversion into oncogenes that give rise to cancer [81]. Namely, *ErbB2* proto-oncogene is found to be expressed in aggressive tumor types such as invasive ductal carcinomas, and up to 30% of all breast tumors due to its amplification and over-expression [82,83].

1.2.1 Pioneer Factors vs. Dynamic Assisted Loading

Pioneer factors are the special class of transcription factors that have the ability to penetrate through closed chromatin and access DNA target sites, which is followed by the binding of other transcription factors [84,85]. Contrary to those "pioneering abilities" that is specific to only certain TFs, later, it is shown that there are more dynamic factor/chromatin interactions for many factors instead of static binding events that are only associated with pioneering activity [86–88]. A subset of FOXA1, whose

protein family is one of the first described pioneer factors [89], has shown to be dependent on estrogen and glucocorticoid receptors, where those receptors open closed chromatin for FOXA1 binding, similar behavior to pioneer factors [90]. Those dynamic binding events allow the transcription factors to facilitate the indirect binding of other transcription factors. This results in the enhancement of each other's DNA binding via making the chromatin more accessible; this phenomenon is called as "dynamic assisted loading" [91,92]

1.2.2 Forkhead Box A1 (FOXA1)

Forkhead box A1 (FOXA1), also known as HNF3A or TCF3A, is one of the members of FOX transcription factors that have a role in regulating gene expression, cell differentiation during embryogenesis [93–95], controlling tissue-specific gene expression and regulation of gene expression in differentiated tissues [96], development of several congenital pathologies [97] as well as cancer development [98]. DNA binding domain of FOX proteins is known as the forkhead (FH) domain that belongs to a much larger superfamily, the winged helix [99–101]. This domain folds into a winged helix motif and gives the ability to function as a transcription factor to the *Fox* protein family [99,100]. FOXA1 is shown to be highly expressed in breast cancer cells and correlated with tumor differentiation [102], and a good prognosis in breast cancer [103,104]. Besides, FOXA1 mRNA expression is also shown in a subset of ER-negative tumors [105]. FOXA1 is required for almost all ERα binding events in breast cancer cells [106] as being a well-established pioneer factor in the regulation of ERα along in primary breast tumors and cell lines [106–109]. For example, the transcription of *CCND1* (cyclin D1) [110,111], a well-known oncogene, is induced by ERα through the chromatin remodeling activities as a pioneering function of FOXA1.

1.2.3 Gata Binding Protein 3 (GATA3)

GATA3 is a member of the GATA transcription factor family of zinc finger DNA binding proteins that recognize the consensus motif T/A GATA A/G [112,113]. Crucial functions of GATA3 include mammary gland development [114,115] and T cell differentiation [116], along with inducing DNA accessibility with chromatin remodeling events as pioneer factor features [117,118]. Additionally, GATA3 is involved in breast

cancer progression as being one of the most mutated cancer genes in breast cancer [5,119]. However, it is not certain if GATA3 acts as a tumor suppressor or as an oncogene. Large-scale genomic sequencing studies reveal that 10-15% of breast tumors comprise GATA3 mutations frequently in invasive ductal carcinoma [4,5], which utilize a good mark of luminal breast cancer. More than 70% of those mutations consist of small nucleotide deletions or insertions that result in protein truncation or extension [120]. These mutations with high frequency suggest a driver role of GATA3 in breast cancer, while GATA3 also displays a role in the suppressor role in epithelial to mesenchymal transition of breast cancer [121]. The expression levels of GATA3, ERα, and FOXA1 are highly correlated in luminal breast cancer due to their interplay in transcriptional regulation [108,117,122,123]. Thus, FOXA1 and GATA3 motifs are enriched in ERα binding regions [108,123], leading to GATA3-dependent chromatin localization of FOXA1 and ERα [124].

1.2.4 Estrogen Receptor Alpha (ERα)

ERα is a transcription factor encoded by the *ESR1* gene located in the long arm of chromosome 6. ERα expression is tissue-specific and a high level of ERα expression is observed in the uterus, liver, ovary, muscle, mammary gland, pituitary gland, adrenal gland, spleen, and heart, while prostate, testis, adipose tissue, thyroid gland, lymph nodes, and spleen reveal lower levels of ERα expression [125–127]. ERα has an upregulating activity for cell growth-related genes [128,129], and relatedly, in breast cancer, ERα activation by estrogens is mainly considered for enhanced proliferation resulting in tumorigenesis and progression of the disease [129–131]. This activity proceeds around 10% in normal tissues, while in breast tumors persists up to 80% with the augmenting activity in tumorigenesis and progression of breast cancer [130]. Activation of ERα is mainly mediated by its ligand, estradiol (E2), and the activation of ERα is followed by either a direct nuclear interaction through the estrogen-response elements or protein-protein interaction with other TFs to act on the estrogen target genes [132]. ERα-mediated gene regulation depends on several cofactors, including coactivators and corepressors and chromatin remodeling enzymes [133–135]. ERα can recruit those cofactors and regulate the interplay between other TFs such as FOXA1 and GATA3. The interplay provides remodeling of chromatin architecture for ERα binding

and elicits gene expression [124] . The complexes formed by the recruitment of possible proteins in the different promoter and enhancer regions and the distances suggest that various mechanisms are possible to regulate ERα binding [136]. Additionally, ERα confers distinct binding sites in different tissues through not only FOXA1 and GATA3 but other TFs. So, to define ERα binding sites (ERα cistrome), a more complex picture should be considered [137].

1.2.5 ERα - GATA3 – FOXA1 Network

Transcriptional regulation through the ERα, FOXA1, and GATA3 network serves a critical function in the ER-positive and estrogen-reliant breast cancers, which constitute approximately 60% of breast cancers [138]. This interplay implicates FOXA1, and GATA3 essentiality with functioning as pioneer transcription factors for the most effective ERα binding to DNA [84,108,124,139] as FOXA1 and GATA3 DNA motifs are enriched in the ERα binding regions [108,123]. Those binding events include closed chromatin accessibility and inducing chromatin opening, which leads to recruitment of other transcription factors, remodeling factors, and histone modifiers followed by gene activation [84,139–143]

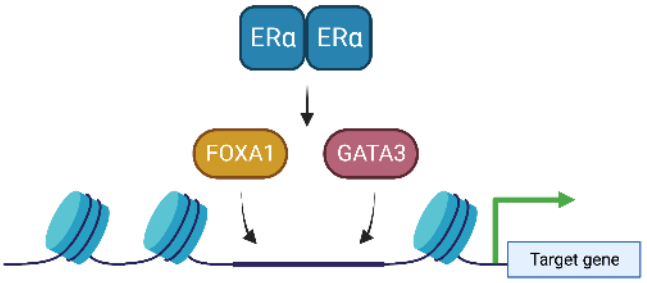


Figure 1.2: ERα, FOXA1, and GATA3 interplay in the context of breast cancer. ERα, GATA3 and FOXA1 interact together in the ERα binding events in breast cancer. FOXA1 and GATA3 display pioneering activity. *(Adapted from [144], created with Biorender)*

1.3 Next Generation Sequencing (NGS)

DNA sequencing started with the development of Sanger sequencing, which is based on the chain-termination and dideoxy synthesis [145,146], and Maxam-Gilbert chemical cleavage methods [147]. The emergence of these techniques depends on the development of polymerase chain reaction (PCR) [148,149], the widespread availability of high-quality nucleic acid–modifying enzymes, and the development of fluorescent automated DNA sequencing [150]. Those advancements resulted in the finalization of the first human genome draft in 2001 [151,152], and the first complete human genome was announced three years after that [153]. Since the first draft of the human genome was revealed, genomic studies have evolved rapidly and quantitatively, and bioinformatics has been the major scientific and training discipline of NGS technology. NGS can be used to sequence the whole genome or specific areas of interest, such as the whole exome or all coding genes as well as individual genes [154].

NGS technologies are classified according to the read length: short-read and long-read. Short-read sequencing provides higher-accuracy data that are useful for population-level research and clinical variant discovery at a lower cost [155].On the other hand, long-read sequencing is suitable for *de novo* genome assembly applications and full-length isoform sequencing [155,156]. Different methodologies based on the technology are used for short-read and long-read sequencing types. For example, 454 Pyrosequencing [157], ion torrent semiconductor sequencing, Sequencing by ligation (SOLiD), and Reversible terminator sequencing (Illumina) are the applications of short-read sequencing, while dominating technologies for long-read sequencing are Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) nanopore sequencing. Illumina is the dominating technology for short-read sequencing due to its maturity as a technology, a high level of cross-platform compatibility, and its wide range of platforms [156].

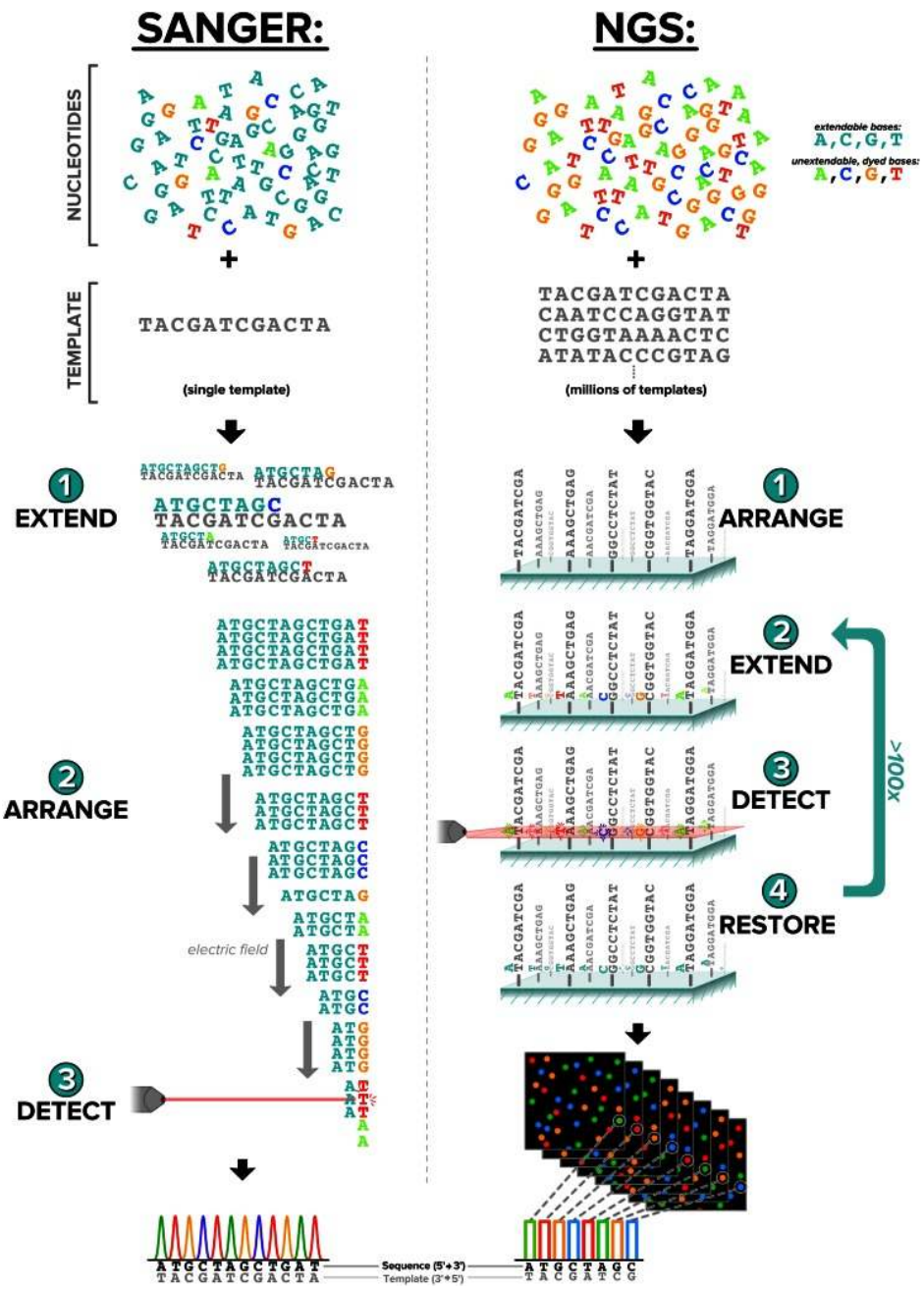


Figure 1.3: Overview of the Sanger and next generation sequencing. *(Taken from [158])*

NGS starts with the library preparation. This step includes: DNA fragmentation by enzymes or sonication and adaptor ligation to the fragmented DNA. Next, clonal library amplification is conducted by PCR in order to generate a strong and robust signal. Finally, sequencing by synthesis takes place in different steps. First,

fluorescently-labeled deoxynucleoside triphosphate (dNTP) is added to the fragment by a polymerase chain reaction to line up tens of millions of clusters on the flow cell surface in parallel. The label of the nucleotide plays a role as a terminator for polymerization. After each dNTP joining, the fluorescent signal is detected, and the base is identified. The enzymatic cleavage of dNTP is followed by the incorporation of the next nucleotide. This process results in an unbiased signal coming from separate molecules of each nucleotide. This technology provides reduced error rates compared to other technologies [155,156,159]. Sanger and NGS represent similar profiles fundamentally; however, the order of the steps additional to the restoration part in the NGS methodology serves the differences and generates an enormous impact on throughput [158]. NGS provides a systematic investigation of cancer genomics and its causative mutations [154]. Related studies have revealed frequently mutated genes in distinct cancer types (e.g., identification of BRAF oncogene) [160,161] as well as activating mutations with their relationship to drug response in cancer therapy such as EGFR mutations providing sensitivity of lung cancers to kinase inhibitor gefitinib [162–164]. NGS sheds light on the fields such as genomic loss and amplification, identification of the cellular pathways that underlie cancer, and improved selection of therapeutic targets with the high-resolution genome-wide studies [165,166]. For example, Isocitrate dehydrogenase 1 and 2, IDH1 and IDH2, are identified as cancer driver genes emerging from glioblastoma multiforme [167]. Additionally, NGS paved the way for large-scale genomic projects, including The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) that is providing the characterization of over 20,000 primary cancer and matched normal samples along 33 cancer types, and the International Cancer Genome Consortium (ICGC; http://www.icgc.org/) providing the mapping of genetic failures among the most important cancer types [166].

1.3.1 Chromatin Immunoprecipitation Sequencing (ChIP-Seq)

Chromatin immunoprecipitation (ChIP) is the main methodology to investigate the DNA fragments bound to specific proteins or certain nucleosomes [168]. ChIP followed by sequencing (ChIP-Seq) method investigates the protein–DNA interactions and epigenetic marks in a genome-wide and in a high-throughput manner with a small amount of starting material [169–172]. Thus, ChIP-seq provides significantly improved fragment reads by enabling precise identification of protein binding sites of transcription factors, histones, and chromatin variants [168,171,173–175]

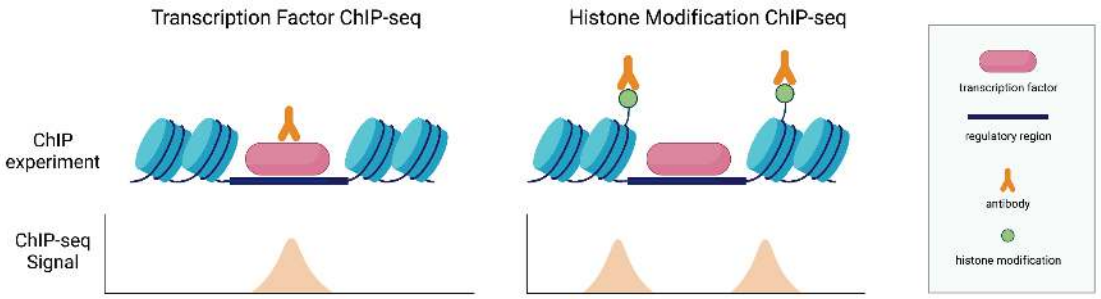


Figure 1.4: Types of ChIP-seq experiments and resultant signals. ChIP-seq can be used to investigate binding sites of DNA associated proteins. Histone modifications and TFs are the subjects of ChIP-seq and their signals display differences. Histone ChIP-seq reveals broad signals, while TFs expose narrow signals. *(Adapted from [176], created with Biorender)*

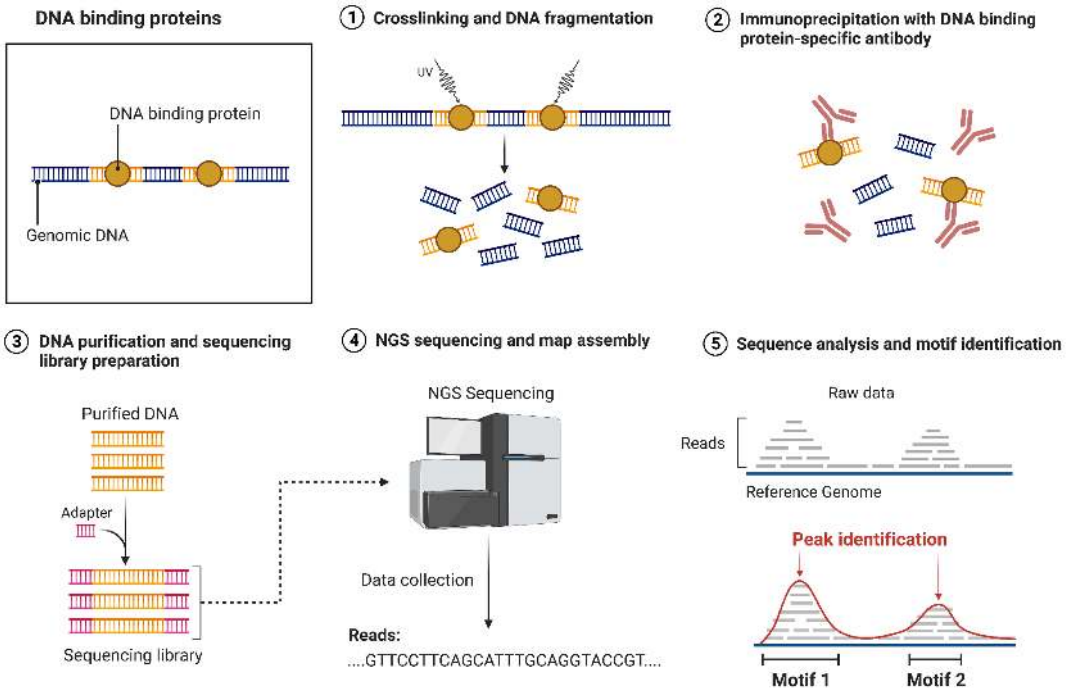


Figure 1.5: Main steps ChIP-seq. ChIP experiment fundamentally includes crosslinking of the protein, fragmentation of the DNA, and immunoprecipitation by using specific antibody for the crosslinked protein. After sequencing, bioinformatic analyzes further used to investigate the protein binding sites. (*Taken from Biorender)*

ChIP-seq starts with *in vivo* crosslinking of DNA to the associated proteins by formaldehyde treatment. Next, once the DNA is isolated, fragmentation, ideally around 200 bp (up to 600 bp) for sequence-specific TFs [177,178] is performed either by using sonication or micrococcal nuclease. Fragmented DNA is immunoprecipitated using a specific antibody. Finally, reverse crosslinking provides only the DNA fragments immunoprecipitated by the antibody of interest which isfollowed by sequencing of ChIPped DNA fragments and obtaining DNA sequencing reads. The main analysis steps include quality control, mapping, peak detection, motif analysis, annotation, and visualization, followed by in-depth bioinformatic analysis [173].

ChIP-seq represents one of the early practices of NGS, with the first being published in 2007 [171,172,179]. Successful ChIP-seq studies include revealing the role

of histone deacetylases in androgen-regulated transcription in prostate cancer [180], and shared transcriptional and demethylation profiles between sensitive and resistant breast tumor cells [181].

1.3.2 RNA Sequencing (RNA-seq)

In order to map the transcriptome and investigate total RNA, pre-mRNA, and noncoding RNA, such as microRNA and long ncRNA, RNA sequencing (RNA-Seq) technique is a widely used application of NGS [182]. The methodology is fundamentally based on the quantification of gene expression abundance. This quantitative approach further sheds light on the discovery of novel transcripts, identification of alternatively spliced genes, and detection of allele-specific expression [183]. RNA-seq methodology steps include: RNA extraction from the cells of interest, purifying the subset of RNA using a specific protocol, such as the poly-A selection protocol to enrich polyadenylated transcripts or a ribo-depletion protocol to remove ribosomal RNAs, conversion of the RNA to its complementary DNA (cDNA) by reverse transcription followed by adaptor ligation to the ends of cDNA, and finally, sequencing after PCR amplification.

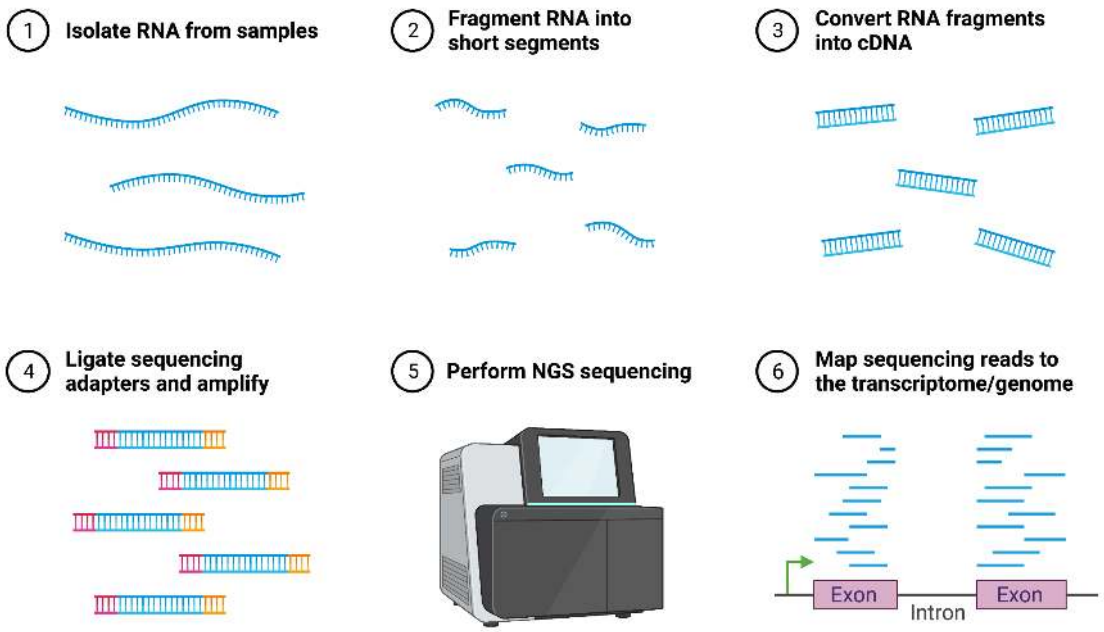


Figure 1.6: Overall look at RNA-seq. RNA-seq analysis include RNA isolation and purification, converting RNA to cDNA, adapter ligation and sequencing, followed by bioinformatic analysis.(*Taken from Biorender)*

The sequencing step itself is generally not different from the DNA sequencing; however, the library preparation step sets apart the methodology from DNA sequencing, and the necessary bioinformatic analysis differs in order to extract gene expression information from the sequencing data [184]. Gene expression patterns vary on many factors such as cell type, stimulus received, time, and loci under specific conditions. Thus, RNA-seq is a great method to investigate gene expression under specific conditions or cell stages. Successful applications include the reveal that >85% of the genome is transcribed, while it was known as only 3% [185] and a lower rate of validated somatic single nucleotide variants (SNVs) that actually occur in breast cancer transcriptome [186]. Furthermore, the integrative investigation of RNA-seq results with other omics data advances the understanding of gene regulation in a more complex context. The Cancer Genome Atlas [187], an integrative large-scale project, combines gene expression data with other -omics data such as transcription factor (TF) binding and histone modifications.

1.4 Genome Editing

The ability to engineer biological systems and organisms holds great potential for applications across basic science, medicine, and biotechnology. Genome editing technologies are a subset of those applications that manipulate the genome or directly genes to alter gene expression [188]. Particularly in medicine, the aim is the investigation of cancer, virus infections, genetic diseases, and the detection of pathogens [189]. Genome engineering mainly uses engineered endonucleases which are zinc-finger nucleases (ZFNs) [190–193], transcription activator-like effector nucleases (TALENs) [193–197], RNA-guided clustered regularly interspaced short palindromic repeats (CRISPR)-Cas (CRISPR-associated proteins) (CRISPR-Cas) nuclease system [198–205]. These genome editing tools are used to create precise double-strand breaks in the genome by recognizing the target DNA sequence. Compared to Zinc fingers and TALENS, CRISPR-Cas9 is the most advanced tool [188] with being easier, more

efficient, and simple while targeting multiple genes at once with a low cost [201,205]. Subsequently, the CRISPR-based systems are shown to target and study cancer-causing mutations in the cancer genome with promising applications in the way of treating cancer [188,206–208].

1.4.1 CRISPR

Archaea and bacteria have evolved to cope with viruses, plasmids, and other mobile genetic elements by an RNA-based adaptive immune system to proceed with their existence in highly dynamic, fast-evolving environments [204]. Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated genes (Cas) are the main components of their defense mechanism [198–200].

CRISPR systems are subtyped mainly based on the *Cas* proteins that serve different genetic, structural, and functional profiles in the related CRISPR system [209]. There are mainly two classes of CRISPR/Cas systems: Class I and Class II. While class I CRISPR systems have Cas protein complexes with multiple subunits, class II systems employ a single Cas protein [210]. Due to the simple structure of the type II CRISPR/Cas-9 system, it has been widely studied and used as a precision genome engineering tool [211]. A subset of the applications of this system includes expediting crop and livestock breeding, engineering new antimicrobials, and controlling disease-carrying insects with gene drives [208].

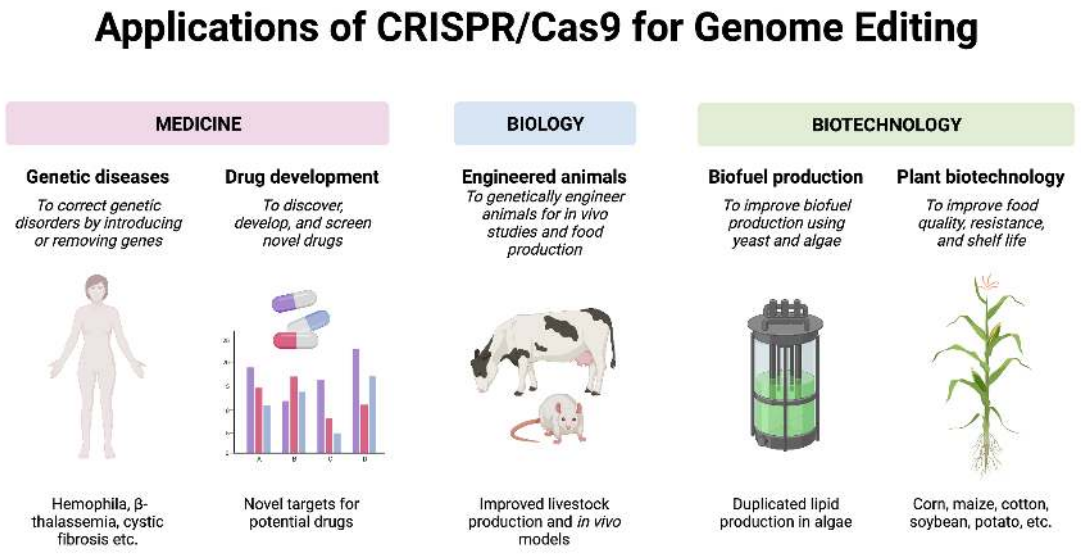


Figure 1.7: Applications of CRISPR-Cas9 genome editing system. CRISPR-Cas9 has applications in distinct fields and not only human genome editing such as animal engineering, biofuel production and plant biotechnology. (*Taken from Biorender)*

The mechanism of the CRISPR/Cas-9 system consists of 20 base pair single guide RNA that matches with the target sequence in the DNA, which is followed by a protospacer adjacent motif (PAM) sequence, a 2–5 base-pair length short conserved DNA sequence [212]. An appropriate PAM sequence is recognized by Cas9 protein and generates a double-stranded break by cleaving the DNA approximately three nucleotides upstream of the PAM [201]. Double-stranded break triggers non-homologous end joining (NHEJ) and homology-directed repair (HDR) DNA repair mechanisms [213] and results in inactivation, activation, or knockout of the targeted genomic part [201,202,205].

CRISPR-based screening is a large-scale and high-throughput assay and is used to map the function of a multitude of genes and/or regulatory regions in parallel [214]. This methodology uses multiple sgRNAs to generate large-scale genomic, transcriptomic, or epigenomic perturbations in a cell population. Differentially expressed genes related to the targeted phenotypes are further investigated. There are different Cas9 variants dependent on the functionality of interest and perturbation that is sought for the study. The function of wild-type Cas9 protein is ideal and used for knockout experiments. Frameshifts as a result of double-stranded breaks in DNA created by wild-type Cas9

cause knockout of the targeted genes, and loss of function of the target gene is observed [215,216]. CRISPR activation (CRISPRa) and CRISPR repression (CRISPRi) systems are used for altering transcription at a particular locus [216–218] as an alternative for knockout screens and depends on the dead Cas9 (dCas9), which do not have a nuclease activity, conjugated with an activator or a repressor. CRISPRa/i systems are ideal for the investigation of the genome function studies for observing genome function at the transcriptional level without permanently modifying genomic structure [217]. Epigenome studies encompass dCas9 fusion constructs that modify histones and change the structure of DNA, altering the accessibility of the genome gene expression as a result of this alteration [219–221]. dCas9 proteins can be fused either to DNA (cytosine-5)-methyltransferase 3A (DNMT3A) and ten-eleven translocation (TET) proteins in order to change the DNA structure [220,221] or histone modifying enzymes [219,222–224] for histones to take place. Those epigenetic modifications alter DNA structure in the long term. With this advantage, structural change effects on the neighboring genes can also be investigated [225]. CRISPR-based base editors are a relatively new sub-application area of the CRISPR system that uses cytidine deaminase or adenosine deaminase fused with dCas9 in order to generate precise single base changes [226], which are C→T and A→G mutations, specifically. CRISPR-based base editing method involves the gain or loss-of-function studies that need precise point mutations [227–229]. Furthermore, combinatorial approaches using the CRISPR system to investigate functional relationships between genes and regulation of gene expression are the subject of current CRISPR studies [230] like using lentiviral products [231], mouse and human orthologous U6 promoters [232,233].

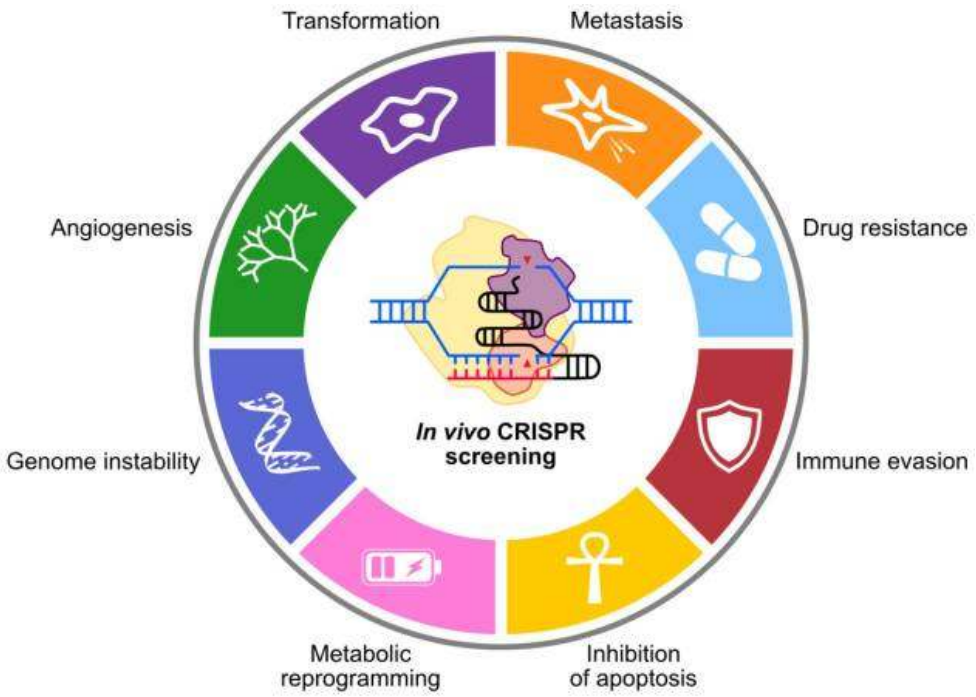


Figure 1.8: Functional genomics applications of in vivo CRISPR screening. *(Taken from [234])*

CRISPR genome editing is essential for cancer studies targeting cancer cells that comprise genetic aberrations by targeting those genes involved in cancer formation and development [234]. CRISPR screens are used in various applications essentially for investigating cancer phenotypes [234]. Initially, CRISPR screening is applied to human and mouse cells to demonstrate targeted DNA cleavage *in vitro* [201,205]. This is followed by a rapid increase in CRISPR-based studies in a short period of time. A genome-wide *in vivo* CRISPR screen is applied to a non-metastatic cancer cell line, and functional lung metastasis gene hits are determined [235]. Followed by this study, CRISPR screens are applied to identify oncogenes [235,236], tumor suppressor genes [237–239] and synthetically lethal genes [240] as well as cancer immunotherapy regulators in co-culture [241] and transplant tumor models [242]. Also, there are CRISPR applied studies on drug resistance [243–245], determination of essential genes [246,247], and synergistic and synthetic lethal interactions [248] that contribute to shedding light on genes involved in cancer. Such applications include the knockout of genes that play a role in cancer cell survival or chemoresistant genes [249].

CRISPR/Cas9 Library Design

In a CRISPR-based screening experiment, the first step would be the identification of where the perturbation will occur in the genome. Targeting specific genomic loci with the CRISPR system needs two main components: Cas9 protein to serve endonuclease activity resulting in dsDNA breaks and sgRNA targeting a specific locus. sgRNA that is used by Cas9 to match the target DNA consists of crRNA (target DNA complementary) and trans-activating crRNA (tracrRNA - binding scaffold for the Cas protein), which are two nuclease active sites to generate DSBs [250,251]. PAM sequence directs sgRNA to bind specific target sequence and Cas9 to serve endonuclease activity [252]. PAM contains an NGG motif where N can be any bases, but two G:C base pairs are needed following it [253].

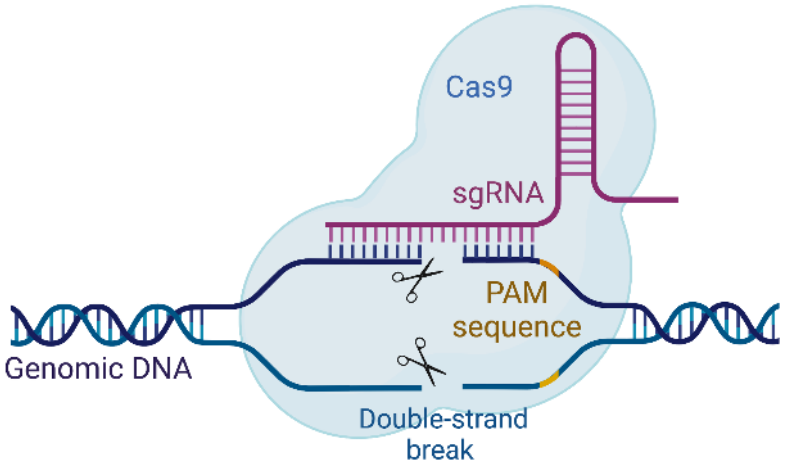


Figure 1.9: CRISPR/Cas9 system. In the CRISPR/Cas9 system, sgRNA finds the target sequence with the presence of PAM. Cas9 recruitment is followed by double-strand break and trigger DNA repair mechanisms that give rise to knock-out or knock-in. (*Created with Biorender)*

Ideal CRISPR library design and sgRNA selection should include efficiency and specificity evaluation [254,255]. Cas9 complex can bind to off-target regions [255,256]. This targeted versus non-targeted binding activity is considered specificity. On the other hand, even if binding occurs on the target region, intended cleavage might not occur efficiently [254,256–258]. Therefore, off-target activity is expected to be lower while

on-target activity to be higher. There are various determinants of off-target activity, such as the number and the position of mismatches, GC-content of the targeting sequence, variants of nuclease as well as the method of delivery [201,250,259–261]. Additionally, efficiency can be affected by the composition of purine/pyrimidine near the 3′ end of the spacer sequence [215], and nucleotide composition downstream of the spacer sequence could be predictive of the sgRNA efficiency [262] . Several algorithms are developed in order to determine sgRNAs along with, evaluating the efficiency and specificity of the sgRNAs, and generate CRISPR libraries [256,262–264].

Chapter 2:

Material and Methods

2.1 File Formats

In this study, there are several different file types are used according to the data stored. These file types are used for different analyzes and different information.

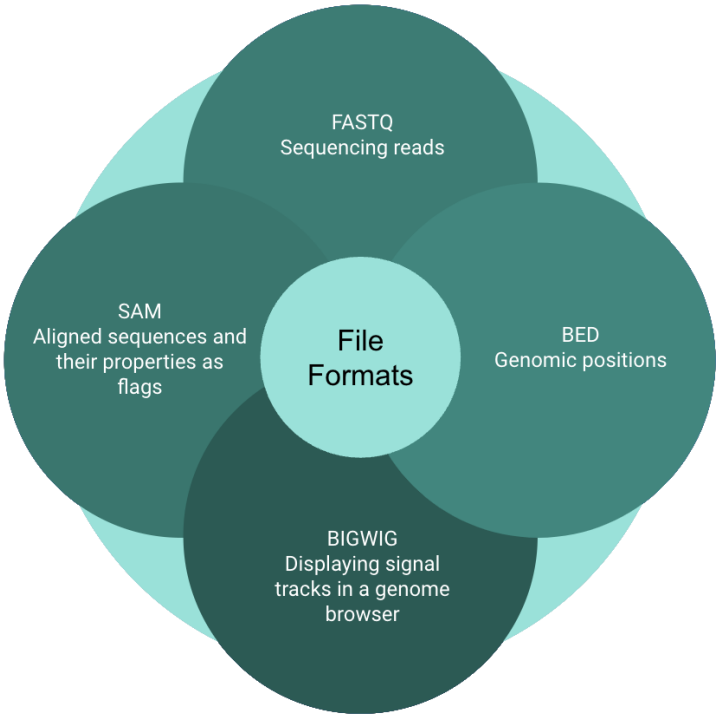


Figure 2.1: File types used in the study. There are different types of files used in the study. The data analysis starts with raw reads in fastq format. BED format is the main file type used for the analyzes.

2.1.1 Fastq File

Fastq is the fundamental text-based format that stores sequencing reads, and their quality scores [265,266].

2.1.2 SAM File

This is the format to store aligned sequences and their properties as *flags*. By using these flags, the aligned sequences can be manipulated (e.g.: duplicate removal) [267].

2.1.3 BigWig File

The bigWig format is the binary format in order to process and represent dense and continuous data. This format is suitable for displaying signal tracks in a genome browser [268].

2.1.4 BED File

BED is the general file format used to store genomic positions [269]. In this study, mainly BED (Browser Extensible Dat) format is used. There are three required fields which are the chromosome number, start and end position of the region. Additionally, nine other fields can be used, but this is optional.

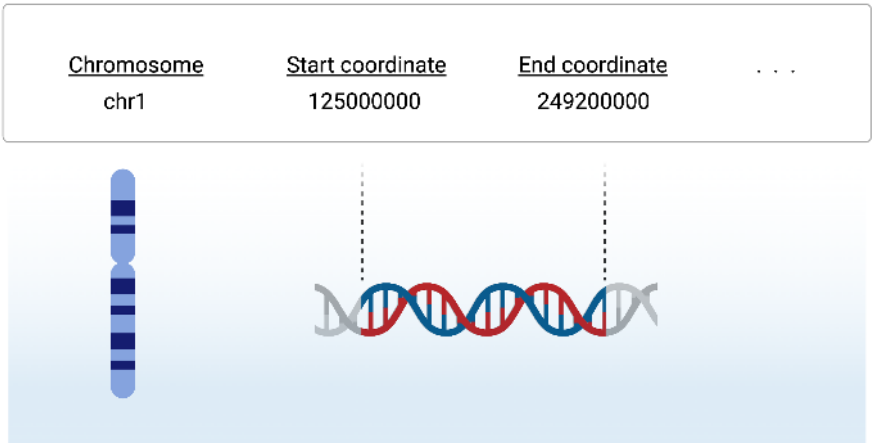


Figure 2.2: BED file format. Storing genomic information requires three fundamental fields in BED format: Chromosome number, start position and end position.

2.2 ChIP-seq Datasets

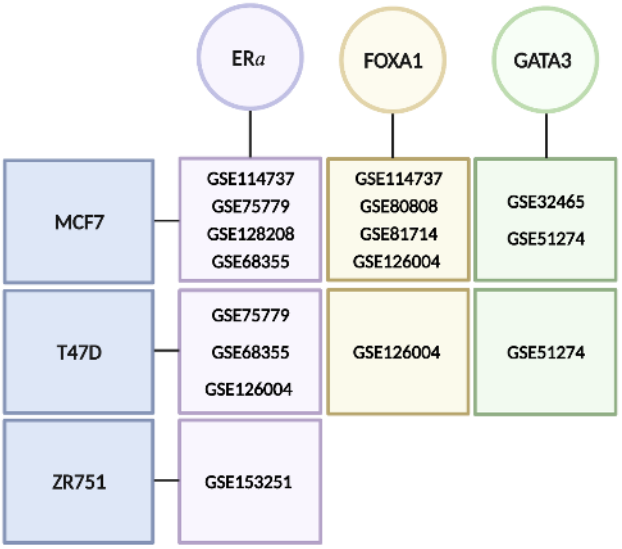


Figure 2.3: GEO series accession numbers of the related transcription factor in the indicated cell line. There are 36 datasets in total collected through the GEO database for TFs in three cell lines.

ChIP-seq datasets analyzed in this research study are obtained from the collection of publicly available datasets, Gene Expression Omnibus (GEO), [269–271] and their unique identification numbers are presented in Figure 2.3. It is aimed to profile 3 transcription factors; ERα, FOXA1, and GATA3 in MCF7, T47D, and ZR75-1 cell lines. All datasets contain single-end reads and biological samples that were not exposed to any treatment. In their original studies, the functional investigation of TFs are conducted. They were used either as control or untreated condition. Each replicate is either coming from the same experiment or distinct experiments/datasets. These are considered as biological replicates and the samples belonging to the same transcription factor in the same cell line are combined. In this complete dataset, there are **8** replicates of ERα in MCF7, **8** replicates of ERα in T47D, **4** replicates of ERα in ZR75-1, **7** replicates of FOXA1 in MCF7, **4** replicates of FOXA1 in T47D, **3** replicates of GATA3 in MCF7 and **2** replicates of GATA3 in T47D. Unfortunately, only ERα ChIP samples were available for ZR75-1.

In conclusion, in total **36** samples are collected for three antibodies in three cell lines. All of the samples have overall good quality which do not require additional analysis steps and experiments, as indicated in the Cistrome data browser [272,273] and their FastQC files were also checked and confirmed (v0.11.9) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, Andrew S., 2010).

2.3 ChIP-seq Analysis

2.3.1 Optimization of the ChIP-seq Analysis

During the optimization process, several tools have been used and either changed or modified. Briefly, collection (fastq-dump), mapping (bowtie2), peak calling (macs2), finally motif finding and annotation (homer), remained the same with adjustments in the parameters and options.

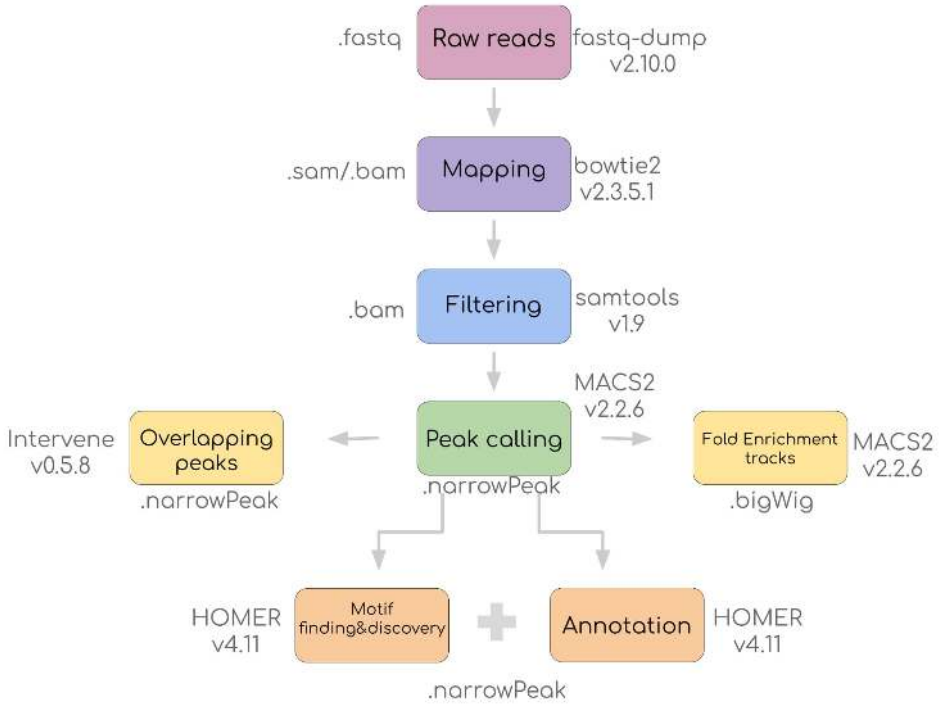


Figure 2.4: Old version of the ChIP-seq analysis. In ChIP-seq data analysis MACS2 and HOMER are used mainly. The changed parts include tools for filtering the duplicates, taking overlaps of the peaks and generating signal tracks.

In the filtering step, *Picard Toolkit (*"Picard Toolkit." 2019. Broad Institute, GitHub Repository. https://broadinstitute.github.io/picard/; Broad Institute*)* and *Sambamba [274]* are used instead of *samtools* for discarding the possible duplicate reads which could be generated in the polymerase chain reaction step of the experimental of ChIP-seq and keeping only the uniquely mapped reads which are indicators of true binding sites. *Picard Toolkit* shares some functions with samtools, but is widely used specifically in filtering steps and is user friendly. Sambamba serves faster BAM file processing. Instead of fold enrichment tracks generated by MACS2, bigWig files for the signal tracks are used. Normalization and correlation using deeptools and combining the peaks with a statistical framework have been added with MSPC. The final version is explained further in the *Optimized ChIP-seq Analysis Workflow* section, the tools are chosen according to their usage feasibility.

2.3.2 Motif Analysis

Motif analysis is performed to determine which transcription factors control the transcription of a set of genes by detecting enrichment of known binding motifs in the genes' regulatory regions. Motif analysis can be mainly separated into two; *de novo* motif discovery which is searching for unknown patterns, and motif finding which is applying existing patterns to the detection of pattern occurrences [275–278]. *HOMER* (version 4.11) [279] is used for conducting the motif analysis. HOMER is suitable to find motifs in the narrowPeak files. P-values obtained by *HOMER* motif analysis should be highly significant (i.e., << 1e-50). Otherwise, it could indicate a failure in one of the steps of the ChIP-seq process. In this study, chr1q motif analysis resulted in high p-values.

2.3.3 Combining the Peaks as Consensus Peaks

In order to combine all the peaks that belong to the same transcription factor in the same cell line *MSPC (Multiple Sample Peak Calling) [280]* is used which is specifically designed for taking peak files of unlimited replicates (either biological or technical) and improving sensitivity and specificity of each one. In this process, the consensus regions identified between the given samples are combined. Finally, the resultant file is a single dataset with its summary statistics. Henceforth, the combined peaks are named

Consensus Peaks. All the samples are considered biological replicates and p-values thresholds are taken as 1e-4 and 1e-8 to define weak and stringent peaks, respectively. The overall peak numbers are representing the binding patterns of ERα, FOXA1, and GATA3 independently from the replicate numbers. FOXA1 has the highest values in both cell lines compared to ERα and GATA3. Overall, a very limited number of direct binding peaks contained ERE, GATA3 and FOXA1 motifs for the relative transcription factor. Thus, approximately the same amount of consensus peaks are obtained for three of the transcription factors (Figure 2.5).

		Chr1q Peaks	Consensus Chr1q Peaks	Direct Chr1q Peaks
MCF7	ERα	6749	2.559	150
T47D		3807	1.692	98
ZR75-1		3528	1226	98

		Chr1q Peaks	Consensus Chr1q Peaks	Direct Chr1q Peaks
MCF7	GATA3	4674	2547	702
T47D		1694	1131	93

		Chr1q Peaks	Consensus Chr1q Peaks	Direct Chr1q Peaks
MCF7	FOXA1	14511	2.707	945
T47D		6670	2.595	978

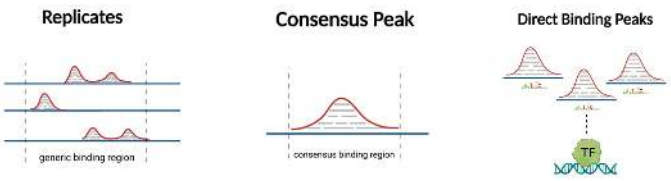


Figure 2.5: Peak numbers for chr1q. Consensus peaks are obtained from individual peaks and direct binding peaks are filtered among consensus peaks. Most of the peaks belong to the FOXA1 independent from the replicate number. For ERα and GATA3, MCF7 cells have more peaks than T47D.

2.3.4 Obtaining Directly Bound Consensus Peaks

ChIP-seq peaks could be detected as a result of either direct TF–DNA interactions or indirect protein-protein interactions with other regulators such as co-factors. Moreover, unspecific binding is also detected. Thus, ChIP-seq experiments are prone to generate artifacts. Consequently, the depiction of bona fide TF-bound regions is still an ongoing challenge. Standard motif enrichment analysis and motif discovery approaches sometimes fail to correctly identify the binding motif for the ChIP'ed factor [281]. In this section, the aim is to obtain the direct binding transcription factor peaks with desired motif occurrences.

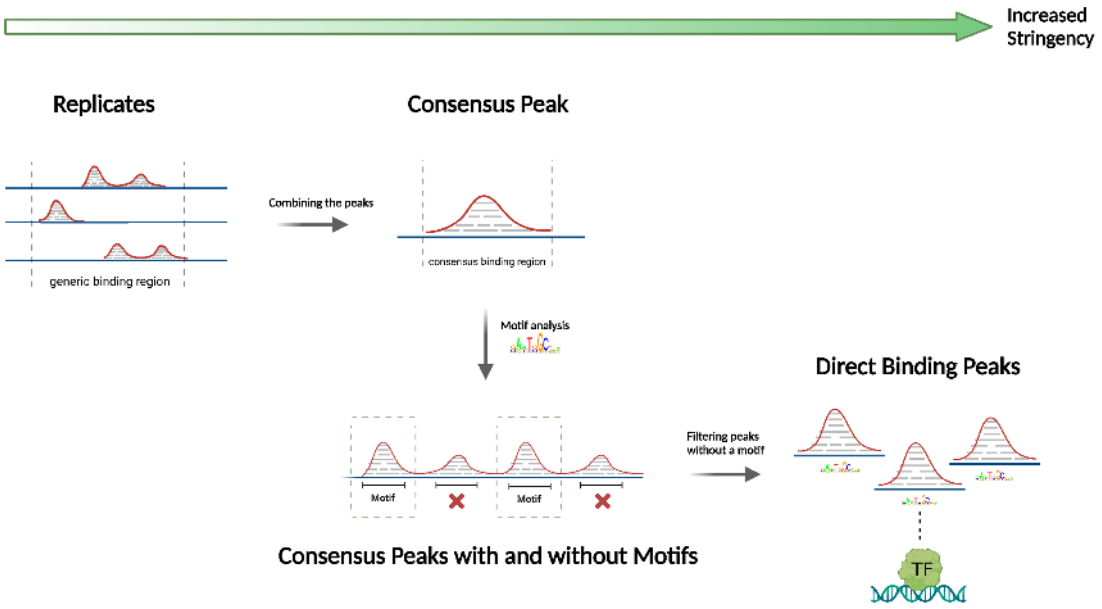


Figure 2.6: Overview of acquiring direct binding sites from the individual peaks. Direct binding peaks are obtained in two fundamental steps after peak calling: consensus peak generation and motif analysis. Only motif exhibiting peaks are considered as direct binding peaks.

For this purpose, firstly peaks are combined with the statistical significance of repeated evidence [280]. This provides the classification of the weak peaks and strong peaks with respect to the given p-value thresholds (explained in the *Combining the Peaks as Consensus Peaks* section). Followingly, weak peaks and false positives are discarded.

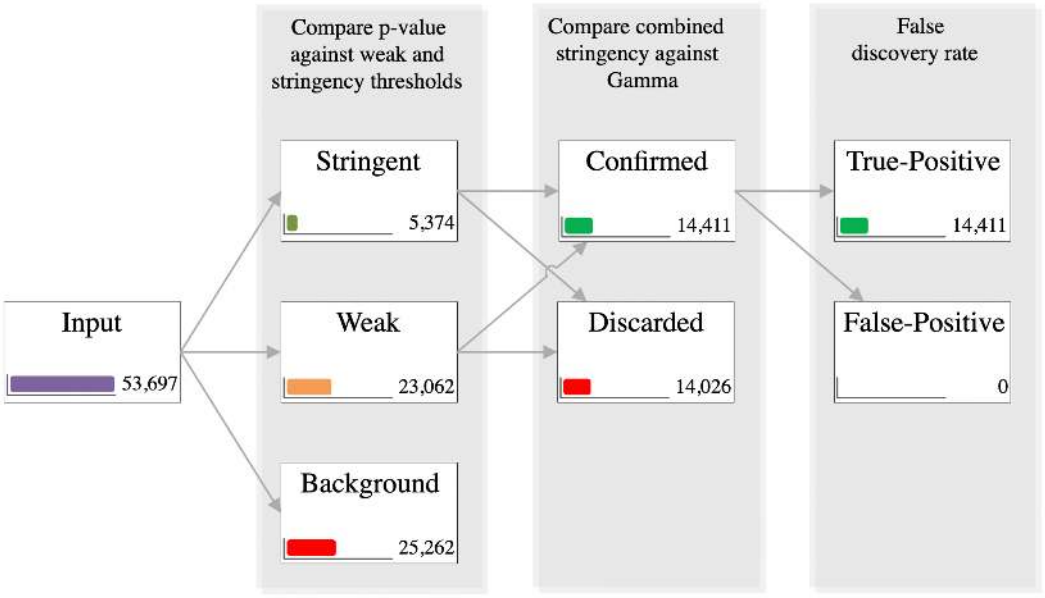


Figure 2.7: Combining the peaks with MSPC. Peaks are classified according to the p-value thresholds. Compared values are used in order to classify peaks as stringent, weak, and background. Background peaks are discarded and weak peaks are evaluated again. Only peaks that are confirmed are kept and considered as True-Positive. (*Taken from [280]*)

Next, motif analysis is performed. This classifies the peaks that contain and do not contain a known motif. Finally, only peaks that are classified as stringent yet contain a known motif are classified as *direct binding* peaks.

2.3.5 Motif Enrichment Based on the Frequency

Motif analyses and the order of the most enriched motifs is based on the p-value. For example, ER*a* ChIP-seq samples resulted in the ERE motif as mostly enriched. However, there are different motifs also found as enriched in the same sample. In this section, the motif analysis result is evaluated based on the frequency of the motif instances. The motif analyses on direct and indirect peaks were conducted by *HOMER*. This means that the analysis is applied for the second time based on the peaks' subset group (direct-indirect). Word cloud representation is made in Rstudio using *wordcloud2* (v. 0.2.1), *RColorBrewer* (v. 1.1-2) and *tm* (v. 0.7-8) packages (also stated in the *Visualization* section).

2.3.6 Annotation of the Consensus Peaks

Annotation is the step in which the resultant peaks are assigned to the genomic regions (i.e., exon, intron) and related genes as well as gene IDs of related databases. Genomic regions that the peaks are located in with motifs are determined according to the annotation analysis using *HOMER*.

2.4 CRISPR Library Design: Benchmarking of CRISPR-DO

CRISPR library design is conducted by using the tool CRISPR-DO (Design and Optimization) [282]. The tool is based on the integration of the sgRNA efficiency prediction model and an off-target scoring function that allows being more selective according to the specificity and sensitivity of the sgRNA. Target sequences are annotated with the PhastCons conservation score, which is the probability of nucleotide matches with the evolutionarily conserved elements. Eventually, looping each interval in a given list for CRISPR-DO is conducted, and the sgRNA list is generated.

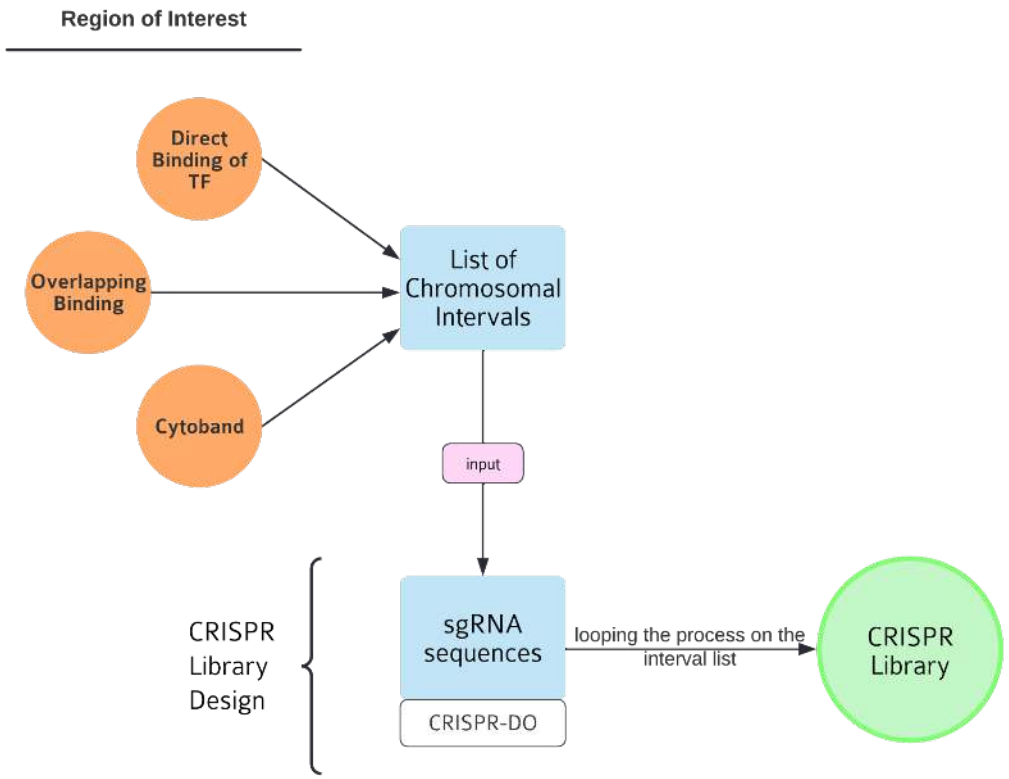


Figure 2.8: sgRNA library design process. Three main types of libraries are generated in this study. Direct and overlapping peaks are used to target the binding sites

of ERα, FOXA1, and GATA3. Cytoband libraries target the specific cytobands that are determined in the previous analysis. All of the libraries need chromosome positions in order to find sgRNAs. Each chromosomal interval is one run and looping for all the intervals is needed to generate a specific library.

2.5 RNA-seq Data and Analysis

Publicly available RNA-seq data for MCF7, T47D, and ZR75-1, luminal A breast cancer cell lines and 184A1, 184B5, MCF10A, MCF12A, immortalized breast cell lines were obtained from GEO database with GSE48213, GSE73526, GSE152908, GSE58135 accession numbers using sra-tools (version 2.10.0). The relevant studies looked at the mRNA levels of breast cancer cell lines that were profiled at baseline such as in the absence of perturbation for the investigation of breast cancer drivers, vulnerabilities, and resistance and revealed possible breast cancer-specific cell surface markers and potential targets. Also, they suggested new treatment strategies for breast tumor subtypes, and revealed new features of insights of breast cancer.

Table 2.1: GEO accession numbers of the complete RNA-seq data set. RNA-seq analysis contains four datasets in total. First three of the four datasets contain both breast cancer and normal-like cell lines, except the last only has for cancer cell lines.

LuminalA Breast Cancer Cell Lines			Normal Breast Cell Lines			
MCF7	T47D	ZR751	184A1	184B5	MCF10A	MCF12A
GSE48213	GSE48213	GSE48213	GSE48213	GSE48213	GSE48213	GSE48213
GSE73526	GSE73526	GSE73526	GSE73526	GSE73526	GSE73526	GSE73526
GSE152908	GSE152908	GSE152908	GSE152908	GSE152908	GSE152908	GSE152908
GSE58135	GSE58135	GSE58135				

The quality control analysis of raw reads were conducted individually by FastQC, and collectively by MultiQC (v1.9) [283]. If adapter sequences are affecting the quality of the data, they were removed using TrimGalore (v0.6.5) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, Krueger F., 2015). Fastq reads were aligned to the reference genome (GRCh37.p13), obtained from GENCODE database, by STAR (v2.6.7a) [284] Next, expression (count) files were created with Salmon (v1.1.0) [285]. The count values of each transcript were combined

at the gene level using the Tximport library [286] in R (v4.0.2). In the identification process of upregulated and downregulated genes shared among subjected breast cancer cell lines, DESeq2 library [287] in R is used. After the control of the Principal Component Analysis (PCA) results between biological replicates are distinguished according to their subtypes.

The two main approaches in RNA-seq analysis are 1) to detect all the differentially expressed genes and 2) identifying genes that exhibit the most significant changes among all. For this purpose, DEG analysis is based on two different parameters: expression value and p-value. In this study, genes having more than 2-fold expression change with p-value less than 0.001 are considered. As a result, 4850 genes are found to be differentially expressed.

2.6 Copy Number Data

Copy number data for MCF7, T47D, and ZR75-1 cell lines were obtained from Cancer Cell Line Encyclopedia (CCLE) [288] in 2020. CCLE is a project to characterize the cancer cell lines with respect to their mutation profiles, histone and epigenetic modifications, fusion calls, RNA expression along with metabolomics and proteomics datasets. The downloaded dataset is called as CCLE_copynumber_byGene_2013-12-03.txt and as the name indicates, there is copy number information of every gene with respect to the cell line. The copy number values are calculated as log2(CN/2) as they indicated.

2.7 Visualization

Visualization of the data as venn diagrams, word clouds, scatter plots, and rug plots is done by using the *ggplot2* (version 3.3.5) R package.

Chapter 3: RESULTS

3.1 Optimized ChIP-seq Analysis Workflow

The analysis of ChIP-seq samples start with the retrieval of the raw reads coming as the result of the sequencing of ChIP experiments. Single-end reads (all the datasets) are downloaded in the fastq file format by using their SRR (run accession for actual sequencing data for the particular experiment) numbers provided from National Center for Biotechnology Information (NCBI), GEO database. To do that, the *fast-dump* subtool of *SRA Toolkit* is used (version 2.10.0) [289]. Acquired raw reads are aligned to the reference genome, which is human genome assembly GRCh37 (hg19). For the alignment step, the *bowtie2* (version 2.3.5.1) tool is used since it is fast, widely used, and memory-efficient tool [290]. The human genome reference file (GRCh37 - hg19) is downloaded from the University of Santa Cruz (UCSC) Genome Browser. *Bowtie2* outputs the aligned files in the format of Sequence Alignment Map (SAM). Those SAM files are converted into Binary Alignment Map (BAM) files using *samtools* (version 1.9) [267]. BAM is the converted version of SAM to a binary format that facilitates the usage of files over shorter times in a memory-efficient manner. Then, reads are sorted and filtered if their MAPQ scores are less than 20. Only unique reads are kept. Next, peak calling is conducted. Peak calling is the fundamental step in the analysis of ChIP-seq data [171,179]. *Peaks* are the enriched genomic regions of the immune-precipitated factor. In this step, previously obtained BAM files are used for peak calling to generate ChIP signals with respect to the input, which is the same DNA sequence that has been cross-linked and sonicated but not immuno-precipitated. In this step, MACS2 (version 2.2.6) is used as it is one of the most commonly used peak callers successful in investigating narrow peaks (transcription factors) [291]. Normalized correlation is calculated in RPKM (Reads Per Kilobase per Million mapped reads) unit by using *deeptools* (version 3.5.0) and signal tracks in bigWig file format. Binary bigWig files allow investigating a continuous signal track on Integrative Genomics Viewer (IGV) [292] with respect to enrichment and chromosomal location. MSPC is used for combining all the samples and generating a uniform bed file for the same transcription factor in the same cell line. Motif analysis and annotation of the

motifs are following combining the peaks. Direct binding peaks are achieved with the result of those last two steps.

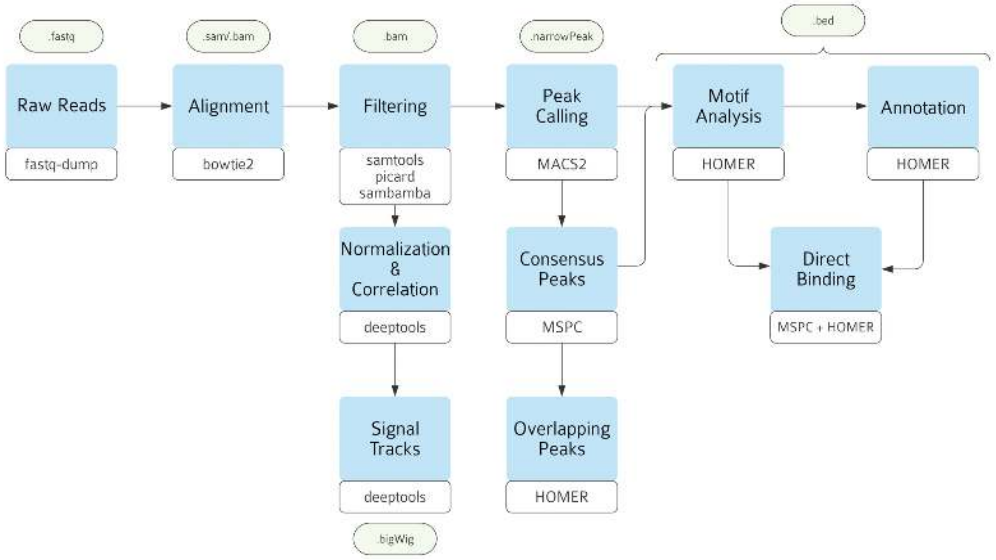


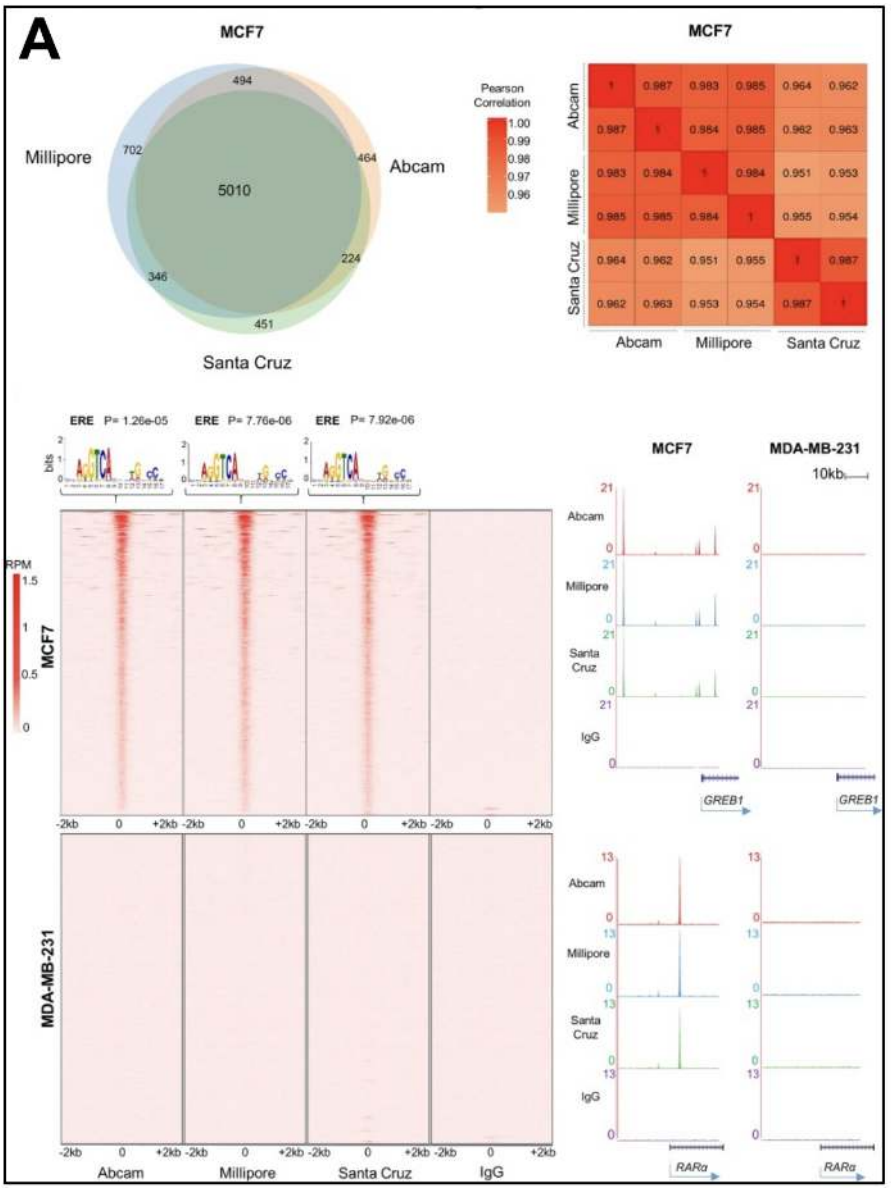
Figure 3.1: Overall look of ChIP-seq analysis. ChIP-seq analysis starts with raw DNA sequence downloading and aligning those read into a reference genome. Certain filters are applied to only leave good quality reads. Fundamental steps, peak calling and motif analysis provides the TF binding sites and their motif occurrences. Biological replicates are combined for proceeding uniformly, and in an easier way for the further steps. Different visualizations are used to represent binding of target TFs and their certain properties. Different data formats are generated for the proceeding steps in order to represent genomic position essentially.

3.2 Reproducibility of the Analysis

In order to demonstrate the accuracy of the optimized methodology, a published dataset (GSE128208) [293] was analyzed and the results were compared. The reference figure represents collective data analysis from Glont et al. (Figure 3.2A) and the regenerated figure by performing the optimized pipeline on the same dataset (Figure 3.2B), respectively. In the reference study [293], the performance of three distinct ERα antibodies was evaluated. Each biological replicate has 3 or 2 runs of fast files. These needed to be merged in order to proceed with the regular analysis. ,Both of the analyses

exhibit overlapping peak consistency. Re-generated Pearson correlation resulted in a higher correlation between biological replicates of 3 antibodies. All the samples resulted in their mostly enriched motif. Peaks' occupancy on DNA portions resulted in the ERE motif as a meaningful part of DNA. It can be seen that the p-value is dramatically different; because we conducted motif analysis on the peaks, not the whole reads located throughout the whole genome. This will be explained in the following sections. Finally, enriched peaks located on the same genes can be seen. Here, input is used as negative control instead of IgG.

Altogether, ChIP-Seq analysis was optimized and its reproducibility was checked and confirmed.



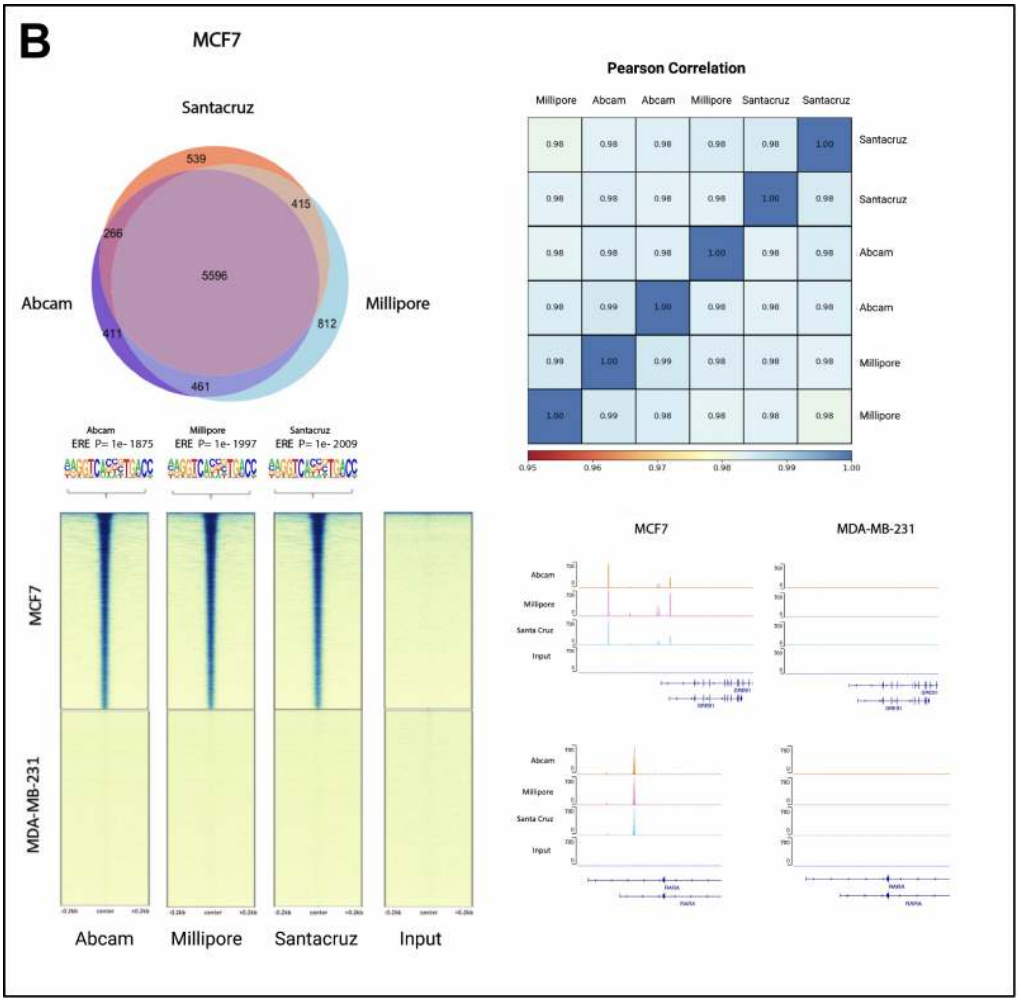


Figure 3.2: Reproducing the reference ChIP-seq analysis with the workflow generated in this study. Results are highly consistent between the reference study (A) and reproduced one (B). Peak numbers and overlapping portions are approximately the same between two studies. Pearson correlation differs slightly while the analysis in this study resulted in constant values for each sample. Heatmap enrichments and signals around the same genes are indicating that results are reproducible.

3.3 Detection of Total Peaks, Consensus Peaks and Direct Binding Peaks

Indicated GEO numbers are used to obtain datasets. They are further analyzed using the optimized workflow. Final numbers of individual peaks (summed), consensus peaks (obtained from individual peaks), and direct binding peaks.

Table 3.1: Results of the ChIP-seq analysis. First, all the individual samples are analyzed and consensus peaks are obtained. Motif analysis is performed to detect direct binding peaks. FOXA1 has the highest peak number in all three sections relative to the cell line.

Cell line	ChIP-AB	GEO NUMBER	Chr1q Peaks	Consensus Chr1q Peaks	Direct(MSPC) Chr1q Genome
MCF7	ERα	GSE75779 GSE68355 (3) GSE126004 (4) (8 replicates in total)	6749	2.559	150
T47D		GSE75779 GSE68355 (3) GSE126004 (4) (8 replicates in total)	3807	1.692	98
ZR75.1		GSE153251 (4) (4 replicates in total)	3528	1226	98
Cell line	ChIP-AB	GEO NUMBER	Chr1q Peaks	Consensus Chr1q Peaks	Direct(MSPC) Chr1q Genome
MCF7	GATA3	GSE32465 GSE51274 (3 replicates in total)	4674	2547	702
T47D		GSE51274 (2 replicates in total)	1694	1131	93
Cell line	ChIP-AB	GEO NUMBER	Chr1q Peaks	Consensus Chr1q Peaks	Direct(MSPC) Chr1q Genome
MCF7	FOXA1	GSE114737 GSE80808 GSE81714 GSE126004 (4) (7 replicates in total)	14511	2.707	945
T47D		GSE126004 (4 replicates in total)	6670	2.595	978

In the result of the analysis, MCF7 cells consist of the highest peak numbers for all TFs. In the TF manner, FOXA1 has the highest peak number in each cell line. All individual peaks (*Chr1q peaks* in the table) obtained from different samples are combined and the number of peaks are reduced. Finally, significantly less number peaks are remain as *direct binding peaks.*

3.3.1 Motif Analysis on Direct - Indirect Consensus Peaks

Motif analysis summarized all the other factors that could have a role to recruit indicated TF. Figure 3.3 reveals cell type dependent changes in ERα, FOXA1 and GATA3 consensus peaks.

Figure 3.3: The most frequent motifs found in each TF-cell line pair. Motif analysis conducted on subset of the peaks; direct and indirect. Then, the first ten motifs are represented as a wordcloud based on their frequency. Two subset of the peaks revealed different motifs. The first motif based on its frequency of occurrence, differs from the most significant motif. FOXA1 is the dominant motif among three cell lines and three TFs especially for ERα.

FOXA1 is a highly enriched motif in ERα peaks especially in MCF7 and ZR751, showing that FOXA1 exhibits pioneering activity for the ERα binding sites.

Interestingly, COUP-TFII, which is a ligand inducible nuclear receptor acting as a transcriptional activator or repressor dependent on the cell type [294–296], is enriched only in the T47D cell line which indicates the cell line specific role of TFs. Moreover, since GATA3 is mutant in MCF7 cell line the most enriched motif is FOXA1, whereas T47D cell line harbors wild-type GATA3 which reflects itself as the most dominant motif in this cell line.

3.3.2 Annotation of the Consensus Peaks

Annotation of the ERα, GATA3, and FOXA1 display binding to the non-coding regions that display regulatory activity for the genes. Namely, intergenic, intronic and promoter regions.

Figure 3.4: ERα, GATA3, and FOXA1 mainly bind to intronic and intergenic regions. ERα, GATA3, and FOXA1 peaks are annotated according to the peaks' location. They prevelantly bind to intergenic regions of the human genome as well as introns and promoters of the genes.

This analysis could be also considered as a validation of the ChIP-seq analysis with annotating ERα, GATA3 and FOXA1 peaks to the expected regions. Most of the peaks are located at the intergenic regions as well as introns.

3.3.3 Overlapping Consensus Peaks

To determine the shared binding sites of each TF among the indicated cell lines, overlapping peaks are determined from the consensus peaks, using the HOMER mergePeaks function. Shared TF binding sites have a higher probability of being preserved over the evolution and involving an essential function in the cell.

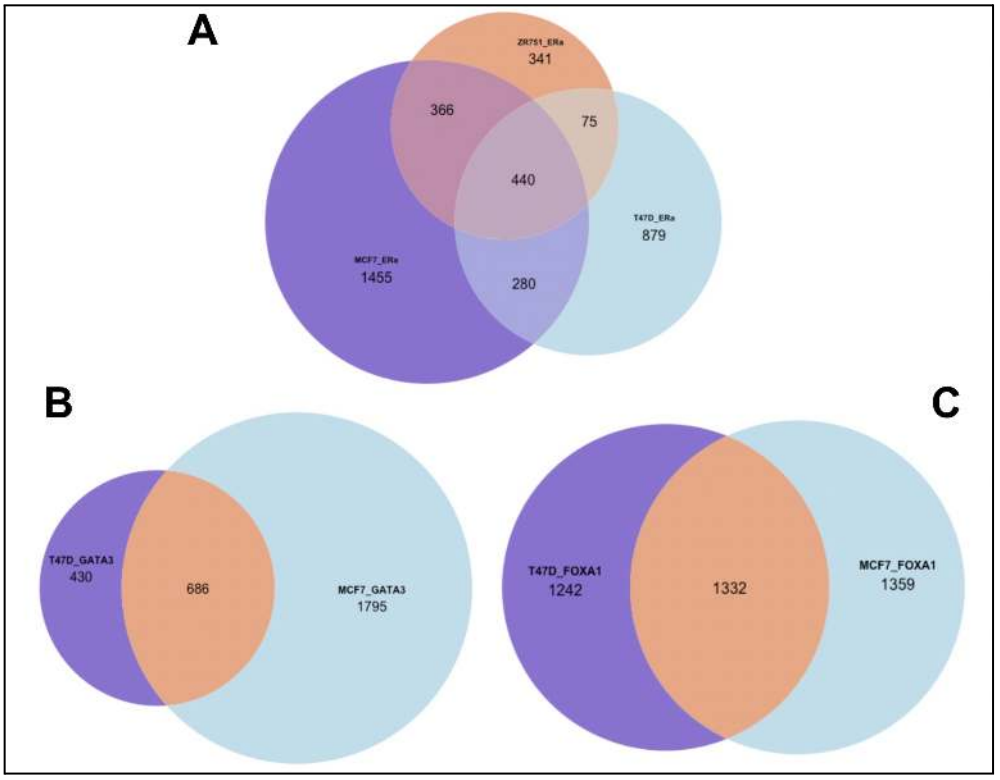


Figure 3.5: The numbers of ERα, GATA3 and FOXA1 overlapping binding sites. A) ERα peaks in MCF7, T47D and ZR75-1 cells. The highest number of peaks present in MCF7 and the lowest number of peaks in ZR75-1 cell line. B) More than half of the GATA3 peaks in T47D are shared with MCF7 while MCF7 has a slightly higher number of peaks. C) FOXA1 has the most consistent binding among two cell lines compared to two other TFs. Also, half of the FOXA1 peaks are shared approximately the same amount.

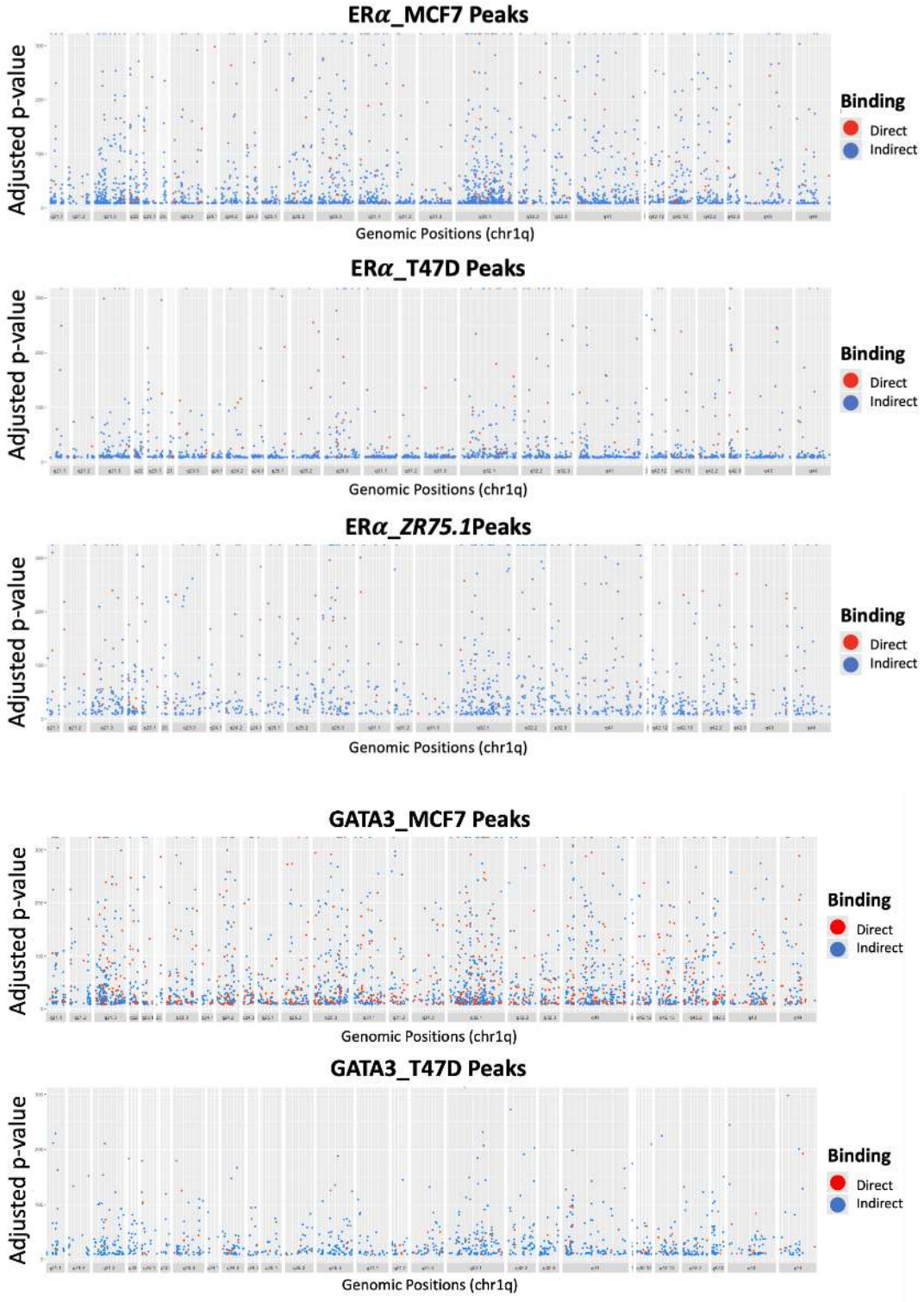
3.4 Distribution of the TF binding sites

3.4.1 According to adjusted p-value

All the direct and indirect binding peaks are mapped to the related cytoband on chromosome 1q with their adjusted p-values. Adjusted p-value is used as a statistical significance of the binding and it is calculated from the input -10log10(p-value) in the individual peaks.

The adjusted p-value is the Benjamini–Hochberg corrected p-value in order to control false discovery rate in multiple testing. Scatter plots are plotted according to their adjusted p-value. For the adjusted p-value, the significance of each peak is adjusted for multiple testing. This forms the combination of the evidence in replicates and increases the number of obtained peaks up to double what is obtained with a single sample at the same stringency [280].

ERα binding in MCF7, strongly, and in ZR75-1, slightly has a distinguishable pattern that accumulates especially in q21.3, q22 and q32.1 bands. Similarly, GATA3 peaks also show intense binding on those cytobands specifically in MCF7. T47D binding is seen as slightly weak and steady compared to MCF7 and ZR75-1. High direct and indirect binding profile is observed in FOXA1 peaks. This is not due to the peak number, but the ability to bind the DNA. For example, even if the replicate number is higher in ERα, the peak numbers are less than FOXA1 in MCF7. Accordingly, the highest direct binding is seen in GATA3 in MCF7 cells and FOXA1 in both cell lines.


ERα_MCF7 Peaks
Adjusted p-value
Genomic Positions (chr1q)
Binding
Direct
Indirect
ERα_T47D Peaks
Adjusted p-value
Genomic Positions (chr1q)
Binding
Direct
Indirect
ERα_ZR75.1Peaks
Adjusted p-value
Genomic Positions (chr1q)
Binding
Direct
Indirect
GATA3_MCF7 Peaks
Adjusted p-value
Genomic Positions (chr1q)
Binding
Direct
Indirect
GATA3_T47D Peaks
Adjusted p-value
Genomic Positions (chr1q)
Binding
Direct
Indirect

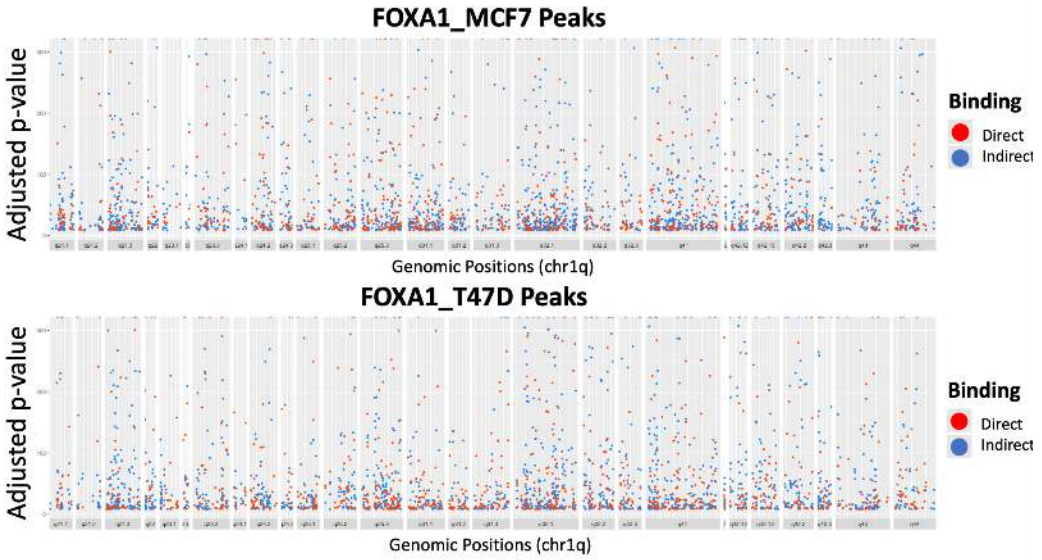


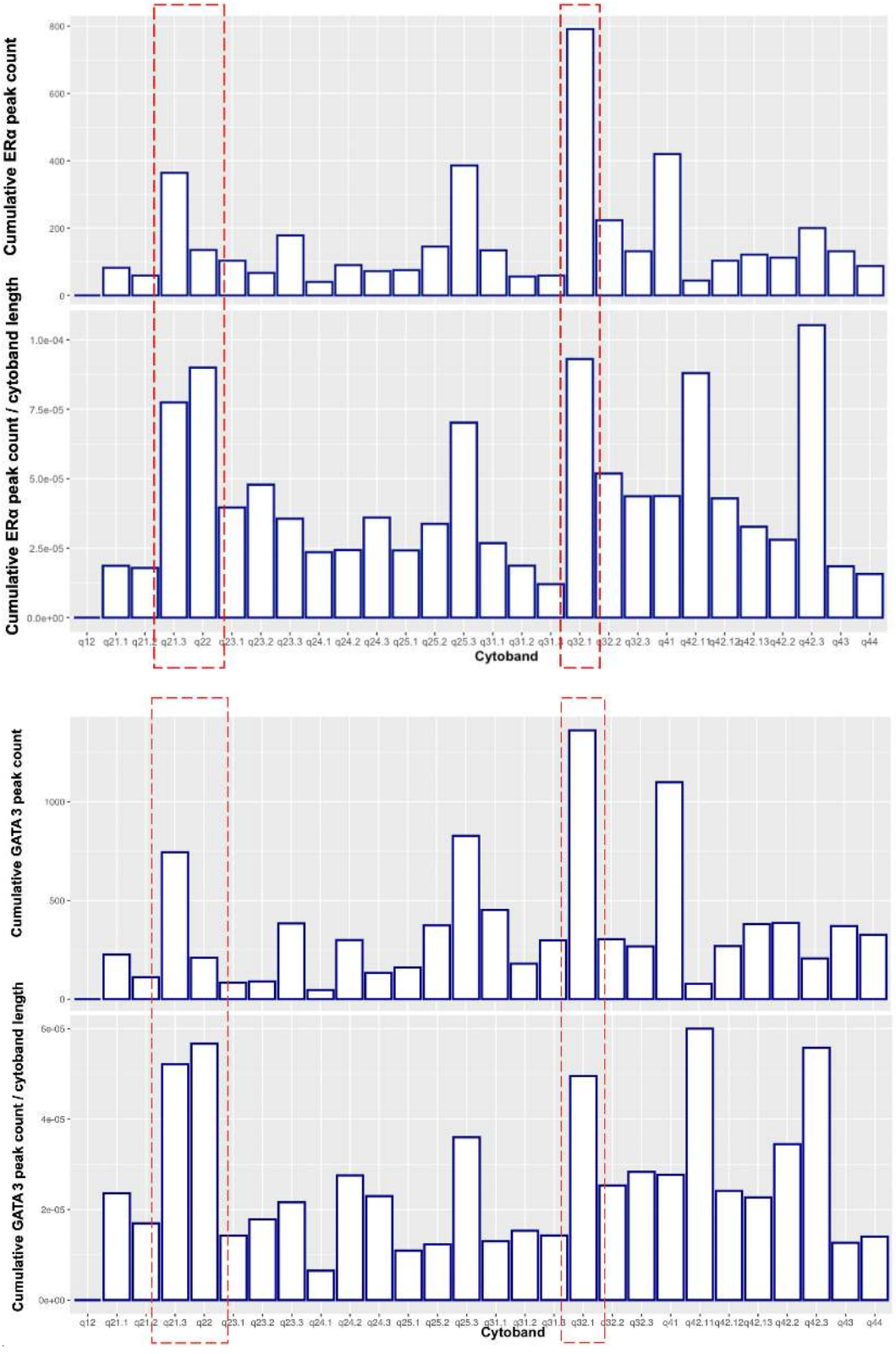
Figure 3.6: Scatter plot of binding of TFs along chr1q. Direct (red) and indirect (blue) peaks (Cytobands are represented relative to their base pair lengths and each point in the plot represents a peak). ERα and GATA3 peaks can be seen accumulated in the q21.3, q22 and q32.1 regions in three cell lines but most evidently in MCF7 cells. Also, the binding intensity is looking different due to the replicate numbers and consequently, peak numbers. FOXA1 has a dispersed binding throughout 1q. Additionally, FOXA1 peaks serve the most direct binding pattern than ERα and GATA3.

Eventually, ERα and GATA3 peaks prominently present in q21.3, q22 and q32.1 cytobands have more intensive binding while FOXA1 has a generic binding referring to its pioneering abilities for the ERα binding.

3.4.2 According to peak numbers

Next, to further investigate the binding quantitatively, instead of adjusted p-value, the peak numbers (obtained with peak calling and consensus peaks) corresponding to the relative cytoband are taken. Cytoband length has an influence over the total number of peaks present in a given cytoband. As an example, q22 cytoband is seen to have less binding than q21.3 especially for ERα and GATA3 (Figure 3.7). However, when all the counts are divided by the length in bp of the cytoband, q22 binding is amplified even though it is one of the shortest cytobands across chromosome 1q (1500000 bp). On the

other hand, q32.1 being not the largest band (8500000 bp), still exhibits one of the strongest binding among all cytobands of chr1q.



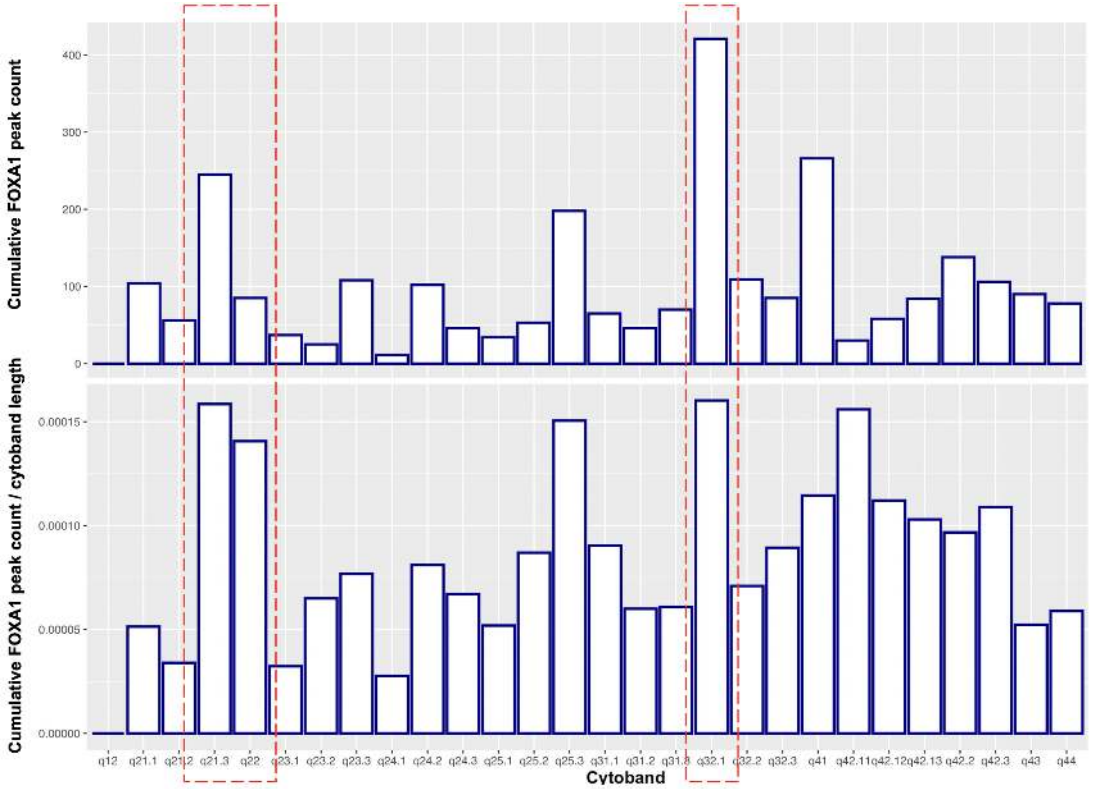


Figure 3.7: Bar plots are representing the sum of the peak counts and it is divided by the cytoband length. Here, two different representations are showing the peak counts on the cytoband and the peak count ratio relative to the cytoband length of the peak. Previously determined cytobands are clearly seen in this analysis too. The highest point is determined to belong q32.1.

Here, the peak counts reveal distinct binding intensities for ERα, GATA3 and FOXA1 peaks across chr1q cytobands. Along with the adjusted p-value evaluation of the peak distribution, there is a significant evidence that q21.3, q22 and q32.1 cytobands confer specific ERα and GATA3 binding that could harbor abnormal gene regulation in MCF7, T47D and ZR75-1 cells.

3.5 CRISPR Library design

In order to assess the functional TF binding sites and the cellular phenotypes of distinct binding patterns, CRISPR-based functional genetic screening approach to perturb decisive binding sites of this study will be followed. This part of the study aims

to detect altered genomic sites by copy number gains in chr1q which hopefully will shed light on the regions having role in breast cancer progression and metastasis. For this purpose, CRISPR-DO algorithm is benchmarked and used as it is discussed in the materials methods section.

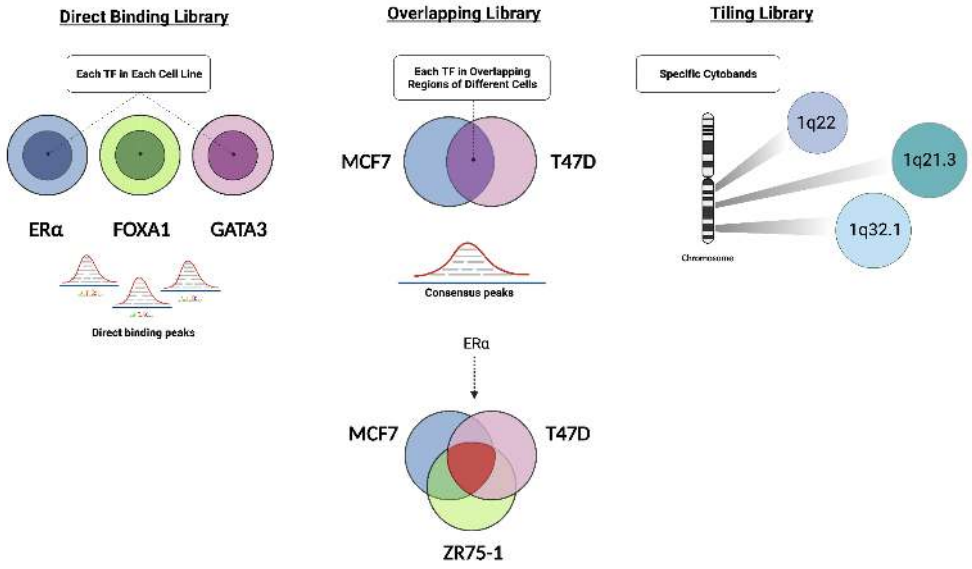


Figure 3.8: Representation of different libraries designed in order to investigate distinct genomic functions. Distinct sources of chromosomal locations are used in the generation of libraries. For the direct binding and overlapping sgRNA designs, peak coordinates are used directly. For the cytoband specific targeting, determined cytoband coordinates are used after discarding the gene locations.

In order to understand the role of three cytobands determined as a result of having both significant binding patterns and higher peak number, CRISPR-based libraries targeting those cytobands are also generated. Cytoband coordinates are obtained from UCSC Table Browser and sgRNA design is conducted only on specific cytobands demonstrating distinguished binding intensity.

3.5.1 Direct Binding sgRNA Design

One of the CRISPR libraries is targeting the binding regions of transcription factors of interest in the individual cell lines. For this purpose, the direct binding

locations stated in the section *Obtaining Directly Bound Consensus Peaks* are firstly extended into 100 bp regions. All the locations are iterated for CRISPR-DO and gRNA sequences are obtained for each location separately. Then, they are pooled for the same TF but different cell lines, and a filter is applied. For filtering, the efficiency cut-off score is set to 0.3 which is the recommended minimum score, and the specificity cut-off score is set to 60. Then, for obtaining the locations that are lost after applying cut-off filtering, all the files are labeled if they have results for that specific location. Followingly, the locations are determined if they have results within the 100 bp region. As the next step, the initial regions which resulted in either 0 sgRNA or sgRNAs that are not passing through the cut-off are extended into 200 bp regions and the process is applied again. After cutting the output sequence from its PAM sequence, the resultant 20 bp sgRNAs pooled, and duplicates are removed.

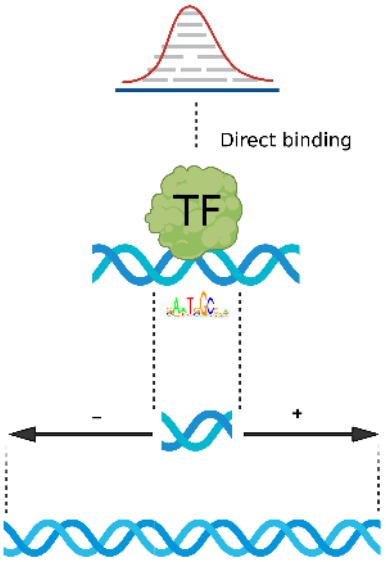


Figure 3.9: Location determination of the sgRNA design on direct binding peaks. Direct binding peak locations are determined with their motif bearing. Next, peak is extended to 100 bp with aiming strict targeting on the ERα, GATA3 and FOXA1 binding sites.

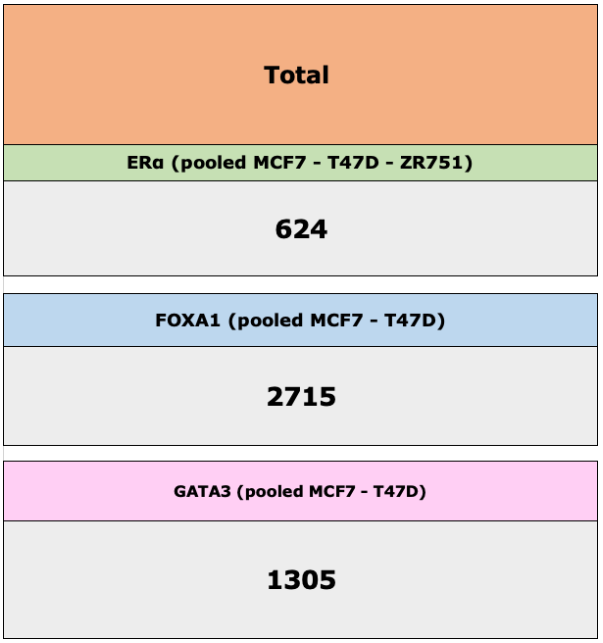
Table 3.2: sgRNA numbers in the first part of the direct binding analysis. The sgRNAs are pooled with respect to the TF. Filtering is done multiple times. Highest sgrRNA number is obtained for FOXA1, as expected.

100 bp						
ERα (pooled MCF7 - T47D - ZR751)						
total gRNA (duplicates removed)	efficiency cutoff 0.3	specificity cutoff 50	specificity cutoff 60	specificity cutoff 70	specificity cutoff 75	specificity cutoff 80
1431	738	591	530	433	376	292
FOXA1 (pooled MCF7 - T47D)						
total gRNA (duplicates removed)	efficiency cutoff 0.3	specificity cutoff 50	specificity cutoff 60	specificity cutoff 70	specificity cutoff 75	specificity cutoff 80
4889	2552	1892	1667	1329	1109	834
GATA3 (pooled MCF7 - T47D)						
total gRNA (duplicates removed)	efficiency cutoff 0.3	specificity cutoff 50	specificity cutoff 60	specificity cutoff 70	specificity cutoff 75	specificity cutoff 80
2363	1224	945	837	674	557	416

Table 3.3: sgRNAs after the extension of the region into 200 bp. These sgRNAs are the "rescued" sgRNAs coming from the regions with zero sgRNAs in the 10 bp region. FOXA1 acquired a high number of sgRNAs, followed by GATA3.

200 bp		
ERα (pooled MCF7 - T47D - ZR751)		
total gRNA (duplicates removed)	efficiency cutoff 0.3	specificity cutoff 60
489	200	94
FOXA1 (pooled MCF7 - T47D)		
total gRNA (duplicates removed)	efficiency cutoff 0.3	specificity cutoff 60
4851	2251	1048
GATA3 (pooled MCF7 - T47D)		
total gRNA (duplicates removed)	efficiency cutoff 0.3	specificity cutoff 60
2032	893	468

Table 3.4: Total number of the sgRNAs. Highest sgRNA number belong to FOXA1 as expected, then GATA3 and finally ERα.



Two interval lengths are used to obtain direct binding sgRNAs. In order to observe the losses in specificity and efficiency cutt-offs. For the efficiency, 0.3, and for the specificity, 60 determined as the thresholds.

3.5.2 Overlapping sgRNA Design

In this section, the overlapping peaks of the same transcription factor (FOXA1 and GATA3) across two cell lines (in the case of ERα, three cell lines) are subjected to sgRNA design.

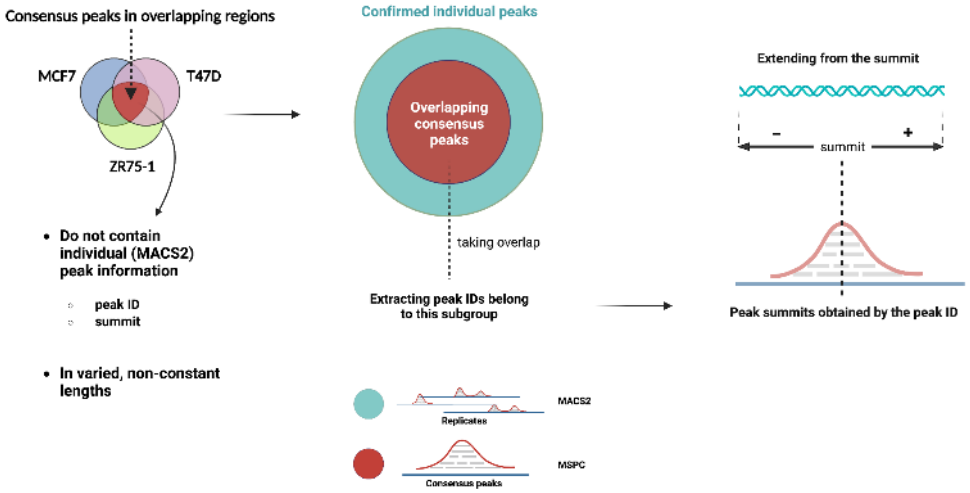


Figure 3.10: Scheme of location determination of overlapping sgRNA design. Summit locations are obtained in a reversal approach using consensus peaks and individual peak IDs. Consensus overlapping peaks are overlapped with confirmed individual peaks. After obtaining the individual peaks along with their IDs, their summits are acquired and extended. Resultant intervals are used in the library design.

For this purpose, firstly, overlapping regions are obtained as portions of consensus peaks. Consensus peaks consist of individual replicates, and identities of the individual peaks are not available in consensus peaks. For obtaining the individual peaks, a reverse methodology is designed by taking the overlap of consensus peaks and the individual peaks that belong to one TF in one cell line. Since individual peaks have their unique peak IDs along with their summit location, summits are extracted for each peak using the individual peaks' IDs. In order to scan the 100 bp region, 50 bp is extended to the left and right, and in total 101 bp region is obtained. Design analysis followed by duplicate removal is conducted on those regions. Since peaks are considered individually, duplicates remain a significant portion of the total sgRNAs and they are removed.

Table 3.5: Total number of sgRNAs obtained from overlapping regions. Duplicate sgRNAs exhibit significant noise compared with the total. After removal of duplicate sgRNAs, ERα lost more than two thirds of its number while still having sgRNAs more than GATA3. FOXA1 sgRNAs decreased approximately two-fold.

MCF7 - T47D - ZR751				
ERalpha				
Consensus Overlapping Regions	Total Regions (coming from individual peaks through consensus regions)	Total sgRNA	Cut-offs	sgRNAs after Duplicates Removed after Crop (20bp)
440	4408	27543		8134
			Efficiency cut-off 0.3	4625
			Specificity cut-off 60	3235
			Specificity cut-off 70	2693
			Specificity cut-off 75	2335
			Specificity cut-off 80	1898
			Specificity cut-off 85	1392

MCF7 - T47D				
FOXA1				
Consensus Overlapping Regions	Total Regions (coming from individual peaks through consensus regions)	Total sgRNA	Cut-offs	sgRNAs after Duplicates Removed after Crop (20bp)
1332	9681	34870		11557
			Efficiency cut-off 0.3	6175
			Specificity cut-off 60	4308
			Specificity cut-off 70	3515
			Specificity cut-off 75	3019
			Specificity cut-off 80	2387
			Specificity cut-off 85	1635

MCF7 - T47D				
GATA3				
Consensus Overlapping Regions	Total Regions (coming from individual peaks through consensus regions)	Total sgRNA	Cut-offs	sgRNAs after Duplicates Removed after Crop (20bp)
686	2750	10444		5082
			Efficiency cut-off 0.3	2724
			Specificity cut-off 60	1946
			Specificity cut-off 70	1570
			Specificity cut-off 75	1334
			Specificity cut-off 80	1063
			Specificity cut-off 85	741

In the end of the library generation, efficiency cut-off is selected 0.3 again, to obtain the possible sgRNAs present in the loci. Number of sgRNAs based on different specificity cut-off calculated.

3.5.3 Tiling Library Design

Designing tailor-made libraries targeting 1q22, 1q21.3, and 1q32.1 starts with taking the genomic coordinate of each cytoband. Next, the locations of the genes, as well as their 250 bp upstream and downstream regions, are discarded from cytobands in order to eliminate targeting gene-related regulatory elements. The remaining regions were used for the sgRNA design analysis.

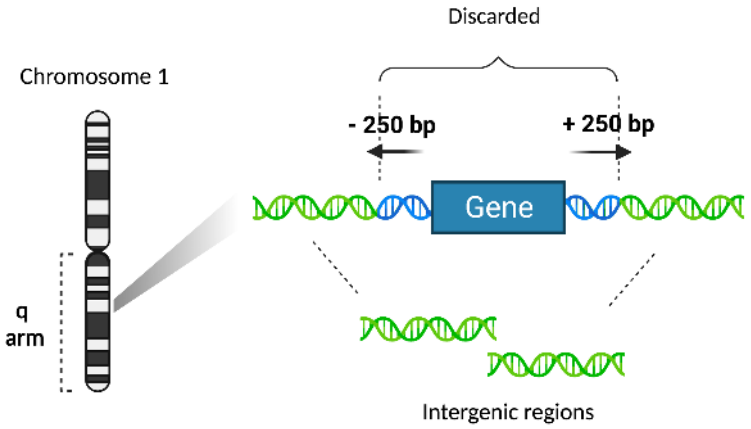


Figure 3.11: Location determination of the sgRNA design on cytobands targeting intergenic regions. The cytoband chromosome positions are retrieved. Gene positions located on the relative cytoband are extended 250 bp both from upstream and downstream. These extended intervals are subtracted from the cytobands and library design is conducted on the remaining regions.

Table 3.6: Total sgRNA numbers obtained cytobands after cutt-off filterings. Efficiency cut-off is selected as lowest, in order to not to eliminate possible sgRNAs that might be cutting the target loci. Whereas specificity score is considered more important to cut the target loci instead of any other functional region that can form an undesired phenotype.

q22									
Total sgRNAs	Gene Locations Substracted	Duplicates Removed	efficiency cutoff 0.7	specificity cutoff 50	specificity cutoff 60	specificity cutoff 70	specificity cutoff 75	specificity cutoff 80	specificity cutoff 85
113627	17666	15670	2195	1061	906	724	585	462	328
Total sgRNAs	Gene Locations Substracted	Duplicates Removed	efficiency cutoff 0.3	specificity cutoff 50	specificity cutoff 60	specificity cutoff 70	specificity cutoff 75	specificity cutoff 80	specificity cutoff 85
113627	17666	15670	8596	4146	3573	2801	2304	1789	1218

q21.3									
Total sgRNAs	Gene Locations Substracted	Duplicates Removed	efficiency cutoff 0.7	specificity cutoff 50	specificity cutoff 60	specificity cutoff 70	specificity cutoff 75	specificity cutoff 80	specificity cutoff 85
293641	101539	89537	11265	5656	4830	3614	2906	2136	1363
Total sgRNAs	Gene Locations Substracted	Duplicates Removed	efficiency cutoff 0.3	specificity cutoff 50	specificity cutoff 60	specificity cutoff 70	specificity cutoff 75	specificity cutoff 80	specificity cutoff 85
293641	101539	89537	47334	24867	21487	16645	13515	9851	6148

q32.1									
Total sgRNAs	Gene Locations Substracted	Duplicates Removed	efficiency cutoff 0.7	specificity cutoff 50	specificity cutoff 60	specificity cutoff 70	specificity cutoff 75	specificity cutoff 80	specificity cutoff 85
516227	184255	165969	22375	11005	9412	7227	5841	4289	2807
Total sgRNAs	Gene Locations Substracted	Duplicates Removed	efficiency cutoff 0.3	specificity cutoff 50	specificity cutoff 60	specificity cutoff 70	specificity cutoff 75	specificity cutoff 80	specificity cutoff 85
516227	184255	165969	87799	48789	42315	32845	26696	19726	12682

Thresholds are again applied multiple times. Efficiency is determined as 0.3 in this study as explained in the previous library design sections. Each cytoband has a total number of sgRNAs with respect to their bp lengths. Namely, q21.3 - 4700000, q22 - 1500000 and q32.1 - 8500000. Expectedly, highest number of sgRNAs present in the q32.1 arm.

3.6 Associating ChIP-seq peaks to the nearest gene expression

After the analysis of RNA-seq datasets, PCA for normal breast and breast cancer resulted qith 61% variance between the groups and 11% between the replicates. Next, the reliable RNA-seq and ChIP-seq data is integrated by the *HOMER annotatePeaks.pl* tool using "-gene" option. This step includes the matching of Ensembl IDs of the genes that are near to a TF binding site, and the gene expression of that gene having the same Ensembl ID (Figure 3.13).

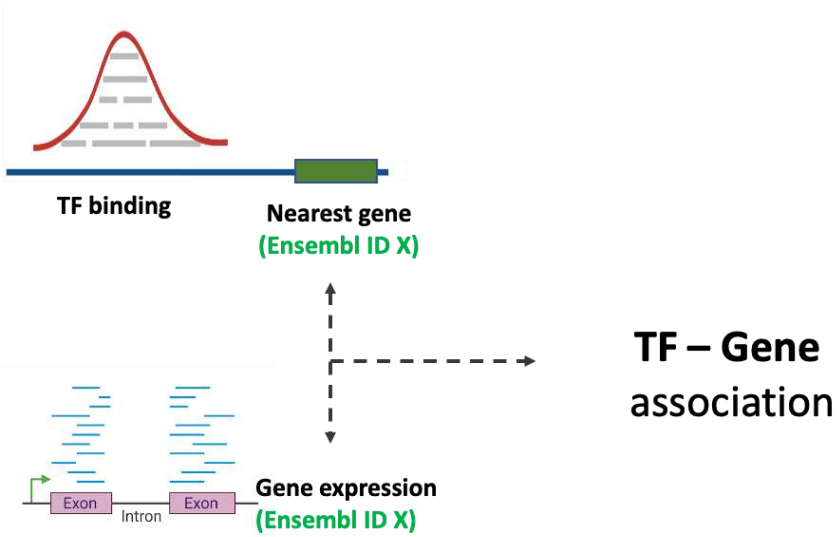


Figure 3.12: TF binding and gene expression association. TF binding site could be annotated to a gene, and the gene's Ensembl ID is obtained. Similarly, the Ensembl ID of the same gene is obtained. The association is conducted by matching the IDs.

Table 3.7: Average counts of each cell line. Each cell line containing four replicates is averaged. Also, averages of breast cancer and normal breast cell lines are shown. Avgarege breast cancer count is not used in this step. There is no outlier cell line count that changes the average dramatically.

Avg MCF7 Count	Avg ZR751 Count	Avg T47D Count	Avg Breast Cancer count	Avg Normal count	Avg 184A1 Count	Avg 184B5 Count	Avg MCF10A Count	Avg MCF12A Count
2025,421571	864,4284039	1922,127149	1603,992375	52,60770941	29,6933973	93,75563089	33,36548829	53,61632114
3514,127278	1863,374932	2792,0292	2723,177137	384,6100222	139,6129025	674,2184457	92,99603477	631,6127058
615,9114592	1230,42152	1120,477943	988,9369741	194,3079857	123,8768742	305,5780776	120,4144319	227,3625591
207,3091863	181,3221997	109,2935243	165,9749701	25,92990403	4,970280272	35,68925937	53,22691652	9,833159966
6524,627075	6360,032385	7289,168743	6724,609401	1477,463536	594,8859916	2747,671653	1793,166813	774,1296883
2329,241694	4529,344628	3006,025685	3288,204002	860,3958762	549,2595369	1124,800502	947,7538708	819,7695953
1433,426339	71,61563667	269,3162185	591,4527314	17,31839582	9,472304349	23,84304736	13,61805202	22,34017955
392,8916674	726,3666242	129,7628348	416,3403755	4924,031312	985,776838	9576,239636	4299,874481	4834,234293
4745,720563	3573,189966	5538,851793	4619,254107	18071,18593	28419,55175	21628,21335	14848,77478	7388,203854

The approach is similar to the peak annotation. The annotation step includes associating TF peaks to the nearby gene with its Ensembl ID. In this step, user-provided gene expression data is associated with the relative ChIP-seq peaks by taking Ensembl IDs as reference. As the gene expression data, average counts of MCF7, T47D and ZR75-1 cell lines containing four replicates each are used. Since the ChIP-seq peaks are in MCF7, T47D and ZR75-1 cell lines, normal breast cell line counts are not used in this step. As a result, **180, 121, 133, 178, 178, 177** and **105** genes are found to be associated with MCF7-ERα, T47D-ERα, ZR75-1-ERα, MCF7-FOXA1, T47D-FOXA1, MCF7-GATA3 and T47D-GATA3 peaks, respectively.

3.6.1 Determination of common genes

In this section, only cytobands exhibiting a distinguished TF binding pattern are considered for the selection of the possible genes. For this, we focused on 1q21.3, 1q22 and 1q32.1 as indicated in *Overall binding on chr1q distinguishes on specific cytobands* section. Next, genes located on those cytobands are further filtered according to their commonality in MCF7, T47D and ZR75-1 for ERα and MCF7 and T47D for GATA3 ChIP-seq – RNA-seq data. FOXA1 is not examined in this section due to its more generic binding pattern. Remaining genes are also filtered according to their gene count as being greater than 10. After the filtering, nine genes are determined: *ANXA9, CGN, EFNA3, GOLT1A, PLEKHA6, SOX13, GPR37L1*, *S100A9* and *S100A10*.

Table 3.8: Common genes determined after filtering. The decisive genes and their locations can be seen in the table.

Gene	Cytoband	Gene Alias	
ANXA9	q21.3	Annexin A9	Calcium-dependent phospholipid-binding protein
CGN	q21.3	Cingulin	Cytoplasmic scaffold protein localized at tight junctions
EFNA3	q22	Ephrin A3	Cell surface GPI-bound ligand for tyrosine kinase ephrin receptors
GOLT1A	q32.1	Golgi Transport 1A	ER-to-Golgi transport
PLEKHA6	q32.1	Pleckstrin Homology Domain Containing A6	Trafficking and retention of transmembrane proteins at basolateral membranes and apicolateral adherens junctions
SOX13	q32.1	SRY-Box Transcription Factor 13	A member of the SOX family of transcription factors
GPR37L1	q32.1	G Protein-Coupled Receptor 37 Like 1	Family A orphan G protein-coupled receptors (GPCR)
S100A9	q21.3	S100 Calcium Binding Protein A9	Found as calprotectin (S100A8/A9), regulates inflammatory processes and immune response
S100A10	q21.3	S100 Calcium Binding Protein A10	Binding partner of ANXA2, regulates ion channels and GPCRs

3.6.2 Enrichment of ERα and GATA3 binding around the common genes

Next, the signal of ERα and GATA3 around the common genes is investigated using IGV (Figure: 3.12). There are distinct patterns in the enrichment. *ANXA9, CGN, EFNA3, GOLT1A, PLEKHA6, SOX13, GPR37L1*, *S100A9* and *S100A10*.

Although signal strength is changing from gene to gene, the general scheme of the binding signals are indicative of regulatory roles of ERα and GATA3 for the determined genes.

GPR37L1: Significant enrichment of ERα binding is observed in both cell lines, whereas GATA3 has a higher binding affinity in MCF7.

PLEKHA6: *PLEKHA6* gene is in close proximity to another common gene, *GOLT1A*. They both have signals in intergenic regions which are strongly seen in each track, except for GATA3 - T47D.

EFNA3: GATA3 binding reveals a stronger signal in both cell lines than ERα peaks located in the intronic region on ZR75.1 and T47D and in the upstream of the *EFNA3* gene in MCF7.

ANXA9: Strong binding profile is observed at the promoter of *ANXA9*.

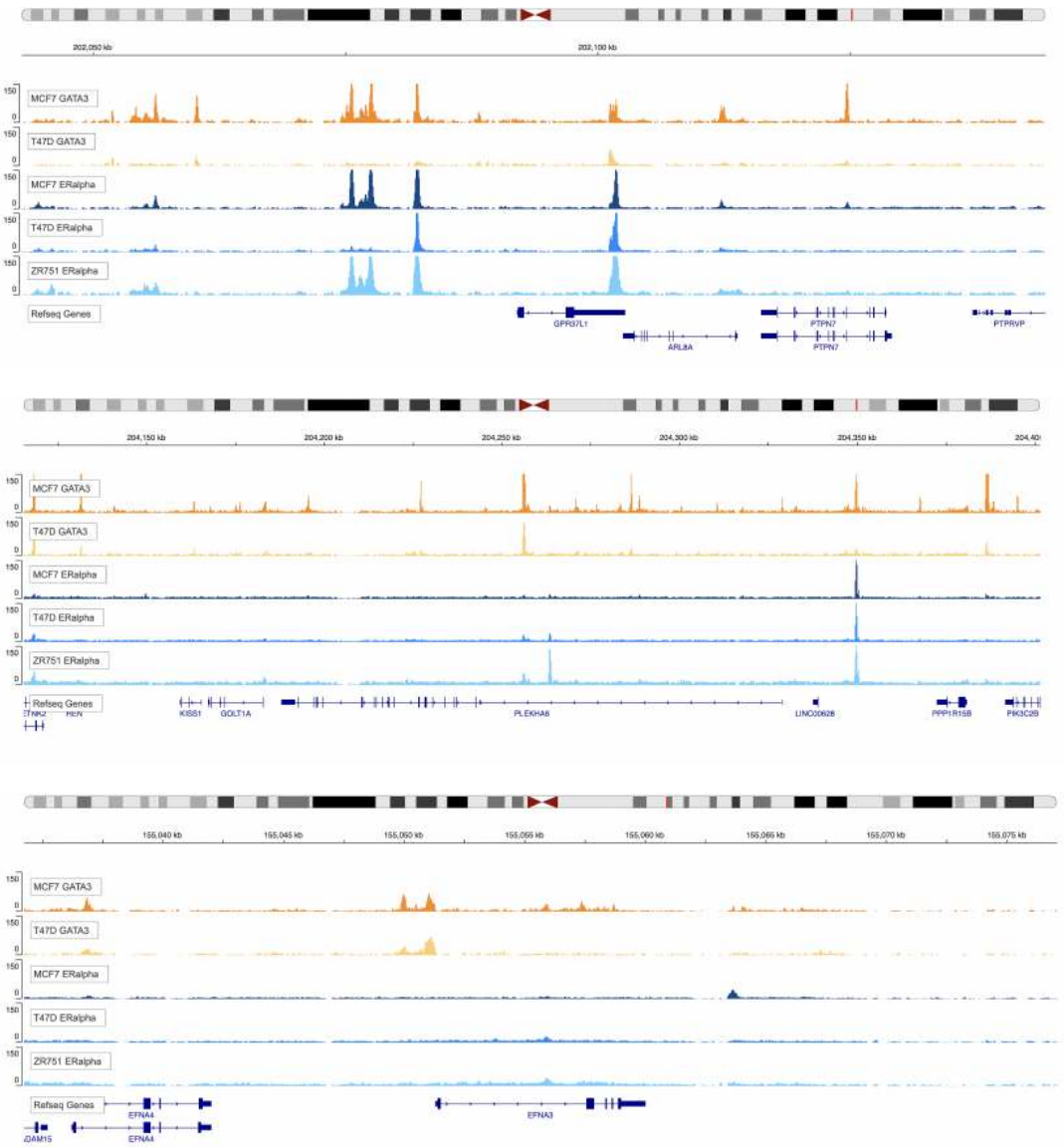
GOLT1A: Compared to other genes, *GOLT1A* has weaker signal strength.

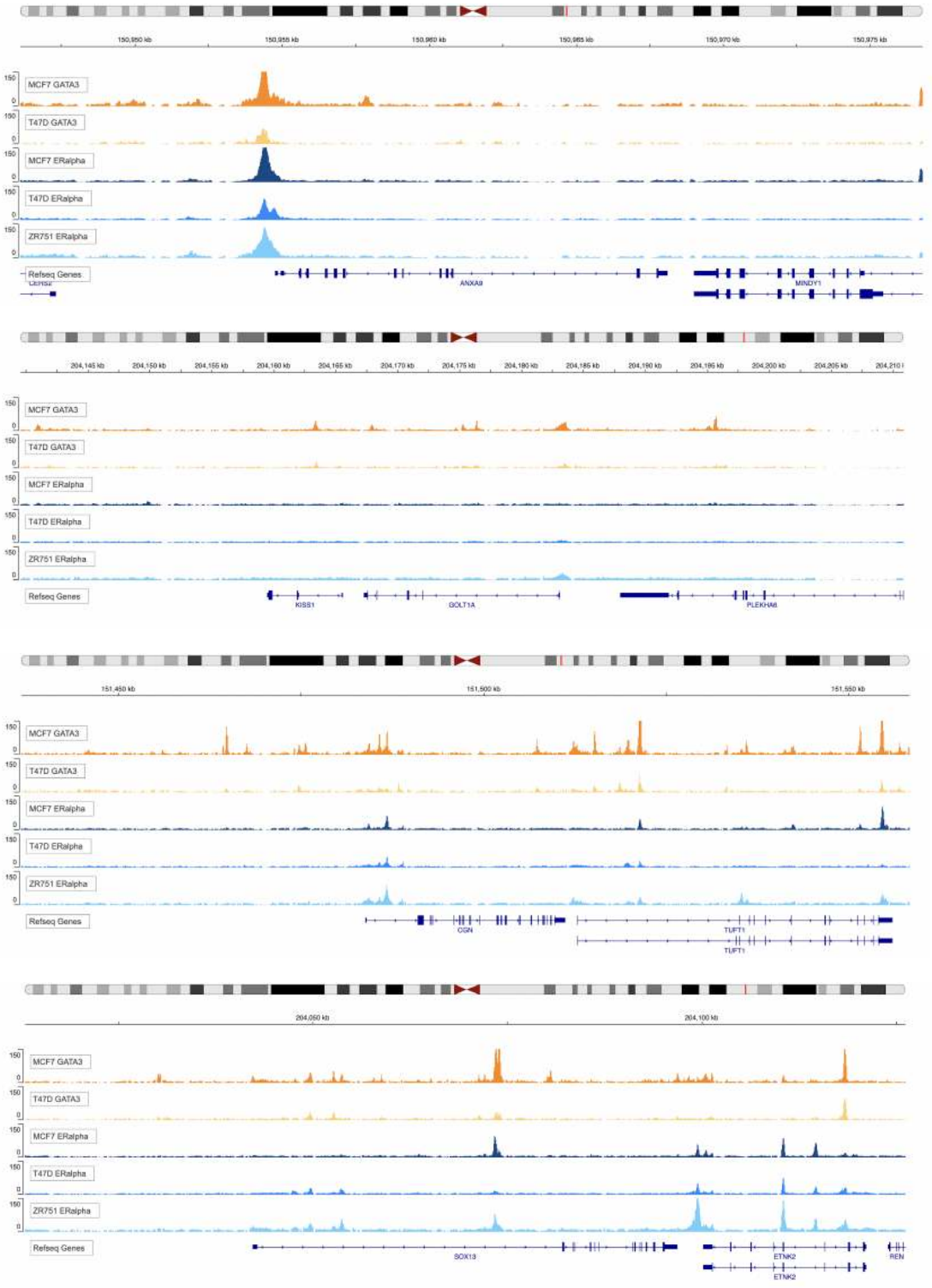
CGN: *CGN* has a weak binding of ERα compared to other genes, but it is consistent in 3 cell lines. GATA3 binding is slightly more in MCF7 cell line.

SOX13: SOX13 displays intronic peaks and it is more significant in ZR75-1 and MCF7 cells.

S100A9: There is binding of ERα and GATA3 in all cell lines and it is located at the downstream of S100A9 gene

S100A10: Similar to S100A9, S100A10 has also binding particularly downstream and upstream in distant regions.




150,950 kb
150,955 kb
150,960 kb
150,965 kb
150,970 kb
150,975 kb
MCF7 GATA3
T47D GATA3
MCF7 ERalpha
T47D ERalpha
ZR751 ERalpha
Refseq Genes
ANXA8
MINDY1
204,145 kb
204,150 kb
204,155 kb
204,160 kb
204,165 kb
204,170 kb
204,175 kb
204,180 kb
204,185 kb
204,190 kb
204,195 kb
204,200 kb
204,205 kb
KISS1
GOLT1A
PLEKHA6
151,450 kb
151,500 kb
151,550 kb
CSN
TUFT1
204,050 kb
204,100 kb
SOX13
ETNK2
REN

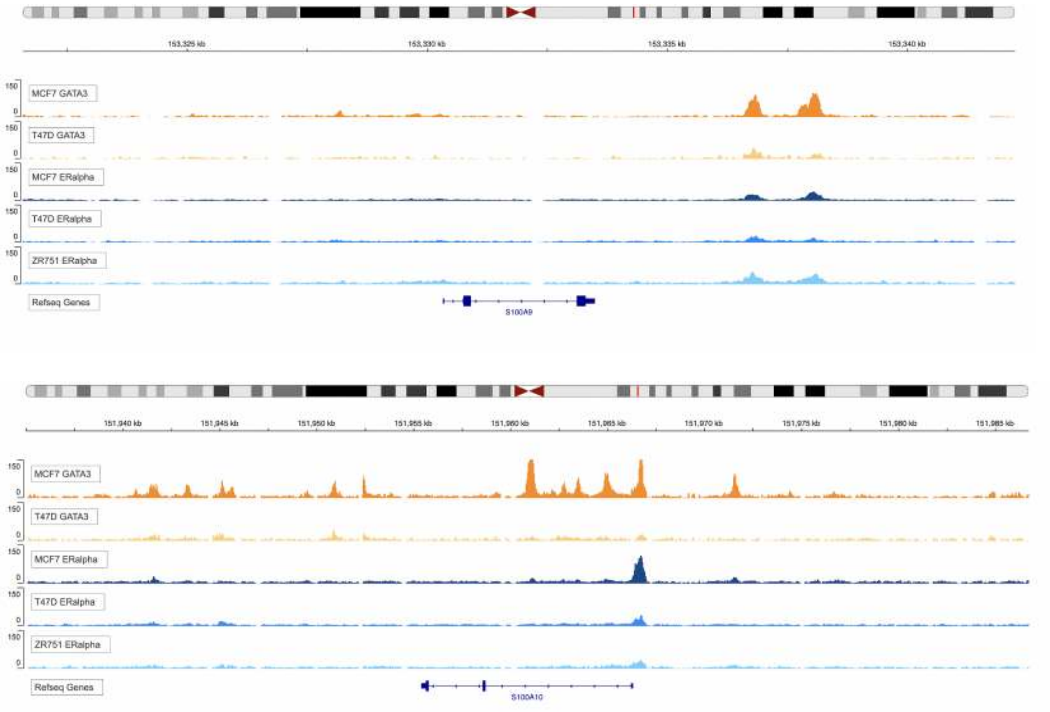


Figure 3.13: ERα and GATA3 exhibit high binding in the intergenic regions of all the nine genes. All the genes display enrichment of ERα and GATA3 around their locations.

3.7 Evaluation of transcription factor binding, gene expression and copy number change of the common genes

9 genes are further investigated in the context of copy number change. It is seen that *ANXA9, GOLT1A, PLEKHA6, SOX13, S100A9*, *S100A10* and *GPR37L1* genes have significantly high copy number in three of the cell lines except only *EFNA3* has copy number loss in MCF7 cell line. Among three cell lines, ZR75-1 implies the highest copy number gain for those genes followed by MCF7 (Table 3.9).

Table 3.9: Expression and copy number information of the genes. Ensembl IDs of the common genes and the cytobands they are located are shown. Cytobands are determined with ChIP-seq analysis and genes are filtered according to their RNA-seq counts and presence in all cell lines. Average gene expression count, log2 fold change of the breast cancer vs. normal-like cell lines, numbers present in the RNA-seq data, and log2 copy numbers are obtained from CCLE.

	Ensembl ID	Gene	Gene expression log2FC	Cytoband	MCF7 log2 Copy number	T47D log2 Copy number	ZR75-1 log2 Copy number	Avg Breast Cancer count	Avg Normal count
Upregulated	ENSG00000143412	ANXA9	4,76	q21.3	1,0563	0,489	1,0386	1603,992375	52,60770941
	ENSG00000143375	CGN	2,54	q21.3	1,0563	0,489	1,01	2723,177137	384,6100222
	ENSG00000143590	EFNA3	2,21	q22	-0,3317	0,4912	1,0297	988,9369741	194,3079857
	ENSG00000174567	GOLT1A	2,44	q32.1	0,8123	0,5055	1,232	165,9749701	25,92990403
	ENSG00000143850	PLEKHA6	2,03	q32.1	0,8123	0,5055	1,232	6724,609401	1477,463536
	ENSG00000143842	SOX13	1,86	q32.1	0,8123	0,5055	1,232	3288,204002	860,3958762
	ENSG00000170075	GPR37L1	4,85	q32.1	0,8123	0,5071	1,232	591,4527314	17,31839582
Downregulated	ENSG00000163220	S100A9	-3,23	q21.3	0,1072	0,4912	1,0297	416,3403755	4924,031312
	ENSG00000197747	S100A10	-1,81	q21.3	1,0563	0,7833	1,01	4619,254107	18071,18593

Average gene expression counts of normal-like and breast cancer cell lines showing difference for all the genes (Figure 3.14). Average breast cancer count is higher than normal breast count for *ANXA9, CGN, EFNA3, GOLT1A, PLEKHA6, SOX13, GPR37L1* except for *S100A9* and *S100A10*. Consistently, log2 fold change values indicate downregulation of those two genes and other seven implicit an upregulation in their expression levels. The most significantly upregulated gene is *GPR37L1* followed by *ANXA9*.

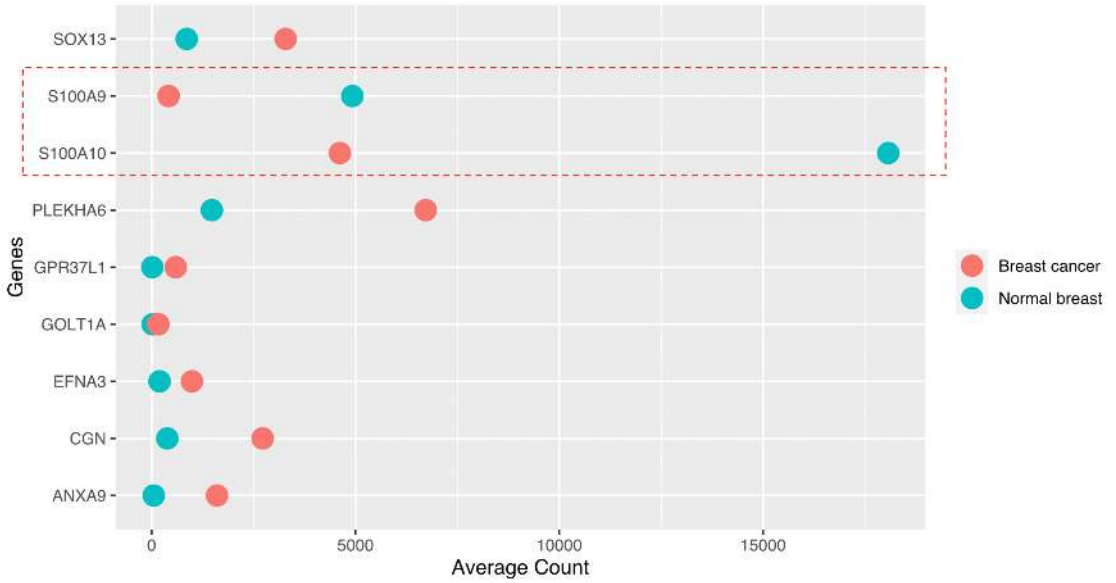


Figure 3.14: Average gene expression of the genes in breast cancer versus normal-like cell lines. The comparison of breast cancer and normal cell lines for nine genes indicating upregulation except for S100A9 and S100A10 (red rectangle) with the change of their count differences in breast cancer and normal-like cell lines.

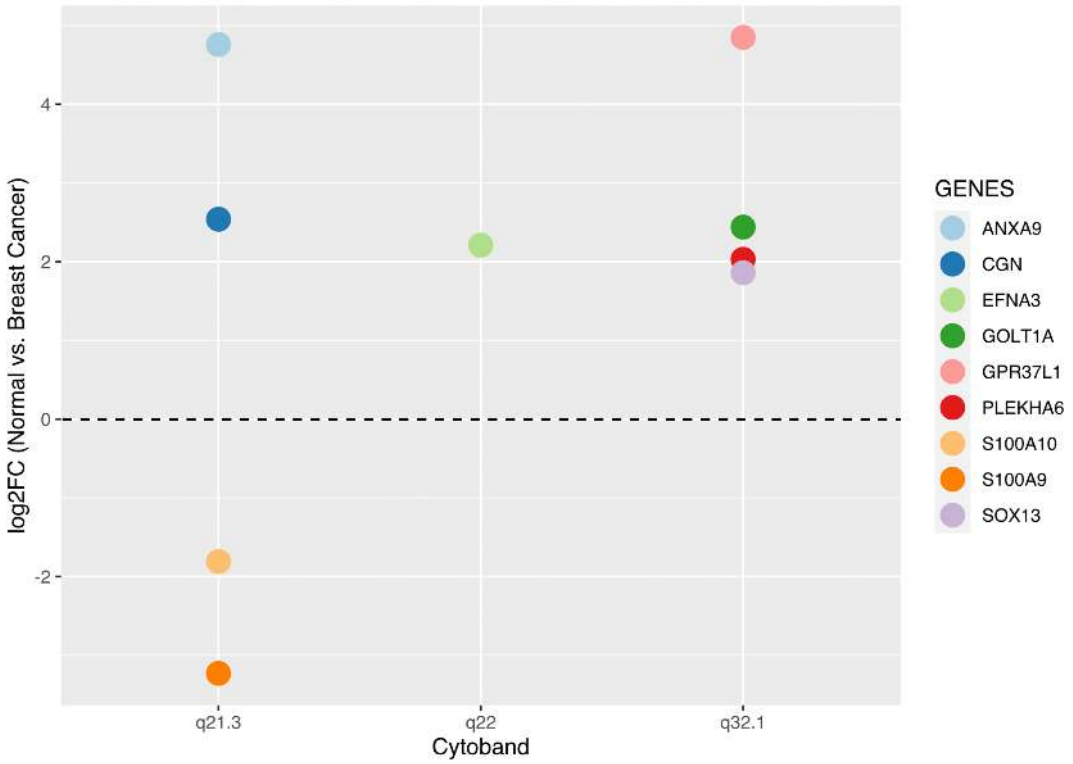


Figure 3.15: Scatter plot showing log2 fold change of gene expression of each common gene in y axis and their cytoband locations. Upregulation can be seen in the seven of the genes, namely, *ANXA9, CGN, EFNA3, GOLT1A, PLEKHA6, SOX13* and *GPR37L1*. Among them, the **upregulated** genes, *GPR37L* and *ANXA9* genes have the highest and *SOX13* has the lowest log2 fold change values. Moreover, *S100A10* has the lowest value indicating downregulation, followed by *S100A9*.

3.7.1 Expression of the common in genes in breast tissues

For investigating the expression profile of the genes in the normal, cancerous, and metastatic tissues, TNMplot.com [297] is used, which is an online analysis portal that enables the comparison of gene expression changes across all genes and multiple platforms by mining the GEO, GTEx, TCGA and TARGET databases (Figure 3.16).

All the determined genes, consistent with the results of this study, serve higher expression in tumor and metastatic tissues (with a lower number of samples) than normal tissues with the exception of S100A9 and S100A10.

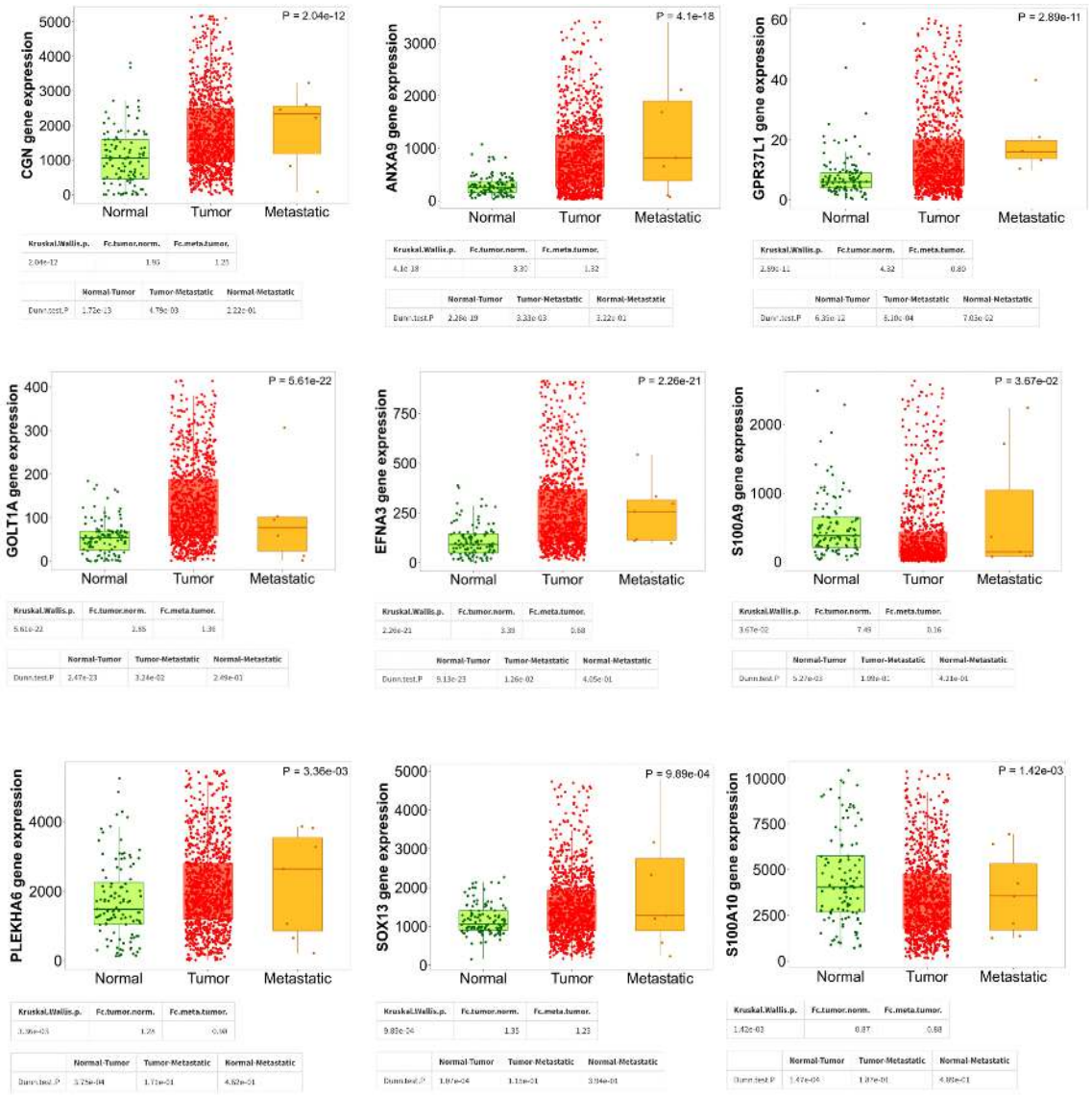


Figure 3.16: Boxplots showing the expression of the genes in normal, tumor and metastatic tissues. All of the genes have the highest number of samples for the tumor tissue. Fold change values indicate that *NXA9, CGN, EFNA3, GOLT1A, PLEKHA6, SOX13* and *GPR37L1* genes have upregulated in the tumor and metastatic tissues. Also, S100A9 has an association with metastatic tissues.

According to this analysis done in TNMplot, there is a significant upregulation in *ANXA9, CGN, EFNA3, GOLT1A, PLEKHA6, SOX13* and *GPR37L1* genes in tumor tissues with the evidence of metastasis. Finally, these results indicate that ***ANXA9***, ***CGN***, ***EFNA3***, ***GOLT1A***, ***PLEKHA6***, ***SOX13*** and ***GPR37L1*** genes are the promising candidates associated with luminal breast cancer.

Chapter 4:

DISCUSSION

Somatic and structural alterations trigger the tumorigenesis and also affect the treatment strategies and related resistance mechanisms. Thus, it is vital to understand the interaction between these alterations and their link with the gene regulation, and ultimately with the phenotype. Within the scope of this thesis, one of the most commonly observed copy number alterations, ***amplification of chromosome 1q***, was investigated. Cytobands exhibiting the higher copy number changes were determined. In order to link the copy number amplification profile to the transcriptional profile, binding of 3 important and luminal breast cancer specific transcriptional factors namely ERa, FOXA1, and GATA3, were examined through publicly available datasets. Finally, benchmarking of CRISPR-library design algorithm CRISPR-DO was performed to generate libraries to target consensus peaks and the defined cytobands.

Optimization of ChIP-seq analysis was performed as a first step and reproducibility of the optimized protocol was confirmed. As a next step, GEO database was visited and the optimized ChIP-seq protocol was applied to detect ERα peaks in MCF7, T47D and ZR75.1, and FOXA1 and GATA3 peaks in MCF7 and T47D. Consensus peaks were identified and as a result of the motif analysis direct peaks were also listed. Those peaks were plotted according to their chromosomal locations and a pattern could be clearly seen; ***some of the cytobands hold a greater number of stronger (significant) peak bindings.*** In this part, the scatter plotting resulted with extremely inflated p-values. This should be further investigated with p-value histogram and qq-plot to show the deviation of residuals. This future analysis is going to give an idea about the reliability of the gathering of points around zero. Besides the adjusted p-value representation, when the number of peaks was checked in those cytobands, they outnumbered the average. Thus, 1q21.3, 1q22 and 1q32.1 cytobands still contain significant and abundant peaks, and according to thorough analysis of CCLE data (data not shown) those cytobands were also detected as having a higher copy number profile.

A list containing the location of the ERα and GATA3 peaks specifically found at the 1q21.3, 1q22 and 1q32.1 was generated and the nearest gene that is located at each peak was identified. Normalized count data of MCF7, T47D and ZR75.1 was used to determine the expression of the identified genes and shared genes were pinpointed.

Total number of 9 genes were shared between all conditions and 7 out of 9 also revealed a higher expression level in breast cancer cell lines compared to normal-like breast cell lines. Moreover, analysis of the expression of the genes in normal, tumor and metastatic tissues indicates a higher expression profile for those 7 genes. *ANXA9* gene is shown to be associated with bone metastasis in breast cancer [298], also, *EFNA3* gene is suggested to contribute to metastatic spread of breast cancer via HIF-mediated induction of EFNA3 lncRNAs and subsequent Ephrin-A3 protein accumulation [299]. *GPR37L* gene is associated with glioblastoma [300] and could be related with breast cancer also. *GOLT1A,* a fusion gene, is shown to be highly correlated with metastatic cancer cell lines [301]. *CGN* gene is shown to be downregulated during epithelial-mesenchymal transition in breast cancer cells [302]. *PLEKHA6* and *SOX13* genes and their association with breast cancer is not shown. Their characteristics should also be further investigated, such as whether they have a circulation in the blood, or have an expression in the breast tumor, or what is the methylation status; all these questions will shed light to distinguish their prognostic and predictive properties. In this study one of the limitations is the tumor and metastatic sample size imbalance. The analyzes were performed in TNMplot.com [297]. These analyzes could be performed by selecting the right analysis model and the data regarding this analysis can be retrieved manually from related databases and creating the ideal sample size for tumor and metastasis analyzes.

To understand the functional TF binding sites, CRISPR-DO algorithm was used to design custom made CRISPR-Cas9 libraries to target ERα, FOXA1, and GATA3 binding sites that are shared among the luminal breast cancer cells. In addition to that, custom made CRISPR-Cas9 libraries targeting the determined cytobands were also generated. Further investigation will be performed to identify binding sites and intergenic sites that have a functional role in breast cancer. Mainly, CRISPR-based functional genetic screening will be conducted.

The determined genes are bioinformatically revealed and they should be further investigated both bioinformatically and experimentally. The network analysis of the genes would reveal the common pathways and co-expression of them that could be related with each other or different gene sets. Also, integration of different sequencing analyzes would give better and in depth understanding. For example, histone ChIP-seq would be added to acquire the information of histone modifications of DNA related to those genes. In the integration, complex tools such as Multi Omics Factor Analysis

(MOFA) [303] can be used in order to interpret the driving sources of variation, thus facilitating the identification of cellular states or subgroups of more multi-omic datasets in an unsupervised fashion. Also, the investigation of the somatic mutations of those genes should be investigated. One of the future directions of this study is also the RNA-seq data analysis part, PCA should be done for RNA-seq and ChIP-seq data.

Altogether, ChIP-seq and RNA-seq data association focused on specific cytobands revealed upregulated genes that might be promising candidates for further studies.

BIBLIOGRAPHY

1. Kanavos P. The rising burden of cancer in the developing world. Ann Oncol. 2006;17 Suppl 8: viii15–viii23.

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71: 209–249.

3. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406: 747–752.

4. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486: 346–352.

5. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490: 61–70.

6. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer. 2012;12: 323–334.

7. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, et al. Breast cancer. Nat Rev Dis Primers. 2019;5: 66.

8. Struski S, Doco-Fenzy M, Cornillet-Lefebvre P. Compilation of published comparative genomic hybridization studies. Cancer Genet Cytogenet. 2002;135: 63–90.

9. Teixeira MR, Pandis N, Heim S. Cytogenetic clues to breast carcinogenesis. Genes Chromosomes Cancer. 2002;33: 1–16.

10. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov. 2022;12: 31–46.

11. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016;534: 47–54.

12. Pereira B, Chin S-F, Rueda OM, Vollan H-KM, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun. 2016;7: 11479.

13. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature. 2012;486: 353–360.

14. Hakem R. DNA-damage repair; the good, the bad, and the ugly. EMBO J. 2008;27: 589–605.

15. Stephens P, Edkins S, Davies H, Greenman C, Cox C, Hunter C, et al. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nat Genet. 2005;37: 590–592.

16. Lei JT, Gou X, Seker S, Ellis MJ. alterations and metastasis in estrogen receptor positive breast cancer. J Cancer Metastasis Treat. 2019;5. doi:10.20517/2394-4722.2019.12

17. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev

Genet. 2006;7: 85–97.

18. Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. Complex SNP-related sequence variation in segmental genome duplications. Nat Genet. 2004;36: 861–866.

19. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16: 172–183.

20. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444: 444–454.

21. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10: 451–481.

22. Lupski JR, Stankiewicz P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet. 2005;1: e49.

23. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009;10: 691–703.

24. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453: 56–64.

25. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007;318: 420–426.

26. Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, et al. Mobile elements create structural variation: analysis of a complete human genome. Genome Res. 2009;19: 1516–1526.

27. Aravind L, Koonin EV. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. Genome Res. 2001;11: 1365–1374.

28. MacArthur JAL, Spector TD, Lindsay SJ, Mangino M, Gill R, Small KS, et al. The rate of nonallelic homologous recombination in males is highly variable, correlated between monozygotic twins and independent of age. PLoS Genet. 2014;10: e1004195.

29. Beckmann JS, Estivill X, Antonarakis SE. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet. 2007;8: 639–646.

30. Maxwell PH, Burhans WC, Curcio MJ. Retrotransposition is associated with genome instability during chronological aging. Proc Natl Acad Sci U S A. 2011;108: 20376–20381.

31. Lee JA, Carvalho CMB, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell. 2007;131: 1235–1247.

32. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100: 57–70.

33. Liu Y, Hu X, Han C, Wang L, Zhang X, He X, et al. Targeting tumor suppressor genes for cancer therapy. Bioessays. 2015;37: 1277–1286.

34. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al.

Signatures of mutational processes in human cancer. Nature. 2013;500: 415–421.

35. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339: 1546–1558.

36. Woodward AM, Davis TA, Silva AGS, Kirk JA, Leary JA, kConFab Investigators. Large genomic rearrangements of both BRCA2 and BRCA1 are a feature of the inherited breast/ovarian cancer phenotype in selected families. J Med Genet. 2005;42: e31.

37. Lahti-Domenici J, Rapakko K, Pääkkönen K, Allinen M, Nevanlinna H, Kujala M, et al. Exclusion of large deletions and other rearrangements in BRCA1 and BRCA2 in Finnish breast and ovarian cancer families. Cancer Genet Cytogenet. 2001;129: 120–123.

38. Yoshida K, Miki Y. Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. Cancer Sci. 2004;95: 866–871.

39. Duncan JA, Reeves JR, Cooke TG. BRCA1 and BRCA2 proteins: roles in health and disease. Mol Pathol. 1998;51: 237–247.

40. Brewster AM, Thompson P, Sahin AA, Do K, Edgerton M, Murray JL, et al. Copy number imbalances between screen- and symptom-detected breast cancers and impact on disease-free survival. Cancer Prev Res . 2011;4: 1609–1616.

41. Endesfelder D, Burrell R, Kanu N, McGranahan N, Howell M, Parker PJ, et al. Chromosomal instability selects gene copy-number variants encoding core regulators of proliferation in ER+ breast cancer. Cancer Res. 2014;74: 4853–4863.

42. van Beers EH, Nederlof PM. Array-CGH and breast cancer. Breast Cancer Res. 2006;8: 210.

43. Murakami F, Tsuboi Y, Takahashi Y, Horimoto Y, Mogushi K, Ito T, et al. Short somatic alterations at the site of copy number variation in breast cancer. Cancer Sci. 2021;112: 444–453.

44. Gray JW, Collins C. Genome changes and gene expression in human solid tumors. Carcinogenesis. 2000;21: 443–452.

45. Bautista S, Theillet C. CCND1 and FGFR1 coamplification results in the colocalization of 11q13 and 8p12 sequences in breast tumor nuclei. Genes Chromosomes Cancer. 1998;22: 268–277.

46. Volik S, Raphael BJ, Huang G, Stratton MR, Bignel G, Murnane J, et al. Decoding the fine-scale structure of a breast cancer genome and transcriptome. Genome Res. 2006;16: 394–404.

47. Atkin NB. Chromosome 1 aberrations in cancer. Cancer Genet Cytogenet. 1986;21: 279–285.

48. Dawson S-J, Rueda OM, Aparicio S, Caldas C. A new genome-driven integrated classification of breast cancer and its implications. EMBO J. 2013;32: 617–628.

49. Baslan T, Kendall J, Volyanskyy K, McNamara K, Cox H, D'Italia S, et al. Novel insights into breast cancer copy number genetic heterogeneity revealed by single-cell genome sequencing. Elife. 2020;9. doi:10.7554/eLife.51480

50. Orsetti B, Nugoli M, Cervera N, Lasorsa L, Chuchana P, Rougé C, et al. Genetic profiling of chromosome 1 in breast cancer: mapping of regions of gains and losses and identification of candidate genes on 1q. Br J Cancer. 2006;95: 1439–1447.

51. Loo LWM, Grove DI, Williams EM, Neal CL, Cousens LA, Schubert EL, et al. Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. Cancer Res. 2004;64: 8541–8549.

52. Schmidt TM, Fonseca R, Usmani SZ. Chromosome 1q21 abnormalities in multiple myeloma. Blood Cancer J. 2021;11: 83.

53. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, et al. Copy number variation at 1q21.1 associated with neuroblastoma. Nature. 2009;459: 987–991.

54. Voutsadakis IA. The Landscape of Chromosome Instability in Breast Cancers and Associations with the Tumor Mutation Burden: An Analysis of Data from TCGA. Cancer Invest. 2021;39: 25–38.

55. Courjal F, Theillet C. Comparative genomic hybridization analysis of breast tumors with predetermined profiles of DNA amplification. Cancer Res. 1997;57: 4368–4377.

56. Tirkkonen M, Tanner M, Karhu R, Kallioniemi A, Isola J, Kallioniemi OP. Molecular cytogenetics of primary breast cancer by CGH. Genes Chromosomes Cancer. 1998;21: 177–184.

57. Li Z, Zhang X, Hou C, Zhou Y, Chen J, Cai H, et al. Comprehensive identification and characterization of somatic copy number alterations in triple‐negative breast cancer. Int J Oncol. 2020;56: 522–530.

58. Saunus JM, Quinn MCJ, Patch A-M, Pearson JV, Bailey PJ, Nones K, et al. Integrated genomic and transcriptomic analysis of human brain metastases identifies alterations of potential clinical significance. J Pathol. 2015;237: 363–378.

59. Hawthorn L, Luce J, Stein L, Rothschild J. Integration of transcript expression, copy number and LOH analysis of infiltrating ductal carcinoma of the breast. BMC Cancer. 2010;10: 460.

60. Silva GO, He X, Parker JS, Gatza ML, Carey LA, Hou JP, et al. Cross-species DNA copy number analyses identifies multiple 1q21-q23 subtype-specific driver genes for breast cancer. Breast Cancer Res Treat. 2015;152: 347–356.

61. Rodrigues-Peres RM, de S Carvalho B, Anurag M, Lei JT, Conz L, Gonçalves R, et al. Copy number alterations associated with clinical features in an underrepresented population with breast cancer. Mol Genet Genomic Med. 2019;7: e00750.

62. Okkema PG, Krause M. Transcriptional regulation. WormBook. 2005; 1–40.

63. Struhl K. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell. 1999;98: 1–4.

64. Mitchell PJ, Tjian R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science. 1989;245: 371–378.

65. Ptashne M. How eukaryotic transcriptional activators work. Nature. 1988;335: 683–689.

66. Latchman DS. Transcription factors: an overview. Int J Biochem Cell Biol. 1997;29: 1305–1312.

67. Doane AS, Elemento O. Regulatory elements in molecular networks. Wiley Interdiscip Rev Syst Biol Med. 2017;9. doi:10.1002/wsbm.1374

68. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 2015;527: 384–388.

69. Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat Biotechnol. 2015;33: 555–562.

70. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10: 252–263.

71. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152: 327–339.

72. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, et al. Diversity and complexity in DNA recognition by transcription factors. Science. 2009;324: 1720–1723.

73. Mohibullah N, Donner A, Ippolito JA, Williams T. SELEX and missing phosphate contact analyses reveal flexibility within the AP-2[alpha] protein: DNA binding complex. Nucleic Acids Res. 1999;27: 2760–2769.

74. Kurokawa R, Yu VC, Näär A, Kyakumoto S, Han Z, Silverman S, et al. Differential orientations of the DNA-binding domain and carboxy-terminal dimerization interface regulate binding site selection by nuclear receptor heterodimers. Genes Dev. 1993;7: 1423–1435.

75. Emery P, Strubin M, Hofmann K, Bucher P, Mach B, Reith W. A consensus motif in the RFX DNA binding domain and binding domain mutants with altered specificity. Mol Cell Biol. 1996;16: 4486–4494.

76. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489: 75–82.

77. Robinson DR, Wu Y-M, Vats P, Su F, Lonigro RJ, Cao X, et al. Activating ESR1 mutations in hormone-resistant metastatic breast cancer. Nat Genet. 2013;45: 1446–1451.

78. Toy W, Weir H, Razavi P, Lawson M, Goeppert AU, Mazzola AM, et al. Activating Mutations Differentially Affect the Efficacy of ER Antagonists. Cancer Discov. 2017;7: 277–287.

79. Toy W, Shen Y, Won H, Green B, Sakr RA, Will M, et al. ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. Nat Genet. 2013;45: 1439–1445.

80. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013;152: 1237–1251.

81. Bishop JM. The molecular genetics of cancer. Science. 1987;235: 305–311.

82. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science. 1987;235: 177–182.

83. Allred DC, Clark GM, Molina R, Tandon AK, Schnitt SJ, Gilchrist KW, et al. Overexpression of HER-2/neu and its relationship with other prognostic factors change during the progression of in situ to invasive breast cancer. Hum Pathol. 1992;23: 974–979.

84. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. Genes Dev. 2011;25: 2227–2241.

85. Swinstead EE, Paakinaho V, Presman DM, Hager GL. Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective: Multiple transcription factors can effect chromatin pioneer functions through dynamic interactions with ATP-dependent chromatin remodeling factors. Bioessays. 2016;38: 1150–1157.

86. Morisaki T, Müller WG, Golob N, Mazza D, McNally JG. Single-molecule analysis of transcription factor binding at transcription sites in live cells. Nat Commun. 2014;5: 4456.

87. Sugo N, Morimatsu M, Arai Y, Kousoku Y, Ohkuni A, Nomura T, et al. Single-Molecule Imaging Reveals Dynamics of CREB Transcription Factor Bound to Its Target Sequence. Sci Rep. 2015;5: 10662.

88. Chen J, Zhang Z, Li L, Chen B-C, Revyakin A, Hajj B, et al. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. Cell. 2014;156: 1274–1285.

89. Gualdi R, Bossard P, Zheng M, Hamada Y, Coleman JR, Zaret KS. Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. Genes Dev. 1996;10: 1670–1682.

90. Swinstead EE, Miranda TB, Paakinaho V, Baek S, Goldstein I, Hawkins M, et al. Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. Cell. 2016;165: 593–605.

91. Goldstein I, Baek S, Presman DM, Paakinaho V, Swinstead EE, Hager GL. Transcription factor assisted loading and enhancer dynamics dictate the hepatic fasting response. Genome Res. 2017;27: 427–439.

92. Voss TC, Schiltz RL, Sung M-H, Yen PM, Stamatoyannopoulos JA, Biddie SC, et al. Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. Cell. 2011;146: 544–554.

93. Jin Y, Liang Z, Lou H. The Emerging Roles of Fox Family Transcription Factors in Chromosome Replication, Organization, and Genome Stability. Cells. 2020;9. doi:10.3390/cells9010258

94. Weigel D, Jürgens G, Küttner F, Seifert E, Jäckle H. The homeotic gene fork head encodes a nuclear protein and is expressed in the terminal regions of the Drosophila embryo. Cell. 1989;57: 645–658.

95. Grossniklaus U, Pearson RK, Gehring WJ. The Drosophila sloppy paired locus encodes two proteins involved in segmentation that show homology to mammalian transcription factors. Genes Dev. 1992;6: 1030–1051.

96. Kaestner KH. The FoxA factors in organogenesis and differentiation. Curr Opin Genet

Dev. 2010;20: 527–532.

97. Mears AJ, Jordan T, Mirzayans F, Dubois S, Kume T, Parlee M, et al. Mutations of the forkhead/winged-helix gene, FKHL7, in patients with Axenfeld-Rieger anomaly. Am J Hum Genet. 1998;63: 1316–1328.

98. Bernardo GM, Keri RA. FOXA1: a transcription factor with parallel functions in development and cancer. Biosci Rep. 2012;32: 113–130.

99. Carlsson P, Mahlapuu M. Forkhead transcription factors: key players in development and metabolism. Dev Biol. 2002;250: 1–23.

100. Kaestner KH. The hepatocyte nuclear factor 3 (HNF3 or FOXA) family in metabolism. Trends Endocrinol Metab. 2000;11: 281–285.

101. Marsden I, Jin C, Liao X. Structural changes in the region directly adjacent to the DNA-binding helix highlight a possible mechanism to explain the observed changes in the sequence-specific binding of winged helix proteins. J Mol Biol. 1998;278: 293–299.

102. Wolf I, Bose S, Williamson EA, Miller CW, Karlan BY, Koeffler HP. FOXA1: Growth inhibitor and a favorable prognostic factor in human breast cancer. Int J Cancer. 2007;120: 1013–1022.

103. Badve S, Turbin D, Thorat MA, Morimiya A, Nielsen TO, Perou CM, et al. FOXA1 expression in breast cancer--correlation with luminal subtype A and survival. Clin Cancer Res. 2007;13: 4415–4421.

104. Thorat MA, Marchio C, Morimiya A, Savage K, Nakshatri H, Reis-Filho JS, et al. Forkhead box A1 expression in breast cancer is associated with luminal subtype and good prognosis. J Clin Pathol. 2008;61: 327–332.

105. Doane AS, Danso M, Lal P, Donaton M, Zhang L, Hudis C, et al. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. Oncogene. 2006;25: 3994–4008.

106. Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS. FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat Genet. 2011;43: 27–33.

107. Bernardo GM, Lozada KL, Miedler JD, Harburg G, Hewitt SC, Mosley JD, et al. FOXA1 is an essential determinant of ERalpha expression and mammary ductal morphogenesis. Development. 2010;137: 2045–2054.

108. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell. 2005;122: 33–43.

109. Clark KL, Halay ED, Lai E, Burley SK. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. Nature. 1993;364: 412–420.

110. Eeckhoute J, Carroll JS, Geistlinger TR, Torres-Arzayus MI, Brown M. A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer. Genes Dev. 2006;20: 2513–2526.

111. Buckley MF, Sweeney KJ, Hamilton JA, Sini RL, Manning DL, Nicholson RI, et al.

Expression and amplification of cyclin genes in human breast cancer. Oncogene. 1993;8: 2127–2133.

112. Ko LJ, Engel JD. DNA-binding specificities of the GATA transcription factor family. Mol Cell Biol. 1993;13: 4011–4022.

113. Orkin SH. GATA-binding transcription factors in hematopoietic cells. Blood. 1992;80: 575–581.

114. Asselin-Labat M-L, Sutherland KD, Barker H, Thomas R, Shackleton M, Forrest NC, et al. Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. Nat Cell Biol. 2007;9: 201–209.

115. Kouros-Mehr H, Slorach EM, Sternlicht MD, Werb Z. GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. Cell. 2006;127: 1041–1055.

116. Hattori N, Kawamoto H, Fujimoto S, Kuno K, Katsura Y. Involvement of transcription factors TCF-1 and GATA-3 in the initiation of the earliest step of T cell development in the thymus. J Exp Med. 1996;184: 1137–1147.

117. Eeckhoute J, Keeton EK, Lupien M, Krum SA, Carroll JS, Brown M. Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer. Cancer Res. 2007;67: 6477–6483.

118. Lee HJ, Takemoto N, Kurata H, Kamogawa Y, Miyatake S, O'Garra A, et al. GATA-3 induces T helper cell type 2 (Th2) cytokine expression and chromatin remodeling in committed Th1 cells. J Exp Med. 2000;192: 105–115.

119. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012;486: 400–404.

120. Takaku M, Grimm SA, Wade PA. GATA3 in Breast Cancer: Tumor Suppressor or Oncogene? Gene Expr. 2015;16: 163–168.

121. Yan W, Cao QJ, Arenas RB, Bentley B, Shao R. GATA3 inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition. J Biol Chem. 2010;285: 14042–14051.

122. Kong SL, Li G, Loh SL, Sung W-K, Liu ET. Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state. Mol Syst Biol. 2011;7: 526.

123. Lin C-Y, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, et al. Whole-genome cartography of estrogen receptor alpha binding sites. PLoS Genet. 2007;3: e87.

124. Theodorou V, Stark R, Menon S, Carroll JS. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. Genome Res. 2013;23: 12–22.

125. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics. 2014;13: 397–406.

126. Matthews J, Wihlén B, Tujague M, Wan J, Ström A, Gustafsson J-A. Estrogen receptor (ER) beta modulates ERalpha-mediated transcriptional activation by altering the

recruitment of c-Fos and c-Jun to estrogen-responsive promoters. Mol Endocrinol. 2006;20: 534–543.

127. Bondesson M, Hao R, Lin C-Y, Williams C, Gustafsson J-Å. Estrogen receptor signaling during vertebrate development. Biochim Biophys Acta. 2015;1849: 142–151.

128. Abba MC, Hu Y, Sun H, Drake JA, Gaddis S, Baggerly K, et al. Gene expression signature of estrogen receptor alpha status in breast cancer. BMC Genomics. 2005;6: 37.

129. Jia M, Dahlman-Wright K, Gustafsson J-Å. Estrogen receptor alpha and beta in health and disease. Best Pract Res Clin Endocrinol Metab. 2015;29: 557–568.

130. Huang B, Omoto Y, Iwase H, Yamashita H, Toyama T, Coombes RC, et al. Differential expression of estrogen receptor α, β1, and β2 in lobular and ductal breast cancer. Proc Natl Acad Sci U S A. 2014;111: 1933–1938.

131. Dickson RB, Stancel GM. Estrogen receptor-mediated processes in normal and cancer cells. J Natl Cancer Inst Monogr. 2000; 135–145.

132. Gruber CJ, Tschugguel W, Schneeberger C, Huber JC. Production and actions of estrogens. N Engl J Med. 2002;346: 340–352.

133. Parker MG. Transcriptional activation by oestrogen receptors. Biochem Soc Symp. 1998;63: 45–50.

134. Torchia J, Glass C, Rosenfeld MG. Co-activators and co-repressors in the integration of transcriptional responses. Curr Opin Cell Biol. 1998;10: 373–383.

135. Shibata H, Spencer TE, Oñate SA, Jenster G, Tsai SY, Tsai MJ, et al. Role of co-activators and co-repressors in the mechanism of steroid/thyroid receptor action. Recent Prog Horm Res. 1997;52: 141–64; discussion 164–5.

136. Krivega I, Dean A. Enhancer and promoter interactions-long distance calls. Curr Opin Genet Dev. 2012;22: 79–85.

137. Droog M, Mensink M, Zwart W. The Estrogen Receptor α-Cistrome Beyond Breast Cancer. Mol Endocrinol. 2016;30: 1046–1058.

138. Siersbæk R, Kumar S, Carroll JS. Signaling pathways and steroid receptors modulating estrogen receptor α function in breast cancer. Genes Dev. 2018;32: 1141–1154.

139. Takaku M, Grimm SA, Shimbo T, Perera L, Menafra R, Stunnenberg HG, et al. GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler. Genome Biol. 2016;17: 36.

140. Iwafuchi-Doi M, Zaret KS. Pioneer transcription factors in cell reprogramming. Genes Dev. 2014;28: 2679–2692.

141. Takaku M, Grimm SA, De Kumar B, Bennett BD, Wade PA. Cancer-specific mutation of GATA3 disrupts the transcriptional regulatory network governed by Estrogen Receptor alpha, FOXA1 and GATA3. Nucleic Acids Res. 2020;48: 4756–4768.

142. King HW, Klose RJ. The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. Elife. 2017;6. doi:10.7554/eLife.22631

143. Iwafuchi-Doi M. The mechanistic basis for chromatin regulation by pioneer transcription factors. Wiley Interdiscip Rev Syst Biol Med. 2019;11: e1427.

144. Carroll JS. Mechanisms of oestrogen receptor (ER) gene regulation in breast cancer. Eur J Endocrinol. 2016;175: R41–9.

145. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94: 441–448.

146. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–5467.

147. Maxam AM, Gilbert W. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. 1980;65: 499–560.

148. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230: 1350–1354.

149. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988;239: 487–491.

150. Levy SE, Boone BE. Next-Generation Sequencing Strategies. Cold Spring Harb Perspect Med. 2019;9. doi:10.1101/cshperspect.a025791

151. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921.

152. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291: 1304–1351.

153. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431: 931–945.

154. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed. 2013;98: 236–238.

155. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107: 1–8.

156. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17: 333–351.

157. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science. 1998;281: 363, 365.

158. Muzzey D, Evans EA, Lieber C. Understanding the Basics of NGS: From Mechanism to Variant Calling. Curr Genet Med Rep. 2015;3: 158–165.

159. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009;55: 641–658.

160. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002;417: 949–954.

161. Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S, et al. High frequency of mutations of the PIK3CA gene in human cancers. Science. 2004;304: 554.

162. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 2004;350: 2129–2139.

163. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304: 1497–1500.

164. Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, et al. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci U S A. 2004;101: 13306–13311.

165. Wheeler DA, Wang L. From human genome to cancer genome: the first decade. Genome Res. 2013;23: 1054–1062.

166. Nagahashi M, Shimada Y, Ichikawa H, Kameyama H, Takabe K, Okuda S, et al. Next generation sequencing-based gene panel tests for the management of solid tumors. Cancer Sci. 2019;110: 6–15.

167. Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008;321: 1807–1812.

168. Solomon MJ, Larsen PL, Varshavsky A. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell. 1988;53: 937–947.

169. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009;457: 854–858.

170. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007;448: 553–560.

171. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007;4: 651–657.

172. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129: 823–837.

173. Mundade R, Ozer HG, Wei H, Prabhu L, Lu T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle. 2014;13: 2847–2852.

174. Ricci WA, Levin L, Zhang X. Genome-Wide Profiling of Histone Modifications with ChIP-Seq. Methods Mol Biol. 2020;2072: 101–117.

175. Singh AA, Schuurman K, Nevedomskaya E, Stelloo S, Linder S, Droog M, et al. Optimized ChIP-seq method facilitates transcription factor profiling in human tumors. Life Sci Alliance. 2019;2: e201800115.

176. Ryan GE, Farley EK. Functional genomic approaches to elucidate the role of enhancers during development. Wiley Interdiscip Rev Syst Biol Med. 2020;12: e1467.

177. Keller CA, Wixom AQ, Heuston EF, Giardine B, Hsiung CC-S, Long MR, et al. Effects of sheared chromatin length on ChIP-seq quality and sensitivity. G3 . 2021. doi:10.1093/g3journal/jkab101

178. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10: 669–680.

179. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316: 1497–1502.

180. Chng KR, Chang CW, Tan SK, Yang C, Hong SZ, Sng NYW, et al. A transcriptional repressor co-regulatory network governing androgen response in prostate cancers. EMBO J. 2012;31: 2810–2823.

181. Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nat Genet. 2019;51: 1060–1066.

182. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320: 1344–1349.

183. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015;2015: 951–969.

184. Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther. 2012;22: 271–274.

185. Hangauer MJ, Vaughn IW, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 2013;9: e1003569.

186. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486: 395–399.

187. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010;11: 476–486.

188. Chen M, Mao A, Xu M, Weng Q, Mao J, Ji J. CRISPR-Cas9 for cancer therapy: Opportunities and challenges. Cancer Lett. 2019;447: 48–55.

189. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014;157: 1262–1278.

190. Porteus MH, Baltimore D. Chimeric nucleases stimulate gene targeting in human cells. Science. 2003;300: 763.

191. Miller JC, Holmes MC, Wang J, Guschin DY, Lee Y-L, Rupniewski I, et al. An improved zinc-finger nuclease architecture for highly specific genome editing. Nat Biotechnol. 2007;25: 778–785.

192. Sander JD, Dahlborg EJ, Goodwin MJ, Cade L, Zhang F, Cifuentes D, et al.

Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). Nat Methods. 2011;8: 67–69.

193. Wood AJ, Lo T-W, Zeitler B, Pickle CS, Ralston EJ, Lee AH, et al. Targeted genome editing across species using ZFNs and TALENs. Science. 2011;333: 307.

194. Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, et al. Targeting DNA double-strand breaks with TAL effector nucleases. Genetics. 2010;186: 757–761.

195. Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. Nat Biotechnol. 2011;29: 149–153.

196. Moscou MJ, Bogdanove AJ. A simple cipher governs DNA recognition by TAL effectors. Science. 2009;326: 1501.

197. Sanjana NE, Cong L, Zhou Y, Cunniff MM, Feng G, Zhang F. A transcription activator-like effector toolbox for genome engineering. Nat Protoc. 2012;7: 171–192.

198. Deveau H, Garneau JE, Moineau S. CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol. 2010;64: 475–493.

199. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science. 2010;327: 167–170.

200. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011;9: 467–477.

201. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339: 819–823.

202. Cho SW, Kim S, Kim JM, Kim J-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol. 2013;31: 230–232.

203. Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. RNA-programmed genome editing in human cells. Elife. 2013;2: e00471.

204. Bhaya D, Davison M, Barrangou R. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet. 2011;45: 273–297.

205. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013;339: 823–826.

206. Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014;346: 1258096.

207. Charpentier E, Marraffini LA. Harnessing CRISPR-Cas9 immunity for genetic engineering. Curr Opin Microbiol. 2014;19: 114–119.

208. Barrangou R, Doudna JA. Applications of CRISPR technologies in research and beyond. Nat Biotechnol. 2016;34: 933–941.

209. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 2015;13:

722–736.

210. Makarova KS, Koonin EV. Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol. 2015;1311: 47–75.

211. Liu Z, Dong H, Cui Y, Cong L, Zhang D. Application of different types of CRISPR/Cas-based systems in bacteria. Microb Cell Fact. 2020;19: 172.

212. Ceasar SA, Rajan V, Prykhozhij SV, Berman JN, Ignacimuthu S. Insert, remove or replace: A highly advanced genome editing system using CRISPR/Cas9. Biochim Biophys Acta. 2016;1863: 2333–2344.

213. Liu M, Rehman S, Tang X, Gu K, Fan Q, Chen D, et al. Methodologies for Improving HDR Efficiency. Front Genet. 2018;9: 691.

214. Ford K, McDonald D, Mali P. Functional Genomics via CRISPR-Cas. J Mol Biol. 2019;431: 48–65.

215. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014;343: 80–84.

216. Dominguez AA, Lim WA, Qi LS. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. Nat Rev Mol Cell Biol. 2016;17: 5–15.

217. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell. 2013;154: 442–451.

218. Polstein LR, Perez-Pinera P, Kocak DD, Vockley CM, Bledsoe P, Song L, et al. Genome-wide specificity of DNA binding, gene regulation, and chromatin remodeling by TALE- and CRISPR/Cas9-based transcriptional activators. Genome Res. 2015;25: 1158–1169.

219. Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, et al. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol. 2015;33: 510–517.

220. Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, et al. Editing DNA Methylation in the Mammalian Genome. Cell. 2016;167: 233–247.e17.

221. Vojta A, Dobrinić P, Tadić V, Bočkor L, Korać P, Julg B, et al. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. Nucleic Acids Res. 2016;44: 5615–5628.

222. Laufer BI, Singh SM. Strategies for precision modulation of gene expression by epigenome editing: an overview. Epigenetics Chromatin. 2015;8: 34.

223. Kwon DY, Zhao Y-T, Lamonica JM, Zhou Z. Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. Nat Commun. 2017;8: 15315.

224. Kearns NA, Pham H, Tabak B, Genga RM, Silverstein NJ, Garber M, et al. Functional annotation of native enhancers with a Cas9-histone demethylase fusion. Nat Methods. 2015;12: 401–403.

225. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.

Cell. 2015;161: 1012–1025.

226. Hess GT, Tycko J, Yao D, Bassik MC. Methods and Applications of CRISPR-Mediated Base Editing in Eukaryotic Genomes. Mol Cell. 2017;68: 26–43.

227. Hess GT, Frésard L, Han K, Lee CH, Li A, Cimprich KA, et al. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. Nat Methods. 2016;13: 1036–1042.

228. Ma Y, Zhang J, Yin W, Zhang Z, Song Y, Chang X. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. Nat Methods. 2016;13: 1029–1035.

229. Kuscu C, Parlak M, Tufan T, Yang J, Szlachta K, Wei X, et al. CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. Nat Methods. 2017;14: 710–712.

230. Najm FJ, Strand C, Donovan KF, Hegde M, Sanson KR, Vaimberg EW, et al. Orthologous CRISPR-Cas9 enzymes for combinatorial genetic screens. Nat Biotechnol. 2018;36: 179–189.

231. Wong ASL, Choi GCG, Cui CH, Pregernig G, Milani P, Adam M, et al. Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM. Proc Natl Acad Sci U S A. 2016;113: 2544–2549.

232. Shen JP, Zhao D, Sasik R, Luebeck J, Birmingham A, Bojorquez-Gomez A, et al. Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. Nat Methods. 2017;14: 573–576.

233. Han K, Jeng EE, Hess GT, Morgens DW, Li A, Bassik MC. Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. Nat Biotechnol. 2017;35: 463–474.

234. Chow RD, Chen S. Cancer CRISPR Screens In Vivo. Trends Cancer Res. 2018;4: 349–358.

235. Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X, et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. Cell. 2015;160: 1246–1260.

236. Braun CJ, Bruno PM, Horlbeck MA, Gilbert LA, Weissman JS, Hemann MT. Versatile in vivo regulation of tumor phenotypes by dCas9-mediated transcriptional perturbation. Proc Natl Acad Sci U S A. 2016;113: E3892–900.

237. Katigbak A, Cencic R, Robert F, Sénécha P, Scuoppo C, Pelletier J. A CRISPR/Cas9 Functional Screen Identifies Rare Tumor Suppressors. Sci Rep. 2016;6: 38968.

238. Kodama M, Kodama T, Newberg JY, Katayama H, Kobayashi M, Hanash SM, et al. In vivo loss-of-function screens identify KPNB1 as a new druggable oncogene in epithelial ovarian cancer. Proc Natl Acad Sci U S A. 2017;114: E7301–E7310.

239. Song C-Q, Li Y, Mou H, Moore J, Park A, Pomyen Y, et al. Genome-Wide CRISPR Screen Identifies Regulators of Mitogen-Activated Protein Kinase as Suppressors of Liver Tumors in Mice. Gastroenterology. 2017;152: 1161–1173.e1.

240. Yau EH, Kummetha IR, Lichinchi G, Tang R, Zhang Y, Rana TM. Genome-Wide

CRISPR Screen for Essential Cell Growth Mediators in Mutant KRAS Colorectal Cancers. Cancer Res. 2017;77: 6330–6339.

241. Patel SJ, Sanjana NE, Kishton RJ, Eidizadeh A, Vodnala SK, Cam M, et al. Identification of essential genes for cancer immunotherapy. Nature. 2017;548: 537–542.

242. Manguso RT, Pope HW, Zimmer MD, Brown FD, Yates KB, Miller BC, et al. In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. Nature. 2017;547: 413–418.

243. Hou P, Wu C, Wang Y, Qi R, Bhavanasi D, Zuo Z, et al. A Genome-Wide CRISPR Screen Identifies Genes Critical for Resistance to FLT3 Inhibitor AC220. Cancer Res. 2017;77: 4402–4413.

244. Sanjana NE, Wright J, Zheng K, Shalem O, Fontanillas P, Joung J, et al. High-resolution interrogation of functional elements in the noncoding genome. Science. 2016;353: 1545–1549.

245. Ruiz S, Mayor-Ruiz C, Lafarga V, Murga M, Vega-Sendino M, Ortega S, et al. A Genome-wide CRISPR Screen Identifies CDC25A as a Determinant of Sensitivity to ATR Inhibitors. Mol Cell. 2016;62: 307–313.

246. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015;350: 1096–1101.

247. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, et al. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 2015;163: 1515–1526.

248. Wang T, Yu H, Hughes NW, Liu B, Kendirli A, Klein K, et al. Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. Cell. 2017;168: 890–903.e15.

249. Valletta S, Dolatshad H, Bartenstein M, Yip BH, Bello E, Gordon S, et al. ASXL1 mutation correction by CRISPR/Cas9 restores gene function in leukemia cells and increases survival in mouse xenografts. Oncotarget. 2015;6: 44061–44071.

250. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337: 816–821.

251. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A. 2012;109: E2579–86.

252. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014;507: 62–67.

253. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011;471: 602–607.

254. Haeussler M, Concordet J-P. Genome Editing with CRISPR-Cas9: Can It Get Any Better? J Genet Genomics. 2016;43: 239–250.

255. Zhang X-H, Tee LY, Wang X-G, Huang Q-S, Yang S-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Mol Ther Nucleic Acids. 2015;4: e264.

256. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013;31: 827–832.

257. Brinkman EK, Chen T, Amendola M, van Steensel B. Easy quantitative assessment of genome editing by sequence trace decomposition. Nucleic Acids Res. 2014;42: e168.

258. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. Nat Protoc. 2013;8: 2281–2308.

259. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol. 2013;31: 233–239.

260. Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A. 2011;108: 10098–10103.

261. Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res. 2011;39: 9275–9282.

262. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014;32: 1262–1267.

263. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;34: 184–191.

264. Xu H, Xiao T, Chen C-H, Li W, Meyer CA, Wu Q, et al. Sequence determinants of improved CRISPR sgRNA design. Genome Res. 2015;25: 1147–1157.

265. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85: 2444–2448.

266. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38: 1767–1771.

267. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

268. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010;26: 2204–2207.

269. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12: 996–1006.

270. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30: 207–210.

271. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41:

D991–5.

272. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res. 2017;45: D658–D662.

273. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res. 2019;47: D729–D735.

274. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31: 2032–2034.

275. Holmes I, Bruno WJ. Finding regulatory elements using joint likelihoods for sequence and expression profile data. Proc Int Conf Intell Syst Mol Biol. 2000;8: 202–210.

276. Stormo GD, Hartzell GW 3rd. Identifying protein-binding sites from unaligned DNA fragments. Proc Natl Acad Sci U S A. 1989;86: 1183–1187.

277. Frith MC, Fu Y, Yu L, Chen J-F, Hansen U, Weng Z. Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res. 2004;32: 1372–1381.

278. Boeva V. Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. Front Genet. 2016;7: 24.

279. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38: 576–589.

280. Jalili V, Matteucci M, Masseroli M, Morelli MJ. Using combined evidence from replicates to evaluate ChIP-seq peaks. Bioinformatics. 2015;31: 2761–2769.

281. Gheorghe M, Sandve GK, Khan A, Chèneby J, Ballester B, Mathelier A. A map of direct TF-DNA interactions in the human genome. Nucleic Acids Res. 2019;47: e21.

282. Ma J, Köster J, Qin Q, Hu S, Li W, Chen C, et al. CRISPR-DO for genome-wide CRISPR design and optimization. Bioinformatics. 2016;32: 3336–3338.

283. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048.

284. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.

285. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14: 417–419.

286. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. 2015;4: 1521.

287. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550.

288. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483: 603–607.

289. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011;39: D19–21.

290. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10: R25.

291. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137.

292. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29: 24–26.

293. Glont S-E, Papachristou EK, Sawle A, Holmes KA, Carroll JS, Siersbaek R. Identification of ChIP-seq and RIME grade antibodies for Estrogen Receptor alpha. PLoS One. 2019;14: e0215340.

294. Shiau AK, Barstad D, Loria PM, Cheng L, Kushner PJ, Agard DA, et al. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. Cell. 1998;95: 927–937.

295. Kruse SW, Suino-Powell K, Zhou XE, Kretschman JE, Reynolds R, Vonrhein C, et al. Identification of COUP-TFII orphan nuclear receptor as a retinoic acid-activated receptor. PLoS Biol. 2008;6: e227.

296. Li Y, Suino K, Daugherty J, Xu HE. Structural and biochemical mechanisms for the specificity of hormone binding and coactivator assembly by mineralocorticoid receptor. Mol Cell. 2005;19: 367–380.

297. Bartha Á, Győrffy B. TNMplot.com: A Web Tool for the Comparison of Gene Expression in Normal, Tumor and Metastatic Tissues. Int J Mol Sci. 2021;22. doi:10.3390/ijms22052622

298. Smid M, Wang Y, Klijn JGM, Sieuwerts AM, Zhang Y, Atkins D, et al. Genes associated with breast cancer metastatic to bone. J Clin Oncol. 2006;24: 2261–2267.

299. Gómez-Maldonado L, Tiana M, Roche O, Prado-Cabrero A, Jensen L, Fernandez-Barral A, et al. EFNA3 long noncoding RNAs induced by hypoxia promote metastatic dissemination. Oncogene. 2015;34: 2609–2620.

300. Coleman JLJ, Ngo T, Smythe RE, Cleave AJ, Jones NM, Graham RM, et al. The N-terminus of GPR37L1 is proteolytically processed by matrix metalloproteases. Sci Rep. 2020;10: 19995.

301. Zhang L, Cheng A, Yu Y, Zou N, Wang W, Lv L, et al. GOLT1A-KISS1 fusion is associated with metastasis in adenoid cystic carcinomas. Biochem Biophys Res Commun. 2020;526: 70–77.

302. Papageorgis P, Lambert AW, Ozturk S, Gao F, Pan H, Manne U, et al. Smad signaling is required to maintain epigenetic silencing during breast cancer progression. Cancer Res. 2010;70: 968–978.

303. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14: e8124.